High-dimensional propensity scores for data-driven confounder adjustment in UK electronic health records

John Ross Tazare

Thesis submitted in accordance with the requirements for the degree of

Doctor of Philosophy of the University of London

July 2021

Department of Medical Statistics

Faculty of Epidemiology and Population Health

London School of Hygiene & Tropical Medicine

Funded by a Medical Research Council London Intercollegiate Doctoral Training Partnership Studentship

Grant code MR/N013638/1

Research group affiliation: Electronic Health Records Group

Declaration

Statement of Own Work

I, John Tazare, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, this has been indicated in the thesis. I have read and understood the School's definition of plagiarism and cheating given in the Research Degrees Handbook.

John Tazare,

[redacted]

July 2021

The copyright of this thesis rests with the author and no quotation from it or information derived from it may be published without the prior written consent of the author.

Abstract

Electronic health record (EHR) databases are increasingly used to investigate the effect of medications. When the aim is to answer causal questions surrounding the benefits and harms of medications, a key methodological issue is confounder adjustment. Furthermore, successful mitigation of confounding effects often relies on capturing hard to measure markers of frailty, disease severity or health seeking behaviour. This can be especially hard in this context since these data are not collected for research purposes.

The high-dimensional propensity score (HDPS) algorithm is a semi-automated data driven approach for confounder identification, prioritisation and adjustment tailored for use in large healthcare databases. The HDPS is increasingly applied in pharmacoepidemiological studies amid growing evidence supporting the benefit of these approaches in comparison to standard covariate adjustment methods. Developed in administrative claims databases, there has been little exploration of how best to translate the algorithm beyond this setting.

In this thesis, I propose modifications for implementing HDPS principles in UK primary care EHRs that aim to better characterise features of these data. These modifications are applied to case studies where residual confounding is a key concern. In addition, I propose diagnostic tools and guidance for the reporting of HDPS approaches in general. Furthermore, the developed HDPS approaches are implemented in the Stata statistical software package. Finally, I extend the existing HDPS framework to support the incorporation of laboratory test result information.

Whilst the HDPS is not a panacea, the collective findings of this thesis demonstrate the utility of HDPS approaches for overcoming intractable confounding in UK EHRs. Future work will further explore the use of test result data within the HDPS framework.

Acknowledgements

This thesis has only been possible thanks to the help and support of many people.

Firstly, I would like to thank my supervisors Fizz Williamson and Ian Douglas. Thank you for your encouragement, expertise and generosity. This has been such a rewarding experience and I am extremely grateful for all you have helped me achieve. Thank you to Stephen Evans and Liam Smeeth, your insightful advice and comments have helped to develop and improve much of this work. Finally, thanks to everyone in the Electronic Health Records Research Group, it has been a privilege working with such an inspiring and supportive team.

I would like to thank the Medical Research Council for funding this PhD studentship and a placement at GlaxoSmithKline. Thanks also to all the administrative staff who have helped me, including Lara Crawford, Lauren Dalton and Jenny Fleming.

During this PhD, I have been fortunate enough to undertake placements that have improved this work and developed me as a researcher. Thanks to Josh Gagne, Sebastian Schneeweiss and the Division of Pharmacoepidemiology at Harvard University for being so generous with their time and expertise. I would also like to thank the Real World Analytics team at GlaxoSmithKline. Thanks in particular to Dan Gibbons and John Logie for being so encouraging and supportive, I have greatly enjoyed working with you both. Finally, my thanks to the Datalab team at the University of Oxford for being so patient and motivating, even in the midst of a global pandemic. I would especially like to thank Alex Walker and Ben Goldacre for their trust and enthusiasm.

I am very lucky to have amazing friends who have supported me throughout this PhD.

Thanks to Ali and Alex for much needed coffee and chats. To Dave and Rob, thank you for being such consistently reliable friends. Thanks to Josh for much needed distractions and making our place in Highgate a home. Above all, thank you to Ellie for your persistent encouragement, positivity and support.

I owe an enormous debt of gratitude to my family. Thank you to my parents for your unwavering support and motivation.

Finally, thank you to the millions of individuals who allow their data to be shared within the Clinical Practice Research Datalink, this work would not have been possible without you.

Acronyms

ACS Acute Coronary Syndrome

AKP Alkaline Phosphatase

ALP Alkaline Phosphatase Level

ALT Alanine Aminotransferase

ARR Apparent Relative Risk

ASD Absolute Standardised Difference

AST Aspartate Aminotransferase

ATT Average Effect of Treatment in the Treated

BMI Body Mass Index

BNF British National Formulary

CHD Coronary Heart Disease

CKD Chronic Kidney Disease

COPD Chronic Obstructive Pulmonary Disease

COVID-19 2019 Novel Coronavirus

COX-2I Cyclooxygenase-2 Inhibitors

CPRD Clinical Practice Research Datalink

DAG Directed Acyclic Graph

EHR Electronic Health Record

GERD Gastro-Oesophageal Reflux Disease

GFR Glomerular Filtration Rate

GORD Gastro-Oesophageal Reflux Disease

GP General Practitioner

GSK GlaxoSmithKline

H2RA H2 Receptor Antagonists

HDPS High Dimensional Propensity Score

HES APC Hospital Episode Statistics Admitted Patient Care

HIV Human Immunodeficiency Virus

HR Hazard Ratio

IBD Inflammatory Bowel Disease

IBM International Business Machines

ICD-10 International Classification of Diseases, 10 ed.

ICPE International Conference on Pharmacoepidemiology

IMD Index of Multiple Deprivation

IPTW Inverse Probability of Treatment Weight

IQR Interquartile Range

ISAC Independent Scientific Advisory Committee

IV Instrumental Variable

LSHTM London School of Hygiene and Tropical Medicine

LSOA Lower Layer Super Output Area

MCH Mean Corpsucular Haemoglobin

MCV Mean Corpsucular Volume

MI Multiple Imputation

MINAP Myocardial Ischaemia National Audit Project

ML Machine Learning

NHS National Health Service

NSAID Non-Steroidal Anti Inflammatory Drugs

OCS Oral Corticosteroid

ONS Office of National Statistics

OTC Over The Counter

PPI Proton Pump Inhibitor

PR Prevalence Ratio

PS Propensity Score

PVD Peripheral Vascular Disease

QOF Quality and Outcomes Framework

RALES Randomized Aldactone Evaluation Study

RBC Red Blood Count

RCT Randomised Controlled Trial

RR Relative Risk

SCCS Self Controlled Case Series

SD Standard Deviation

SE Standard Error

SMD Standardised Mean Difference

SSRI Selective Serotonin Reuptake Inhibitors

T2DM Type 2 Diabetes Mellitus

UGIB Upper Gastrointestinal Bleeding

UK United Kingdom

WBC White Blood Count

WHO World Health Organisation

Contents

1 Introduction **22**

1.1 Motivation 23

1.2 Aim 26

1.3 Objectives 26

1.4 Thesis structure 26

2 Background **29**

2.1 Overview 30

2.2 Confounding in pharmacoepidemiology 31

2.2.1 Motivation 31

2.2.2 Theoretical perspective on confounding control 32

2.3 Types of healthcare databases 33

2.3.1 Administrative claims data 34

2.3.2 Electronic health records 36

2.4 Data sources 36

2.4.1 Clinical Practice Research Datalink 37

2.4.2 Linkages 37

2.5 Propensity score analysis 39

2.5.1 Causal inference and the potential outcomes framework 39

2.5.2 Definition and assumptions 40

2.5.3 Estimation 41

2.5.4 Analysis 41

2.5.5 Variable selection 43

2.5.6 Comparison with multivariable adjustment 43

2.6 High-dimensional propensity scores 45
2.6.1 Proxy adjustment 45
2.6.2 Description 47
2.6.3 Properties 49
2.6.4 Critique 53
3 Paper A: Implementing HDPS principles in UK EHRs 57
3.1 Overview 58
3.2 Abstract 61
3.3 Introduction 62
3.4 Propensity scores 63
3.5 Description of the HDPS approach and underlying principles 63
3.5.1 Preliminary steps 63
3.5.2 Identification of most relevant covariates 63
3.5.3 Prioritisation 64
3.5.4 Estimation of the HDPS 65
3.6 Proposed implementation of HDPS principles to UK EHRs 65
3.6.1 Principle 1: Identification of dimensions 67
3.6.2 Principle 2: Code granularity 67
3.6.3 Principle 3: Code recurrence 68
3.6.4 Principle 4: Selected number of variables 69
3.7 Application to example in CPRD 69
3.7.1 Data 69
3.7.2 Design 70
3.7.3 Statistical analysis 71
3.8 Results 72
3.9 Discussion 77
3.10 Ethics statement 79
3.11 Supporting information 80
4 Paper B: HDPS diagnostic tools and reporting considerations . . . 85
4.1 Overview 86
4.2 Abstract 90

4.3 Introduction 91
4.4 High-dimensional propensity scores 92
4.5 Considerations for reporting 93
4.6 Data for illustration 95
4.6.1 Background 95
4.6.2 Summary of HDPS analysis 97
4.7 Diagnostic & visualisation tools 99
4.7.1 Model summaries 100
4.7.2 Comparison of PS distributions 100
4.7.3 Covariate balance 103
4.7.4 Identification of potentially influential covariates 104
4.8 Sensitivity analyses 113
4.8.1 Varying number of covariates selected 113
4.8.2 Quantifying impact of potentially influential covariates 116
4.9 Discussion 118
4.10 Ethics statement 119
4.11 Supporting information 119
5 Paper C: The HDPS suite of commands in Stata 143
5.1 Overview 144
5.2 Abstract 147
5.3 Introduction 147
5.4 High-dimensional propensity scores 149
5.5 The hdps commands 153
5.5.1 Installation 153
5.5.2 Data formats 153
5.5.3 The hdps setup command 154
5.5.4 The hdps prevalence command 156
5.5.5 The hdps recurrence command 157
5.5.6 The hdps priortize command 157
5.5.7 The hdps graphics command 158

5.6 Example using simulated data . . . 160
5.6.1 Simulated data . . . 160
5.6.2 High-dimensional propensity score procedure . . . 162
5.6.3 Investigator propensity score analysis . . . 168
5.6.4 High-dimensional propensity scores analysis . . . 170
5.7 Discussion . . . 171
5.8 Acknowledgments . . . 172
5.9 Supporting information . . . 172
6 HDPS analysis of GI bleed risk in NSAID and COX-2I users . . . 184
6.1 Overview . . . 185
6.2 Introduction . . . 187
6.3 Methods . . . 188
6.3.1 Data source . . . 188
6.3.2 Study population . . . 188
6.3.3 Exposure . . . 189
6.3.4 Covariates . . . 189
6.3.5 Outcomes . . . 191
6.3.6 Statistical analysis . . . 191
6.4 Results . . . 194
6.4.1 Investigator-led traditional PS analysis . . . 194
6.4.2 HDPS analysis . . . 195
6.5 Discussion . . . 206
6.6 Supporting information . . . 207
7 Paper D: PPIs and risk of all-cause and cause-specific mortality . . . 240
7.1 Overview . . . 241
7.2 Abstract . . . 245
7.3 Introduction . . . 246
7.4 Methods . . . 247
7.4.1 Data source . . . 247
7.4.2 Study population . . . 248
7.4.3 Exposure . . . 249

7.4.4 Covariates 249
7.4.5 Outcomes 250
7.4.6 Statistical analysis 251
7.5 Results 253
7.5.1 Risk of mortality relative to H2RA users 253
7.5.2 Risk over different time periods 259
7.5.3 Non-user comparison 259
7.5.4 Risk of mortality relative to non-users 259
7.5.5 Sensitivity analysis 260
7.6 Discussion 261
7.7 Acknowledgments 267
7.8 Ethics statement 267
7.9 Supporting information 267
8 Incorporating test result information within the HDPS framework 292
8.1 Overview 293
8.2 Introduction 295
8.3 PPI-Mortality study 295
8.3.1 Data summary 296
8.3.2 Results summary 296
8.3.3 Re-analysis using HDPS modifications 297
8.4 Types of test result information 300
8.4.1 Overview in UK EHRs 300
8.4.2 Test requested 301
8.4.3 Continuous test results 301
8.5 Data analysis 302
8.5.1 Tests requested 303
8.5.2 Cleaning of continuous blood test results 304
8.5.3 Cut-offs 305
8.5.4 Continuous modelling 309
8.6 Results 310
8.6.1 Tests requested 312

8.6.2 Cleaning . . . 314
8.6.3 Cut-offs . . . 315
8.6.4 Continuous modelling . . . 316
8.7 Discussion . . . 320
8.8 Ethics statement . . . 322
8.9 Supporting information . . . 323
8.9.1 A: Cleaned test results . . . 323
8.9.2 B: Continuous blood test results incorporated . . . 342
9 Discussion . . . 343
9.1 Overview . . . 344
9.2 Summary of findings . . . 345
9.2.1 Obj. 1: Describe UK EHRs and relevant PS methodology . . . 345
9.2.2 Obj. 2: Propose modifications for implementing HDPS principles in UK EHRs . . . 346
9.2.3 Obj. 3: Apply the HDPS and proposed modifications in UK EHRs 348
9.2.4 Obj. 4: Provide guidance surrounding diagnostic tools and reporting of HDPS analyses . . . 350
9.2.5 Obj. 5: Develop reusable software to implement HDPS approaches in Stata . . . 352
9.2.6 Obj. 6: Investigate extensions for incorporating labratory test information . . . 353
9.3 Strengths . . . 355
9.3.1 Application of proposed approaches to applied studies . . . 355
9.3.2 Accessibility of methods . . . 355
9.4 Limitations . . . 358
9.4.1 Comparison with machine learning approaches . . . 358
9.4.2 Generalisability of results . . . 359
9.5 Future work . . . 361
9.5.1 Incorporating additional data available in UK EHRs . . . 361
9.5.2 HDPS R package . . . 362
9.5.3 CPRD Aurum . . . 362

9.5.4 Empirical studies 362
9.5.5 Prediction modelling 363
9.6 Concluding remarks 364
Appendix A LSHTM Ethical approval for PPI-Clopidogrel study . . . 366
Appendix B ISAC application & approval for PPI-Clopidogrel study . 368
Appendix C LSHTM Ethical approval for NSAID-COX2i study 385
Appendix D ISAC application & approval for NSAID-COX2i study . 387
Appendix E LSHTM Ethical approval for PPI-Mortality study 417
Appendix F ISAC application & approval for PPI-Mortality study . . 419
Appendix G License Agreement for Papers A & D 444
Bibliography 444

List of Tables

2.1 Summary of information typically available in healthcare databases . . . 35

3.1 Summary of dimensions for UK electronic health records 67
3.2 Baseline characteristics by proton pump inhibitor status amongst clopidogrel and aspirin users 73
3.3 Estimated treatment effect of proton pump inhibitor use on myocardial infarction risk 74
3.4 Top 100 unmapped Read codes from Read to ICD-10 cross-map procedure 81

4.1 Reporting considerations for key features and decisions of the HDPS approach 96
4.2 Summary of Clinical Research Practice Datalink study used for illustration 98
4.3 Summary of established and proposed diagnostic tools for HDPS models 99
4.4 Comparison of the mean absolute standardised differences in the unweighted, pre-defined and pre-defined and HDPS weighted populations . 112
4.5 Sensitivity analyses exploring the impact of influential covariates 117

6.1 Characteristics of NSAID and COX-2 inhibitor users in unmatched and matched samples 199
6.2 Results from primary analysis comparing investigator and HDPS models 201
6.3 Sensitivity analyses for the HDPS analysis 201

7.1 Covariates adjusted for in statistical models 250
7.2 Absolute standardised differences between PPI and H2RA users before and after weighting 255

8.1 Association between PPI prescription and COPD-mortality applying HDPS modifications 298
8.2 Proposed cut-offs for generating binary test result HDPS covariates . . . 307
8.3 Comparison of methods for incorporating laboratory test information in the HDPS framework 311
8.4 Read codes for the top 50 tests requested in the PPI-Mortality cohort . 312
8.5 Summary of the 35 clean blood test results 319

List of Figures

1.1 Number of publications using the CPRD 25

2.1 Example of proxy adjustment 46

2.2 Summary of HDPS algorithm steps 48

2.3 Illustration of pre-exposure identification of features for HDPS 49

3.1 Flowchart depicting HDPS steps, underlying principles and adaptations 66

3.2 Empirical performance of HDPS across our implemented adaptations . 75

3.3 Comparison of the estimated propensity score from investigator and HDPS approaches 76

4.1 Summary of high-level concepts captured in HDPS covariates 101

4.2 Overlap plot comparing propensity score distributions between pre-defined and primary HDPS analysis 102

4.3 Prevalence of the top 500 Bross-prioritised HDPS pre-exposure covariates by treatment group and by data dimensions 105

4.4 Comparison of absolute standardised differences between unweighted and HDPS weighted sample under the primary analysis 106

4.5 Comparison of absolute standardised differences in a set of key covariates 107

4.6 Comparison of absolute standardised differences in the pre-defined and top 250 HDPS covariates 108

4.7 Distribution of absolute log Bross bias values for top 500 HDPS covariates110

4.8 Comparison of the covariate-exposure and covariate-outcome associations for the top 500 bias-based HDPS covariates 111

4.9 Sensitivity analysis assessing the impact of selecting 100, 250 and 750 HDPS covariates 114
4.10 Sensitivity analysis assessing the impact of incrementally adjusting for the top 750 HDPS covariates 115

5.1 Summary of a generic implementation of the high dimensional propensity score (HDPS) algorithm 152
5.2 Example cohort study illustrating the setting in which the HDPS algorithm is traditionally applied. 161
5.3 Distribution of absolute log Bross bias values for each of the top 100 HDPS covariates. 164
5.4 Prevalence of the top 100 HDPS covariates by treatment group 166
5.5 Comparison of the strength of covariate-exposure and covariate-outcome associations for the top 100 bross ranked HDPS covariates 167

6.1 Schematic showing active comparator study design NSAID and COX-2 inhibitor use on upper GI bleeding risk 190
6.2 Prescribing trends for NSAIDs and COX-2 inhibitors across the study period 197
6.3 Comparison of estimated propensity score distributions in the investigator-matched sample 198
6.4 Prevalence of the top 500 Bross-prioritised covariates by treatment group 202
6.5 Strength of covariate-exposure and covariate-outcome associations for the top 500 HDPS covariates 203
6.6 Comparison of absolute standardised differences (ASDs) between unmatched and HDPS matched samples 204
6.7 Comparison of estimated propensity score distributions in the HDPS-matched sample 205

7.1 PPI-Mortality Study flow chart 254
7.2 Forest plot for HRs between PPIs and both all-cause and broad-level cause-specific mortality 257
7.3 Forest plot for HRs between PPIs and mortality from individual causes 258

7.4 Forest plot for HRs between PPIs both all-cause and broad-level cause-specific mortality stratified by time 262
7.5 Forest plot for HRs between PPIs and mortality from individual causes stratified by time 263

8.1 Overlap plot comparing propensity score distributions between investigator and modified HDPS analysis 299
8.2 Overlap plot comparing propensity score distributions between HDPS analyses 317
8.3 Prevalence of the top 500 Bross-prioritised HDPS covariates by treatment group and by data dimensions 318
8.4 Comparison of the covariate-exposure and covariate-outcome associations for the top 500 bias-based HDPS covariates 318

Chapter 1

Introduction

1.1 Motivation

Pharmacoepidemiology is a branch of epidemiology concerned with the application of epidemiological methods to study the benefits and harms of drugs in human populations (*Strom et al.*, 2013). The randomised controlled trial (RCT) is usually considered the gold-standard for studies of this nature, however, they are typically inadequate for addressing important questions surrounding the long-term and rare effects of medications. Recent legislation mandates pharmaceutical companies to conduct safety and effectiveness studies in routine care and large healthcare databases can provide the best opportunity to obtain powerful estimates of these effects (*Council of European Union*, 2010; *Toh*, 2017; *US FDA*, 2011).

The proliferation of electronic health record (EHR) data, such as the UK Clinical Practice Research Datalink (CPRD) (*Herrett et al.*, 2015; *Wolf et al.*, 2019), has lead to increased optimism surrounding the possibility such data provide for obtaining affordable, reliable and timely answers to important causal questions surrounding the effects of medications. Figure 1.1 highlights the increased use of the CPRD in published research articles over the last 30-years (from *CPRD* (2021)); a pattern seen across large healthcare databases in general. Whilst Figure 1.1 indicates a drop in publications in 2020, this was likely due to the coronavirus (COVID-19) pandemic where there was delayed access to up-to-date data meaning researchers relied on additional UK data sources, e.g. OpenSAFELY (*Williamson et al.*, 2020). Therefore, it is likely that the use of UK EHR data will continue to increase in the future.

The use of these data within the field of pharmacoepidemiology has developed rapidly and reliable estimates of treatment effects have been obtained, in part due to an emphasis on the formulation of questions within a causal framework and careful planning of suitable study designs (*Hernán and Robins*, 2020; *Hernan and Robins*, 2016; *Schneeweiss and Avorn*, 2005; *Wettermark*, 2013). However, examples of inconsistent and incorrect conclusions being drawn highlight the necessity to study and explore potential issues and biases further (*de Vries et al.*, 2006; *Douglas et al.*, 2012; *Freemantle et al.*, 2013; *Ray*, 2003). Whilst information bias and selection bias are important areas

of concern, adjustment for confounding often remains the key issue (*Schneeweiss and Avorn*, 2005; *Strom et al.*, 2013; *Suissa*, 2009).

Introduced in 1983 by Rosenbaum and Rubin, the propensity score (PS) has become an important method for confounder-adjustment in pharmacoepidemiology (*Jackson et al.*, 2017; *Rosenbaum and Rubin*, 1983). PSs have several advantages which mean that they are often preferred to outcome regression in this setting. For example, PS can readily convert a large amount of confounder information into a single number and they explicitly force investigators to consider indications for treatment use (*Glynn et al.*, 2006). The popularity of these approaches has motivated developments in PS methodology, such as the high-dimensional propensity score (HDPS) algorithm (*Schneeweiss et al.*, 2009); a data-driven approach that attempts to optimise confounding control by harnessing the full volume of data available within a healthcare database (*Schneeweiss*, 2019).

Whilst these novel methods are becoming widely adopted in a diverse range of settings, it is difficult to establish whether a particular method has uniform superiority since healthcare databases vary widely in complexity and the information available. This thesis contributes to the developing literature surrounding the HDPS by investigating the potential of these methods for improving confounder adjustment in UK EHRs. Futhermore, I develop graphical diagnostic tools for assessing these models and investigate how to incorporate laboratory test information within the HDPS framework.

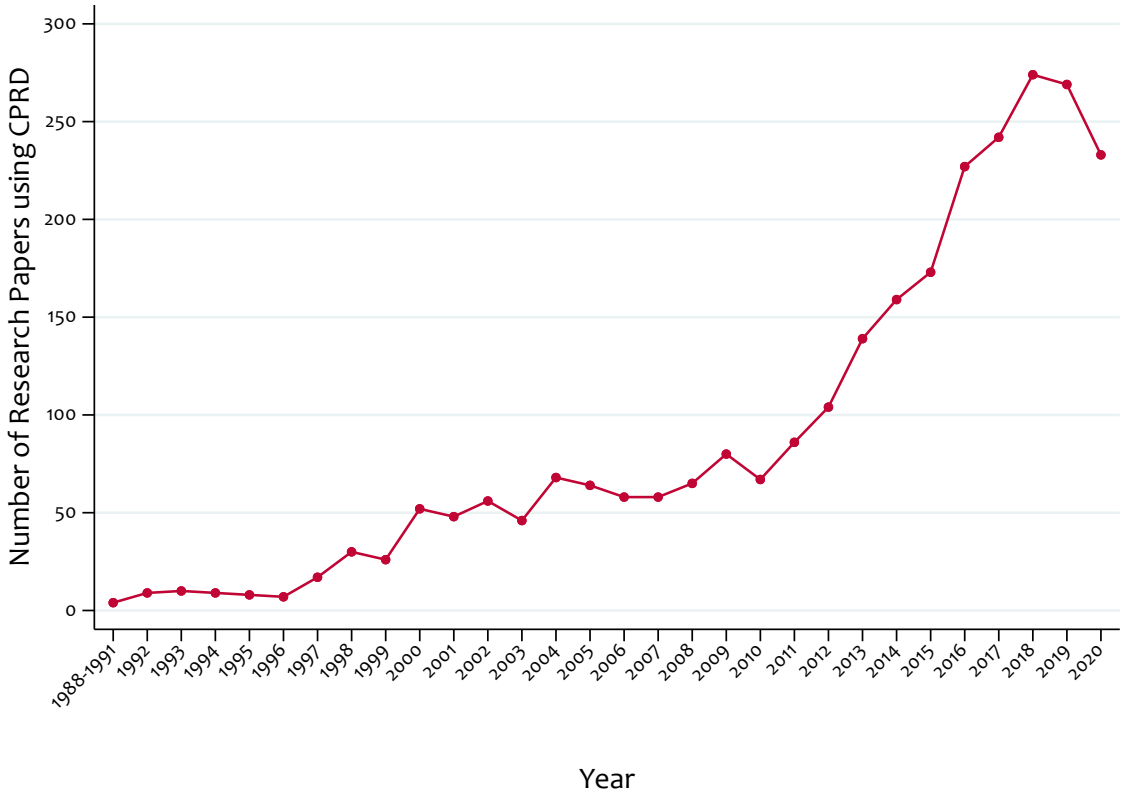


Figure 1.1: *Number of publications using the Clinical Practice Research Datalink from 1988 to 2020*

1.2 Aim

To investigate the use of data-driven approaches, within the HDPS framework, for confounder adjustment in UK EHRs with particular emphasis on providing practical guidance and improving accessibility.

1.3 Objectives

The aim will be addressed by the following objectives:

1. Describe UK EHRs and review relevant propensity score methodology.
2. Propose modifications for implementing the underlying principles of the HDPS in UK EHRs.
3. Apply the HDPS and proposed modifications in the context of UK EHRs.
4. Provide guidance surrounding diagnostic tools and reporting of HDPS analyses.
5. Develop reusable software to implement HDPS approaches in the Stata statistical software package.
6. Investigate extensions to the HDPS framework that allow for the incorporation of laboratory test information.

1.4 Thesis structure

This is a *research paper style thesis* comprising of chapters formatted in the style of a journal article (prefixed with "Paper") and more traditional thesis style chapters.

Chapter 1 provides a short introduction motivating the PhD and outlining the aims and objectives. Chapters 2 describes relevant contextual information referenced throughout the thesis, including a) a description of healthcare databases used for conducting

pharmacoepidemiological research and, b) a review of PS methodology, focusing on the use of these methods in pharmacoepidemiological research and the development of the HDPS framework. Finally, I present a critique of the HDPS approach.

In Chapter 3, I present the first HDPS analysis in this thesis. I describe the identified HDPS principles and discuss each of them in the context of UK EHR data, highlighting specific considerations when applying this approach in this setting. These modifications are illustrated by performing a re-analysis of a study by *Douglas et al.* (2012). I develop code applying the HDPS in the Stata software package allowing for implementation of the proposed modifications. This work was published in *Pharmacoepidemiology and Drug Safety* in September 2020 (*Tazare et al.*, 2020).

Chapter 4 provides graphical diagnostic tools and reporting considerations for HDPS analyses. I review existing PS diagnostic tools and discuss the suitability of these in the context of HDPS analyses. I develop and extend diagnostic tools specifically tailored for use in HDPS analyses. In particular, I contribute novel diagnostic tools surrounding the presentation of HDPS models and assessment of covariate balance. Furthermore, I present reporting considerations highlighting the importance of transparently describing key decision made when applying HDPS approaches. This work was partly developed under the supervision of Joshua Gagne during a research visit to the Division of Pharmacoepidemiology and Pharmacoeconomics at Brigham and Women's Hospital, Harvard University. This work is currently under review in *Pharmacoepidemiology and Drug Safety*.

In Chapter 5, I present a suite of commands implementing HDPS approaches (including those developed in Chapter 3) in the Stata statistical software package. I describe each command and illustrate a HDPS analysis using simulated data. Finally, I detail how to install the developed commands and help-files, highlighting example code and data freely available on GitHub. This work is currently under review in *The Stata Journal*.

In Chapter 6, I apply the HDPS modifications developed in Chapter 3 in the context of a study investigating the risk of upper gastrointestinal bleeding in users of selective cyclooxygenase-2 inhibitors and traditional non-steroidal anti-inflammatory drugs. I

undertook this work with additional supervision from the Real World Analytics team at GlaxoSmithKline and as part of a larger project investigating prevalent new user designs.

Chapters 7 and 8 focus on a study investigating the association between proton pump inhibitor use and all-cause and cause-specific mortality, published in the *British Journal of Clinical Pharmacology* in January 2021 (*Brown et al.*, 2021). Chapter 7 presents the study by *Brown et al.* (2021), where I conducted the HDPS analysis. Chapter 8 re-analyses the chronic pulmonary obstructive disease specific-mortality outcome studied in the article, applying the modifications described in Chapter 3 and investigating methods for incorporating test result information within the HDPS framework. I describe the availability of test result information in UK EHRs and highlight issues surrounding data management and missing data. I propose HDPS data dimensions capturing tests-requested and continuous test values. To include continuous test value information in HDPS models, I consider approaches using cut-offs (aligned with traditional HDPS covariates) and continuous variables.

Finally, in Chapter 9 I synthesise the findings of the PhD, framing the key original contributions in the context of the existing literature. There is a discussion of the strengths and limitations of this research as well as an outline of the possible directions for future work.

Whilst much of the development of the HDPS has been in the context of administrative claims databases, this thesis provides an in-depth examination of HDPS approaches in UK EHRs.

Chapter 2

Background

2.1 Overview

Summary

In Chapter 1, I provided an overview of the motivation, aim and objectives of this thesis. In this chapter, I describe the use of large healthcare databases for conducting pharmacoepidemiological research and summarise the data sources used throughout this thesis. Additionally, I review relevant propensity score (PS) methodology, including the high-dimensional propensity score algorithm.

Thesis objective addressed

This chapter addresses the following objective of the overall thesis (Section 1.3):

1. Describe UK EHRs and review relevant propensity score methodology.

2.2 Confounding in pharmacoepidemiology

2.2.1 Motivation

Confounding bias is the systematic difference between a group of patients receiving treatment and a relevant comparator group (*Brookhart et al.*, 2010). In clinical trials, the random assignment of treatment received is a key strength that minimises confounding bias in these studies (*Strom et al.*, 2013). However, there is growing recognition of the importance of conducting pharmacoepidemiological studies, both to supplement findings from clinical trials and contribute evidence surrounding the long-term and rare effects of medications. In these non-interventional studies treatment allocation is not random and investigators therefore need to understand and measure key underlying differences between patients receiving different therapies to mitigate confounding bias (*Brookhart et al.*, 2010).

Importantly, recent studies have highlighted inconsistencies in the results between randomised clinical trials and non-interventional studies and these differences are often hypothesised to be driven by residual confounding (*Douglas et al.*, 2012; *Freemantle et al.*, 2013). One such example was the attempted replication of the Randomized Aldactone Evaluation Study (RALES) in UK EHR data (*Freemantle et al.*, 2013). The RALES trial studied patients with severe heart failure and observed reduced mortality in those who received spironolactone (an aldosterone inhibitior) compared to those who did not, a finding replicated by two other independent trials (*Pitt et al.*, 1999, 2003; *Zannad et al.*, 2011). Conversely, in the observational study, use of spironolactone appeared to be associated with a substantial increase in mortality, despite adjustment for a large set of relevant patient demographics, comorbidities, and medications (*Freemantle et al.*, 2013). The authors concluded that the discrepancy in results was likely due to important factors that they were unable to fully adjust for and, in particular, relating to severity of heart failure and the clinical decision to treat (*Freemantle et al.*, 2013).

These examples motivate the need to understand and overcome issues surrounding confounding bias, particularly in the context of increased discussions surrounding the use

of evidence from large healthcare databases in the regulatory decision making process (*Franklin et al.*, 2019; *Schneeweiss and Glynn*, 2018; *Toh*, 2017).

Finally, contradictory results obtained in non-interventional studies have led some to question whether we can ever reliably trust evidence generated using these data (*Collins et al.*, 2020). However, this implies a false dichotomy and the results from both should be considered complementary to our understanding of the effects of medications. (*Avorn*, 2007).

2.2.2 Theoretical perspective on confounding control

The use of mediciations in a particular healthcare system is often determined by a complex array of factors relating both to the clinician prescribing the medication and patient-level variables (*Brookhart et al.*, 2010). Therefore, to successfully control for confounding bias we are looking to identify a set of variables that, when appropriately adjusted for in a statistical analysis (for example, via multivariable outcome regression or propensity score (PS) analysis), will obtain an unbiased answer to a causal question of interest (*Brookhart et al.*, 2010).

Directed acyclic graphs (DAGs) are an increasingly popular framework for encoding assumptions surrounding the relationships between study variables (*Greenland et al.*, 1999; *Hernán and Robins*, 2020; *Krieger and Davey Smith*, 2016). Furthermore, using graph theory, DAGs allow investigators to identify a minimal set of covariates required to remove confounding bias for a given causal question (*Greenland et al.*, 1999; *Hernán and Robins*, 2020; *Pearl*, 1995). However, this requires the investigators specifying all relationships between study variables, which in the context of healthcare databases can be challenging for many reasons. For example, investigators rarely are able to specify these relationships *a priori* and many variables are not directly measurable in the data available (*Brookhart et al.*, 2010).

In this thesis, when referring to a variable as a 'confounder' this refers to the more formal definition of this variable being a member of the aforementioned minimally suf-

ficient adjustment set of covariates needed to eliminate confounding bias (*VanderWeele and Shpitser*, 2013). Unfortunately, there is no statistical test for identifying whether a variable is a confounder (*Brookhart et al.*, 2010; *Greenland and Robins*, 1986; *Robins*, 2001; *Schneeweiss*, 2019). This has important implications for the data-adaptive methods for confounding control used throughout this thesis and is discussed in Section 2.6.4.

2.3 Types of healthcare databases

The growing focus on pharmacoepidemiological evidence is partly driven by the ubiquity of computerised healthcare databases. Whilst these databases can generally be categorised as either health record databases or administrative claims databases, there is considerable variability between different databases (*Schneeweiss and Avorn*, 2005). The main reason for this variability is that these are secondary data sources, primarily conducted for administrative rather than research purposes and generated to capture relevant information from an underlying healthcare system. Key sources of variation include: 1) differential rates of patients entering and leaving the databases, 2) data quality and completeness and, 3) the ability to link to other data sources (for example, specific disease registries) (*Schneeweiss and Avorn*, 2005). Given the variability between databases, it important to carefully assess a data source to ensure that a specific research question can be adequately answered (*Hennessy*, 2006).

Healthcare databases have several general strengths that make them useful for answering a wide array of pharmacoepidemiological questions (*Hennessy*, 2006; *Schneeweiss and Avorn*, 2005):

- **Affordability:** Compared to clinical trials and epidemiological studies prospectively collecting data, these data are made available at a relatively low cost.
- **Linkage:** It is often possible to link to additional datasets (for example, death certificate data) which can considerably expand the depth and type of research questions investigated.

- **Representativeness:** Databases are often representative of key patient populations, for example, the general population receiving clinical care in a particular healthcare system. Additionally, some databases have good representativeness of specific patient groups about whom we often lack evidence of drug effects, for example, elderly patients, children, and those in care (or nursing) homes.

- **Velocity:** These databases are typically generated in an automated way that avoids extended delays surrounding data access. Furthermore, data are often provided to researchers in a fixed format, allowing for analytic and data management code to be efficiently recycled and answers to be obtained in a timely manner.

- **Volume:** The large size of these databases often allows investigators to obtain powerful estimates of treatment effects, especially important for rare events and to look at important subgroups.

In the following sections, I briefly outline the key similarities and differences between administrative claims databases and EHRs. In the context of high-dimensional propensity scores (HDPS), this is relevant for considering the types of data available to the algorithm in a particular setting. The information typically available in each type of database is summarised in Table 2.1.

2.3.1 Administrative claims data

Administrative insurance claims databases arise as a result of financial transactions between the healthcare system and an insurance carrier (*Strom et al.*, 2013). For example, if a patient is admitted to hospital, the insurance carrier will be billed for the cost of the care received and this will need to be supported by the recording of a diagnosis.

Whilst information relating to these transactions, such as prescriptions dispensed, referrals, and primary diagnoses are reliably recorded, lifestyle information (for example, alcohol use and smoking status) is rarely present in these data (*Schneeweiss and Avorn*, 2005). A further limitation of claims data surrounds high patient turnover rates, which

Table 2.1: *Summary of information typically available in administrative claims databases and electronic health records.* ***Abbreviations:*** *EHR, electronic health record.*

Type of information	Administrative claims	EHR
Clinical diagnoses	✓	✓
Referrals to specialists	✓	✓
Medications	✓	✓
Laboratory test result values	Infrequently available	✓
Lifestyle Information: Smoking status, alcohol use and body mass index	-	✓
Physical activity & diet	-	-

can often restrict the long-term follow up of patients. Common reasons for high turnover rates include patients changing jobs and employers' moving health care providers (*Strom et al.*, 2013). Finally, claims data can be less representative of the broader population since they tend to cluster around socioeconomic strata (*Strom et al.*, 2013).

2.3.2 Electronic health records

To borrow an analogy from *Hennessy* (2006), if claims data provide an "accountant's eye view" of a patient, then EHRs provide a "doctor's eye view". EHRs typically arise from the computerisation of paper medical records, documenting medical information recorded as part of consultations with healthcare professionals (*Strom et al.*, 2013).

In comparison to claims data, these data are more likely to contain some patient lifestyle and laboratory test result information. However, this information will only be available if it is requested and recorded by a healthcare professional. Therefore, there is missing data for a subset of the patient population and this has implications for any statistical analysis (*Farmer et al.*, 2018; *Petersen et al.*, 2019).

Additionally, the completeness of data recording can be more variable in comparison to claims data. Whilst the recording of certain information might be incentivised (for example, the Quality and Outcomes Framework in the UK system (*Lester*, 2008)), the information recorded at a consultation is, at least in part, likely to be driven by healthcare professional or site preference (as opposed to claims data where complete information is required for legal and auditing reasons) (*Strom et al.*, 2013).

2.4 Data sources

In this section, I describe the relevant databases and linkages used throughout this thesis.

2.4.1 Clinical Practice Research Datalink

The United Kingdom (UK) Clinical Practice Research Datalink (CPRD) is one of the largest de-identified longitudinal General Practice (GP) based electronic health record databases in the world and is broadly representative of patients registered at GP practices in the UK (*Herrett et al.*, 2015; *Wolf et al.*, 2019). The CPRD captures information from practices using the Vision and EMIS IT systems that agree to contribute patient data (*Wolf et al.*, 2019). These data are delivered to researchers via the CPRD GOLD (Vision practices) and CPRD Aurum databases (EMIS practices) (*Herrett et al.*, 2015; *Wolf et al.*, 2019). Throughout this thesis, CPRD GOLD data is used to illustrate and apply HDPS methods but the work could easily be applied in CPRD Aurum too.

CPRD GOLD data are delivered to researchers through extract files and contain information relating to: 1) patient demographics and lifestyle information, 2) clinical symptoms and diagnoses, 3) therapy prescriptions, 4) referrals to specialists, and 5) laboratory test results.

Patient-level data from the CPRD can be linked to many other data sources using unique National Health Service (NHS) numbers (*Herrett et al.*, 2015; *Padmanabhan et al.*, 2019). Relevant linkages used in this thesis are described below.

2.4.2 Linkages

Hospital Episode Statistics

Hospital Episode Statistics (HES) is a records based system providing information on hospital admissions, outpatient appointments and accident and emergency (A&E) attendances per period of care at NHS hospitals in England (*Herbert et al.*, 2017).

Index of Multiple Deprivation

The English indices of deprivation provides information on relative deprivation for constituencies across England using key indicators such as income and education (*Herrett et al.*, 2015). Patient level Index of Multiple Deprivation (IMD) will be used to obtain baseline levels of socioeconomic status.

Myocardial Ischaemia National Audit Project

The Myocardial Ischemia National Audit Project (MINAP) audits the quality of care for patients with acute coronary syndrome (ACS) in England and Wales, recording episodes of care for those admitted to acute NHS hospitals with ACS (*Herrett et al.*, 2010a).

Office for National Statistics Mortality data

The Office for National Statistics (ONS) is the largest independent provider of official statistics in the UK. ONS mortality data will be used to accurately ascertain date and cause of death from death certificates (*Herrett et al.*, 2015).

Rural-Urban Classification

The rural-urban classification allows investigators to identify rural and urban areas, which can be important for capturing socioeconomic characteristics (*CPRD*). In this thesis, rural-urban classification is used at the Lower Layer Super Output Area (LSOA) level, a geographic hierarchy designed to capture small areas in England and Wales (*NHS*, 2020).

2.5 Propensity score analysis

In this section, I provide background information surrounding causal inference and the potential outcomes framework. Next, I review relevant propensity score (PS) methodology with a focus on pharmacoepidemiological research.

2.5.1 Causal inference and the potential outcomes framework

In pharmacoepidemiology, we often aim to answer causal questions surrounding the effects of medications, for example, "How does a patient's risk of an outcome Y change if they initiate a new therapy A?" (*Strom et al.*, 2013). When discussing "causes", informally we are referring to the following (from *Pearl et al.* (2016)):

> *"Variable A is a cause of a variable Y if Y in any way relies on A for its value. . . . [Similarly], A is a cause of Y if Y listens to A and decides its value in response to what it hears."*

Efforts to formalise this language have resulted in the widely used Neyman-Rubin counterfactual framework of causality (*Guo and Fraser*, 2010). This framework is based on the idea of potential outcomes which define, given an outcome Y and an intervention A, $Y(a)$ the value Y would take if A were set to a. For a study investigating the effects of a study drug compared to a comparator, each patient will have two potential outcomes, the value of the outcome if they received the study drug ($Y(1)$) and the value if they received the comparator ($Y(0)$). In practice, only one of these outcomes is observed and the other is referred to as the counterfactual outcome (the outcome that would have been observed if, counter to fact, the patient had received the alternative therapy). We conclude that the treatment has a causal effect on a patient's outcome if $Y(1) \neq Y(0)$ (*Guo and Fraser*, 2010). However, the inability to observe both potential outcomes and perform this comparison has been described by *Holland* (1986) as the 'fundamental problem of causal inference'.

Since we are unable to make inferences based on these individual causal effects, we

focus on a group of individuals we want to make inferences about. More formally, we refer to this quantify as an estimand (*Hernán and Robins*, 2020). In the case studies presented in this thesis, two estimands are considered; the Average Treatment Effect (ATE) and the Average Treatment Effect on the Treated (ATT). The ATE represents the difference between two hypothetical mean outcomes, namely contrasting the mean outcome if everyone was treated and the mean outcome if everyone was not treated (*Williamson et al.*, 2012). Alternatively, the ATT refers to the difference between the mean outcome of all treated patients in the population and the mean outcome of these same patients had they not received treatment (*Williamson et al.*, 2012). That is, the ATT refers to the population of patients who ultimately received the treatment (*Austin*, 2011). Whether the ATE or ATT is suitable for a given study will depend on the research question (*Austin*, 2011; *Williamson et al.*, 2012).

2.5.2 Definition and assumptions

Rosenbaum and Rubin introduced the PS in a seminal paper highlighting the potential for these methods to obtain unbiased treatment effects in non-randomised settings (*Rosenbaum and Rubin*, 1983).

We define a sample of individuals $i = \{1, \ldots n\}$ where each is assigned a treatment $A_i = \{0, 1\}$, has two potential outcomes defined by $Y(a)$, $a = 0, 1$, and has a vector of p observed covariates $\mathbf{X_i} = \{X_{1i}, X_{2i}, \ldots, X_{pi}\}$.

The PS (e_i) is defined as the conditional probability of being treated given $\mathbf{X_i}$.

$$e_i = \Pr(A_i = 1 | \mathbf{X_i})$$

The following four assumptions are required to obtain unbiased treatment effects using PSs (*Austin*, 2011; *Williamson et al.*, 2012).

- Positivity: Each individual must have a nonzero probability of being either treated or untreated; i.e. receiving a particular treatment is guaranteed or impossible

(*Hernán and Robins*, 2020).

- Consistency: Given a patient's set of potential outcomes, the observed outcome will be equal to the potential outcome under the treatment received (*Hernán and Robins*, 2020):

$$Y_i = A_i Y_i(1) + (1 - A_i) Y_i(0)$$

- SITA: The Strongly Ignorable Treatment Assignment (SITA) assumption states that treatment assignment and the potential outcomes are conditionally independent given the observed covariates (*Rosenbaum and Rubin*, 1983). The assumption is also referred to as conditional exchangeability since, if the assumption holds, the two groups are 'exchangeable' based on the observed confounders. Informally, this means that there are no unobserved confounders (*Williamson et al.*, 2012). This is a particularly strong assumption in the non-interventional setting.

$$(Y_i(1), Y_i(0)) \perp\!\!\!\perp A_i | \mathbf{X_i}$$

- SUTVA: The Stable Unit Treatment Value Assumption (SUTVA) states that the two potential outcomes for an individual are unaffected by any other individual's treatment status (*Williamson et al.*, 2012).

2.5.3 Estimation

In non-randomised studies the PS is unknown and needs to be estimated from the data (*Austin*, 2011). The most common approach uses a logistic regression model with treatment as the outcome and a set of observed confounders as covariates.

Alternatives to logistic regression for PS estimation include: boosted and bagged classification and regression trees, random forests and neural networks (*Austin*, 2011).

2.5.4 Analysis

The four main methods based on the PS for removing confounding effects are described below.

Covariate adjustment

The outcome is regressed on treatment and the estimated PS; under this model, the coeffient for the treatment variable is the estimated treatment effect (*Williamson et al.*, 2012). However, since the propensity score is a summary measure of many variables, a key challenge is correctly specifying the functional form (*Webster-Clark et al.*, 2020). This approach can target both the ATE and ATT (*Williamson et al.*, 2012).

Stratification

Stratification (or subclassification) involves dividing the distribution of the PS into strata (*Williamson et al.*, 2012), e.g. quartiles or deciles (*Jackson et al.*, 2017). Stratum-specific treatment effects are estimated before being combined using a weighted average to give an overall estimated treatment effect (weights for each stratum are equal to the fraction of the sample within the stratum) (*Williamson et al.*, 2012). This implementation estimates the ATE, however, the ATT can be estimated by updating the weights used (*Williamson et al.*, 2012). Finally, replacing the PS by the strata in the covariate adjustment method gives an approximation to the stratified estimator.

Matching

Each treated individual is matched to an (or many) untreated individual(s) with a similar value of the PS (*Austin*, 2011; *Rassen et al.*, 2012). In practice, calipers are used to restrict eligible matches to help ensure similarity *Lunt* (2014). In the matched sample, the estimated treatment effect is given by comparing the outcomes between treated and untreated subjects (*Austin*, 2011). This approach estimates the ATT, however, the ATE can be estimated by matching each subject in the sample (this will result in some subjects appearing multiple times in the matched sample) (*Williamson et al.*, 2012).

Weighting

PS weighting methods can be used to target various treatment effects, depending on the underlying question of interest (*Desai and Franklin*, 2019; *Webster-Clark et al.*, 2020). One popular method is inverse probability of treatment weighting (IPTW). IPTW uses weights to construct 2 synthetic samples representing the scenarios in which everyone had been treated and everyone had been untreated (*Austin*, 2011). Weights are typically defined as (*Austin*, 2011):

$$w_i = A_i e_i + (1 - A_i)(1 - e_i)$$

For all weighting methods, estimated treatment effects can be obtained by comparing the outcomes between treated and untreated subjects in the weighted sample. IPTW as defined above estimates the ATE. The ATT can be estimated by updating the definition of the weights (*Austin and Stuart*, 2015; *Williamson et al.*, 2012).

2.5.5 Variable selection

The goal of PS modelling is successful confounding adjustment, not perfect prediction of treatment allocation (*Brookhart et al.*, 2006; *Williamson et al.*, 2012). Therefore, all confounders must be included in the PS model to allow consistent estimation of the treatment effect estimates (*Williamson and Forbes*, 2014). Variables related to the outcome (i.e. risk factors) should also be included in a PS model since they decrease the variance of the treatment effect estimate, irrespective of being related to treatment *Brookhart et al.* (2006). The inclusion of variables unrelated to the outcome but predictive of treatment (i.e. instrumental variables), will result in increased variance for the estimated treatment effect and should not be included (*Williamson and Forbes*, 2014).

2.5.6 Comparison with multivariable adjustment

Multivariable adjustment has historically been the favoured approach for reducing the effects of confounding bias. In this paradigm, a single outcome regression model is

fitted adjusting for a treatment variable and a set of covariates.

Theoretically, in certain settings outcome regression models and PS methods should obtain identical results (*Austin*, 2011). In practice, differences in the results obtained are usually minimal and largely driven by issues relating to unmeasured confounding, non-collapsibility of the effect measure of interest and model misspecification (*Austin*, 2011). Furthermore, efforts to compare the results of studies applying PS analysis and multivariable regression have highlighted that the two approaches often yield similar results (*Glynn et al.*, 2006; *Shah et al.*, 2005; *Sturmer et al.*, 2006).

PS analysis has several advantages over regression adjustment in pharmacoepidemiology.

- The separation of PS modelling from treatment effect estimation forces investigators to explicitly consider confounding by indication (*Glynn et al.*, 2006; *Jackson et al.*, 2017).
- PS-based methods are particularly valuable in settings where the outcome is rare but exposures are not (a common scenario in pharmacoepidemiological research). In these settings, PS models can often successfully adjust for a larger set of covariates.
- Investigators can easily assess the ability of the estimated PS to balance measured covariates between the two treatment groups, whereas outcome regression models can be considered more opaque by this metric (*Austin*, 2009b, 2011).

There are also general drawbacks to the PS approach:

- By summarising confounder information into a single score, investigators lose information on the individual coefficients of covariates in the outcome model (*Austin*, 2011)
- Unmeasured confounding is still an issue when using PS methods since achieving good balance in measured covariates does not guarantee balance in unmeasured

covariates (*Williamson et al.*, 2012). Furthermore, despite its importance, covariate balance is often poorly reported (*Granger et al.*, 2020).

Finally, PS methods and traditional outcome regression can be combined using "doubly robust" approaches, which have been shown to have attractive theoretical properties surrounding model misspecification (*Funk et al.*, 2011). However, these methods are beyond the scope of this thesis.

2.6 High-dimensional propensity scores

A key limitation of large healthcare databases surrounds the absence or imperfect recording of information surrounding confounding factors (*Hennessy*, 2006). Furthermore, since successful mitigation of confounding can often rely on capturing hard to measure concepts this can lead to residual confounding (*Schneeweiss et al.*, 2009). Developed in the setting of administrative claims data, the HDPS attempts to overcome this issue by harnessing the full volume of data available to generate and empirically rank data-driven covariates capturing the health status of patients (*Schneeweiss et al.*, 2009). The algorithm selects a large number of important covariates with the overall aim of minimising residual confounding and has become an established method in pharmacoepidemiological research (*Cadarette et al.*, 2017; *Schneeweiss*, 2019; *Schneeweiss et al.*, 2009)

2.6.1 Proxy adjustment

The HDPS relies on the concept of proxy adjustment to optimise confounding control in a given healthcare database (*Schneeweiss et al.*, 2009). Since these data are generated by a healthcare system and not with research in mind, the HDPS conceptualises the information stored within a database as proxies to key underlying confounders (or constructs). Some of these proxies are likely to be strongly correlated with the variables typically included in a PS analysis and others will glean information about patients that

would otherwise be unmeasured.

As an example, we consider capturing the concept of 'frailty' in a database (summarised in Figure 2.1). Frailty is often a key confounder in database studies, however, it can be difficult to accurately measure, even in controlled settings (e.g. clinical trials) (*Brookhart et al.*, 2010; *Schneeweiss et al.*, 2009). Whilst we may struggle to define this concept individually, frailty is likely to be strongly related to concepts we can capture, for example, prescriptions for oxygen canisters, referral for home support or a history of fractures. By adjusting for these surrogates we can attempt to adjust for 'frailty' by proxy. Our ability to successfully adjust for these concepts will depend on how closely related the surrogates are to the unobserved or imperfectly observed confounder (*Greenland*, 1980; *Schneeweiss et al.*, 2009).

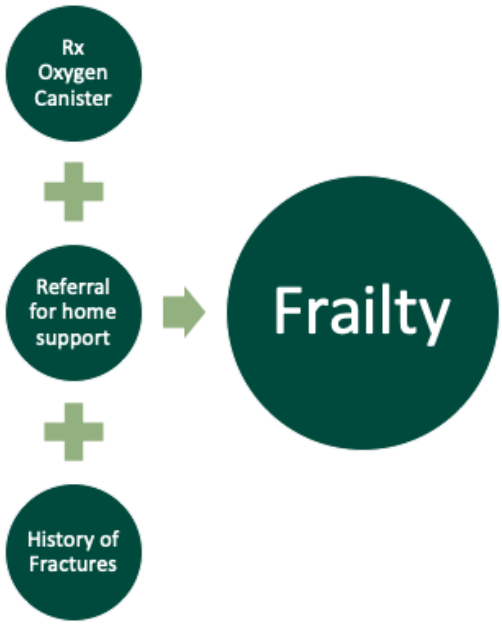


Figure 2.1: *Example of proxy adjustment.*

2.6.2 Description

The generic steps for implementing the HDPS algorithm are as follows (summarised in Figure 2.2) (*Rassen et al.*, 2011b; *Schneeweiss et al.*, 2009; *Wyss et al.*, 2018a).

- **Step 1:** Identify p data dimensions capturing specific aspects of care within the healthcare database. The data dimensions should contain pre-exposure features (often stored as codes) identified in a covariate assessment period (typically during the year prior to study entry), as illustrated in Figure 2.3. Investigator-identified (pre-defined) covariates, including demographics and specific conditions or concepts are also specified. Finally, investigators may choose to *a priori* exclude instruments and other features (dependent on study question) from consideration by the HDPS.

- **Step 2:** Within each of the p data dimensions, sort codes by their prevalence and retain the top n most common codes for the next steps (typically $n = 200$).

- **Step 3:** Assess how frequently each code is recorded per patient during the covariate assessment period. Three binary covariates are generated for each code indicating how often the code occurred: 1) Once: $\geq$ once, 2) Sporadic: $\geq$ median number of times, and 3) Frequent: $\geq$ upper quartile number of times.

- **Step 4:** Prioritise covariates. Steps 1-3 generates as many as $p \times n \times 3$ covariates. These are usually prioritised and ranked univariately, based on the Bross formula (*Bross*, 1966) or strength of association with the treatment (*Rassen et al.*, 2011b) (more details surrounding prioritisation are given in Chapter 3). Machine-learning methods have also been applied for prioritisation in the context of HDPS (*Schneeweiss*, 2018).

- **Step 5:** Select the top k covariates for inclusion in the HDPS model (often $k = 200$ or 500).

Finally, having selected a set of HDPS covariates, a propensity score model is fitted containing both the pre-defined and HDPS covariates. Propensity scores and treatment

effects can be estimated using methods described in Section 2.5.

Throughout this thesis, I refer to the p data dimensions, n most prevalent codes and k selected covariates as the key 'tuning parameters' for the HDPS procedure.

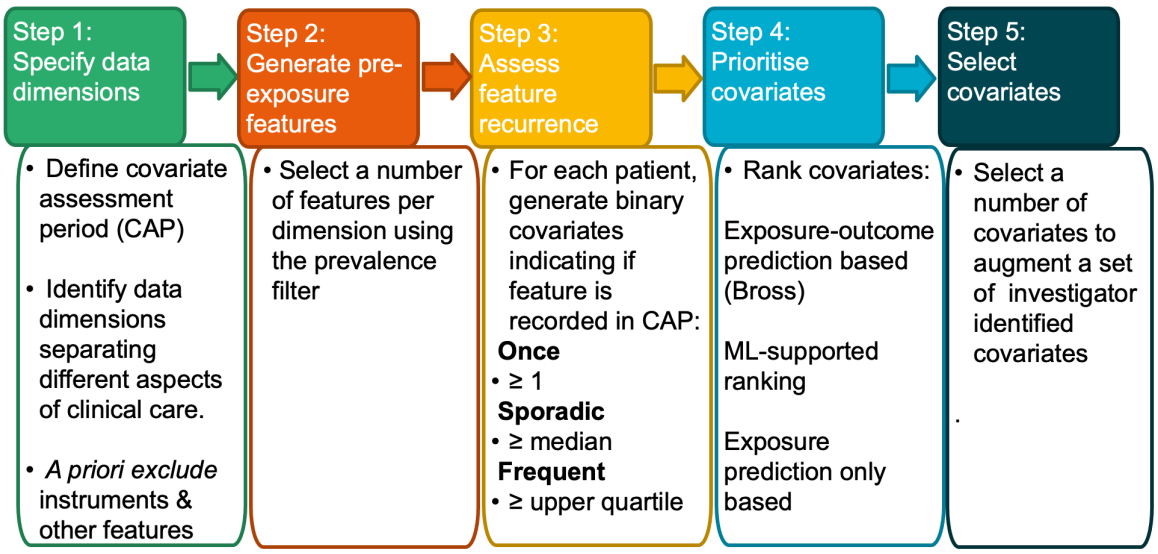


Figure 2.2: *Summary of HDPS algorithm steps.*

Abbreviations: *CAP, covariate assessment period; ML, Machine Learning*

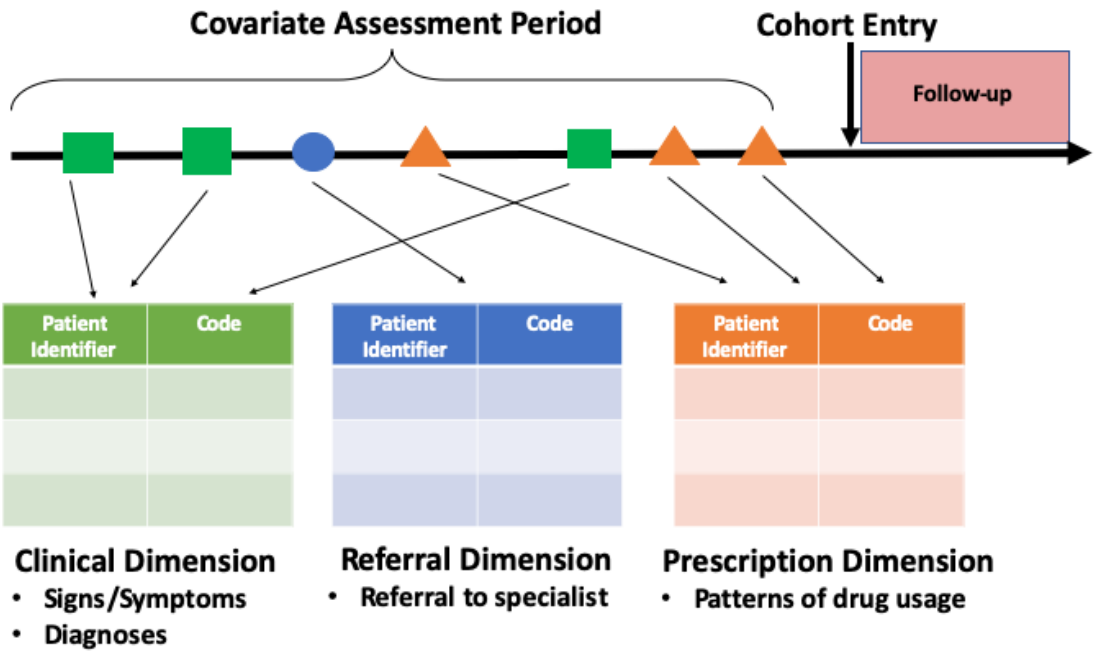


Figure 2.3: *Illustration of pre-exposure identification of features for HDPS.*

2.6.3 Properties

Assessing the performance and properties of proxy adjustment methods in large healthcare databases is a key challenge in pharmacoepidemiological research. Whilst the performance of statistical methods is often compared using fully simulated data based on realistic data-generating mechanisms (*Morris et al.*, 2019), the complexity of these databases make such an approach challenging (*Franklin et al.*, 2014).

Instead, the performance of these methods is typically assessed based on a combination of empirical and theoretical evidence, with the latter often utilising plasmode simulations (*Franklin et al.*, 2014). These simulations are based on an existing cohort from a large healthcare database and attempt to preserve the complex relationships between variables whilst enforcing a known causal effect (*Franklin et al.*, 2014).

Empirical performance

Empirical evidence surrounding the potential benefit of HDPS approaches for achieving improved confounder control has been highlighted in a diverse range of data sources, including databases in the US, Canada, Germany, UK and Denmark (*Schneeweiss*, 2018).

When empirically assessing the performance of the HDPS, results have often been benchmarked against gold-standard studies (such as randomised controlled trials) (*Schneeweiss*, 2018). One prominent example in the literature surrounds the relationship between non-steroidal anti-inflammatory drug and COX-2 inhibitor use on the risk of upper gastrointestinal bleeding (UGIB) (*Schneeweiss*, 2018). Evidence from trials suggests that COX-2 inhibitors reduce the risk of UGIB by between 10 and 25% (*Bombardier et al.*, 2000; *Silverstein et al.*, 2000). In non-interventional studies, a key issue surrounds successful capture of potentially subtle risk factors for GI complications; important in context of potential channelling of COX-2 inhibitors towards patients with higher risk of GI toxicity (*MacDonald et al.*, 2003; *Schneeweiss*, 2018). This has motivated the use of HDPS in this setting and it has been studied in a number of data sources (*Garbe et al.*, 2013; *Hallas and Pottegard*, 2017; *Schneeweiss et al.*, 2009; *Toh et al.*, 2011). In all studies the HDPS successfully obtained results similar to those from the randomised controlled trials, with the HDPS often improving on an analysis adjusting for only the set of investigator covariates (*Schneeweiss*, 2018).

Use of the HDPS has also successfully replicated results obtained from randomised trials investigating the use of oral antidiabetic drugs (glyburide versus glipizide) on the risk of hypoglycemic events (*Gangji et al.*, 2007). A study by *Zhou et al.* (2017) only replicated results from the trials when incorporating both the HDPS and investigator identified covariates. Furthermore, review of the included HDPS covariates identified several important confounders surrounding pregnancy and gestational diabetes which had previously not been specified the study team (*Zhou et al.*, 2017). This highlights the potential ability of HDPS to identify and adjust for key drivers of treatment decisions which might previously be omitted from the investigator set of covariates (*Schneeweiss*,

2018).

Finally, it is important to highlight that whilst the case studies in this thesis focus on application of HDPS in pharmacoepidemiological studies, the potential benefit of HDPS has also been shown in other areas, for example health services research (*Enders et al.*, 2018; *Polinski et al.*, 2012; *Schneeweiss*, 2018).

Theoretical properties and simulation studies

As described in Section 2.6.2, the steps of the HDPS procedure can be classified as either relating to data management (Steps 1 and 3) or analytical choices (Steps 2, 4 and 5). For steps requiring analytical choices, there is a growing literature investigating the properties and impact of investigator decisions on the robustness of results (*Schneeweiss*, 2018).

Step 2 selects the top n most prevalent codes from each of the data dimensions specified. In many studies investigators select $n = 200$, as proposed in the original application of the HDPS (*Schneeweiss et al.*, 2009). However, theoretical results studied by *Schuster et al.* (2015) highlight that the application of a prevalence filter can result in codes with a low marginal prevalence being discarded despite them potentially being highly influential in terms of successful confounder adjustment. In practice, the likely occurrence of variables with these prevalence properties is unclear and future research is needed to properly understand the practical consequences of these findings (*Schuster et al.*, 2015).

The Bross formula is the default method of prioritisation in the HDPS procedure (Step 4) (*Schneeweiss et al.*, 2009). This is a simple method relying on univariate associations, capturing the covariate-outcome and covariate-exposure relationships. However, the Bross formula ignores the non-independence of HDPS covariates, which are likely to be related in a complex way. Furthermore, despite empirical evidence surrounding the performance of the HDPS, it is important to highlight that the theoretical properties of the Bross formula do not guarantee this (*VanderWeele*, 2019). In particular, even if the set of candidate HDPS covariates is sufficient for successful confounder adjustment, the

resulting set of covariates (after prioritisation by the Bross formula) is not guaranteed to have this property (*VanderWeele*, 2019).

Despite the relative simplicity of the Bross formula, in practice, the HDPS often performs comparably to machine learning methods (*Karim et al.*, 2018; *Schneeweiss*, 2018; *Schneeweiss et al.*, 2017). One study by *Schneeweiss et al.* (2017) reanalysed results from five cohort studies and investigated a range of machine learning methods for prioritising HDPS covariates, including Lasso regression, ridge regression, bayesian logistic regression and principal component analysis. This study highlighted that Lasso regression can offer a promising alternative to the Bross formula for prioritising variables. In particular, having identified a set of HDPS covariates (Step 3), the approach modelled the HDPS covariates using Lasso regression to, firstly, prioritise them by their outcome relationship before then including the covariates whose coefficients were not shrunk to zero in a PS analysis (*Schneeweiss et al.*, 2017). This approach has also been found to perform well compared to the HDPS in simulations (*Franklin et al.*, 2015). Despite these findings, the Bross formula is often still preferred in practice, potentially due to being relatively easy to implement in any setting and often less computationally intensive compared to machine learning approaches.

Simulations by *Rassen et al.* (2011a) have highlighted that, in settings with few outcome events, prioritising covariates by the strength of confounder-exposure association can outperform prioritisation by the Bross formula.

The decision surrounding how many covariates to adjust for is a key tuning parameter in the HDPS procedure (Step 5), especially given studies highlighting that results are not always robust to this decision (*Patorno et al.*, 2014).

Early simulation studies by *Rassen et al.* (2011a) highlighted that adjusting for approximately 300 HDPS covariates was likely to be sufficient for successful confounder control in moderate to large samples. However, more recent work has investigated the use of machine learning approaches, such as the SuperLearner and collaborative targeted maximum likelihood estimation (CTMLE), to optimise the number of covariates chosen in a given setting (*Ju et al.*, 2019; *Schneeweiss*, 2018; *Wyss et al.*, 2018b). These simulation

studies highlight the potential for these methods to help avoid overfitting of the HDPS model and are likely to be most useful when sample sizes are small or exposures and outcomes are rare (*Ju et al.*, 2019; *Schneeweiss*, 2018; *Wyss et al.*, 2018b). However, a key drawback from combining these approaches with the HDPS is the potential for substantial additional computational burden (*Ju et al.*, 2019; *Schneeweiss*, 2018).

Finally, there have been methodological developments to the HDPS algorithm, relevant to settings beyond those presented in this thesis, which may give insights into the properties of HDPS approaches more generally. The first surrounds implementation of the HDPS in settings of time-varying treatment exposures with time-varying confounding via marginal structural models (*Neugebauer et al.*, 2015). The second surrounds generalisation of the HDPS to settings where there are more than 2 treatment levels via multinomial HDPS models (*Eberg et al.*, 2020).

2.6.4 Critique

Given the added complexities surrounding covariate generation, prioritisation and selection, the HDPS is not a straightforward extension to PS methodology (*Austin et al.*, 2020; *Schneeweiss et al.*, 2009). I outline some of the key methodological and practical issues in the following paragraphs, highlighting common criticisms of the HDPS approach.

Separation of design and analysis

The separation of design and analysis is considered an advantage of PS methodology, however, ranking by the Bross formula explicitly uses information on the outcome to prioritize covariates (*Austin et al.*, 2020; *Garbe et al.*, 2013; *Schneeweiss et al.*, 2009).

Historically, this feature of PS analysis was discussed in the context of settings where confounders are known and measured (*Rubin*, 2004). However, as highlighted previously, when using large healthcare databases this is far from certain.

Since the tuning parameters can be pre-specified, in theory, the HDPS is an automated and reproducible process with limited scope for cherry picking covariates. Furthermore, despite using outcome information, the treatment effect is still blinded; this maintains some separation by allowing investigators to build and assess covariate balance for a particular PS model before estimating treatment effects. Given the potential benefit for HDPS approaches to include otherwise omitted and important covariates, this deviation is usually accepted for pragmatic reasons in this context (*Schneeweiss*, 2018).

Importantly, the same criticism can also apply to machine learning approaches that might alternatively be used for covariate regularization or selection (many of which similarly incorporate outcome information) (*Franklin et al.*, 2015).

HDPS and 'principled' confounder selection

A key advantage of DAGs is that they allow investigators to identify a set of variables necessary to isolate direct effects of treatment (*Greenland et al.*, 1999; *Hernán and Robins*, 2020).

In large healthcare databases there are often a large number of covariates, making the construction of a complete causal diagram challenging (*VanderWeele*, 2019). Furthermore, the knowledge required to specify causal relationships between all possible covariates is usually unavailable (*VanderWeele*, 2019).

Instead, when using large healthcare databases, literature reviews and prior clinical knowledge are a common starting point and this knowledge may be encoded in a DAG (*Schneeweiss*, 2019). However, it can be difficult to pre-specify key (and potentially subtle) constructs or confounders necessary for successful confounder control, e.g. markers of healthcare utilisation, frailty or disease severity.

VanderWeele (2019) describes the following two approaches, relevant for summarising the principles of confounder selection in HDPS analyses:

- **Common cause approach:** investigators adjust for all pre-treatment variables

thought to be common causes of treatment and outcome.

- **Disjunctive cause approach:** investigators adjust for all pre-treatment variables which are a cause of the treatment, or the outcome or both.

Conceptually, the HDPS lies somewhere between these two approaches (*Schneeweiss*, 2019). Furthermore, when using the HDPS we do not fully understand the causal diagram; instead we aim to select confounders based on these described principles. It is important to acknowledge that this might inadvertently lead to adjustment for variables, such as instruments and colliders, that we would typically want to avoid adjusting for. However, in scenarios realistic of those typically observed in large healthcare databases, the improvement in confounder control often outweighs any bias induced through the inclusion of such variables (*Liu et al.*, 2012; *Myers et al.*, 2011).

'Black-box'

The semi-automated nature of the HDPS, which often results in adjustment for several hundred empirically-derived covariates, has led some to label the approach a black-box (*Rassen and Schneeweiss*, 2012). These concerns can be exacerbated when investigators under-report implementation details and information on the types of variables selected. Ultimately, this can make it difficult for readers to properly scrutinise HDPS analyses.

Conversely, the black-box nature of the HDPS has lead some to see it as a silver bullet and consequently the results from these approaches can be given undue prominence. An article by *Rafi and Greenland* (2020) highlights a recent example of this in the context of serotonergic antidepressant use during pregnancy and the risk of autism spectrum disorder in children (*Brown et al.*, 2017). In the study, the primary HDPS analysis was reported as the key finding despite remaining imbalances (post-HDPS adjustment) in a number of potentially important covariates and discrepancies in the results of sensitivity analyses (*Brown et al.*, 2017). Aside from the issues surrounding the statistical interpretation of results (discussed by (*Rafi and Greenland*, 2020)), the study highlights the need for HDPS approaches to be carefully applied and results

interpreted in the context of sensitivity analyses. As described in the seminal paper by *Schneeweiss et al.* (2009), the HDPS does not guarantee successful mitigation of confounding bias and should not be assumed to have superiority over an analysis based only on investigator-specified covariates.

Issues surrounding the transparency of the HDPS, including diagnostic tools and the reporting of these analyses, are considered in Chapter 4.

Chapter 3

Paper A: Implementing high-dimensional propensity score principles to improve confounder adjustment in UK electronic health records

John Tazare[1], Liam Smeeth[1,2], Stephen JW Evans[1], Elizabeth Williamson[1,2], Ian J Douglas[1,2]

1. London School of Hygiene and Tropical Medicine, London, UK.
2. Health Data Research (HDR) UK, London, UK.

3.1 Overview

Summary

The previous chapter reviewed relevant background information and introduced the high-dimensional propensity score for confounder adjustment in pharmacoepidemiological research. In this chapter, I outline the principles underlying each step of the HDPS and propose modifications for better characterising UK EHR data. I apply the HDPS to a recent study in the CPRD where the results obtained strongly suggested residual confounding between treatment groups. Initially, the work was presented as a poster at the *35th International Conference on Pharmacoepidemiology & Therapeutic Risk Management (2019)*. This paper was published in September 2020 in *Pharmacoepidemiology and Drug Safety*.

Thesis objectives addressed

This chapter addresses the following objectives of the overall thesis (Section 1.3):

2. Propose modifications for implementing the underlying principles of the HDPS in UK EHRs.
3. Apply the HDPS and proposed modifications in the context of UK EHRs.

Role of candidate

I conducted the statistical analysis (including developing code applying these methods) and drafted the paper, Liam Smeeth (LS) provided input on how to characterise GP recording practice in UK primary care. Ian Douglas (ID) provided access to the original study data. I created a procedure for mapping Read codes to ICD-10 codes with review and input from ID and Elizabeth Williamson (EW). The paper was finalised after suggestions, comments and guidance from LS, Stephen Evans, EW and ID.

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

London School of Hygiene & Tropical Medicine
Keppel Street, London WC1E 7HT

T: +44 (0)20 7299 4646
F: +44 (0)20 7299 4656
www.lshtm.ac.uk

RESEARCH PAPER COVER SHEET

Please note that a cover sheet must be completed for each research paper included within a thesis.

SECTION A – Student Details

Student ID Number	LSH1401926	**Title**	Mr
First Name(s)	John		
Surname/Family Name	Tazare		
Thesis Title	High-dimensional propensity scores for data-driven confounder adjustment in UK electronic health records		
Primary Supervisor	Elizabeth Williamson & Ian Douglas		

If the Research Paper has previously been published please complete Section B, if not please move to Section C.

SECTION B – Paper already published

Where was the work published?	Pharmacoepidemiology & Drug Safety		
When was the work published?	September 2020		
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion	N/A		
Have you retained the copyright for the work?*	**Yes**	Was the work subject to academic peer review?	**Yes**

*If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	-
Please list the paper's authors in the intended authorship order:	-
Stage of publication	Choose an item.

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	I planned and conducted the analysis, and drafted the article with feedback from co-authors.

SECTION E

Student Signature	John Tazare
Date	14/05/2021

Supervisor Signature	Ian Douglas
Date	14/05/2021

3.2 Abstract

Purpose

Recent evidence from US claims data suggests use of high-dimensional propensity score (HDPS) methods improve adjustment for confounding in non-randomised studies of interventions. However, it is unclear how best to apply HDPS principles outside their original setting, given important differences between claims data and electronic health records (EHRs). We aimed to implement the HDPS in the setting of United Kingdom (UK) EHRs.

Methods

We studied the interaction between clopidogrel and proton pump inhibitors (PPIs). While previous observational studies suggested an interaction (with reduced effect of clopidogrel), case-only, genetic and randomised trial approaches showed no interaction, strongly suggesting the original observational findings were subject to confounding. We derived a cohort of clopidogrel users from the UK Clinical Practice Research Datalink linked with the Myocardial Ischaemia National Audit Project. Analyses estimated the hazard ratio (HR) for myocardial infarction (MI) comparing PPI users with non-users using a Cox model adjusting for confounders. To reflect unique characteristics of UK EHRs, we varied the application of HDPS principles including the level of grouping within coding systems and adapting the assessment of code recurrence. Results were compared with traditional analyses.

Results

Twenty-four thousand four hundred and seventy-one patients took clopidogrel, of whom 9111 were prescribed a PPI. Traditional PS approaches obtained a HR for the association between PPI use and MI of 1.17 (95% CI: 1.00-1.35). Applying HDPS modifications

resulted in estimates closer to the expected null (HR 1.00; 95% CI: 0.78-1.28).

Conclusions

HDPS provided improved adjustment for confounding compared with other approaches, suggesting HDPS can be usefully applied in UK EHRs.

3.3 Introduction

Electronic Health Records (EHRs) are increasingly used for research investigating the effects of medications (*Council of European Union*, 2010; *US FDA*, 2011). Adequate adjustment for confounding remains a key issue and incorrect conclusions can be drawn amid concerns of residual or unmeasured confounding (*Douglas et al.*, 2012; *Freemantle et al.*, 2013).

Developed in US claims data to improve confounder adjustment, the high-dimensional propensity score (HDPS) approach treats information stored within healthcare databases as proxies for key underlying confounders (*Schneeweiss et al.*, 2009). Some proxies may be strongly correlated with variables typically included in a traditional propensity score (PS) analysis; others may represent information about patients that is otherwise unmeasured e.g. frailty (*Schneeweiss et al.*, 2009).

Despite application in various settings (including UK EHRs) (*Schneeweiss*, 2018; *Suissa et al.*, 2017a,b; *Toh et al.*, 2011), detailed guidance on how to apply the HDPS outside US claims data is lacking. Important differences between data sources mean that careful consideration is needed when implementing HDPS principles to ensure source-specific characteristics are handled appropriately.

We propose a series of modifications to the HDPS that aim to characterise key features of UK EHRs whilst adhering to the underlying principles (*Schneeweiss*, 2018; *Schneeweiss et al.*, 2009).

3.4 Propensity scores

The PS is the conditional probability of being treated given a set of observed covariates (*Austin*, 2011; *Jackson et al.*, 2017; *Williamson et al.*, 2012).

PSs model the treatment allocation process and therefore offer advantages over multivariable analysis in EHRs, since investigators are forced to consider indications for treatment use and can convert large amounts of confounder information into a single number (*Freemantle et al.*, 2013).

At a particular value of the PS, the distribution of observed covariates is balanced between treated and untreated individuals, allowing consistent estimation of treatment effects, assuming all confounders are included in the model (*Williamson and Forbes*, 2014).

3.5 Description of the HDPS approach and underlying principles

3.5.1 Preliminary steps

Demographics (d) and clinical factors believed to be important confounders (l) are forced into the PS model (*Schneeweiss et al.*, 2009). A baseline time-window for assessing patient confounder information is established (often 1 year before study entry date).

3.5.2 Identification of most relevant covariates

Relevant information in the database is separated into p dimensions (*Schneeweiss et al.*, 2009). The underlying principle is that each dimension should represent a different aspect of care relevant to the healthcare system under investigation (principle 1). For

example, in US claims data, it is typical to separate information pertaining to diagnoses, procedures and prescribing (*Schneeweiss et al.*, 2009).

Healthcare databases typically store information in the form of thousands of discrete codes which vary by database. To avoid sparsity, information is often grouped at a granularity level set by the investigator that captures related aspects of health status and care (principle 2). We illustrate this using an example from the International Classification of Diseases (ICD-10) (*World Health Organisation*, 2019). The ICD-10 coding system is hierarchical meaning that all information pertaining to one concept, for example type 2 diabetes mellitus (T2DM), begins with the same 3-character code (E11 for T2DM).

Code groups are ranked by prevalence and investigators pre-specify a number to be selected from each dimension (*Schneeweiss et al.*, 2009).

Code frequency is then assessed for each individual; measuring the recurrence of identified codes in the baseline time-window. This is summarised by three indicator variables:

- Once: Code is recorded $\geq$ once.
- Sporadic: Code is recorded $\geq$ the median
- Frequent: Code is recorded $\geq$ the 75th percentile

This classification assumes that frequency of recording relates to the importance of a code as a descriptor a patient's health status (principle 3).

3.5.3 Prioritisation

The steps so far generate a large pool of potential confounders. Attempting to include all of these variables in the PS model would often lead to concerns of overfitting therefore a variable selection step is necessary to ensure statistical stability.

The HDPS uses the Bross formula to prioritise covariates across dimensions by their potential to bias the treatment-outcome relationship (*Bross*, 1966; *Schneeweiss et al.*,

2009; *Wyss et al.*, 2018a). This has three components. Firstly, it takes the confounded apparent relative risk (ARR) for a particular binary covariate as a function of the relative risk (RR) in the absence of confounding by this covariate. Secondly, the imbalance in prevalence amongst the exposed (P_{C1}) and unexposed (P_{C0}) patients. Thirdly, the independent association between a confounder and the study outcome (RR_{CD}):

$$\text{ARR} = \text{RR} \times \text{bias}_\text{M}, \text{ where } \text{bias}_\text{M} = \frac{P_{C1}(RR_{CD} - 1) + 1}{P_{C0}(RR_{CD} - 1) + 1} \text{ for all } RR_{CD}$$

Each dimension is sorted in descending order by the magnitude of $|\log(\text{bias}_\text{M})|$. This bias term takes a larger value the greater the potential a covariate has to bias the relationship of interest. Therefore, the top k empirical covariates are included in the PS. Typically several hundred covariates are selected.

3.5.4 Estimation of the HDPS

The selected empirical covariates are added to the predefined variables before estimating the PS. Traditional PS methods are then used to estimate the treatment effect (*Williamson et al.*, 2012). The final principle is that after accounting for the top k empirically selected covariates, residual confounding effects are assumed to be negligible (principle 4).

3.6 Proposed implementation of HDPS principles to UK EHRs

In this section, issues surrounding the translation of HDPS principles to UK EHRs are discussed alongside our proposed modifications (summarised in Figure 3.1).

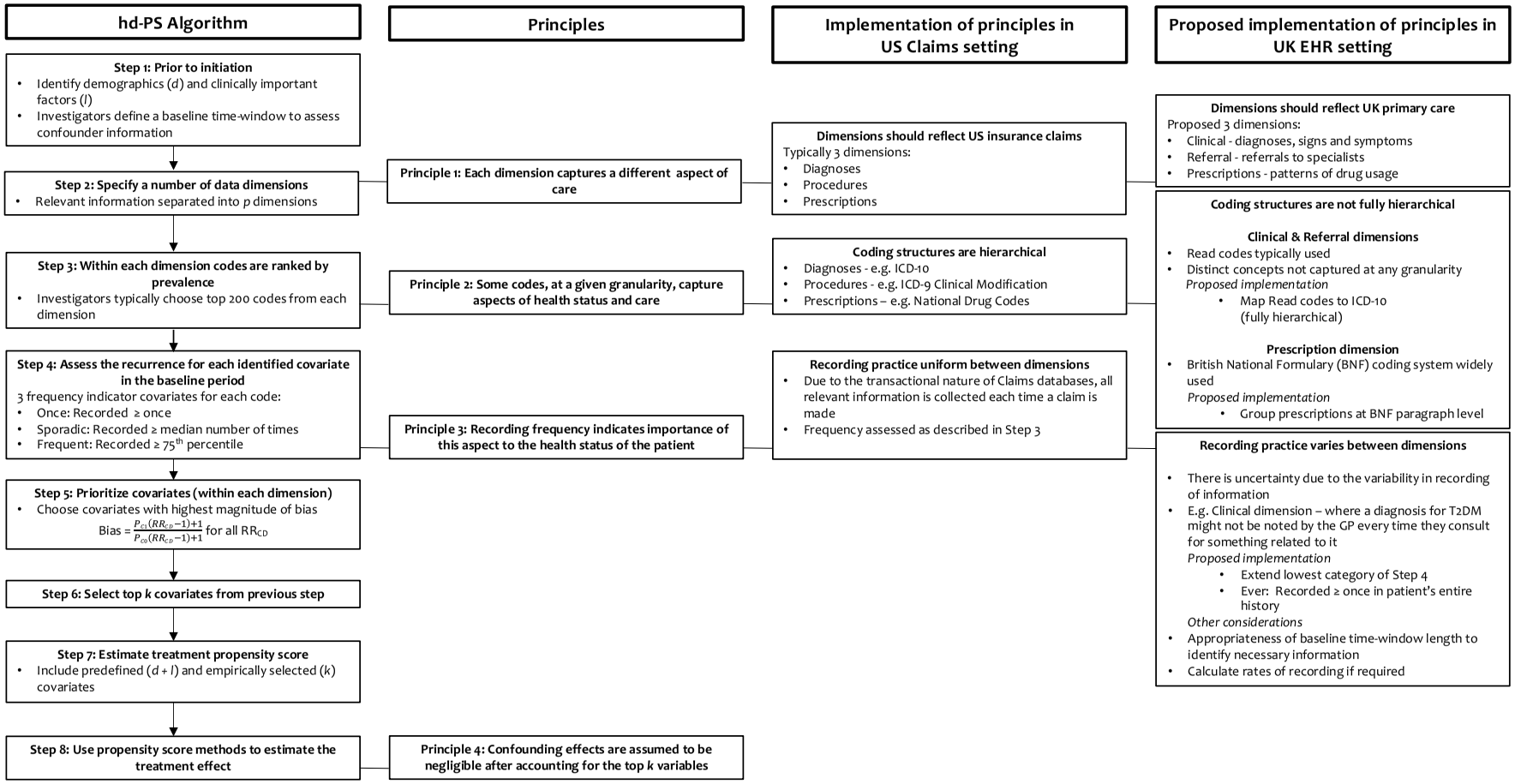


Figure 3.1: *Flowchart depicting HDPS steps, underlying principles and adaptations for translating to UK electronic health records.* ***Abbreviations:*** *HDPS, high-dimensional propensity score; ICD-10, International Classification of Disease; T2DM, Type 2 diabetes mellitus; GP, General Practitioner*

3.6.1 Principle 1: Identification of dimensions

There are important differences between insurance claims and EHR data in terms of data availability, structure and the reasons for data recording.17, 18 This necessitated the identification of clinically relevant dimensions based on patient contact with primary care services in the UK. Since previous applications of HDPS in UK EHRs have not reached a consensus about what these dimensions should be, we drew on general practitioner (GP) experience within our research team (*Azoulay et al.*, 2015; *Toh et al.*, 2011). We proposed three dimensions separating clinical, referral and prescription information (summarised in Table 3.1).

Table 3.1: *Summary of dimensions for UK electronic health records*

Dimension	Information included	Health status and care
Clinical	Diagnoses, signs and symptoms*	Indicates underlying health of patient and frequency of contact with healthcare system
Referral	Referrals to specialists	Indicates escalation in care or investigation
Prescriptions	Drug prescriptions issued in primary care	Frequency and patterns of drug usage

* The clinical dimension also contains information relating to administrative codes or references to measurements that occurred without results.

3.6.2 Principle 2: Code granularity

Data in the clinical and referral dimension are recorded using the Read code system (*Herrett et al.*, 2015). Read codes are less structured than coding systems used in claims databases (eg, ICD-10 (*World Health Organisation*, 2019)). Consequently, the Read

coding system does not fully capture distinct concepts at any level of granularity. For example, whilst the Read code 1434.00 relates to history of diabetes mellitus, grouping codes at the three-digit level (eg, 143) would capture concepts in addition to diabetes such as codes relating to thyroid disorder. Therefore, two codes with the same three-digit Read code may capture disparate clinical concepts, whereas conversely, two codes capturing similar concepts may have different three-digit Read codes.

A manual solution to group all Read codes at a level capturing distinct medical concepts is not practical, therefore we mapped Read codes to the ICD-10 coding system. This was achieved using cross maps developed by NHS Digital (*NHS Digital*, 2019a) and allowed replication of the approach taken by *Schneeweiss et al.* (2009) which hierarchically grouped distinct medical concepts at a certain granularity level.

For the prescription dimension the British National Formulary (BNF) coding system is used. We classified prescriptions at the BNF paragraph level which typically groups prescriptions by indication rather than mechanism of action (*NHS Digital*, 2019b).

3.6.3 Principle 3: Code recurrence

Code frequency is assessed by the HDPS to provide an indicator of a patient's underlying health (*Schneeweiss et al.*, 2009). In claims data all relevant information is recorded at each instance a claim is completed which leads to an intrinsic link between disease severity and code frequency.

EHRs exist for clinical record keeping which means that such a link is harder to discern since all relevant information will not necessarily be recorded at each consultation. Frequency of recording is instead likely to be a function of several factors including severity of illness, frequency of consultation and GP preference.

We classified the frequency of codes in a pre-specified baseline time-window, 1 year prior to study entry. Recognising the variability in recording we replaced the "Once" indicator with an "Ever" indicator which captured whether a code had been recorded during a patient's entire history. The remaining frequency indicators were assessed

during the baseline time-window.

We hypothesised that the degree to which information is recorded at each consultation was likely to vary by dimension, with more complete recording likely in the prescription and referral dimensions. However, in the clinical dimension relevant information is often not re-recorded at each consultation. For example, a patient receiving prescriptions relating to a diagnosis of T2DM will have this diagnosis recorded but not necessarily at each relevant consultation.

To investigate whether this information was likely to be overlooked when assessing information in a narrow time-window we extended the baseline time-window for the Clinical dimension. Acknowledging the fact that patients will have varying lengths of baseline information available we classified the frequency of codes by assessing rates instead of counts. We used three indicators to classify our revised frequency assessment (see Figure 3.1 for full definition).

3.6.4 Principle 4: Selected number of variables

The capacity of the HDPS to control for confounding can be sensitive to the number of covariates selected (*Garbe et al.*, 2013; *Wyss et al.*, 2018b). Whilst in claims data investigators typically specify 500 empirical covariates it is unclear if this is appropriate in UK EHRs. We investigated the impact of selecting 100, 250, 500 and 750 covariates.

3.7 Application to example in CPRD

3.7.1 Data

The Clinical Practice Research Datalink (CPRD) is a de-identified primary care database broadly representative of patients registered at GPs in the UK. It includes data pertaining to prescribing, diagnosis, referrals and some lifestyle factors for approximately 9% of the UK population (*Herrett et al.*, 2015).

A recent cohort study using the CPRD linked with the Myocardial Ischaemia National Audit Project (MINAP) investigated the combined use of proton pump inhibitors (PPI) with clopidogrel and aspirin. A possible interaction whereby PPIs may reduce the conversion of clopidogrel to its active metabolite had been suggested, raising concerns that combined use may lead to a reduction in clopidogrel effectiveness and an increased risk of vascular events. The cohort analysis found that combined use was indeed associated with an increased risk of myocardial infarction (MI) (*Douglas et al.*, 2012).

The pattern of associations found strongly suggested that residual confounding between patients may have explained the results as they were not specific to MI and were found for both strong and weak inhibitors of cytochrome P450 3A4 (the mechanism proposed for the drug interaction). Furthermore, a self-controlled case series (SCCS) analysis conducted on the same data found no evidence of increased risk (*Whitaker et al.*, 2005).

The authors concluded that the results from the cohort study reflect confounding in the cohort estimate. In addition, unconfounded studies based on genetic instrumental variable approaches using genetic effects on drug metabolism pathways also suggested no evidence of increased risk (*Holmes et al.*, 2011). A randomised double-blind trial has subsequently also suggested a lack of clinical effect of PPIs on MI risk, when used in combination with clopidogrel (HR = 0.92; 95% CI: 0.44-1.90) (*Bhatt et al.*, 2010).

3.7.2 Design

We summarise the original study design conducted by *Douglas et al.* (2012). Patients had to be present in the CPRD with at least 12 months of prior registration before first prescription for clopidogrel. Study entry was defined as the latest of first recorded clopidogrel prescription in combination with aspirin or 1 January 2003. Patients were censored at the earliest of stopping treatment for aspirin or clopidogrel, death, transferring out of the practice, last data collection date for the practice, 31 July 2009 or an occurrence of MI. Exposure was defined as any prescription for a PPI. We focus on the incident MI outcome which was ascertained using the MINAP database.

3.7.3 Statistical analysis

The original study analysed the hazard ratio (HR) for the association between PPI treatment and MI using Cox models, adjusting for 10 selected confounders. Missing data for body mass index, smoking and alcohol consumption were handled using missing categories. These conditions were applied consistently across all analyses.

We reanalysed the original data taking an intent-to-treat approach that classified patients according to original exposure status and incorporated baseline confounder information using PSs. We estimated the PS using multivariable logistic regression to model the relationship between treatment and potential confounders. Inverse probability of treatment weights (IPTW) were calculated from the PS which essentially constructs two synthetic samples representing the scenarios in which everyone had been treated and everyone had been untreated (*Austin*, 2011). A weighted Cox model incorporating the IPTWs was used to model the outcome.

Unless otherwise stated, all HDPS analyses defined the three aforementioned dimensions and assessed patient confounder information recorded in the year prior to cohort entry. The top 200 most prevalent codes were selected from each dimension and 500 covariates were included in the PS model.

We performed a standard HDPS analysis which implemented the algorithm using Read codes (classified at three-character Read code granularity) for the clinical and referral dimensions. All Read codes were included regardless of whether they map to ICD-10 to represent the default position of applying the method wholesale to the coded data in these dimensions. We then applied our modifications: mapping the clinical and referral dimensions to ICD-10 and extending the frequency assessment.

A sensitivity analysis extended the baseline time-window to 3 and 5 years for the Clinical dimension. We also investigated the impact of selecting 100, 250 and 750 covariates on confounding control.

All HR results are presented with 95% confidence intervals in parentheses. Analyses were conducted using Stata 14 (*StataCorp*, 2015).

3.8 Results

Demographics and clinical characteristics for the cohort study are summarised in Table 3.2. Twenty-four thousand four hundred and seventy-one patients took clopidogrel, of whom 9111 were prescribed a PPI. Of PPI users, 313 (3.4%) had an incident MI vs 421 (2.7%) in the non-users. Users of PPIs were older and were more likely to have had a history of cancer, diabetes or peripheral vascular disease compared to non-users (Table 3.2).

For the modified analyses, we mapped the clinical and referral dimensions from Read code to ICD-10. A large number of Read codes represent non-clinical information, for example, codes relating to administrative procedures. Since the aim of the mapping procedure is solely to capture clinically relevant information unmapped Read codes were expected. Upon inspection, the resulting unmapped codes could generally be categorised as either administrative information (eg, a letter), an indicator of a completed test without the result (eg, "blood pressure reading was taken") or coarse information we would typically include more granularly in the pre-defined covariates (eg, broad smoking terms). We include a sample of the most frequently occurring unmapped Read codes in the Supporting Information.

Results for all analyses are presented in Table 3. Using the confounders originally identified by *Douglas et al.* (2012) we obtained a HR for the association between PPI use and MI of 1.17 (1.00-1.35).

Applying our modifications reduced the HR for the association between PPI use and MI moving it towards a null result (Figure 3.2). The fully modified hd-PS obtained an HR of 1.00 (0.78 to 1.28).

In sensitivity analyses, extending the baseline time-window for the Clinical dimension lead to point estimates further from the null. Varying the number of covariates did not meaningfully alter point estimates. However, selecting fewer than 500 variables did improve the precision of effect estimates (Table 3.3).

Table 3.2: *Baseline characteristics by proton pump inhibitor status amongst clopidogrel and aspirin users.* ***Abbreviations****: PPI, proton pump inhibitor*

	Clopidogrel and aspirin users	
	No PPI	PPI
	N (%)	N (%)
Total	15360 (62.8)	9111 (37.2)
Median age (years)	68.9	71.1
Sex	N (%)	N (%)
Male	10007 (65.1)	5323 (58.4)
Body mass index (kg/m^2)		
<20	480 (3.1)	429 (4.7)
20-25	3987 (26.0)	2339 (25.7)
>25	10004 (65.1)	5809 (63.8)
Missing	889 (5.8)	534 (5.9)
Smoking status		
Non-smoker	4781 (31.1)	2780 (30.5)
Current	2760 (18.0)	1503 (16.5)
Ex-smoker	7777 (50.6)	4799 (52.7)
Missing	42 (0.3)	29 (0.3)
Alcohol status		
Non-drinker	1528 (9.9)	1080 (11.9)
Ex-drinker	938 (6.1)	687 (7.5)
Amount not specified	399 (2.6)	254 (2.8)
<2 units/day	3060 (19.9)	1908 (20.9)
3–6 units/day	7488 (48.8)	4106 (45.1)
>6 units/day	1180 (7.7)	606 (6.7)
Status unknown	767 (5.0)	470 (5.2)
History of:		
Diabetes	4404 (28.7)	3090 (33.9)
Peripheral vascular disease	1629 (10.6)	1095 (12.0)
Coronary heart disease	12198 (79.4)	7292 (80.0)
Ischaemic stroke	1571 (10.2)	954 (10.5)
Cancer	2038 (13.3)	1381 (15.2)

Table 3.3: *Estimated treatment effect of proton pump inhibitor use on myocardial infarction risk by variations in high-dimensional propensity score approach.* ***Abbreviations:*** *d, number of demographics; k, number of variables empirically selected by the algorithm; l, number of predefined covariates.*

Model	Dimension code granularity	Baseline assessment period	Most prevalent codes selected by dimension	Code frequency assessment	Covariates included in propensity score model	Total covariates in propensity score model	Outcome model HR (95% CI)	log(HR) SE
1	-	-	-	-	Unadjusted	-	1.23 (1.06 to 1.42)	0.08
2	-	-	-	-	Demographics + predefined*	d =2, l =8	1.17 (1.00 to 1.35)	0.10
3	3-digit Read† + BNF‡	1 year	200	Counts	+ Empirical covariates	d =2, l =8, k =500	1.07 (0.86to 1.34)	0.11
4	3-digit ICD-10§ + BNF	1 year	200	Counts	+ Empirical covariates	d =2, l =8, k =500	1.15 (0.91 to 1.45)	0.12
5	3-digit ICD-10 + BNF	1 year	200	Ever category + counts	+ Empirical covariates	d =2, l =8, k =500	1.00 (0.78 to 1.28)	0.13
6	3-digit ICD-10 + BNF	3 years	200	Ever category + counts + rates (clinical dimension)	+ Empirical covariates	d =2, l =8, k =500	1.12 (0.91 to 1.39)	0.11
7	3-digit ICD-10 + BNF	5 years	200	Ever category + counts + rates (clinical dimension)	+ Empirical covariates	d = 2, l = 8, k = 500	1.10 (0.90 to 1.36)	0.11
8	3-digit ICD-10 + BNF	1 year	200	Ever category + counts	+ Empirical covariates	d =2, l =8, k =100	1.07 (0.87 to 1.32)	0.10
9	3-digit ICD-10 + BNF	1 year	200	Ever category + counts	+ Empirical covariates	d =2, l =8, k =250	1.02 (0.81 to 1.27)	0.12
10	3-digit ICD-10 + BNF	1 year	200	Ever category + counts	+ Empirical covariates	d =2, l =8, k =750	1.03 (0.79 to 1.28)	0.13

* Demographics: age, sex; predefined covariates: smoking status, alcohol status, categorised body mass index, peripheral vascular disease, coronary heart disease, ischaemic stroke, cancer.

† Clinical terms are defined using Read codes in the Clinical Practice Research Datalink.

‡ British National Formulary (BNF) code at paragraph level.

§ International Classification of Disease (ICD-10).

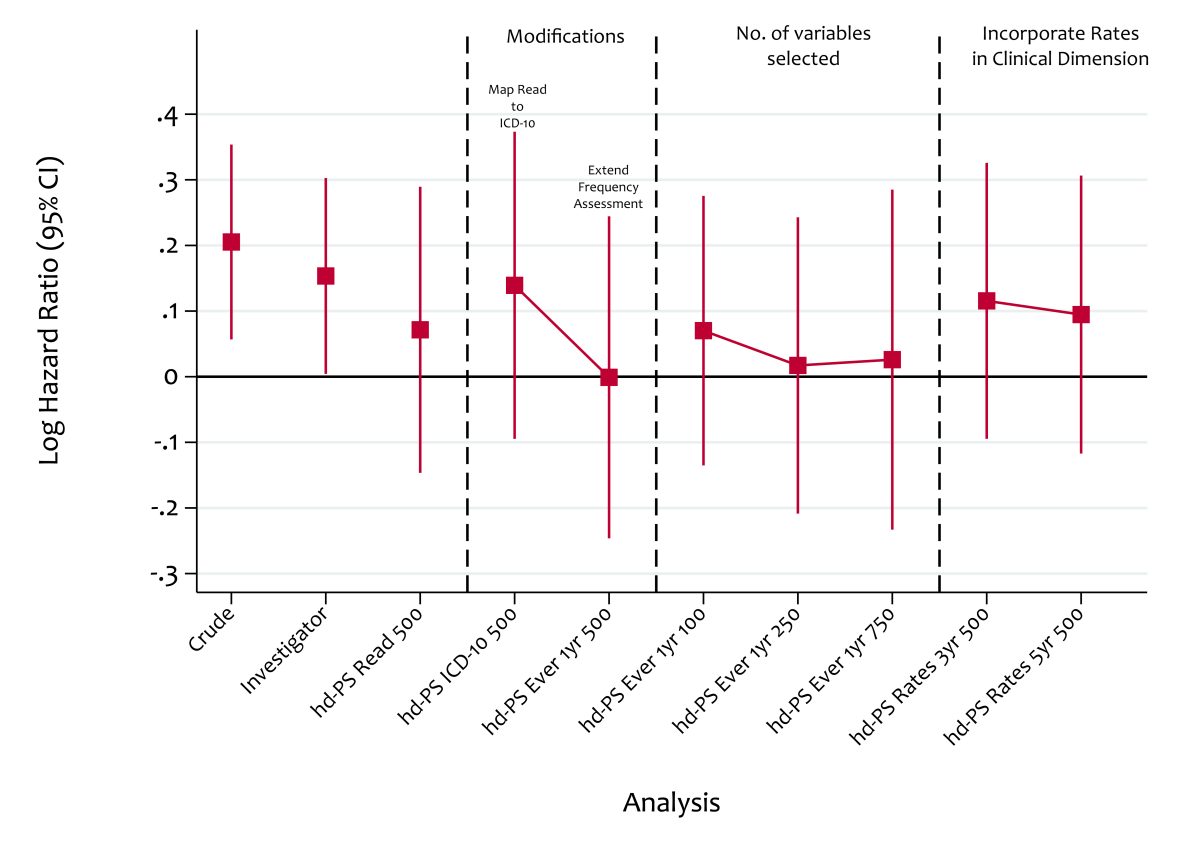


Figure 3.2: *Empirical performance of HDPS across our implemented adaptations.*

Abbreviations: *HDPS, high-dimensional propensity score; ICD-10, International Classification of Disease*

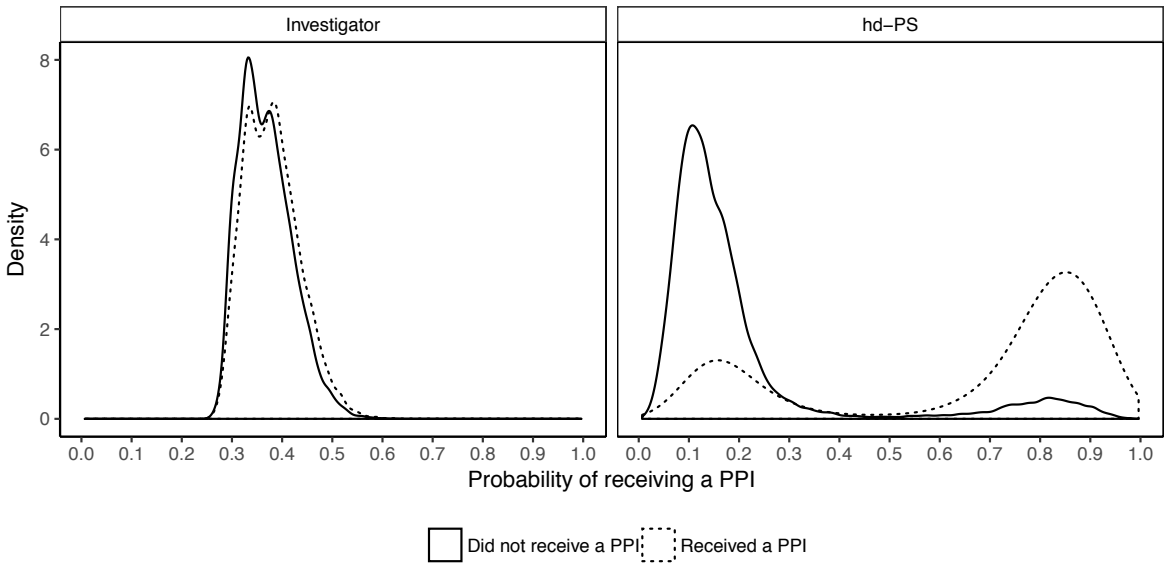


Figure 3.3: *Comparison of the estimated propensity score from investigator and HDPS approaches.* ***Abbreviations:*** *hd-PS, high-dimensional propensity score; PPI, proton pump inhibitor*

We investigated the estimated PS distributions by treatment group obtained from investigator led and HDPS analyses (Figure 3.3). These distributions compare the characteristics of patients in the populations under investigation. Compared to the investigator led approach, the HDPS exposed greater variation between the treatment groups and captured extra predictors of prescribing which were also causing confounding bias.

3.9 Discussion

In this study, we aimed to optimise the application of HDPS principles in UK EHR data. To investigate the potential of the HDPS to account for residual confounding we took a study where the authors were confident the result obtained was subject to strong between patient confounding. We aimed to get an improved point estimate, closer to the expected null result, with similar precision to the original study. After mapping Read to ICD-10 codes, changing the frequency assessment, selecting 500 variables for inclusion and having a 1 year assessment period for covariates, our final hd-PS model obtained an HR for the association between MI and PPI use of 1.00 (0.78-1.28), compared to 1.17 (1.00-1.35) using confounders selected using an investigator led approach. Our modifications therefore achieved results closer to those obtained by a randomised double-blind trial, although the precision does not rule out results obtained from other studies (*Bhatt et al.*, 2010; *Douglas et al.*, 2012). Sensitivity analyses suggested that extending the covariate assessment period for the Clinical dimension to 3 or 5 years might not be helpful in this setting.

The authors of the original study had suspected unmeasured frailty or comorbidity severity was different between PPI users and non-users. Here, we have demonstrated that differences between PPI users and non-users are more apparent when using HDPS than with traditional approaches. This highlights the potential for HDPS approaches to include proxies for influential but unmeasured information regarding a patient's underlying health status.

Our adaptations aimed to tailor the HDPS to UK EHRs and should be considered when

applying the HDPS in UK EHR data. The mapping of clinical and referral information to ICD-10 allows for the identification of homogeneous clinically meaningful proxies to be included in the HDPS, although we acknowledge that information contained in the unmapped codes is lost in this process. The inclusion of an Ever category to the frequency assessment of the HDPS also more accurately captures recording practice in EHRs. Selecting 500 variables for inclusion in the final HDPS model performed well, however selecting fewer variables obtained a very similar result with improved precision. The framework we have built could also be extended to include laboratory test results and free text information, the latter of which has been previously explored (*Schneeweiss*, 2018).

Whilst there have been several developments to the HDPS since its inception (*Schneeweiss*, 2018), there has been little exploration of how to translate the algorithm beyond claims data. Much of this development work for HDPS has been focussed on demonstrating it obtains known associations, such as the effect of non-steroidal anti-inflammatory drugs on the risk of gastrointestinal bleed (*Garbe et al.*, 2013; *Hallas and Pottegard*, 2017; *Schneeweiss et al.*, 2009; *Toh et al.*, 2011). However, these results have also been obtained through traditional methods of confounder adjustment. In the case study we present, a HDPS approach has removed a known confounded association discovered using traditional methods.

Future applications of the HDPS in this context will benefit from updates to the cross-map between Read and ICD-10. In the literature accompanying these cross-maps NHS Digital state that not every concept in one coding system can or should be represented in another (*NHS Digital*, 2019a). NHS Digital's intention was to map clinically meaningful terms only, and it was reassuring to observe that the majority of unmapped Read codes were clinically uninformative and would typically be discarded in an investigator analysis (see Supporting Information).

When calculating the SEs for treatment effects we have ignored variable selection or estimation of the PS. Theoretically, this is likely to result in narrower confidence intervals (*Greenland*, 2008), although the practical consequences are yet to be fully explored. We obtained a bias-corrected bootstrap 95% CI based on 1000 replications for our final

model of 0.70 to 1.30 (final model: HR = 1.00; 95% CI: 0.78-1.28).

Our results highlight the potential benefit of employing HDPS approaches in EHR studies, especially to overcome intractable confounding. However, the HDPS is not a panacea and we acknowledge that in studies where the confounding structure is relatively simple, the robustness of results is unlikely to differ between traditional and HDPS methods. We recognise the need for further exploration of the HDPS in this setting, via both controlled conditions and case studies. One outstanding issue surrounds the transparency of reporting when using HDPS approaches and there is a need for tools to better communicate proxies included in the final HDPS model.

This study has shown that the application of HDPS methods outside the context of claims data requires careful consideration of how to optimally apply HDPS principles. By adapting HDPS principles to the UK EHR setting we have demonstrated the potential for HDPS to improve confounder adjustment in EHRs.

3.10 Ethics statement

Scientific approval was obtained to use CPRD data by the Independent Scientific Advisory Committee (ISAC) (Protocol 17_194) and ethical approval from the London School of Hygiene & Tropical Medicine ethics committee (see Appendices A & B for details).

3.11 Supporting information

Summary of the unmapped Read codes from Read to ICD-10 cross-map procedure ranked by occurrence in Clinical & Referral files

The top 100 unmapped Read codes are displayed in Table 3.4 and are based on code recurrence in the CPRD GOLD Clinical and Referral files. The vast majority of unmapped Read codes are administrative codes or references to measurements that occurred without results.

Unmapped codes can either occur during the Read to SNOMED or SNOMED to ICD-10 stage of the mapping procedure. Whilst these 3 coding systems are designed for different purposes, they overlap in capturing clinically relevant information. Since the aim of the mapping is to capture clinical information, only a subset of Read codes are likely to translate to ICD-10. Furthermore, a key feature of ICD-10 is that it tends to focus on the presence of a disease, symptom or exposure rather than the absence.

It is important to highlight that some unmapped, but clinically meaningful, Read codes such as 1371.00 (Never smoked tobacco) and 1361.00 (Teetotaller) are already incorporated in the predefined covariates in a more granular form. Therefore, whilst these specific codes are not mapped, for reasons discussed in the previous paragraph, confounder information on smoking status and alcohol consumption is still adjusted for.

Table 3.4: *Top 100 unmapped Read codes from Read to ICD-10 cross-map procedure ranked by occurrence in Clinical & Referral files*

Rank	Read code	Read Code Description	Number of Clinical & Referral Events
1	246..00	O/E - blood pressure reading	57756204
2	22A..00	O/E - weight	26618496
3	9N31.00	Telephone encounter	16345242
4	229..00	O/E - height	12928844
5	9N36.00	Letter from specialist	11827848
6	6A...00	Patient reviewed	11787798
7	ZZZZZ00	_Converted code	11305996
8	1371.00	Never smoked tobacco	9837835
9	4K22.00	Cervical smear: negative	9065939
10	9N11.00	Seen in GP's surgery	8148080
11	8CB..00	Had a chat to patient	7747239
12	136..00	Alcohol consumption	7106240
13	9344.00	Notes summary on computer	6690411
14	9N42.00	Did not attend - no reason	5719630
15	679..11	Advice to patient - subject	5719099
16	138..00	Exercise grading	5605114
17	246..11	O/E - BP reading	5553348
18	61...00	Contraception	5197948
19	6781.00	Health education offered	5056122
20	9D1..00	MED3 - doctor's statement	4963513
21	8B3H.00	Medication requested	4917311
22	13A..00	Diet - patient initiated	4632141
23	9....00	Administration	4582678
24	9Z...00	Administration NOS	4296705
25	9N19.00	Seen in hospital casualty	4042024

Continued on next page

Rank	Read code	Read Code Description	Number of Clinical & Referral Events
26	8H...00	Referral for further care	3562720
27	9NDZ.00	Incoming mail NOS	3472821
28	4145.00	Blood sample ->Lab NOS	3346549
29	8CAL.00	Smoking cessation advice	3231319
30	81H..00	Dressing of wound	2998774
31	9N1C.11	Home visit	2846846
32	137..00	Tobacco consumption	2752703
33	9ND6.00	Communication from:	2596479
34	137L.00	Current non-smoker	2532705
35	93A..00	Discharge summary	2469843
36	9N4..00	Failed encounter	2434238
37	9N33.00	Letter encounter from patient	2420590
38	681..00	Screening - general	2360277
39	242..00	O/E - pulse rate	2355607
40	6791.00	Health ed. - smoking	2308064
41	9b04.00	Comment note	2179771
42	662..12	Hypertension monitoring	2101170
43	9N1p.00	Seen in orthopaedic clinic	2097740
44	663..11	Asthma monitoring	2095518
45	9N33.11	Letter encounter	2083833
46	8B31400	Medication review	1987681
47	9c0C.00	Result	1985824
48	8CA..00	Patient given advice	1985406
49	1361.00	Teetotaller	1901909
50	8C1B.00	Nursing care blood sample taken	1810315
51	2126.00	Patient's condition improved	1800310
52	7L17200	Blood withdrawal for testing	1775905

Continued on next page

Rank	Read code	Read Code Description	Number of Clinical & Referral Events
53	9ND..11	Incoming mail	1773142
54	0....00	Occupations	1749732
55	1151.00	No known allergies	1713811
56	9314.00	Lloyd George record received	1683381
57	2128.00	Patient's condition the same	1657396
58	Z4A..00	Discussion	1645977
59	7L17.00	Blood withdrawal	1631924
60	9N3D.00	Letter received	1626076
61	614D.00	Oral contraceptive prescribed	1618060
62	6896.00	Depression screening using questions	1616593
63	8HE..00	Discharged from hospital	1593309
64	9OL..00	Diabetes monitoring admin.	1563845
65	7305011	Syringe ear to remove wax	1556272
66	6637.00	Inhaler technique observed	1547508
67	66U..11	Hormone replacement therapy	1516641
68	9N3A.00	Telephone triage encounter	1504371
69	9N1K.00	Seen in ophthalmology clinic	1445181
70	66A..00	Diabetic monitoring	1443955
71	68R..00	New patient screen	1443089
72	66YJ.00	Asthma annual review	1438236
73	67E..00	Foreign travel advice	1414733
74	677B.00	Advice about treatment given	1385323
75	8B31100	Medication given	1371240
76	9N1A.00	Seen in hospital out-pat.	1354710
77	1226.00	No FH: Ischaemic heart disease	1341930
78	13l4.00	Main spoken language English	1336582
79	1362.12	Drinks occasionally	1201169

Continued on next page

Rank	Read code	Read Code Description	Number of Clinical & Referral Events
80	1....00	History / symptoms	1200084
81	9R8..00	Date records held from	1174167
82	22K..00	Body Mass Index	1172043
83	1225.11	No FH: CVA/Stroke/TIA	1170044
84	9D11.00	MED3 issued to patient	1168474
85	424..00	Full blood count - FBC	1150794
86	9OX6.00	Influenza vaccination invitation letter sent	1146923
87	6859.00	Ca cervix - screen done	1144825
88	662..00	Cardiac disease monitoring	1118590
89	2B6..00	O/E - visual acuity R-eye	1115920
90	2B7..00	O/E - visual acuity L-eye	1113622
91	66AS.00	Diabetic annual review	1096146
92	9OW..00	New patient screen admin.	1081374
93	62N..00	Antenatal examinations	1080387
94	9NZ..00	Patient encounter data NOS	1075948
95	9877.11	Injection given	1063498
96	65E..00	Influenza vaccination	1058065
97	9S10.00	White British	1045953
98	1992.00	Vomiting	1024773
99	9NC3.00	Letter sent to patient	1020591
100	9N35.00	Letter encounter to patient	1014891

Chapter 4

Paper B: Transparency of high-dimensional propensity score analyses: guidance for diagnostics and reporting

John Tazare[1], Richard Wyss[2], Jessica M Franklin[2], Liam Smeeth[1,3], Stephen J W Evans[1], Shirley V Wang[2], Sebastian Schneeweiss[2], Ian J Douglas[1,3], Joshua J Gagne[2], Elizabeth Williamson[1,3]

1. London School of Hygiene and Tropical Medicine, London, UK.
2. Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA.
3. Health Data Research (HDR) UK, London, UK.

4.1 Overview

Summary

Chapter 3 highlighted the need for careful consideration when applying the HDPS to UK EHRs. The HDPS has grown in popularity across a number of settings and several hundred covariates are often included to augment an investigator-identified set of confounders. However, in the literature, applications of HDPS often fail to perform or report diagnostic checks and sensitivity analyses surrounding the covariates selected. Furthermore, despite the semi-automated nature of the approach and the need for investigators to specify tuning parameters (e.g. the number of covariates selected), reporting of HDPS analyses is inconsistent. In this chapter, I present diagnostic tools, sensitivity analyses and reporting considerations for improving the transparency of HDPS analyses. The work was initially presented as an oral presentation at the *ICPE All Access 2020* online conference. This paper has been submitted to *Pharmacoepidemiology and Drug Safety* and is currently under review.

Thesis objective addressed

This chapter addresses the following objective of the overall thesis (Section 1.3):

4. Provide guidance surrounding diagnostic tools and reporting of HDPS analyses.

Role of candidate

I reviewed existing propensity score diagnostic tools and considered their relevance in the context of HDPS analyses. Furthermore, I developed and extended diagnostic tools, especially surrounding the presentation of covariates in HDPS models and assessmenet of covariate balance. I conducted the statistical analysis and implemented the diagnostic tools and sensitivity analyses. I drafted the reporting considerations and paper draft.

After initial input from my supervisory team, I further developed these ideas under the supervision of Joshua Gagne and Sebastian Schneeweiss during a research visit at The Division of Pharmacoepidemiology and Pharmacoeconomics at Brigham and Women's Hospital Department of Medicine and Harvard Medical School. The paper was finalised after suggestions, comments and guidance from all co-authors.

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

London School of Hygiene & Tropical Medicine
Keppel Street, London WC1E 7HT

T: +44 (0)20 7299 4646
F: +44 (0)20 7299 4656
www.lshtm.ac.uk

RESEARCH PAPER COVER SHEET

Please note that a cover sheet must be completed for each research paper included within a thesis.

SECTION A – Student Details

Student ID Number	LSH1401926	**Title**	Mr
First Name(s)	John		
Surname/Family Name	Tazare		
Thesis Title	High-dimensional propensity scores for data-driven confounder adjustment in UK electronic health records		
Primary Supervisor	Elizabeth Williamson & Ian Douglas		

If the Research Paper has previously been published please complete Section B, if not please move to Section C.

SECTION B – Paper already published

Where was the work published?	-		
When was the work published?	-		
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion	-		
Have you retained the copyright for the work?*	Choose an item.	Was the work subject to academic peer review?	Choose an item.

*If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	Pharmacoepidemiology & Drug Safety
Please list the paper's authors in the intended authorship order:	John Tazare, Richard Wyss, Jessica M Franklin, Liam Smeeth, Stephen JW Evans, Shirley V Wang, Sebastian Schneeweiss, Ian J Douglas, Joshua J Gagne, Elizabeth Williamson
Stage of publication	**Submitted**

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	I planned and conducted the analysis, and drafted the article with feedback from co-authors.

SECTION E

Student Signature	John Tazare
Date	14/05/2021

Supervisor Signature	Elizabeth Williamson
Date	14/05/2021

4.2 Abstract

Purpose

The high-dimensional propensity score (HDPS) is a semi-automated procedure for confounder identification, prioritisation, and adjustment in large healthcare databases that requires investigators to specify data dimensions, prioritisation strategy, and tuning parameters. In practice, reporting of these decisions is inconsistent and this can undermine the transparency, and reproducibility of results obtained. We illustrate reporting tools, graphical displays, and sensitivity analyses to increase transparency and facilitate evaluation of the robustness of analyses involving HDPS.

Methods

Using a study from the UK Clinical Practice Research Datalink that implemented HDPS we demonstrate the application of the proposed recommendations.

Results

We identify 7 considerations surrounding the implementation of HDPS, such as the identification of data dimensions, method for code prioritisation and number of variables selected. Graphical diagnostic tools include assessing the balance of key confounders before and after adjusting for empirically-selected HDPS covariates and the identification of potentially influential covariates. Sensitivity analyses include varying the number of covariates selected and assessing the impact of covariates behaving empirically as instrumental variables. In our example, results were robust to both the number of covariates selected and the inclusion of potentially influential covariates. Furthermore, our HDPS models achieved good balance in key confounders.

Conclusions

The data-adaptive approach of HDPS and the resulting benefits have led to its popularity as a method for confounder adjustment in pharmacoepidemiological studies. Reporting of HDPS analyses in practice may be improved by the considerations and tools proposed here to increase the transparency and reproducibility of study results.

4.3 Introduction

Bias arising from confounding is a key concern for pharmacoepidemiological studies and its mitigation depends on the ability to identify, measure and adjust for underlying differences between patients receiving different therapies (*Brookhart et al.*, 2010). Successful adjustment for confounding often hinges on capturing hard to measure concepts, such as markers of frailty, disease severity, or health-seeking behaviour.

The high-dimensional propensity score (HDPS) algorithm is a method for variable identification, prioritisation, and adjustment tailored for large healthcare databases (*Schneeweiss et al.*, 2009; *Wyss et al.*, 2018a). The HDPS conceptualises information in these databases as proxies to key underlying constructs; some are likely to be strongly correlated with other measured variables, but others act as proxies for constructs that would otherwise be unmeasured. The procedure treats these features as additional covariates for adjustment with the aim of optimising confounding capture and control.

Whilst the HDPS often incorporates several hundred additional covariates, the types of features included is rarely communicated leading some to label the HDPS a 'black-box' approach. Diagnostic tools can offer important insights into the properties of these features, enhancing our knowledge of the factors driving treatment decisions and checking for possible errors, e.g., relating to linkage error or exclusion criteria.

Despite studies highlighting the potential lack of robustness to investigator decisions (e.g., the number of covariates chosen (*Patorno et al.*, 2014; *Rassen et al.*, 2011a)), reporting of sensitivity analyses remains inconsistent and this can undermine the trans-

parency and reproducibility of HDPS analyses. Recent guidelines surrounding the reporting of pharmacoepidemiological studies state that "high dimensional proxy adjustment" methods should be reported in full; guidance is needed about exactly what this entails (*Langan et al.*, 2018).

Building on existing propensity score (PS) literature we describe and illustrate diagnostic tools and sensitivity analyses for HDPS analyses. We also provide considerations for reporting relevant information.

4.4 High-dimensional propensity scores

The generic five steps of the HDPS procedure are as follows (*Schneeweiss et al.*, 2009):

- Step one, investigators specify the data structure. This can involve declaring data dimensions capturing different aspects of care in the database under investigation.
- Step two, pre-exposure features are generated, and a prevalence filter is typically applied (often selecting the top 200 most common features from each dimension). Features are usually in the form of codes or free-text information and grouped at a specific granularity level. For example, codes might be truncated to the first three digits if they are International Classification of Diseases, 10th edition (ICD-10) codes.
- Step three, the recurrence of features is assessed in a pre-exposure period, creating binary covariates based on a set of frequency-based cut-offs (*Schneeweiss et al.*, 2009).
- Step four, the large pool of covariates generated in the previous step are prioritised. This is typically achieved using the Bross formula, which uses univariate associations of covariates with treatment and outcome, to identify those with the highest potential to bias the treatment-outcome relationship (*Schneeweiss et al.*, 2009; *Wyss et al.*, 2018a).

- Step five, a number of covariates (typically the top 200 to 500 from the covariate prioritisation (*Schneeweiss*, 2018; *Schneeweiss et al.*, 2009)) are selected to augment a set of pre-defined variables (selected by the investigators based on background knowledge) used for estimation of the PS model. Standard PS methods (e.g., matching or weighting (*Austin*, 2011; *Williamson and Forbes*, 2014)) are used to estimate treatment effects based on both sets of covariates.

4.5 Considerations for reporting

We initially conducted a literature search surrounding PS diagnostics and reporting guidance, identifying important gaps in the current literature surrounding the reporting of HDPS models. Utilising the extensive experience and knowledge of HDPS analyses within the research team, we present considerations for reporting features of the HDPS procedure (summarised in Table 4.1).

Item 1: Specify data dimensions

Data dimensions identified should be summarised, indicating which aspects of care they capture and possibly note data quality and completeness metrics.

Item 2: Describe parameters for generating pre-exposure features

Investigators should describe how features are generated, e.g. specifying the code granularity for a particular coding system (e.g., 3-digit ICD-10) or how free-text information has been processed (*Rassen et al.*, 2013).

Ongoing debate in the literature surrounds the use of marginal prevalence for prioritising features in Step 2 of the HDPS procedure (*Schuster et al.*, 2015). The main concern is the possible omission of influential features where despite a low marginal prevalence there exists strong imbalances within exposure group. Investigators should indicate

whether the prevalence filter is used and if so, state the number of features selected per dimension.

Item 3: Describe feature recurrence assessment

Whilst feature recurrence is typically assessed using the cut-offs outlined by Schneeweiss et al, deviations from these cut-offs exist and should be described in full (*Schneeweiss*, 2018; *Tazare et al.*, 2020). One example suggests explicitly considering the proximity to exposure start (*Schneeweiss*, 2018).

Item 4: Specify covariate prioritisation method

Investigators should describe the method of covariate prioritisation used. Whilst ranking is typically based on the Bross formula, exposure-based ranking (prioritising covariates based on the confounder-exposure association) has been employed in settings with few outcome events (*Rassen et al.*, 2011a; *Schneeweiss et al.*, 2009).

Recent evidence indicates the potential for machine-learning methods to enhance the performance of HDPS, both for covariate prioritisation or by reducing the set of covariates prioritised by the Bross formula (*Karim et al.*, 2017; *Schneeweiss et al.*, 2017; *Wyss et al.*, 2018b).

Item 5: Specify total number of covariates to select

The number of HDPS covariates selected for inclusion in the PS model should be reported. Machine learning-based approaches to determine the number of codes selected should be described in full (*Patorno et al.*, 2014; *Rassen et al.*, 2011a; *Wyss et al.*, 2018b).

Item 6: Specify software

Investigators should describe which software was used to implement the HDPS. There are commonly used packages available in R (*Lendle*, 2017), SAS (*Rassen et al.*, 2020), or Aetion.

Item 7: Describe the results of diagnostics

Subsequent sections describe and discuss the interpretation of relevant diagnostic tools and sensitivity analyses that should be routinely conducted and reported.

4.6 Data for illustration

4.6.1 Background

We use a cohort study from the United Kingdom (UK) Clinical Practice Research Datalink (CPRD) linked with the Myocardial Ischaemia National Audit Project (MINAP) (*Douglas et al.*, 2012). The CPRD is a database capturing information pertaining to contacts with primary care services (including clinical diagnoses, referrals and prescriptions) and is broadly representative of patients registered at general practitioners in the UK (*Herrett et al.*, 2015).

The study investigated whether a pharmacokinetic interaction between clopidogrel and use of proton pump inhibitors (PPI) could reduce clopidogrel effectiveness, leading to increased risk of vascular events. Results indicated an increased risk of MI associated with PPI use which was hypothesised to be largely due to residual confounding between treatment groups (*Douglas et al.*, 2012).

A reanalysis using the HDPS obtained results much closer to the hypothesised null association (*Bhatt et al.*, 2010; *Herrett et al.*, 2015; *Holmes et al.*, 2011), suggesting an improved ability to account for between-patient characteristics that were important for

Table 4.1: *Reporting considerations for key features and decisions of the high-dimensional propensity score approach.* ***Abbreviations:*** *ML, machine-learning.*

Item	Description	Aspect(s) to report
1	Specify data dimensions	• Dimensions identified and which aspect of the healthcare system they characterise
2	Describe parameters for generating pre-exposure features	• Describe how features are generated • Number of codes selected per dimension in prevalence filter
3	Describe feature recurrence assessment	• Whether and how recurrence was considered • Whether and how proximity to exposure start was considered
4	Specify covariate prioritisation method	Ranking based on: • Exposure-outcome prediction based (Bross) • ML-supported exposure-outcome prediction • Exposure prediction only
5	Specify total number of covariates to select	• Number of covariates selected • Justification for number of codes selected, e.g. use of simulation-based approaches.
6	Specify software	• Describe which software package was used to implement the HDPS procedure
7	Describe the results of diagnostics and sensitivity analyses	• Describe diagnostic tools used and highlight key insights gained • Describe the results of sensitivity analyses and discuss the possible implications for interpreting the findings from the primary analysis

confounding control (*Tazare et al.*, 2020).

4.6.2 Summary of HDPS analysis

We defined three dimensions assessing clinical, referral, and therapy information in the year prior to cohort entry. We applied a prevalence filter selecting the top 200 features from each dimension and adjusted for the top 500 HDPS covariates (prioritised by the Bross formula) (*Tazare et al.*, 2020).

The PS was estimated using multivariable logistic regression including both pre-defined and HDPS covariates. Hazard ratios (HR) for the treatment effect were obtained using Cox regression weighted by inverse probability of treatment weights.

Table 4.2 summarises the results, including a sensitivity analysis varying the number of HDPS covariates selected.

Analyses were conducted using Stata 15 and R (*R Core Team*, 2020; *StataCorp*, 2017). Code reproducing the figures presented is available at `www.github.com/johntaz/HDPS-Diagnostics` and in the Supporting Information.

Table 4.2: *Summary of Clinical Research Practice Datalink study, investigating the association between proton-pump inhibitor use and myocardial infarction, used for illustration.* ***Abbreviations:*** *HDPS, high-dimensional propensity score; BMI, body mass index; PVD, peripheral vascular disease; CHD, Coronary heart disease.*

Analysis	Number of covariates	Hazard ratio (95% CI)
Crude	0	1.23 (1.06 to 1.42)
Pre-defined only*	10	1.17 (1.00 to 1.35)
Primary HDPS	10 + 500	1.00 (0.78 to 1.28)
Sensitivity	10 + 100	1.07 (0.87 to 1.32)
Analyses	10 + 250	1.02 (0.81 to 1.27)
	10 + 750	1.03 (0.79 to 1.28)

* Pre-defined covariates: age, sex, smoking status, alcohol status categorised BMI, alcohol status, history of PVD, CHD, stroke, cancer.

4.7 Diagnostic & visualisation tools

In this section we illustrate and discuss novel and established PS diagnostics for assessing HDPS models (summarised in Table 4.3).

Table 4.3: *Summary of established and proposed diagnostic tools for high-dimensional propensity score models.* ***Abbreviations:*** *HDPS, high-dimensional propensity score.*

Diagnostic description	Section discussed	Conventional propensity score	HDPS
Propensity score distribution by treatment group	5.2	✓	✓
Prevalence of selected covariates by treatment group	5.3	-	✓
Absolute standardised differences	5.3	✓	✓
Bross-derived prioritisation distribution	5.4	-	✓
Relationship between confounder-exposure and confounder-outcome associations	5.4	-	✓

4.7.1 Model summaries

We recommend simple descriptions for communicating the covariates included in HDPS models, e.g., highlighting the proportion of selected codes that came from each data dimension. Investigators may also summarise high-level clinical concepts captured by the covariates included in the HDPS. Our study categorised codes using British National Formulary (BNF) paragraph level (prescription dimension) and ICD-10 (clinical and referral dimensions). We exploited the hierarchy of these coding systems to investigate codes aggregated by the chapter level. Figure 4.1 shows that in the clinical and referral dimensions, the majority of covariates selected corresponded to codes relating to symptoms, signs and abnormal findings. Additionally, covariates derived from the therapy dimension corresponded most to prescriptions from the cardiovascular system or nutrition and blood BNF chapters.

4.7.2 Comparison of PS distributions

Inspecting the distributions of the estimated PS by treatment group is a common diagnostic highlighting the ability of covariates included in the PS model to predict treatment received in the population being studied.

Whilst this is recommended when applying the HDPS, it is additionally informative to compare the PS distributions before and after inclusion of the HDPS covariates. This requires estimating the PS under models including a) only the pre-defined covariates and b) the pre-defined and selected HDPS covariates. Figure 4.2 compares the estimated PS distributions under these models.

When including only the pre-defined covariates, the estimated PS distributions appear similar between the treatment groups (Figure 4.2). However, when adding the HDPS covariates we observe greater separation of the distributions (Figure 4.2). In this example, the HDPS captured extra predictors of treatment initiation, highlighting important between-patient differences not apparent when only including the pre-defined covariates.

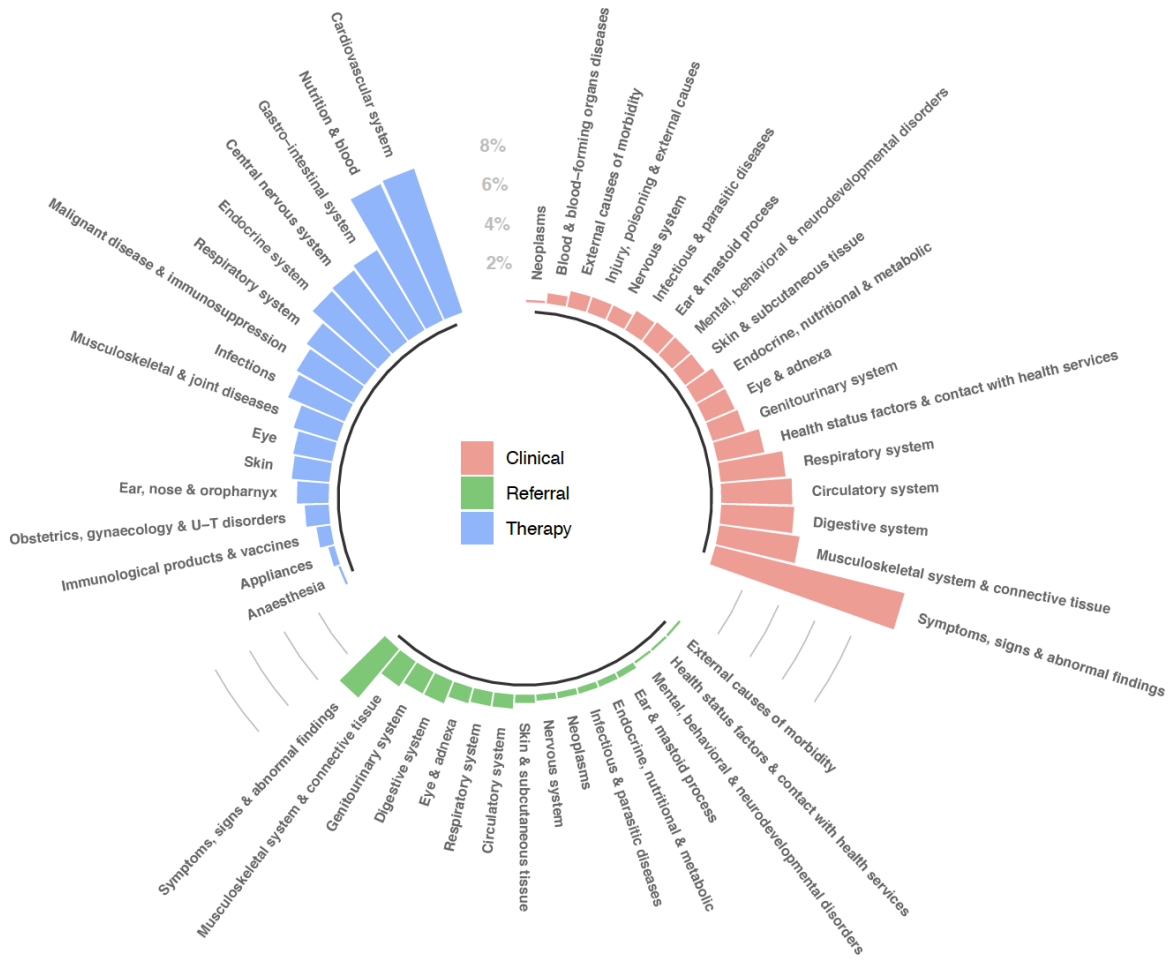


Figure 4.1: *Summary of high-level concepts captured in the top 750 bross-prioritised HDPS pre-exposure covariates separated and colour-coded by data dimension.*

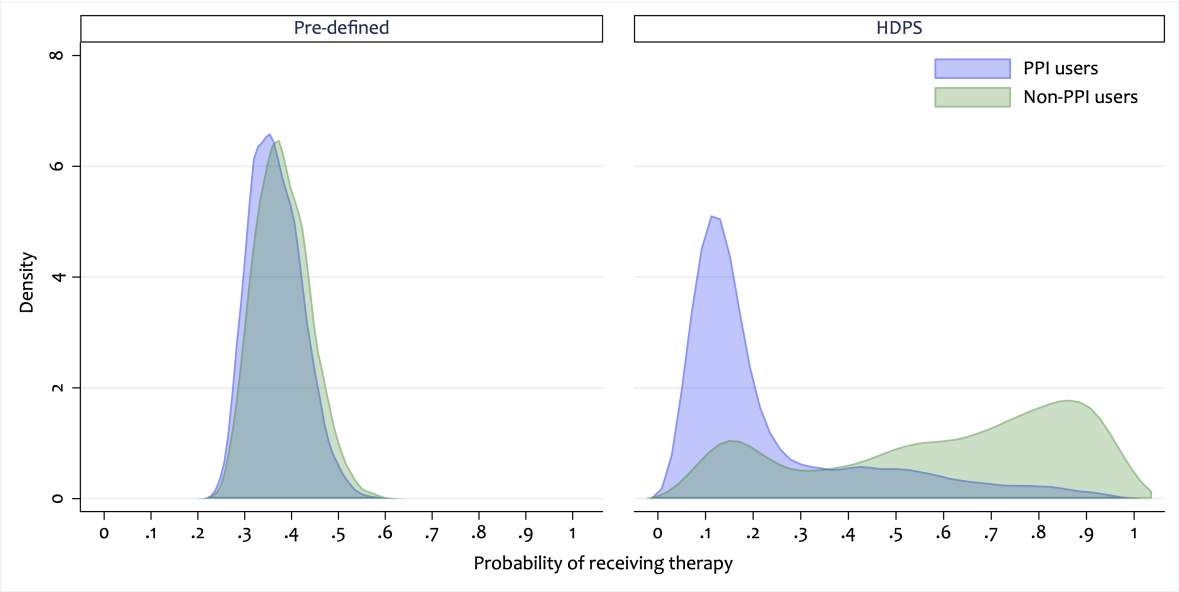


Figure 4.2: *Overlap plot comparing the propensity score distributions including only 10 pre-defined pre-exposure covariates and additionally including the 500 top-ranked HDPS covariates.*

4.7.3 Covariate balance

To investigate the overall balance of HDPS covariates we can plot the prevalence of selected covariates between the two treatment groups (shown in Figure 4.3) (*Franklin et al.*, 2015). Figure 4.3 highlights that for most covariates there is a similar prevalence in both groups, with slightly higher prevalence amongst the PPI users. There are several covariates from the prescription dimension (Figure 4.3, prevalence ratio > 2.0) with moderate to high prevalence amongst PPI users and a low prevalence amongst the non-users.

Measures of covariate balance (e.g., absolute standardised differences) are commonly used when assessing PS models to check for imbalances. In the HDPS setting, investigators should check the balance in the HDPS covariates before and after adjustment. Figure 4.4 indicates some covariates with large imbalances (substantially $> 10\%$) in the unweighted population but all achieve good balance in the HDPS weighted population.

There is a concern that adjusting for many additional HDPS confounders can make achieving balance in pre-defined confounders more difficult, as the PS model tries to simultaneously balance many more variables. If the HDPS variables are weak confounders or even not true confounders, addition of these variables can result in unnecessarily increased bias and variance (*Brookhart et al.*, 2006; *Myers et al.*, 2011). Achieving balance is more important in strong confounders compared to weak confounders (*Ho et al.*, 2007). Therefore, we recommend assessing the balance on selected key confounders before and after inclusion of all selected HDPS covariates (*Austin et al.*, 2020).

For illustrative purposes, we assume that all pre-defined covariates are important confounders and Figure 4.5 presents the balance of these covariates under models additionally including 250, 500 and 750 HDPS covariates. We observe that even after adjusting for 750 HDPS covariates, we achieve good balance in the pre-defined covariates, indicating the suitability of any of these models for preserving balance in the pre-defined covariates.

Another approach investigates the covariate balance in both the pre-defined and a set of

key HDPS confounders (Figure 4.6); we additionally assume all key HDPS confounders are in the top 250. Figure 4.6 highlights that in the pre-defined weighted population, a number of the top-ranked HDPS covariates remain imbalanced. However, when weighting by our primary HDPS model we achieve good balance in both the pre-defined and top 250 covariates.

In Table 4.4 we present mean absolute standardised differences to measure overall covariate balance. For the pre-defined covariates, we observe an increase in imbalance when additionally accounting for the HDPS covariates and this is similar under all HDPS models. Furthermore, we observe that when considering all key confounders (pre-defined and HDPS) the HDPS models perform similarly and achieve better balance than the pre-defined model. In this study, there is little difference in overall balance between the HDPS models, however other studies might see a deterioration in overall balance when including more HDPS covariates. Overall summaries of imbalance could be modified to put more weight on imbalance in covariates thought to be stronger confounders (in which imbalance is more likely to result in confounding bias); Table 4.4 presents one method for achieving this.

The HDPS aims to optimise confounder adjustment but there is a potential trade-off between better adjustment for a broader array of potential confounders versus tighter balance on key confounders. How much imbalance we are willing to permit in key confounders is primarily driven by how strongly these confounders are associated with the outcome. Therefore, a lack of imbalance in pre-defined and HDPS covariates does not necessarily mean all confounding has been removed and key unmeasured confounders may still exist.

4.7.4 Identification of potentially influential covariates

Whilst the full list of covariates selected is sometimes provided (*Schneeweiss et al.*, 2009), this is not easily digestible when interrogating several hundred HDPS covariates. However, manually inspecting the top covariates included can identify groups of codes relating to previously overlooked concepts that are important for minimising

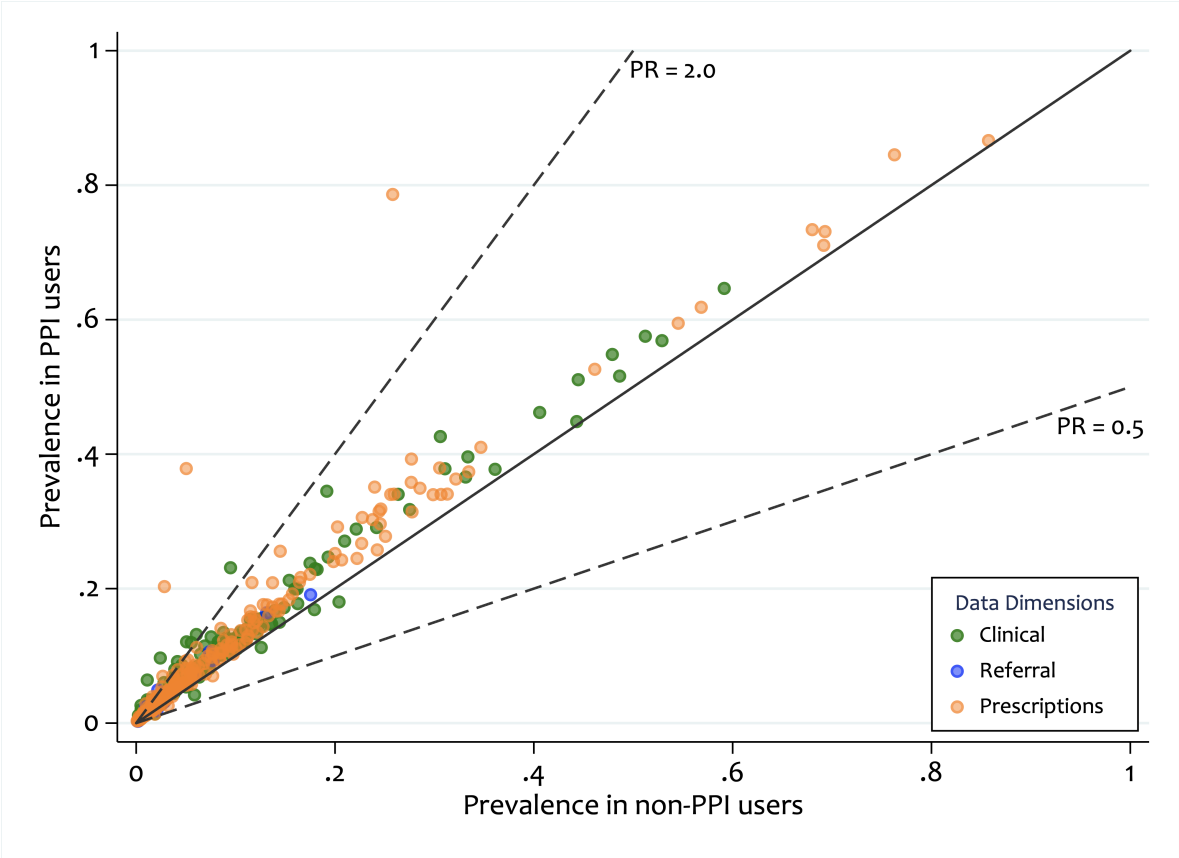


Figure 4.3: *Prevalence of the top 500 Bross-prioritised HDPS pre-exposure covariates by treatment group and by data dimension. The diagonal line indicates equal prevalence in both groups and the dashed lines show prevalence ratios (PR) of 0.5 and 2.0. The colour coding highlights which dimension the covariate was derived from.*

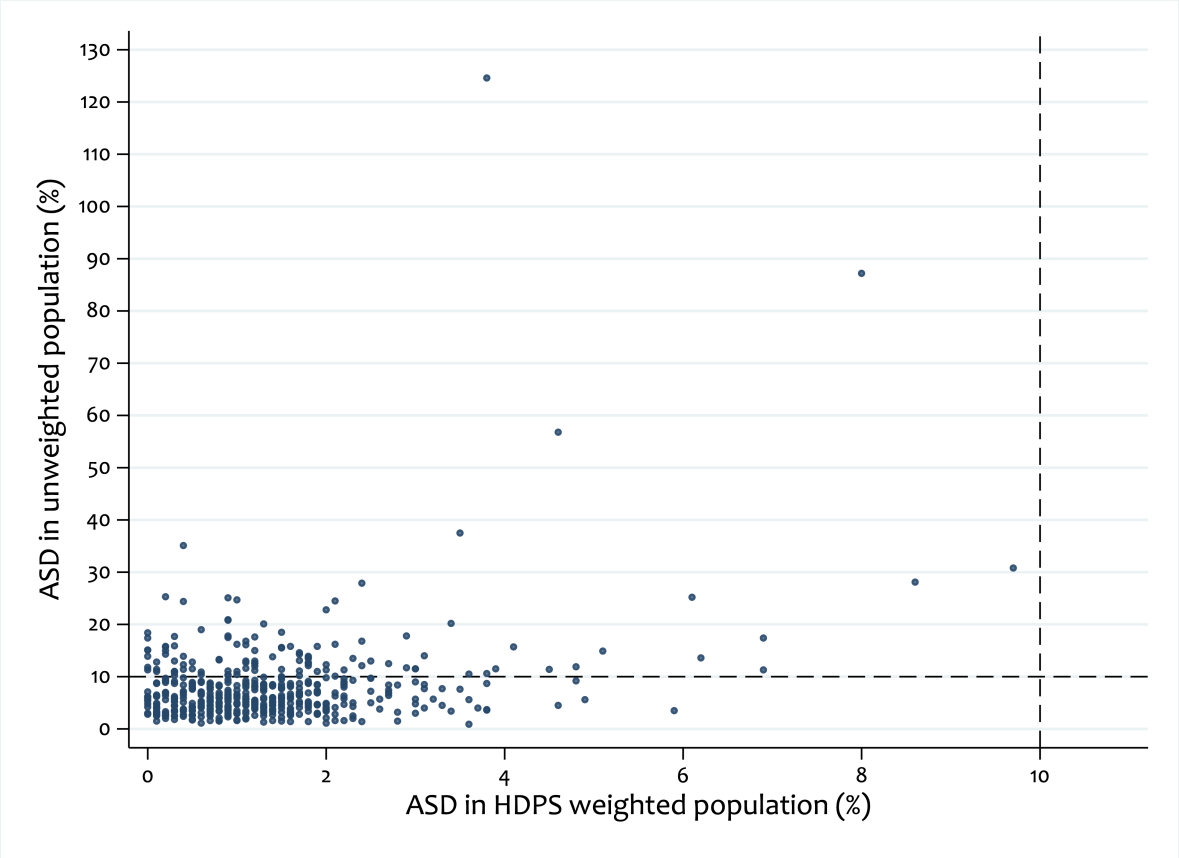


Figure 4.4: *Comparison of absolute standardised differences (ASDs) between unweighted and HDPS weighted sample under the primary analysis, selecting the top 500 HDPS covariates. Dashed lines indicate absolute standardised differences of 10%.*

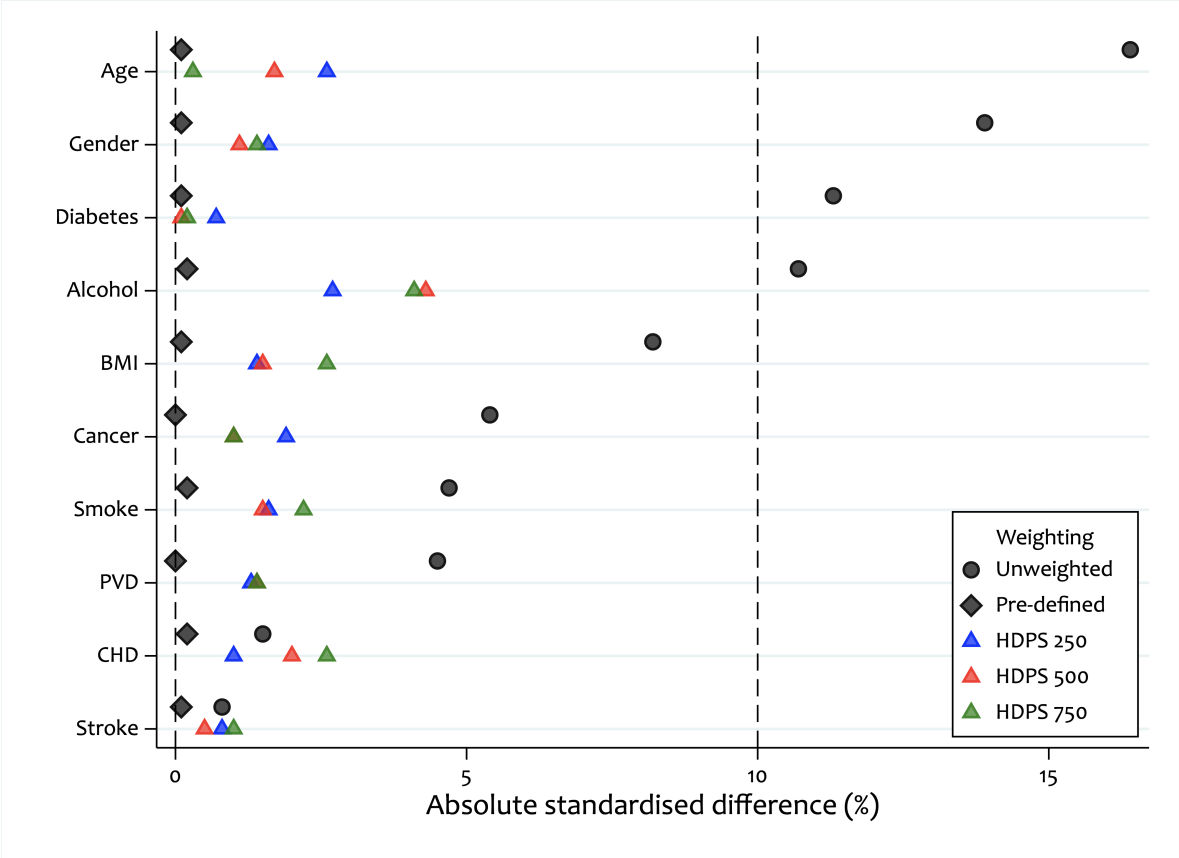


Figure 4.5: *Comparison of absolute standardised differences in a set of key covariates between unweighted, pre-defined covariate weighted, and pre-defined and HDPS covariate weighted samples.*

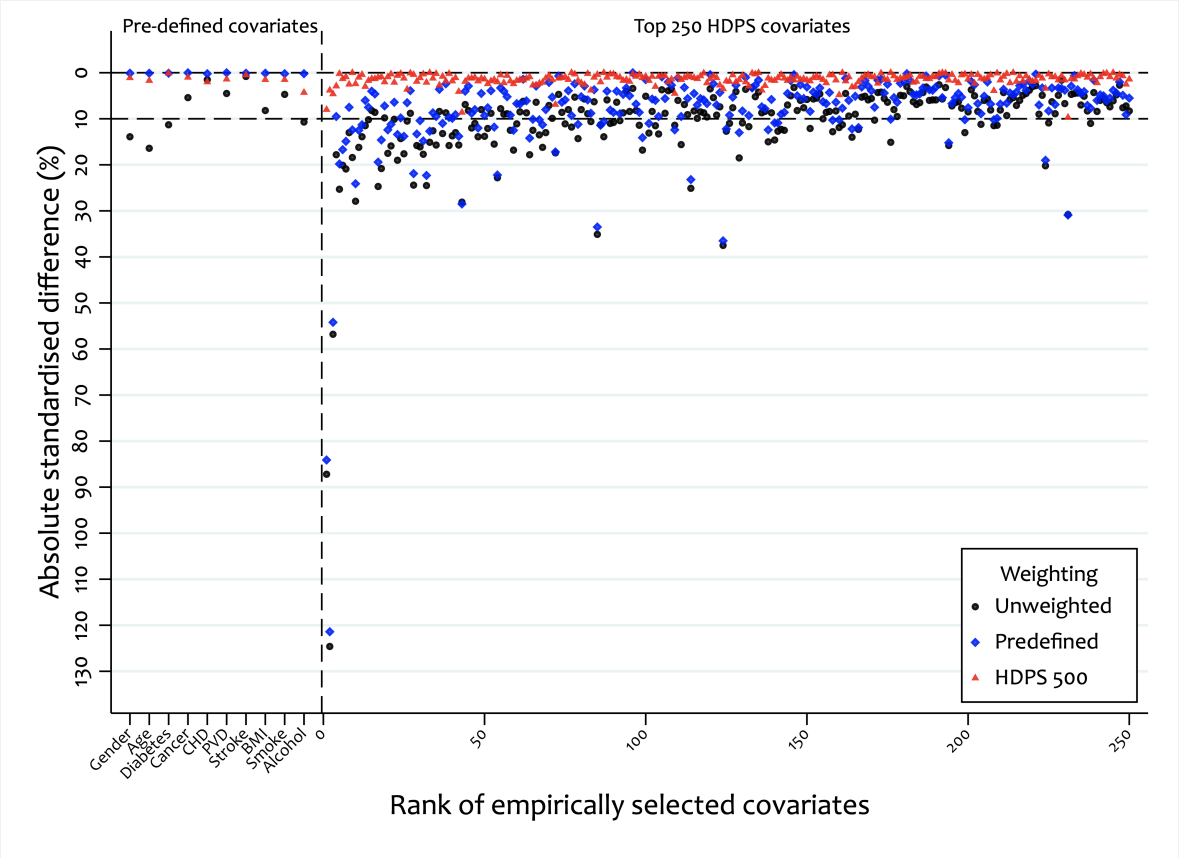


Figure 4.6: *Comparison of absolute standardised differences in the pre-defined and top 250 HDPS covariates between unweighted, pre-defined and HDPS (+500 covariates) weighted samples*

confounding bias (*Zhou et al.*, 2017).

An initial step is to investigate the distribution of Bross-derived bias values; Figure 4.7 shows the ranking score for the top 500 covariates (*Patorno et al.*, 2014). The colour coding indicates which dimension the covariates originated from and highlights that the majority of covariates were from the prescription dimension. Furthermore, this plot allows investigators to observe highly ranked covariates which might have a large amount of influence in the PS model.

The data-driven nature of the HDPS approach does not preclude adjustment for certain variables, such as instrumental variables (IVs) and colliders, which are typically excluded from PS models (*Brookhart et al.*, 2006; *Liu et al.*, 2012; *Myers et al.*, 2011; *Patrick et al.*, 2011). The HDPS uses the Bross formula to try to down weight covariates with these properties. Still, these variables could inadvertently be included, especially if the total number of covariates available is small relative to the proportion selected. However, the potential reduction in confounding bias from the inclusion of these covariates will often outweigh any increase in bias and variance induced (*Liu et al.*, 2012; *Myers et al.*, 2011; *Schneeweiss*, 2019). Whilst there are no statistical tests for classifying these types of variables, we can attempt to identify covariates which behave empirically like IVs. For this purpose, we define a likely IV or near-IV as a variable which is strongly associated with exposure but has a weak association with the outcome. Figure 4.8 describes the relationship between the covariate-exposure and covariate-outcome associations; covariates in the top-left quadrant represent those behaving empirically as IVs. The following empirical cut offs have been proposed to identify covariates behaving like IVs: $|\log(\mathrm{RR_{CE}})| > 1.5$ and $|\log(\mathrm{RR_{CD}})| < 0.5$ and, more restrictively, $|\log(\mathrm{RR_{CE}})| > 1.1$ and $|\log(\mathrm{RR_{CD}})| < 0.5$; where $\mathrm{RR_{CE}}$ and $\mathrm{RR_{CD}}$ are risk ratios for covariate-exposure and covariate-outcome respectively (*Schneeweiss et al.*, 2017).

We explore the sensitivity of results to the inclusion of potentially influential covariates in Section 4.8.

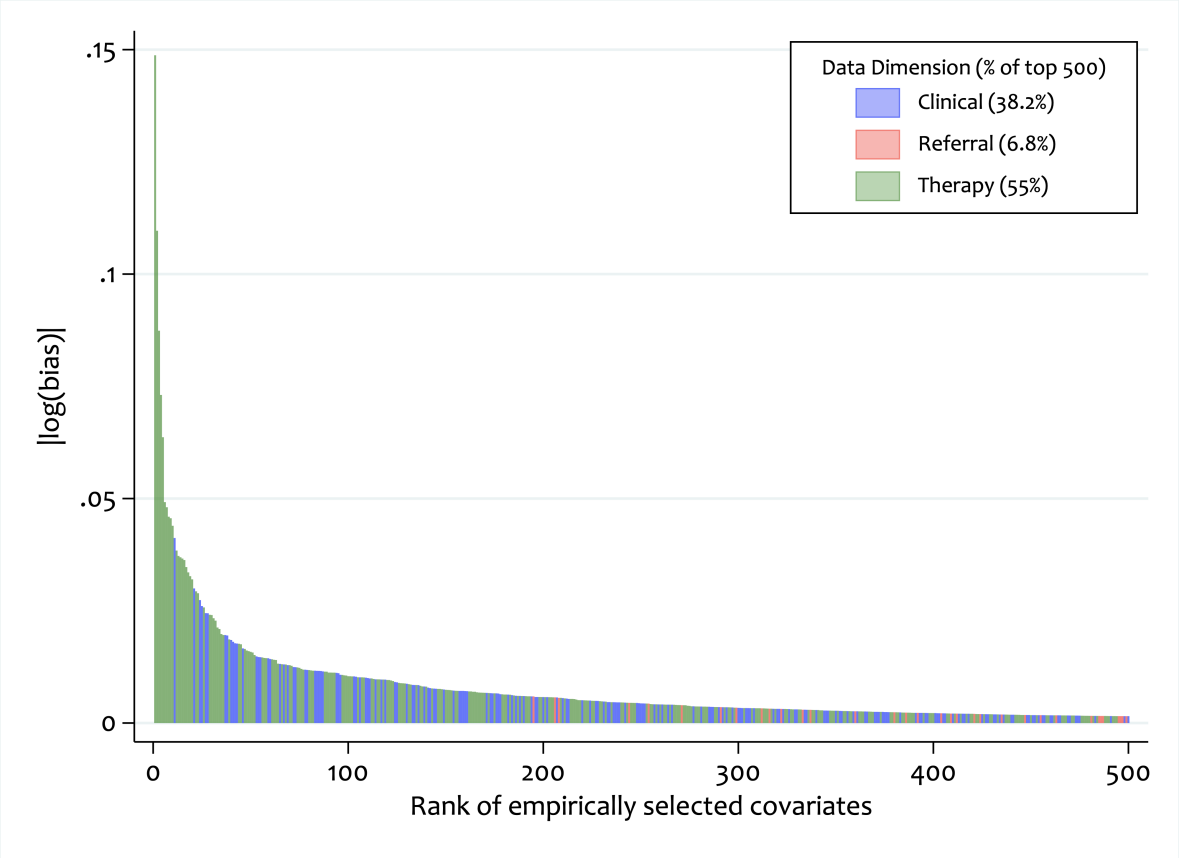


Figure 4.7: *Distribution of absolute log Bross bias values for each of the top 500 HDPS pre-exposure covariates.*

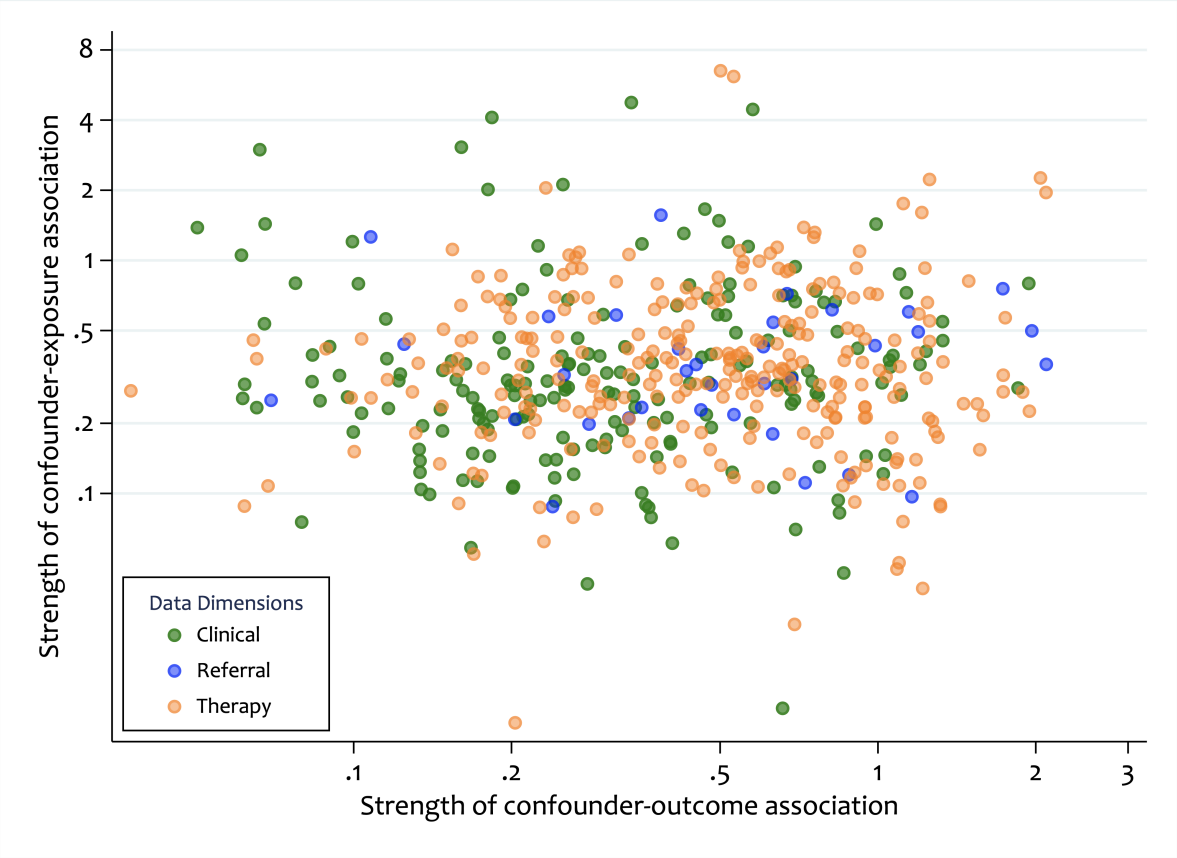


Figure 4.8: *Comparison of the covariate-exposure and covariate-outcome associations for the top 500 bias-based HDPS pre-exposure covariates.*

Table 4.4: *Comparison of the mean absolute standardised differences in the unweighted, pre-defined and pre-defined and HDPS weighted populations.* ***Abbreviations:*** *HDPS, high-dimensional propensity score.*

Set of covariates	Accounting for relative importance of HDPS covariates+	Mean absolute standardised differences				
		Unweighted	Pre-defined only weighted	Top 250 HDPS weighted	Top 500 HDPS weighted	Top 750 HDPS weighted
Pre-defined only	-	7.74	0.11	1.56	1.51	1.68
Top 250 HDPS only	No	10.91	8.15	1.14	1.42	1.51
Pre-defined and top 250 HDPS	Yes	6.73	5.11	0.62	0.77	0.88
	No	10.79	7.84	1.14	1.43	1.51
	Yes*	6.77	4.92	0.64	0.80	0.83

+Given a ranked (e.g., Bross-formula ranking) set of HDPS covariates of size N, importance weights are defined as ((N+1)-rank)/N.

*Predefined covariates are assigned an importance weight of 1.

4.8 Sensitivity analyses

4.8.1 Varying number of covariates selected

A key decision when applying the HDPS surrounds how many covariates to adjust for. Whilst investigators typically choose 200 or 500 variables to augment the pre-defined covariates, this is largely a result of convention. Simulation studies in moderate to large samples by Rassen et al suggest that adjusting for approximately 300 HDPS variables is likely to be sufficient (*Rassen et al.*, 2011a).

In practice, precisely how many HDPS variables to adjust for is likely to be dependent on the question of interest, rarity of outcome and the richness of data available in the database under investigation. Furthermore, previous studies indicate that in settings with few outcome events results can vary greatly depending on the number of covariates selected (*Patorno et al.*, 2014; *Wyss et al.*, 2018b).

Machine learning approaches have been proposed to determine the number of covariates selected for adjustment, but these have not yet been widely adopted (*Franklin et al.*, 2015; *Karim et al.*, 2018; *Schneeweiss et al.*, 2017; *Wyss et al.*, 2018b). Investigators are usually agnostic about how many covariates to select and therefore should assess the sensitivity of results to this decision.

We present two options for varying the number of covariates selected. The first specifics a discrete number of scenarios, for example, a study selecting 500 covariates in the primary analysis might investigate the results obtained from selecting 100, 250 and 750 covariates. Figure 4.9 presents these results next to the primary HDPS analysis, crude model and pre-defined covariates model. Compared to the crude and investigator analysis, varying the number of HDPS covariates selected resulted in consistent, but not monotonic, shifts in our point estimate towards the expected null association.

Another approach investigates the impact of incrementally adjusting for the empirically selected variables (Figure 4.10) (*Patorno et al.*, 2014). Figure 4.10 indicates stabilised results with the inclusion of between 250 and 600 covariates. Where results do not

stabilise, investigators should try to understand the driving factors and avoid undue focus on a specific HDPS analysis. Instead, it may be more suitable to report a range of effect estimates.

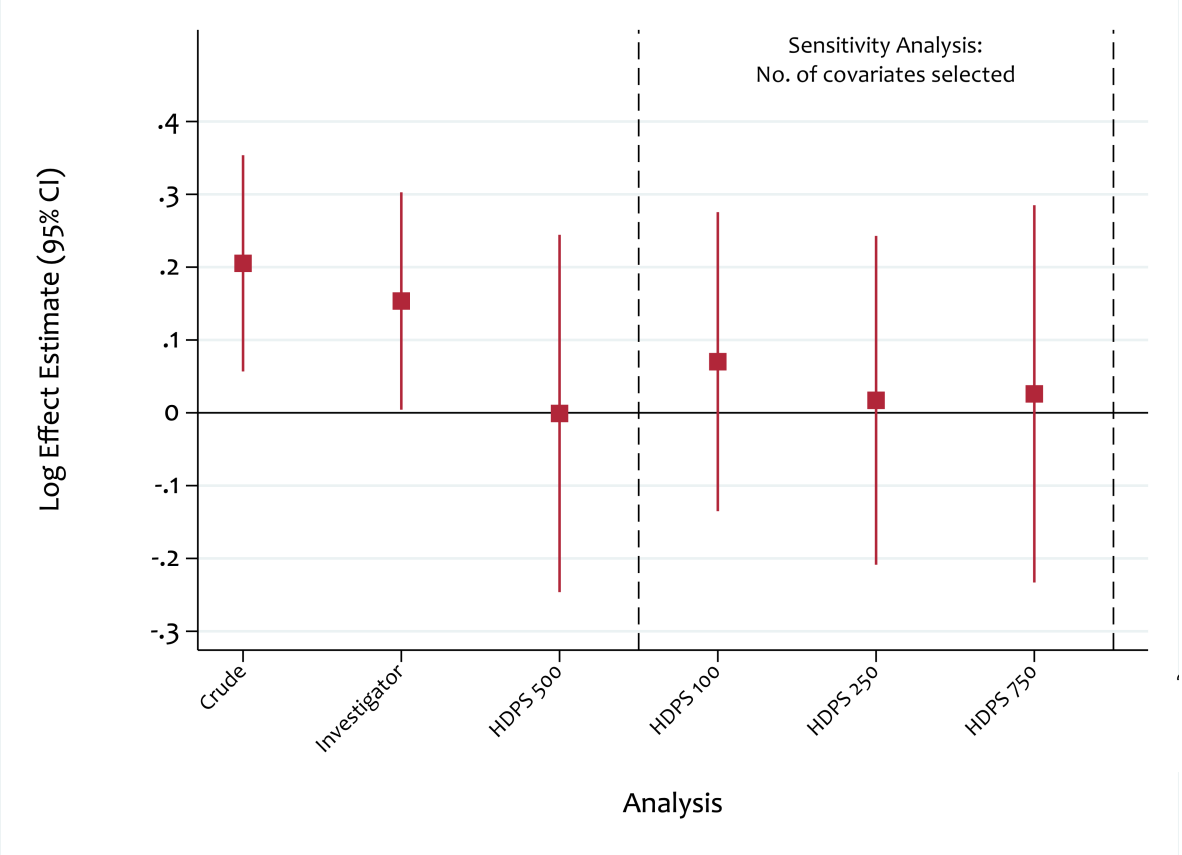


Figure 4.9: *Sensitivity analysis assessing the impact of selecting 100, 250 and 750 HDPS covariates selected on the log effect estimate. Propensity scores were estimated using logistic regression and treatment effects were estimated using an inverse probability of treatment weighted Cox model.*

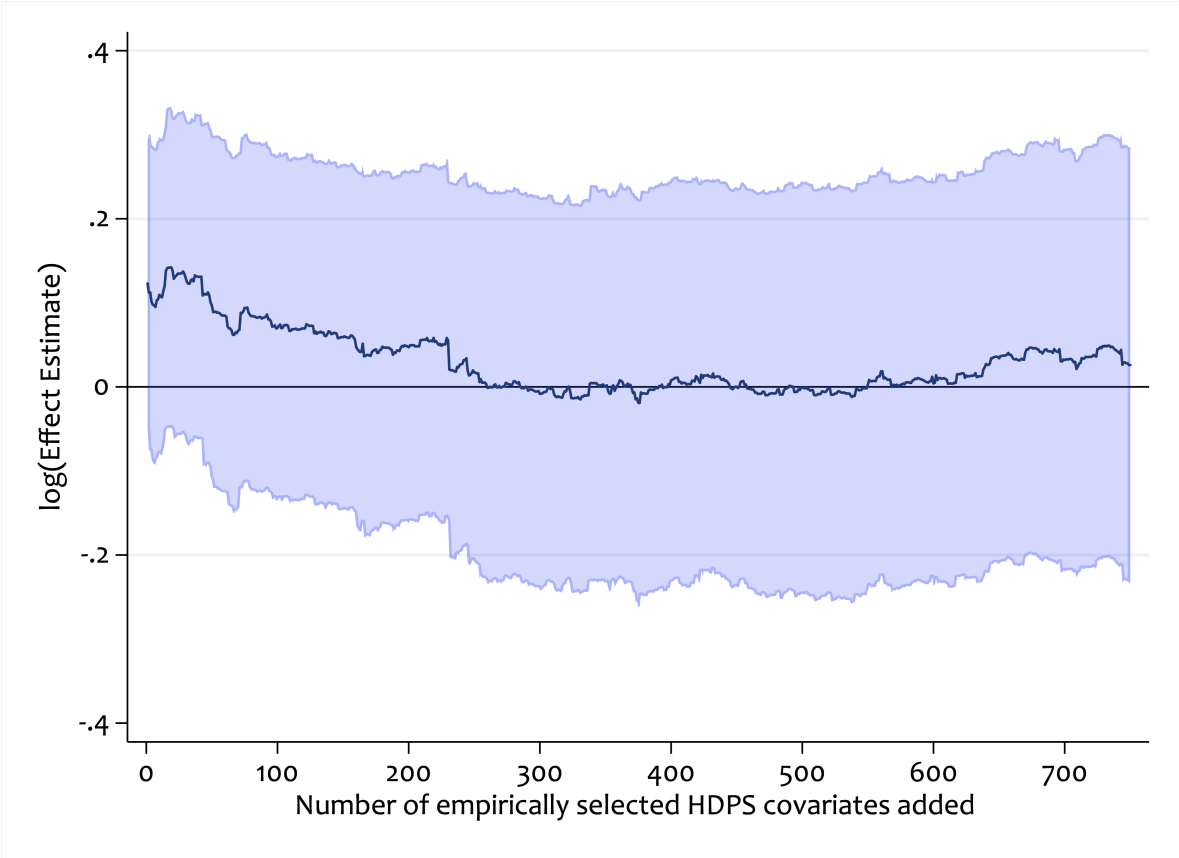


Figure 4.10: *Sensitivity analysis assessing the impact of incrementally adjusting for the top 750 HDPS covariates on the log effect estimate. Propensity scores were estimated using logistic regression and treatment effects were estimated using an inverse probability of treatment weighted Cox model.*

4.8.2 Quantifying impact of potentially influential covariates

In this section we quantify the impact of potentially influential covariates on results obtained in our primary analysis.

The distribution of Bross values (Figure 4.7) highlights that the top 3 ranked HDPS covariates are modestly higher than the rest. To understand the extent to which these covariates explain changes in the point estimates after inclusion of HDPS covariates, we conducted a sensitivity analysis adjusting for the predefined covariates plus only the top 3 ranked covariates (Table 4.5). We obtained a HR of 1.12 (95% CI: 0.93 to 1.34), indicating some residual confounding remained compared to adjustment for the full set of 500 HDPS covariates (HR 1.00; 95% CI: 0.78 to 1.28).

In Section 4.7.4 we identified covariates that behave empirically like IVs. To test the sensitivity of results to their inclusion, we conducted analyses based on Figure 4.8 (removing 7 near-IVs) and the two cut-offs previously described. Removing empirically identified IVs altered results in the 2nd decimal point only, indicating no change in the overall interpretation (Table 4.5).

Table 4.5: *Sensitivity analyses exploring the impact of identified potential influential covariates.* ***Abbreviations:*** *HDPS, high-dimensional propensity score.*

Sensitivity type	Sensitivity conditions	Number of covariates removed	Total number of HDPS covariates	Hazard ratio
Demographics & pre-defined only	-	-	-	1.17 (1.00 to 1.35)
Primary HDPS	-	-	500	1.00 (0.78 to 1.28)
Empirical	Pick the top 3 Bross ranked	497	3	1.12 (0.93 to 1.34)
Empirical	$\lvert\log(\mathrm{RR_{CE}})\rvert > 1.5$ & $\lvert\log(\mathrm{RR_{CD}})\rvert < 0.5$	4	496	1.06 (0.87 to 1.30)
Empirical	$\lvert\log(\mathrm{RR_{CE}})\rvert > 1.1$ & $\lvert\log(\mathrm{RR_{CD}})\rvert < 0.5$	9	491	1.06 (0.89 to 1.26)
Graphically Assess	Figure 4.8	7	493	1.06 (0.86 to 1.30)

4.9 Discussion

The HDPS approach has become a popular and scalable method for augmenting confounder adjustment in a given data source (*Schneeweiss*, 2018). However, as with PS analyses more generally, use of diagnostics and reporting of the details of the implementation is suboptimal (*Ali et al.*, 2015; *Granger et al.*, 2020). Using data from the UK CPRD (*Douglas et al.*, 2012; *Tazare et al.*, 2020), we highlighted diagnostic tools for assessing HDPS models and proposed considerations for reporting key features.

Drawing on established PS methodology, we described the importance of inspecting the estimated PS distributions before and after inclusion of the HDPS covariates. We recommended assessing covariate balance on important key confounders before and after inclusion of the HDPS covariates to investigate the potential impact of adjusting for many covariates on a set of strong confounders. Additionally, we described diagnostic tools more specific to the HDPS setting, e.g., for identifying instrumental-like variables and informing sensitivity analyses surrounding influential covariates.

We recommend that thorough sensitivity analyses should be conducted and reported when applying the HDPS. A key issue surrounds the number of covariates selected for inclusion in the PS model (*Patorno et al.*, 2014; *Wyss et al.*, 2018b), especially since the optimal number in a given setting is often unknown. Where inconsistencies are found, efforts should be made using the tools described to understand the drivers of variability.

Recent HDPS developments have focussed on refining covariate prioritisation and selection, especially using machine learning methods (*Franklin et al.*, 2015; *Karim et al.*, 2018; *Tian et al.*, 2018; *Wyss et al.*, 2018b). Whilst such developments can potentially improve HDPS analyses, no single approach is always optimal and applying the diagnostic tools described here is important to better understand the differences between these approaches.

We hope reporting of these analyses may be improved through more widespread use of the considerations and tools presented here.

4.10 Ethics statement

Scientific approval was obtained to use CPRD data by the Independent Scientific Advisory Committee (ISAC) (Protocol 17_194) and ethical approval from the London School of Hygiene & Tropical Medicine ethics committee (see Appendices A & B for details).

4.11 Supporting information

R and Stata Code for generating graphical tools

```
##########################################################
# R script:    001_highLevelConceptsSummary.R
#
# Author:      John Tazare
#
# Date:        29/04/2021
#
# Description: Plot for summarising high-level concepts in
#              the Top N ranked HDPS covariates separated
#              by data dimension.
#
# Inspired and adapted from:
# https://www.data-to-viz.com/graph/circularbarplot.html
##########################################################

# Load relevant library
library(tidyverse)

##########################################################
# Import data set with high-level concepts at chapter level.
##########################################################

##########################################################
# Variable Descriptions:
# description: chapter label/name
# dim: dimension identifier (e.g. d1 = clinical etc..)
# tot: total number of selected covariates from a particular
#      chapter
# percent: Out of N selected covariates, how many came from
#      this particular chapter

##########################################################

# Load data
data <- read.csv("-insert-data-path-here") %>%
  select(description, dim, tot, percent)
# Order data:
data <- data %>% arrange(dim, tot)

# Create whitespace to separate each dimension by adding empty bars
empty_bar <- 4
to_add <- data.frame(matrix(NA, empty_bar*nlevels(data$dim),
ncol(data)))
colnames(to_add) <- colnames(data)
to_add$dim <- rep(levels(data$dim), each=empty_bar)
data <- rbind(data, to_add)
data <- data %>% arrange(dim)
data$id <- seq(1, nrow(data))

# Get the name, angles and position of dimension chapter
label_data <- data
number_of_bar <- nrow(label_data)
angle <- 90 - 360 * (label_data$id-0.5) /number_of_bar
label_data$hjust <- ifelse( angle < -90, 1, 0)
```

```
label_data$angle <- ifelse(angle < -90, angle+180, angle)

# Make the plot
p <- ggplot(data, aes(x=as.factor(id), y=tot, fill=dim)) +
  geom_bar(stat="identity", alpha=0.5) +
  ylim(-100,120) +
  theme_minimal() +
  theme(
    legend.position = "none",
    axis.text = element_blank(),
    axis.title = element_blank(),
    panel.grid = element_blank(),
    plot.margin = unit(rep(-1,4), "cm")
  ) +
  coord_polar() +
  geom_text(data=label_data, aes(x=id, y=tot+10, label=description,
hjust=hjust), color="black", fontface="bold",alpha=0.6, size=2.5,
angle= label_data$angle, inherit.aes = FALSE )

p

# Save
ggsave("-insert-output-path-here/conceptsPlot.pdf", device = "pdf")
```

```
*****************************************************************
*
*   Do-file:        002_psOverlap.do
*
*   Author:         John Tazare
*
*   Date:           21/04/2021
*
*   Description:    Overlap plot comparing the propensity score
*                   distirbutions under the following models:
*                   1) pre-defined
*                   2) pre-defined + top 500 HDPS covariates
*                      ranked by Bross
*
******************************************************************

global data "insert-path-to-data"
global output "insert-path-to-output"

******************************************************************
* 1) Pre-defined propensity score
******************************************************************
* Note: 'ppi' is the treatment indicator variable

* Load cohort dataset
use "$data/cohort", replace

* Macro containing model specification
local model age_baseline i.gender i.smoke i.bmicat i.alcohol ///
        i.diabetes i.pvd i.chd i.stroke i.cancer

* Logistic regression to estimate propensity score
logit ppi `model' , or

* Predict probabilities
predict pscore, pr

* Plot kdensitys
gen ppi2 = ppi+1
forvalues i=1/2 {
      capture drop x`i' d`i'
      kdensity pscore if ppi2== `i', generate(x`i'  d`i')
  }

gen zero= 0

* Combine for propensity score distribution under this model
#delimit ;
twoway rarea d1 zero x1, color("blue%30")
        ||  rarea d2 zero x2, color("green%30")
        ytitle("")
        xtitle("")
```

```
        ylabel(0(2)8, labsize(medsmall) )
        xlabel(0(0.1)1, labsize(medsmall))
        legend(off)
        plotregion(color(white))
        scheme(uncluttered )
        graphregion(color(white))
        name(investigator, replace)
        title("Pre-defined",
            box
            bexpand
            bcol(none)
            lcol(black)
            size(medsmall)
            )
    ;
#delimit cr

*********************************************************************
* 2) Pre-defined + HDPS covariates propensity score
*********************************************************************

* Load cohort containing top 500 HDPS covariates
use "$data/HDPS_cohort.dta", clear

set matsize 600 // increase matsize for large models

* Logistic regression to estimate propensity score using both
* pre-defined and HDPS covariates

logit ppi `model' d1* d2* d3* , or
* Note: `model' is the same as defined above. d1* d2* d3*
* are the 500 HDPS covariates

* Follow previous steps

predict pscore, pr

gen ppi2 = ppi+1
forvalues i=1/2 {
      capture drop x`i' d`i'
      kdensity pscore if ppi2== `i', generate(x`i'  d`i')
  }

gen zero = 0

#delimit ;
twoway rarea d1 zero x1, color("blue%30")
        ||  rarea d2 zero x2, color("green%30")
        ytitle("")
        yla(, notick labcol(white))
        yscale(lstyle(none))
        xlabel(0(0.1)1, labsize(medsmall))
```

```
        xtitle("")
        legend(
            ring(0)
            pos(2)
            col(1)
            order( 1 "PPI users" 2 "Non-PPI users")
            region(lcolor(white)) size(medsmall)
            )
        plotregion(color(white))
        scheme(uncluttered )
        graphregion(color(white))
        title("HDPS", box
        bexpand
        bcol(none)
        lcol(black)
        size(medsmall))
        name(hdps, replace)
    ;
#delimit cr

*****************************************************************
* Combine the overlap plots
*****************************************************************

#delimit ;
graph combine investigator hdps,
        ycommon
        xcommon
        rows(1)
        plotregion(color(white))
        graphregion(color(white))
        l1(Density, size(medsmall))
        b1(Probability of receiving therapy, size(medsmall))
        ysize(1)
        xsize(2)
        iscale(1)
        imargin(0 0 0 0)
    ;
#delimit cr

graph export "$output/combinedOverlap.png", replace width(2000)

```

```
*****************************************************************
*
*   Do-file:        003_prevalenceByTreatmentGrp.do
*
*   Author:         John Tazare
*
*   Date:           21/04/2021
*
*   Description:    Prevalence of the top 500 bross-prioritised
*                   HDPS covariates by treatment group
*
*****************************************************************

global data "insert-path-to-data"
global output "insert-path-to-output"

*****************************************************************
* Required variables:
*****************************************************************
* pc0   - prevalence of confounder in Drug A group
* pc1   - prevalence of confounder in Drug B group
* rank  - bross-derived ranking
* dim   - data dimension identifier (optional)

* Load dataset with bias information
use "$data/bias_info.dta", clear

* Keep top 500 HDPS covariates
keep if rank <=500

* Data manipulation
gen dim=substr(code_id,1,2)
encode dim, gen(dim2)

*****************************************************************
* Plot
*****************************************************************

#delimit ;
twoway
    // Clinical dimension plot
    (scatter pc1 pc0 if dim2 ==1,
        msize(small) msymbol(circle) mcolor(green%70))
    // Referral dimension plot
    (scatter pc1 pc0 if dim2 ==2,
        msize(small) msymbol(circle) mcolor(blue%50))
    // Therapy dimension plot
    (scatter pc1 pc0 if dim2 ==3,
        msize(small) msymbol(circle) mcolor(orange%50))
    // Prevalence ratio = 0.5
    (function y=x/2, lcol(black*0.8) clpat(dash) range(0 1))
    // Prevalence ratio = 2.0
```

```
    (function y=2*x, lcol(black*0.8) clpat(dash) range(0 0.5))
    // Prevalence ratio = 1
    (function y=x, lcol(black*0.8))
        ,
       ytitle("Prevalence in PPI users" )
       xtitle("Prevalence in non-PPI users" )
       ylabel(,angle(horizontal))
       ylabel(0(0.2)1, labsize(medsmall) angle(horizontal))
       legend(
            order(1 "Clinical" 2 "Referral" 3 "Prescriptions")
            title("Data Dimensions",size(small))
            cols(1)
            rows(3)
            pos(4)
            ring(0)
            symxsize(*0.4)
            size(small)
    )
       plotregion(color(white))
       graphregion(color(white))
       name(prev, replace)
    // Prevalence ratio labels
       text(0.97 0.54  "PR = 2.0" , size(*0.9))
       text(0.44 0.97  "PR = 0.5" , size(*0.9))
;
#delimit cr

graph export "$output/prevPlot.png", width(2000)  replace

```

```
*****************************************************************
*
*   Do-file:        004_stdDiffsHDPS.do
*
*   Author:         John Tazare
*
*   Date:           21/04/2021
*
*   Description:    Absolute standardised differences
*                   between unweighted and HDPS weighted sample
*                   under the primary analysis, selecting the
*                   top 500 HDPS covariates.
*
*****************************************************************

global data "insert-path-to-data"
global output "insert-path-to-output"

*****************************************************************
* Required variables:
*****************************************************************
* stddiff_unwt - absolute standardised difference in unweighted
*                population
* stddiff_wt   - absolute standardised difference in weighted
*                population (top 500 covariates)

* Load dataset with std diffs
use "$data/stdDiffsHDPS.dta", clear

* Keep top 500 HDPS covariates
keep if rank <=500

* Data manipulation
gen dim=substr(code_id,1,2)
encode dim, gen(dim2)

*****************************************************************
* Plot
*****************************************************************

#delimit ;
twoway
    // Plot unweighted vs weighted absolute standardised diff.
    (scatter stddiff_unwt stddiff_wt,
        msymbol(circle) mcolor(navy%70) msize(tiny))
        ,
        xlabel(0(2)10, value angle(0) labsize(small))
        xtitle("ASD in HDPS weighted population (%)")
        ytitle("ASD in unweighted population (%)")
        ylabel(0 (10) 130, labsize(small) angle(0))
        xscale(range(0 11) extend)
        plotregion(color(white))
```

```
        scheme(uncluttered )
        graphregion(color(white))
        legend(off)
     // 10% absolute standardised diff. lines
        yline(10, lwidth(thin) lpattern(dash) lcolor(black))
        xline(10, lwidth(thin) lpattern(dash) lcolor(black))

            ;
#delimit cr
graph export "$output/stdDiffsHDPS.png", width(2000) replace


```

```
*****************************************************************
*
*   Do-file:        005_stdDiffsKeyConfounders.do
*
*   Author:         John Tazare
*
*   Date:           21/04/2021
*
*   Description:    absolute standardised differences in a set
*                   of key covariates between unweighted,
*                   pre-defined covariate weighted, and
*                   pre-defined and HDPS covariate weighted
*                   samples
*
*****************************************************************

global data "insert-path-to-data"
global output "insert-path-to-output"

*****************************************************************
* Required variables:
*****************************************************************
* stddiff_unwt - absolute standardised difference in unweighted
*                population
* stddiff_wt   - absolute standardised difference in weighted
*                population (by pre-defined model)
* stddiff_wt_X - absolute standardised difference in weighted
*                population (by pre-defined + X HDPS covariates
*                model)

* Load dataset with stddiffs
use "$data/stdDiffsHDPS.dta", clear

* Data management
gen order = 10 - (_n) + 1

* Create offset for unweighted and pre-defined weighted popns.
gen orderOffset = order + 0.3

* Label variables
label define orderLab   ///
    10  "Age"       ///
    9  "Gender"     ///
    8  "Diabetes"   ///
    7  "Alcohol"    ///
    6  "BMI"        ///
    5  "Cancer"     ///
    4  "Smoke"      ///
    3  "PVD"        ///
    2  "CHD"        ///
    1 "Stroke"
label values order orderLab
```

```
label values orderOffset orderLab

*****************************************************************
* Plot
*****************************************************************

#delimit ;
graph twoway
    // Unweighted absolute standardised diff.
   (scatter orderOffset stddiff_unwt,
    msymbol(circle) mcolor(black%70))
    // Pre-defined weighted absolute standardised diff.
   (scatter orderOffset stddiff_wt,
    msymbol(D) mcolor(black%70))
    // Top 250 HDPS weighted absolute standardised diff.
   (scatter order stddiff_wt_250,
    msymbol(triangle) mcolor(blue%70))
    // Top 500 HDPS weighted absolute standardised diff.
   (scatter order stddiff_wt_500,
    msymbol(triangle) mcolor(red%70))
    // Top 750 HDPS weighted absolute standardised diff.
   (scatter order stddiff_wt_750,
    msymbol(triangle) mcolor(green%70))
   ,
    ylabel(1(1)10, value angle(0) labsize(small))
    ytitle("")
    xtitle("Absolute standardised difference (%)")
    xlabel(0 (5) 15, labsize(small))
    xscale(range(0 15))
    xline(10, lwidth(thin) lpattern(dash) lcolor(black))
    xline(0, lpattern(dash) lwidth(thin) lcolor(black))
    plotregion(color(white))
    scheme(uncluttered )
    graphregion(color(white))
    legend(
        order(1 "Unweighted" 2 "Pre-defined" 3 "HDPS 250" ///
              4 "HDPS 500" 5 "HDPS 750")
        title("Weighting",size(small) col(black))
        cols(1)
        rows(5)
        pos(5)
        ring(0)
        symxsize(*0.4)
        size(small)
        )
    ;
#delimit cr
graph export "$output/stdDiffs.png", width(2000) replace
```

```
*****************************************************************
*
*   Do-file:        006_stdDiffsPredefinedPlusHDPSs.do
*
*   Author:         John Tazare
*
*   Date:           21/04/2021
*
*   Description:    absolute standardised differences in
*                   pre-defined and top 250 HDPS covariates
*                   between unweighted, pre-defined and HDPS
*                   (+500 covariates) weighted samples
*
*****************************************************************

global data "insert-path-to-data"
global output "insert-path-to-output"

*****************************************************************
* Required variables:
*****************************************************************
* stddiff_unwt   - absolute standardised difference in unweighted
*                  population
* stddiff_wt     - absolute standardised difference in weighted
*                  population (pre-defined only)
* stddiff_wt_500 - absolute standardised difference in weighted
*                  population (pre-defined plus top 500 HDPS)
* rank           - HDPS bross-prioritised rank; negative values
*                  given to pre-defined variables for plotting
*                  purposes

* Load dataset with bias information
use "$data/stdDiffsTop250.dta", clear

*****************************************************************
* Plot
*****************************************************************
#delimit ;
graph twoway
    // Unweighted absolute standardised diff.
    (scatter stddiff_unwt rank ,
        msymbol(circle) mcolor(black%70) msize(tiny))
    // Pre-defined weighted absolute standardised diff.
    (scatter stddiff_wt rank ,
        msymbol(D) mcolor(blue%70) msize(tiny))
    // Top 500 HDPS weighted absolute standardised diff.
    (scatter stddiff_wt_500 rank,
        msymbol(triangle) mcolor(red%70) msize(tiny))
    ,
    xlabel(-60 "Gender"  -54 "Age"  -48 "Diabetes"
           -42 "Cancer"  -36 "CHD"  -30 "PVD"
           -24 "Stroke"  -18 "BMI"  -12 "Smoke"
```

```
            -6 "Alcohol" 0 "0" 50 "50" 100 "100"
            150 "150" 200 "200" 250 "250",
            angle(45) labsize(vsmall)
            )
     xtitle("Rank of empirically selected covariates" )
     ytitle("Absolute standardised difference (%)")
     ylabel(0 (10) 130, labsize(small))
     yscale(reverse extend range(-5 135))
     yline(10, lwidth(thin) lpattern(dash) lcolor(black))
     yline(0, lpattern(dash) lwidth(thin) lcolor(black))
     xline(-0.5, lpattern(dash) lwidth(thin) lcolor(black))
     plotregion(color(white))
     scheme(uncluttered )
     graphregion(color(white))
     legend(
           order(1 "Unweighted" 2 "Predefined" 3 "HDPS 500")
           title("Weighting",size(small) col(black))
           cols(1)
           rows(5)
           pos(5)
           ring(0)
           symxsize(*0.4)
           size(small)
           )
      text( -10 -32  "Pre-defined covariates", size(*0.7))
      text( -10 130 "Top 250 HDPS covariates", size(*0.7))
     ;
#delimit cr
* manually fix labels
graph export "$output/stdDiffs_top250.png", width(2000) replace
```

```
*****************************************************************
*
*   Do-file:        007_brossDistribution.do
*
*   Author:         John Tazare
*
*   Date:           21/04/2021
*
*   Description:    Distribution of absolute log Bross bias
*                   values for each of the top 500 HDPS
*                   covariates
*
******************************************************************

global data "insert-path-to-data"
global output "insert-path-to-output"

******************************************************************
* Required variables:
******************************************************************
* abs_log_bias - bross ranking value
* rank - bross-derived ranking
* dim - data dimension identifier

* Load dataset with bias information
use "$data/bias_info.dta", clear

* Data manipulation
gen dim=substr(code_id,1,2)
encode dim, gen(dim2)

* Label dimensions
label define dimLab 1 "Clinical" 2 "Referral" 3 "Prescription"
label values dim2 dimLab

* Generate counts of codes by dimensions
count if dim2 == 1 // clinical
local dim1 = round(`r(N)'/500*100, 1.0)

count if dim2 == 2 // referral
local dim2 = round(`r(N)'/500*100, 1.0)

count if dim2 == 3 // therapy
local dim3 = round(`r(N)'/500*100, 1.0)

******************************************************************
* Plot
******************************************************************
#delimit ;
twoway
     // Clinical dimension plot
    (bar abs_log_bias rank if rank<=500 & dim2==1,
```

```
          lwidth(vthin) color(blue%40) )
      // Referral dimension plot
     (bar abs_log_bias rank if rank<=500 & dim2==2,
          lwidth(vthin) color(red%40))
      // Therapy dimension plot
     (bar abs_log_bias rank if rank<=500 & dim2==3,
          lwidth(vthin) color(green%40))
             ,
     ytitle("|log(bias)|" )
     ylabel(0(0.05)0.15, labsize(medsmall) angle(horizontal))
     xtitle("Rank of empirically selected covariates" )
     xlabel(0(100)500, labsize(medsmall))
     legend(
     order(1 "Clinical (`dim1'%)"
           2 "Referral (`dim2'%)"
           3 "Therapy (`dim3'%)"
            )
     title("Data Dimension (% of top 500)",
           size(small) col(black)
           )
           cols(1)
           symxsize(*0.4)
           size(small)
           pos(2)
           ring(0)
         )
      plotregion(color(white))
      scheme(uncluttered )
      graphregion(color(white))
      name(bross, replace)
;
#delimit cr

graph export "$output/brossDistribution.png", width(2000) replace
```

```
*****************************************************************
*
*   Do-file:        008_exposureOutcomeStrengths.do
*
*   Author:         John Tazare
*
*   Date:           21/04/2021
*
*   Description:    Comparison of the covariate-exposure and
*                   covariate-outcome associations for the
*                   top 500 HDPS covariates
*
*******************************************************************

global data "insert-path-to-data"
global output "insert-path-to-output"

*******************************************************************
* Required variables:
*******************************************************************
* ce_strength - covariate-exposure strength
* cd_strength - covariate-outcome strength
* rank        - bross-derived ranking
* dim         - data dimension identifier (optional)

* Load dataset with bias information
use "$data/bias_info.dta", clear

* Keep top 500 HDPS covariates
keep if rank <= 500

* Data manipulation
gen dim=substr(code_id,1,2)
encode dim, gen(dim2)

*******************************************************************
* Plot
*******************************************************************
#delimit ;
twoway // Clinical dimension plot
       (scatter ce_strength cd_strength if dim2 ==1,
            msize(small) msymbol(circle) mcolor(green%70))
       // Referral dimension plot
       (scatter ce_strength cd_strength if dim2 ==2,
            msize(small)  msymbol(circle) mcolor(blue%50))
       // Therapy dimension plot
       (scatter ce_strength cd_strength if dim2 ==3,
            msize(small)  msymbol(circle) mcolor(orange%50))
       ,
       ytitle("Strength of confounder-exposure association")
       xlabel(0 0.1 0.2 0.5 1.0 2 3,
            labsize(medsmall) angle(horizontal)
```

```
            )
        xscale(log)
        xtitle("Strength of confounder-outcome association")
        ylabel(0 0.1 0.2 0.5 1.0  2 4 8,
            labsize(medsmall) angle(horizontal)
            )
        yscale(log)
        legend(
            order(1 "Clinical" 2 "Referral" 3 "Therapy")
            title("Data Dimensions",size(small))
            cols(1)
            rows(3)
            pos(7)
            ring(0)
            symxsize(*0.4)
            size(small)
          )
         plotregion(color(white))
         scheme(uncluttered )
         graphregion(color(white))
         name(strength, replace)
;
#delimit cr

graph export "$output/empiricalIV.png", width(2000) replace
```

```
*****************************************************************
*
*   Do-file:        009a_forestPlot.do
*
*   Author:         John Tazare
*
*   Date:           21/04/2021
*
*   Description:    Forest plot for sensitivity analysis
*                   assessing the impact of the number of
*                   HDPS covariates selected
*
*   Note:           Propensity scores were estimated using
*                   logistic regression and treatment effects
*                   were estimated using an inverse probability
*                   of treatment weighted Cox model.
*
******************************************************************

global data "insert-path-to-data"
global output "insert-path-to-output"

******************************************************************
* Required variables:
******************************************************************
* lhr   - Log hazard ratio /effect estimate
* llci  - Log lower confidence interval limit
* luci  - Log upper confidence interval limit

clear all
use "$data/resultsHDPS.dta", replace

* Data management
gen order=1 if _n==1
replace order=4 if _n==2
replace order=7 if _n==3
replace order=10 if _n==4
replace order=13 if _n==5
replace order=16 if _n==6
replace order=19 if _n==7
replace order=22 if _n==8
replace order=25 if _n==9
replace order=28 if _n==10
replace order = order - 6 if order > 4

* Label the analyses
label define orderLab 1 "Crude" 4 "Pre-defined" 7 "HDPS 500" ///
10 "HDPS 100" 13 "HDPS 250" 16 "HDPS 750"
label values order orderLab


******************************************************************
```

```
* Plot
*****************************************************************
#delimit ;
graph twoway
    // Plot crude estimate / confidence interval
    (connected lhr order if order==1,
        mcol(cranberry) lcol(cranberry)
        msize(medium) msymbol(square))

    (rspike llci luci order if order==1,
        lcol(cranberry))

    // Plot pre-defined estimate / confidence interval
    (connected lhr order if order==4,
        mcol(cranberry) lcol(cranberry)
        msize(medium) msymbol(square))

    (rspike llci luci order if order==4,
        lcol(cranberry))

    // Plot HDPS 500 estimate / confidence interval
    (scatter lhr order if order==7,
        mcol(cranberry) lcol(cranberry)
        msize(medium) msymbol(square))

    (rspike llci luci order if order==7,
        lcol(cranberry))

    // Plot HDPS 100,250,750 estimates / confidence intervals
    (scatter lhr order if order<=16 & order>=10,
        mcol(cranberry) lcol(cranberry)
        msize(medium) msymbol(square))

    (rspike llci luci order if order<=16 & order>=10,

        lcol(cranberry))
    ,

    ytitle("Log Effect Estimate (95% CI)" )
    ylabel(-0.3(0.1)0.4,
        labsize(medsmall) angle(horizontal))
    xtitle("Analysis" , margin(t+2) )
    xlabel(1(3)17.8,
    labsize(small) valuelabel angle(45))
    xscale(  range(0.5 17.8) )
    yscale(  range(-0.3 0.5) )
    legend(off)
    plotregion(color(white))
    scheme(uncluttered )
    graphregion(color(white))
    // Add null value line
    yline(0, lcol(black) lpattern(solid) lwidth(thin))
```

```
    // Add vertical separators
    xline(8.5, lpattern(dash) lwidth(thin) lcol(black))
    xline(17.5, lpattern(dash) lwidth(thin) lcol(black))
 ;
#delimit cr
graph export "$output/simpleForestPlot.png", width(2000) replace

```

```
*****************************************************************
*
*   Do-file:        009b_intensiveForestPlot.do
*
*   Author:         John Tazare
*
*   Date:           21/04/2021
*
*   Description:    Forest plot for sensitivity analysis
*                   assessing the impact of the number of
*                   HDPS covariates selected
*
*   Note:           Propensity scores were estimated using
*                   logistic regression and treatment effects
*                   were estimated using an inverse probability
*                   of treatment weighted Cox model.
*
*******************************************************************

global data "insert-path-to-data"
global output "insert-path-to-output"

*******************************************************************
* Generic procedure for obtaining effect estimates from
* incrementally adding HDPS covariates to pre-defined model
*******************************************************************
tempname effectEsts

* Create a postfile to 'post' the number of variables added
* effect estimates and 95% CI bounds
postfile `effectEsts' numVars hr lci uci  ///
        using "intensivePlots.dta", replace


forvalues i = 1/750 {

if mod(`i', 10) == 0 {
noi di " Fitting model `i' out of 750"
}

qui {

* Load dataset with bias information
use "$data/bias_info.dta", clear
gsort - abs_log_bias // sort by ranking metric
keep if rank <= `i' // keep the top `i' codes
qui levelsof code_id, local(final_selection)

* Load overall cohort with HDPS covariates
use "$data/hdpsCohort.dta", replace

* Pre-defined model specification
```

```
local model age_baseline i.gender i.smoke i.bmicat ///
    i.alcohol i.diabetes i.pvd i.chd i.stroke i.cancer

* Add the selected HDPS covariates to this model
foreach item of local final_selection {
    local model `model' `item'
}

* Fit propensity score modoel
logit ppi `model' , or

* Drop any previous pscore/weights
cap drop pscore
cap drop wt

predict pscore, pr

* Generate IPTW weights
gen wt=1/pscore if ppi==1
replace wt=1/(1-pscore) if ppi==0

* Fit outcome model
#delimit ;
stset exit_t [pw=wt],
        origin(dob)
        fail(inc_mi)
        id(anonpatid)
        enter(entry_t)
        scale(365.25)
        ;
#delimit cr

stcox i.ppi, vce(robust)

* Capture and 'post' the HR and 95% CI limits
mat def A = r(table)
local hr = A[1,2]
local lci = A[5,2]
local uci = A[6,2]
post `effectEsts' (`i') (`hr') (`lci') (`uci')

}
    }
postclose `effectEsts'

clear

* Load postfile with effect estimates
use "intensivePlots.dta", replace

* Transform effect estimates
gen llci = log(lci)
```

```
gen luci = log(uci)
gen lhr  = log(hr)

*****************************************************************
* Plot
*****************************************************************
 #delimit ;
twoway
 // Plot effect estimates
 (line lhr numVars, lwidth(medium) color(navy*1.2))

 // Plot confidence interval bounds
 (rarea llci luci numVars, color(blue%20))
 ,
  ytitle("log(Effect Estimate)" )
  ylabel(-0.4(0.2)0.4,  labsize(medsmall)  angle(horizontal))
  xtitle("Number of empirically selected HDPS covariates added")
  xlabel(0(100)750, labsize(medsmall) )
  legend(off)
  yline(0, lcol(black) lpattern(solid) lwidth(thin))
  plotregion(color(white))
  scheme(uncluttered )
  graphregion(color(white))
;
#delimit cr
graph export "$output/incremForestPlot.png", width(2000) replace
```

Chapter 5

Paper C: hdps: a suite of commands for applying high-dimensional propensity score approaches in Stata

John Tazare[1], Liam Smeeth[1,2], Stephen JW Evans[1], Ian J Douglas[1,2], Elizabeth Williamson[1,2]

1. London School of Hygiene and Tropical Medicine, London, UK.
2. Health Data Research (HDR) UK, London, UK.

5.1 Overview

Summary

Work presented in Chapter 3 applying the HDPS required bespoke code to allow for implementation of the proposed modifications. In this chapter, that initial code is converted into a Stata package to facilitate more widespread use of these methods. Whilst there are existing suites for applying the HDPS in the R and SAS statistical software packages, there is no inbuilt or user-written implementation of the HDPS in Stata. Furthermore, Stata is commonly used by investigators using UK EHRs. I present the `hdps` package in Stata for implementing HDPS approaches and generating diagnostic visualisations for the covariates selected. The work was initially presented as an oral presentation at the *2019 UK Stata Conference*. This paper has been submitted to *The Stata Journal* and is currently under review.

Thesis objective addressed

This chapter addresses the following objective of the overall thesis (Section 1.3):

5. Develop reusable software to implement HDPS approaches in the Stata statistical software package.

Role of candidate

I generated a user-friendly set of commands for implementing HDPS approaches in Stata and simulated datasets based on UK EHRs for demonstrating the key features. Elizabeth Williamson (EW) reviewed an initial version of the underlying code and advised on approaches to increase computational efficiency. Tim Morris gave input surrounding the design of the suite of commands in Stata. The paper was finalised after suggestions, comments and guidance from Liam Smeeth, Stephen Evans, Ian Douglas and EW.

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

London School of Hygiene & Tropical Medicine
Keppel Street, London WC1E 7HT

T: +44 (0)20 7299 4646
F: +44 (0)20 7299 4656
www.lshtm.ac.uk

RESEARCH PAPER COVER SHEET

Please note that a cover sheet must be completed for each research paper included within a thesis.

SECTION A – Student Details

Student ID Number	LSH1401926	**Title**	Mr
First Name(s)	John		
Surname/Family Name	Tazare		
Thesis Title	High-dimensional propensity scores for data-driven confounder adjustment in UK electronic health records		
Primary Supervisor	Elizabeth Williamson		

If the Research Paper has previously been published please complete Section B, if not please move to Section C.

SECTION B – Paper already published

Where was the work published?			
When was the work published?			
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion			
Have you retained the copyright for the work?*	Choose an item.	Was the work subject to academic peer review?	Choose an item.

*If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	The Stata Journal
Please list the paper's authors in the intended authorship order:	John Tazare, Liam Smeeth, Stephen JW Evans, Ian J Douglas, Elizabeth J Williamson
Stage of publication	Submitted

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	I planned and wrote the Stata package code and help files, generated simulated data, conducted the analysis, and drafted the article with feedback from co-authors

SECTION E

Student Signature	John Tazare
Date	16/07/2021

Supervisor Signature	Elizabeth J Williamson
Date	16/07/2021

5.2 Abstract

Large healthcare databases are increasingly used for research investigating the effects of medications. However, a key challenge surrounds the ability to capture hard to measure concepts (often relating to fraility and disease severity) that can be crucial for successful confounder adjustment. The high-dimensional propensity score (HDPS) has been proposed as a data-driven method to improve confounder adjustment within healthcare databases and was developed in the context of administrative claims databases. We present `hdps`, a suite of programs implementing this approach in Stata that assesses the prevalence of codes, generates HDPS covariates, performs variable selection and provides investigators with graphical tools for inspecting the properties of covariates selected.

5.3 Introduction

Large healthcare databases, such as electronic health records (EHRs), have become widely used for investigating the benefits and harms of medications (*Sturmer et al.*, 2006). These data have the potential to answer important questions surrounding the long-term and rare effects of medifications, however, confounding bias is often a major concern and can result in misleading conclusions being drawn (*Brookhart et al.*, 2010; *Freemantle et al.*, 2013).

Confounder adjustment is often achieved using outcome regression; modelling the relationship between an outcome variable and a treatment (or exposure) variable conditional on a set of confounders. However, analysis based on the propensity score (PS) is often preferred in the context of large healthcare databases given the ability to summarise a large amount of confounder information in a single score (*Jackson et al.*, 2017; *Rosenbaum and Rubin*, 1983). PS analysis involves modelling the treatment allocation process, using a set of observed variables to estimate the conditional probability of initiating the treatment under investigation. There are several methods for estimating treatment effects based on the estimated propensity scores, for example using weighting

or matching methods. General introductions to the concepts behind PS analysis are given by (*Williamson et al.*, 2012) and (*Austin*, 2011). (*Brookhart et al.*, 2006) provide a discussion surrounding the types of variables to be included in PS models, indicating that all confounders and risk factors should be included. Finally, indications for PS analysis and current practice in pharmacoepidemiology are discussed by (*Jackson et al.*, 2017).

As with outcome regression models, the key assumption of no unmeasured confounding is required to yield unbiased treatment effect estimates from PS methods (*Williamson et al.*, 2012). However, in large healthcare databases successful adjustment for confounding often relies on capturing concepts, such as frailty, which are hard to measure (even in controlled settings, e.g. randomized clinical trials) (*Schneeweiss et al.*, 2009).

The high dimensional propensity score (HDPS) algorithm has been proposed as an extension to propensity score methodology, designed to maximise capture of hard-to-measure or otherwise unmeasured concepts in large healthcare databases (*Schneeweiss et al.*, 2009). The HDPS is a semi-automated data-driven approach for generating and selecting potential features (typically codes captured as part of the routine recording of clinical and administrative information), measured prior to treatment initiation, that are likely to be informative of disease severity and frailty (*Schneeweiss et al.*, 2009). HDPS approaches aim to optimise confounder control in a given setting by adjusting for several hundred of these data-derived covariates. The benefit of these approaches has been illustrated in a diverse range of settings resulting in its popularity as a method for confounder adjustment in pharmacoepidemiological studies (*Schneeweiss*, 2018). Furthermore, whilst implementations of HDPS exist in SAS and R these approaches have yet to be formally implemented in Stata *Lendle* (2017); *Rassen et al.* (2020).

We introduce `hdps`, a suite of commands for performing the HDPS procedure and investigating properties of the selected covariates (*Schneeweiss et al.*, 2009; *Wyss et al.*, 2018a). These commands allow investigators to specify commonly used tuning parameters surrounding key decisions in the HDPS, for example, the method of covariate prioritization and number of covariates selected (*Patorno et al.*, 2014; *Schneeweiss et al.*, 2009; *Wyss et al.*, 2018b). Additonally, recent modifications tailoring the HDPS for

use in UK EHRs are also implemented (*Tazare et al.*, 2020). We demonstrate how to conduct the HDPS procedure and perform a PS analysis with the selected covariates.

5.4 High-dimensional propensity scores

The HDPS is a multi-step algorithm that transforms codes recorded in a healthcare database into covariates to be included within a PS analysis. The codes considered during the HDPS procedure are recorded prior to treatment initiation to avoid inadvertent adjustment for covariates on the causal pathway from treatment to outcome (*Schneeweiss et al.*, 2009). This assessment window is usually defined during the 1-year prior to treatment initiation. The steps of the HDPS are summarised as follows (Figure 5.1):

1. **Data dimensions:** Specify the data to be used for deriving data-driven covariates. Typically, this involves separating information in the healthcare database into multiple datasets, capturing different aspects of clinical care or coding information. For example, in UK EHRs we may separate clinical, referral, hospitalization and prescription information.

2. **Prevalence filter:** Identify the most prevalent codes in each dimension (typically, 200 are chosen) (*Schneeweiss et al.*, 2009). This step is optional and instead all codes can be assessed for potential inclusion.

3. **Assess recurrence:** For each code identified in the previous step, generate up to three binary covariates based on how frequently patients have a particular code recorded in the aforementioned assessment window:

$$\text{Once} = \begin{cases} 1 & \text{if code recorded} \geq \text{once} \\ 0 & \text{otherwise} \end{cases}$$

$$\text{Sporadic} = \begin{cases} 1 & \text{if code recorded} \geq \text{median} \\ 0 & \text{otherwise} \end{cases}$$

$$\text{Frequent} = \begin{cases} 1 & \text{if code recorded} \geq \text{upper quartile} \\ 0 & \text{otherwise} \end{cases}$$

Recent work by (*Tazare et al.*, 2020) implementing the HDPS in UK EHRs extends the bottom frequency category to capture information recorded 'Ever' in a patient's history. For codes originating from data dimensions where this extra information is used, the 'Once' variable is replaced by:

$$\text{Ever} = \begin{cases} 1 & \text{if code recorded anytime in patient's history} \\ & \text{(prior to treatment initiation)} \\ 0 & \text{otherwise} \end{cases}$$

4. **Prioritize covariates:** Prioritize the set of binary covariates to identify those most important for confounder adjustment.

   - **Bross formula:** Typically, this prioritization is performed using the Bross formula to define a multiplicative bias term (*Bross*, 1966; *Schneeweiss et al.*, 2009; *Wyss et al.*, 2018a):

     $$\text{Bias}_{\text{M}} = \frac{\text{P}_{\text{C1}}(\text{RR}_{\text{CD}} - 1) + 1)}{\text{P}_{\text{C0}}(\text{RR}_{\text{CD}} - 1) + 1)}$$

     where RR_{CD} is the covariate-outcome risk ratio and P_{C1} and P_{C0} are the prevalence of the covariate in the treated and untreated, respectively. Covariates are ranked in descending order by $|\log(\text{Bias}_{\text{M}})|$, with higher numbers indicating greater potential for contributing to confounding bias.

   - **Exposure-based:** (*Rassen et al.*, 2011b) have shown that, in studies of few treated patients or few outcome events, prioritizing covariates based solely on the covariate-exposure relationship can perform well compared to the Bross formula.

5. **Select covariates:** From the set of prioritized covariates, a subset is chosen for inclusion in the PS model. This is a key decision in the HDPS procedure and depending on the setting, results can vary considerably (*Patorno et al.*, 2014; *Wyss*

et al., 2018b). Typically, 200 or 500 covariates are selected (*Schneeweiss*, 2018; *Schneeweiss et al.*, 2009), however these numbers are arbitrary and we recommend testing the sensitivity of results to this decision.

6. **Diagnostic tools:** In any PS analysis, it is important to assess covariate balance and perform diagnostics (*Austin*, 2009a; *Granger et al.*, 2020). For HDPS analyses, it is additionally important to understand the covariates selected by identifying potentially influential covariates and investigating covariate balance (*Franklin et al.*, 2015; *Patorno et al.*, 2014).

7. **Propensity score analysis:** The final step surrounds performing a standard PS analysis. The first stage is to estimate the PS, usually via a logistic regression modelling the treatment variable on a set of covariates. In the HDPS setting, this set of covariates includes: 1) a set of 'investigator' covariates identified based on clinical knowledge and, 2) the set of selected HDPS covariates. The second stage involves estimating treatment effects from an outcome model, incorporating the PS using adjustment, matching, weighting or stratification (*Williamson et al.*, 2012).

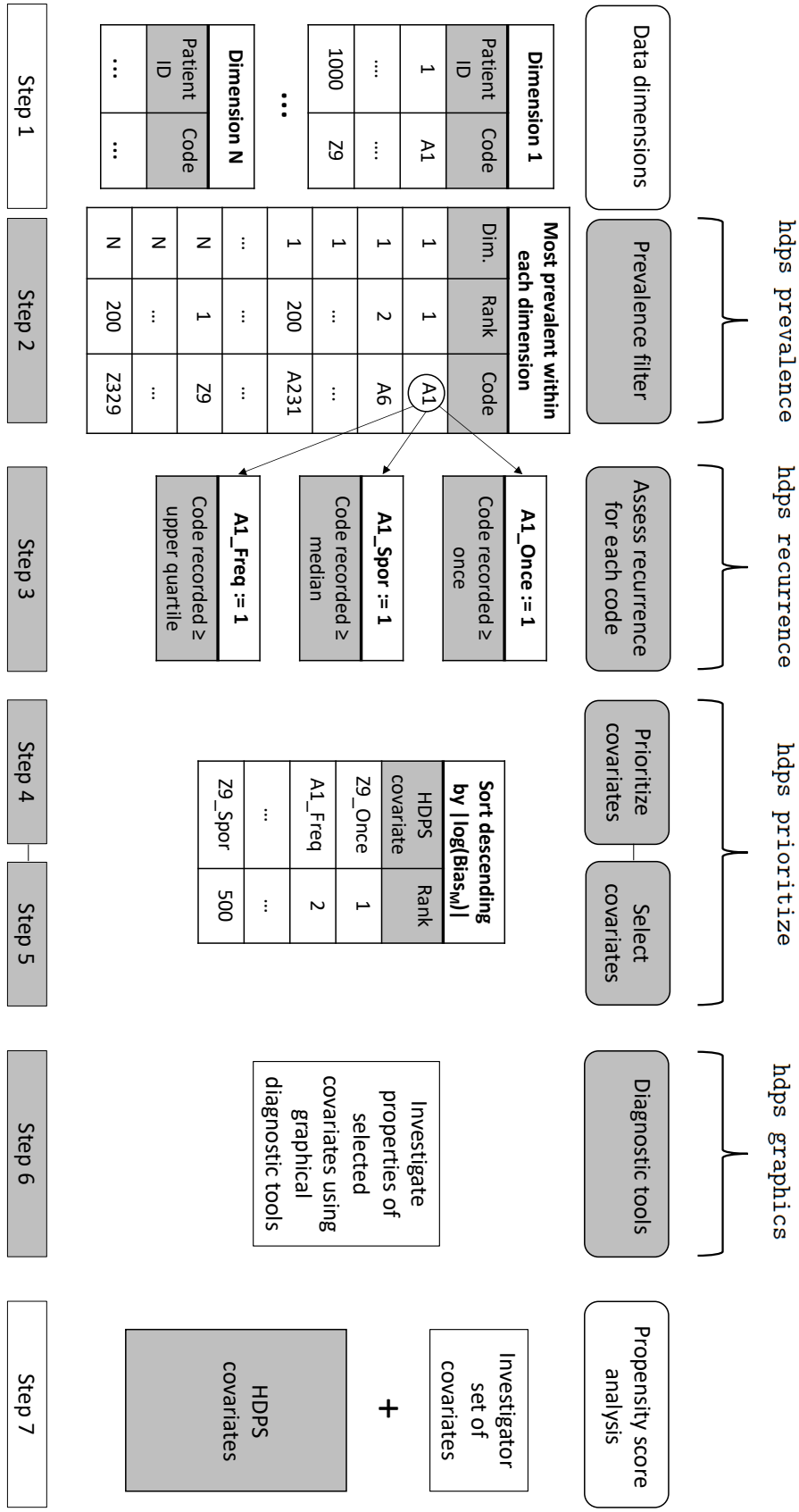


Figure 5.1: *Summary of a generic implementation of the HDPS algorithm, identifying the top 200 most prevalent codes per dimension and selecting the top 500 bross-ranked HDPS covariates. Steps highlighted in grey represent those implemented in the* `hdps` *package.* ***Abbreviations:*** *Dim., dimension.*

5.5 The hdps commands

5.5.1 Installation

The `hdps` suite is hosted and maintained on GitHub (for details, see (*Haghish*, 2020)) and can be installed as follows: 1) install the `github` package, and 2) install `hdps` from the hosted GitHub repository.

```
. net install github, from("https://haghish.github.io/github/")

. github install johntaz/hdps
```

5.5.2 Data formats

The `hdps` suite uses two types of input datasets, a cohort dataset and at least one data dimension.

- **Cohort dataset:** One observation per-patient and including at least patient identifier (stored in all datasets as a *string* variable), binary treatment variable and binary outcome variable (both stored as numeric variables). We show the first 10 observations from an example dataset below:

```
. list patid trt outcome in 1/10
```

	patid	trt	outcome
1.	1000	1	0
2.	1001	1	0
3.	1002	1	1
4.	1003	0	1
5.	1004	1	0
6.	1005	1	1
7.	1006	1	0
8.	1007	1	1
9.	1008	1	1
10.	1009	0	0

- **Data dimension(s):** A long format dataset containing codes recorded during the HDPS assessment window for all patients in the cohort. A separate dataset should

be prepared for each data dimension. This dataset will often be many observations per-patient per-code. We show the first 10 observations for an example patient, highlighting multiple recordings for codes within the assessment window.

```
. list in 1/10
```

	patid	code
1.	1000	M75
2.	1000	R06
3.	1000	I25
4.	1000	M75
5.	1000	L40
6.	1000	I25
7.	1000	R42
8.	1000	K59
9.	1000	K59
10.	1000	R06

- **Ever dimension(s):** If 'Ever' information (as described in Section 5.4, Step 3) is being assessed for a given data dimension, a secondary dataset should be provided. This data will be in long format and contain codes recorded in a patient's entire history (prior to treatment initiation). Since we only want to capture the presence of a specific code, this dataset should be one observation per-patient per-code. Note, to reduce the size of this dataset, users may wish to remove any code already recorded during the assessment window.

```
. list in 1/5
```

	patid	code
1.	1000	B35
2.	1000	D64
3.	1000	E11
4.	1000	R06
5.	1000	V89

5.5.3 The hdps setup command

The `hdps setup` command declares the data dimensions and key variables used throughout the HDPS procedure, further specifying the directory for outputted datasets. Set

the current directory to a folder containing all necessary data and load the cohort dataset into memory.

Syntax

`hdps setup` *dimensions(s)*, `study(`*string*`) save(`*string*`) patid(`*string*`)`
`exposure(`*varname*`) outcome(`*varname*`)`

where a *dimension* term is specified for each of the data dimensions required, using the following syntax:
`(` *filename*, *varname* `[ever] )`

Dimension syntax

- *filename* specifies the file name for the data dimension.
- `varname` specifies the variable in the data dimension containing codes. Note, this is a required option and must be the first option specified.
- `ever` optionally specifies that the recurrence assessment for the dimension should incorporate 'Ever' information. Where `ever` is specified for a particular dimension, the 'Ever' dimension must be named *filename_ever* and the variable containing codes must be named *varname*.

Overall options

- `study(`*string*`)` specifies a study name that serves as a prefix on all output files. `study()` is required.
- `save(`*string*`)` specifies a directory where output files will be saved. `save()` is required.
- `patid(`*string*`)` specifies the variable containing the patient identifiers in the cohort dataset and data dimensions. `patid()` is required.

- `exposure(`*`varname`*`)` specifies the binary treatment or exposure variable. `exposure()` is required.

- `outcome(`*`varname`*`)` specifies the binary outcome variable. `outcome()` is required.

Output

A summary is reported displaying the specifications for the data dimensions declared. `hdps setup` saves a dataset called "study_cohort_info.dta" containing the patient identifier, treatment, and outcome variables.

5.5.4 The hdps prevalence command

`hdps prevalence` performs Step 2 of the HDPS algorithm, identifying the most prevalent codes within each dimension specified and calculating distribution cut-offs used to assess code recurrence. Additionally, for each patient, the command assesses the total frequency of each of the selected codes. To run `hdps prevalence`, data dimensions must have been previously specified using `hdps setup`.

Syntax

`hdps prevalence`, `top(#) nofilter`

Options (one of the following must be specified)

- `top(#)` specifying the number of codes to be selected from each dimension.

- `nofilter` calculates distribution cut-offs and patient frequencies for all available codes. This is following recommendations by *Schuster et al.* (2015) suggesting that a prevalence filter can result in the omission of codes, important for confounder adjustment, with a low marginal prevalence.

Output

The number of codes successfully selected from each dimension is reported in the Results Window. Two datasets are outputted: 1) a summary of the codes selected, reporting the median and upper quartile; used as cut-offs for defining the binary covariates generated ("study_feature_prevalence.dta"), and 2) the per patient code totals for each of the codes selected ("study_patient_totals.dta").

5.5.5 The hdps recurrence command

The `hdps recurrence` command performs Step 3 of the HDPS, creating binary covariates based on the cut-offs described in Section 2. `hdps recurrence` requires the two datasets created by the `hdps prevalence` command. This is presented as a separate command due to the possible computational burden in settings with a large number of dimensions or patients.

Syntax

```
hdps recurrence
```

Output

The total number of binary HDPS covariates generated is returned in the Results Window. The full set of covariates is outputted in a dataset called "study_hdps_covariates.dta".

5.5.6 The hdps priortize command

Finally, the `hdps prioritize` command is used to prioritize and perform variable selection on the set of covariates created by the `hdps recurrence` command (Section 5.4; Steps 4 and 5).

Syntax

`hdps prioritize`, `method(`*`string`*`)` `top(#)` `[zerocell]`

Options

- `method(`*`string`*`)` specifies the method of covariate prioritization. Available methods are 'bross' or 'exposure', as outlined in Section 5.4. `method()` is required.
- `top(#)` specifies the number of covariates to be selected. To obtain multiple datasets varying the number of covariates selected, a list of integers can be provided, e.g. top(200 500). `top()` is required.
- `zerocell` applies a correction of 0.1 to cells used in the calculation of the Bross. As described by (*Rassen et al.*, 2011b), covariates can not be considered for inclusion if the components of the Bross formula are undefined or equal to 0. In settings with few outcomes, this is particularly likely to affect RR_{CD}. Applying this correction therefore allows computation of these values and for covariates to remain under consideration.

Output

The `hdps prioritize` command outputs a dataset containing the data used to calculate the ranking information for each of the HDPS covariates ("study_bias_info.dta"). Additionally, a dataset containing the selected number of covariates (k) for each scenario specified in the `top()` option is outputted in the form "study_hdps_covariates_top_k.dta".

5.5.7 The hdps graphics command

The `hdps graphics` command is a standalone command for graphically assessing the properties of covariates generated and selected by the HDPS procedure. There are three graphical diagnostic tools available (illustrated in Section 5.6).

- **Bross:** inspects the distribution of ranked Bross values used for covariate prioritization (*Patorno et al.*, 2014). This plot requires specifying variables containing the bias ranking values and the numerical rank of covariates (`abs_log_bias` and `rank`, variables available in "study_bias_info.dta").

- **Prevalence:** investigates covariate balance by comparing the prevalence in the two treatment groups (*Franklin et al.*, 2015). This plot requires specifying variables containing these two prevalences (`pc1` and `pc0`, variables available in "study_bias_info.dta

- **Strength:** compares the relationship between the covariate-exposure (`ce_strength`) and covariate-outcome (`cd_strength`) associations, variables available in "study_bias_info.dta

Syntax

`hdps graphics` *varlist* `[if], type(`*string*`) dimension(`*varname*`)`*
`[pr(#)`+ `graph_options]`

where *varlist* corresponds to variables required by a specific plot type, as described above.

Options

- `type(`*string*`)` specifies one of three plot types: 'bross', 'prevalence' or 'strength' (described above). `type()` is required.

- `dimension(`*varname*`)` specifies a numeric variable identifying the dimension a covariate is derived from. * `dimension()` is only required for the 'prevalence' and 'strength' plot types.

- `pr(#)` optionally specifies a prevalence ratio. The prevalence ratio and its reciprocal will be plotted as dashed lines. If not specified, the default is to plot prevalence ratios of 2 and 0.5. + `pr()` is only an option for the 'prevalence' plot type.

- `graph_options` are any of the options documented in [**G-3**] twoway_options.

5.6 Example using simulated data

5.6.1 Simulated data

To illustrate the `hdps` suite we use a simulated cohort study design, representative of pharmacoepidemiological studies that employ HDPS approaches (summarised in Figure 5.2).

We have simulated a cohort dataset containing a patient identifier (`patid`), binary treatment variable (`trt`: 1 "Study Drug" 0 "Comparator Drug"), binary outcome variable (`outcome`: 1 "Yes" 0 "No") and a set of 9 confounders to mimic a priori investigator identified variables. Additionally, two HDPS data dimensions were simulated capturing clinical (International Classification of Disease, 10th edition codes; ICD10) and prescription (British National Formularly codes; BNF) features based on marginal prevalences observed in a previous study applying HDPS in UK EHRs (*Tazare et al.*, 2020). For the clinical data dimension, we have simulated an 'Ever' dimension capturing whether an individual has a record for a particular code in their entire history (i.e. irrespective of whether it occurs in the HDPS covariate assessment window).

These simulated datasets do not attempt to fully capture the complexity of a specific data source. Instead, they have been designed to illustrate the commands and expected data structures. These data have been simulated so that unbiased treatment effect estimation requires the inclusion of several data-derived HDPS covariates, which would be omitted in a standard analysis. After adjustment for the HDPS covariates, we expect the treatment effect to move towards the null.

Throughout the following tutorial we focus on a HDPS analysis with the following tuning parameters: 1) a prevalence filter selecting the top 100 features from each dimension, 2) prioritization using the Bross formula and 3) the top 100 covariates are selected for inclusion in the PS model.

Example data and the analysis code used throughout are available on GitHub: `http://www.github.com/johntaz/HDPS-Stata-Demo/`.

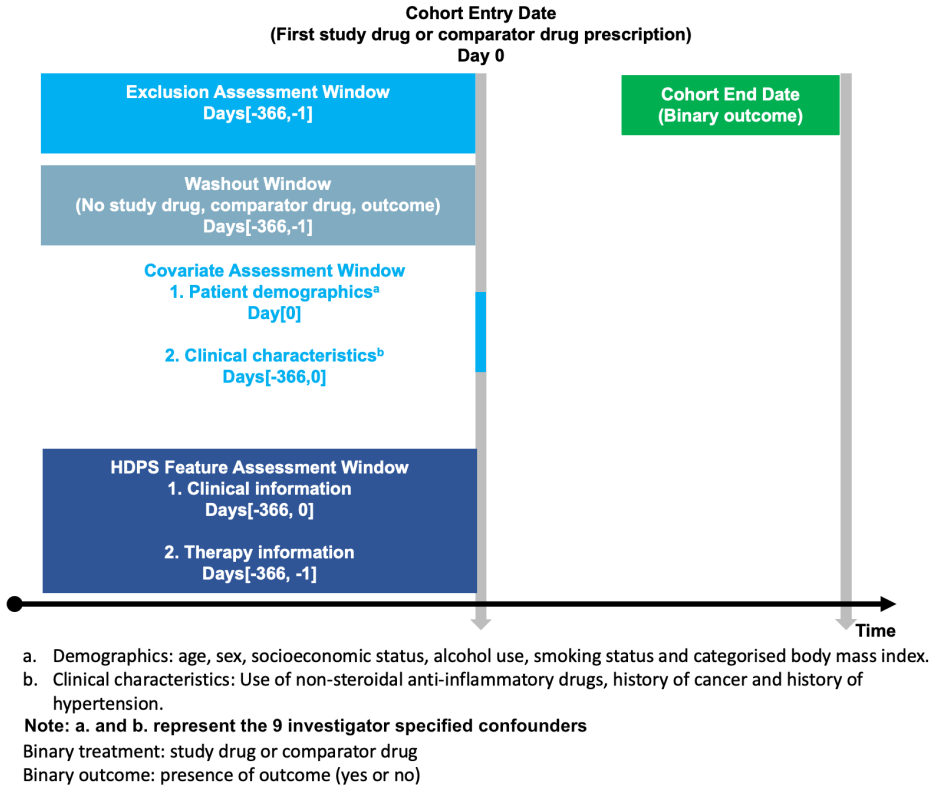


Figure 5.2: *Example cohort study illustrating the setting in which the HDPS algorithm is traditionally applied.*

5.6.2 High-dimensional propensity score procedure

First ensure that the current working directory includes the cohort dataset and relevant data dimensions. Load the cohort dataset containing the `outcome`, `trt` and `patid` variables required for the HDPS procedure. We use the `hdps setup` command to declare these variables and the two data dimensions, specifying the `ever` option for the clinical dimension.

```
. clear all

. use "cohort.dta", replace
(Artificial cohort data for HDPS suite)

. hdps setup (clinical_dim, icd10 ever) ///
>            (therapy_dim, bnf), ///
>            patid(patid) ///
>            exp(trt) ///
>            out(outcome) ///
>            study(example) ///
>            save(../output/)

Data dimensions identified (code variable):
   Dimension 1:     clinical_dim (icd10)
   Dimension 2:     therapy_dim (bnf)

Note: ´ever´ option specified at least once
Ever dimensions:
   Dimension 1:     clinical_dim_ever (icd10)

Output folder:
../output/
```

Next, we use the `hdps prevalence` command to identify the top 100 most prevalent features from each of the data dimensions. Note that we successfully select 100 features from each dimension.

```
. hdps prevalence, top(100)

Identifying most prevalent features:
Selecting top 100 from each dimension
   Dimension 1:     Completed: selected 100 features
   Dimension 2:     Completed: selected 100 features

Incorporating ´ever´ information:
   Dimension 1:     Completed

Output files:
(1) example_feature_prevalence.dta
(2) example_patient_totals.dta
```

We then run the `hdps recurrence` command, which assesses the frequency of patient feature recording to define as many as three binary covariates for each feature, using the cut-offs previously described. Note that the 200 features identified using `hdps`

`prevalence` results in 600 binary HDPS covariates.

```
. hdps recurrence
Loading data:
Completed
Generating HDPS covariates and assessing feature recurrence:
Progress: 0%...20%...40%...60%...80%...Completed
Number of binary HDPS covariates created:
600
Output file:
(1) example_hdps_covariates.dta
```

Next, we use the `hdps prioritize` command to select the most important covariates for confounder adjustment. In this instance, we create two datasets containing the top 50 and 100 covariates based on the Bross formula. Whilst the primary analysis focuses on the model selecting 100 covariates, this shows how easily we can obtain multiple datasets for testing the sensitivity of our results to the number of covariates chosen.

```
. hdps prioritize, method(bross) top(50 100)
Ranking HDPS covariates:
Prioritizing using the Bross formula:
Progress: 0%...20%...40%...60%...80%...Completed
Forming hd-PS cohort(s) based on top ranked covariates:
Selecting: 50, and 100.
Output files:
(1) example_bias_info.dta
(2) example_hdps_covariates_top_50.dta
(3) example_hdps_covariates_top_100.dta
```

We can now use the `hdps graphics` command to investigate the properties of the covariates generated and selected.

Having loaded the "bias_info" dataset, the first step is to investigate the distribution of ranking scores used to prioritize the covariates. This can be achieved by specifying the `bross` option and providing the ranking score variable and rank number variable, as below. We note from Figure 5.3 that there are several high-ranking covariates with relatively larger ranking scores, indicating possible importance for confounder adjustment.

```
. hdps graphics abs_log_bias rank if rank<=100, type(bross)
```

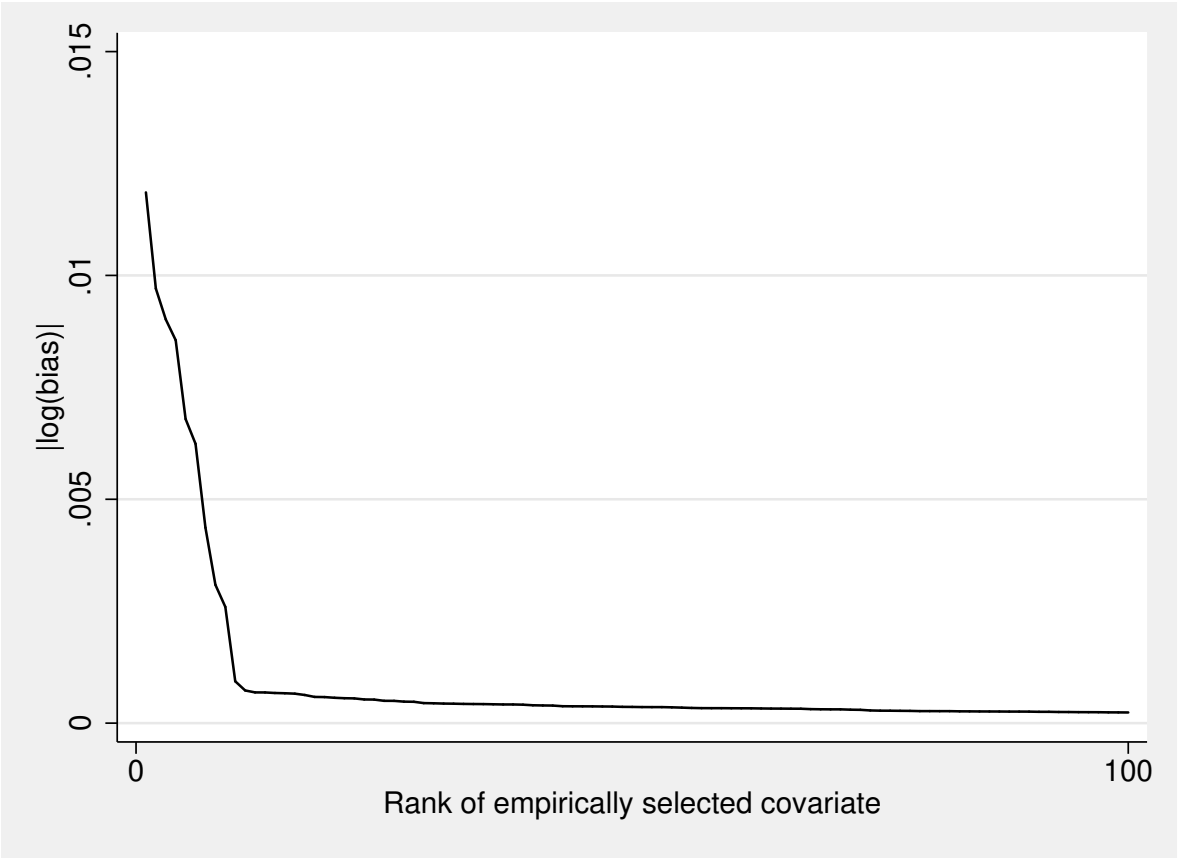


Figure 5.3: *Distribution of absolute log Bross bias values for each of the top 100 HDPS covariates.*

Next we investigate covariate balance by plotting covariate prevalence in the study drug and comparator drug groups (*Franklin et al.*, 2015). Figure 5.4 shows similar prevalence in the two groups whilst also highlighting which dimension covariates were derived from. The dashed lines represent prevalence ratios of 2 and 0.5 to visually highlight covariates with large imbalances between the treatment groups.

```
. hdps graphics pc1 pc0 if rank<=100, type(prevalence) ///
>                                     dim(dim) ///
>                                     legend(order(1 "Clinical" 2 "Prescription") ///
>                                     title("Data Dimensions", size(*0.8)) ///
>                                     cols(3) ///
>                                     rows(1) ///
>                                     ) ///
>                                     ytitle("Prevalence in study drug users") ///
>                                     xtitle("Prevalence in comparator drug users")
.
```

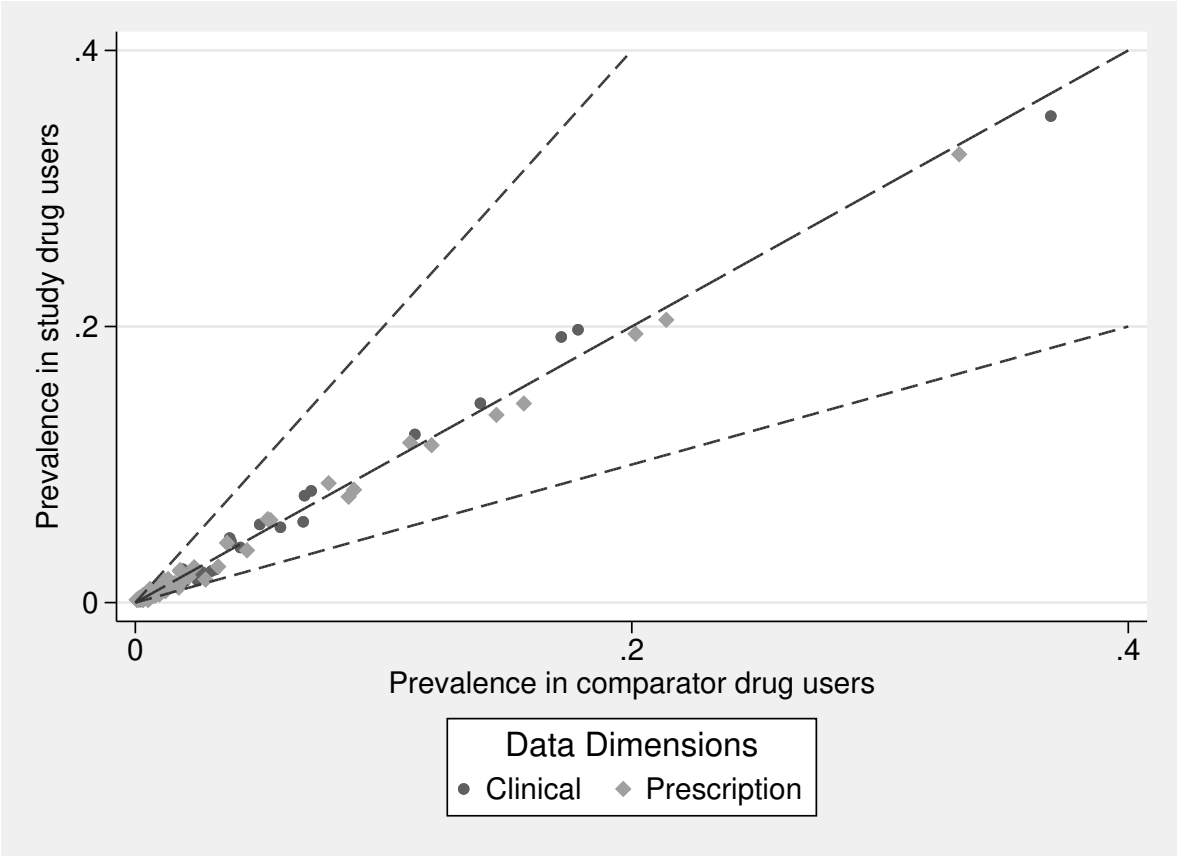


Figure 5.4: *Prevalence of the top 100 HDPS covariates by treatment group. The diagonal line indicates equal prevalence in both groups and the dashed lines show prevalence ratios of 0.5 and 2.0. The different symbols highlight which dimension the covariate was derived from.*

Finally, we inspect the relationship between the strength of covariate-exposure and covariate-outcome associations (defined as the absolute value of the relative association minus 1). In PS analysis the inclusion of covariates strongly related to the treatment but unrelated to the outcome, are known to increase variance (*Brookhart et al.*, 2006). Figure 5.5 can help indicate variables which empirically have these characteristics. Investigators may wish to perform sensitivity analyses assessing the impact of including these variables on the resulting treatment effects and confidence intervals.

```
. hdps graphics ce_strength cd_strength if rank<= 100, ///
>                                 type(strength) ///
>                                 dim(dim) ///
>                                 legend(order(1 "Clinical" 2 "Prescription") ///
>                                 title("Data Dimensions", size(*0.8)) ///
>                                 ) ///
>                                 ytitle("Strength of covariate-treatment association") ///
>                                 xtitle("Strength of covariate-outcome association")
.
```

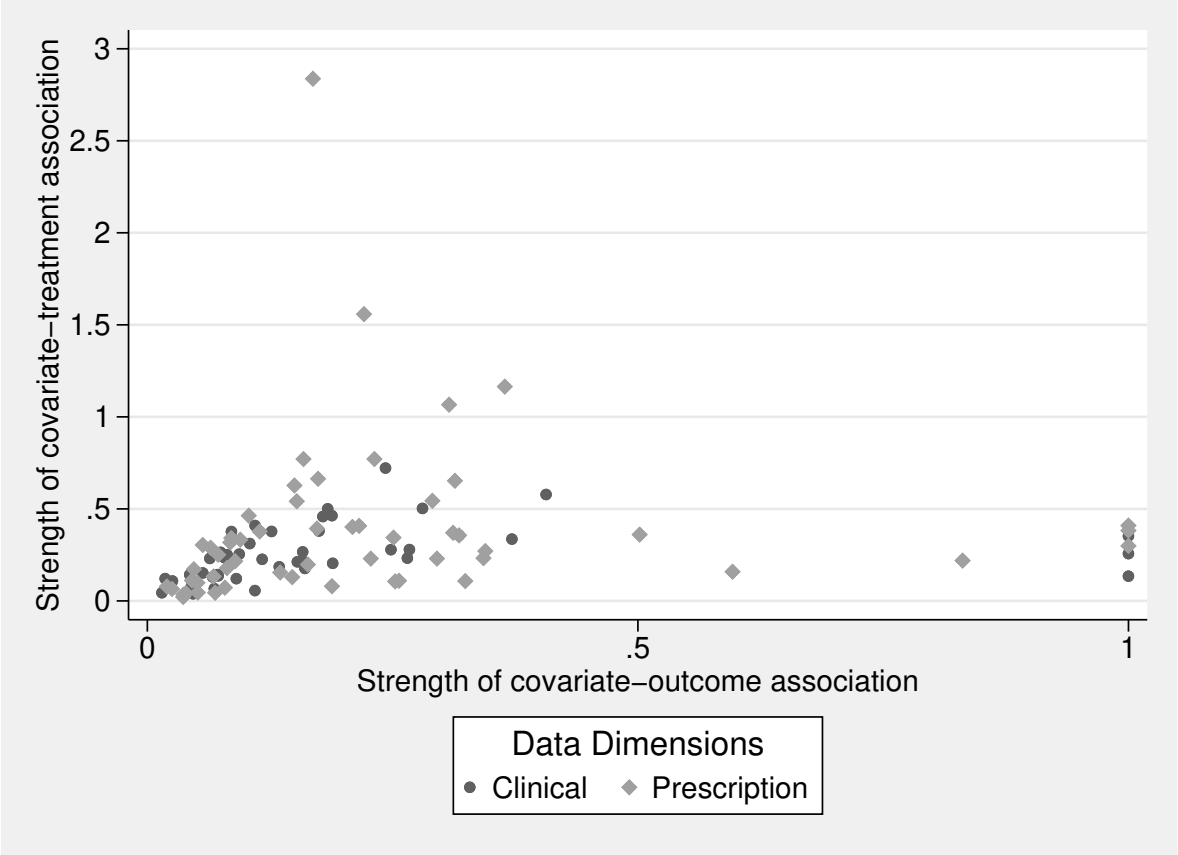


Figure 5.5: *Comparison of the strength of covariate-exposure and covariate-outcome associations for the top 100 bross ranked HDPS covariates. The different symbols highlight which dimension the covariate was derived from.*

5.6.3 Investigator propensity score analysis

In the HDPS literature, investigators often first perform a PS analysis using only the set of covariates identified by the investigators. This provides a useful baseline to compare the performance of subsequent models incorporating the HDPS covariates.

We begin by loading the cohort dataset and describing the variables.

```
. use "cohort.dta", replace
(Artificial cohort data for HDPS suite)

. describe

Contains data from cohort.dta
  obs:        10,000                          Artificial cohort data for HDPS suite
 vars:            12                          9 Apr 2021 18:43
 size:       490,000
-------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
patid           str5    %9s                   Patient Identifier
age             float   %9.0g                 Age at cohort entry
female          float   %9.0g      femalelab
                                               Female
ses             float   %9.0g      lowmedhigh
                                               Socio-Economic Status
smoke           float   %9.0g      smokelab   Smoking status
alc             float   %9.0g      lowmedhigh
                                               Alcohol consumption
bmicat          float   %9.0g      bmilab     Categorised Body Mass Index
nsaid_rx        float   %9.0g      yesno      Previous NSAID prescription
cancer          float   %9.0g      yesno      History of Cancer
hyper           float   %9.0g      yesno      History of Hypertension
trt             float   %9.0g
outcome         float   %9.0g
-------------------------------------------------------------------------------
Sorted by: patid
```

To estimate the PS we fit a logistic regression, modelling the treatment variable on the set of 9 confounders. Whilst other methods, such as matching and stratification are available, we focus on incorporating the PS using inverse probability of treatment weights and these are generated below (*Austin*, 2011; *Williamson and Forbes*, 2014).

Next, we use a weighted logistic regression model to estimate the treatment odds ratio (OR). We apply robust standard errors to acknowledge the lack of independence in the weighted population (*Hernán et al.*, 2000). However, note that theoretically the variance should account for the estimation of the PS. Our models do not account for this. As a result, the confidence intervals will be slightly conservative (*Williamson et al.*, 2012, 2014).

Whilst we have focused on a binary outcome, these methods can similarly be applied

```
. logit trt age female ses smoke alc bmicat nsaid_rx cancer hyper
Iteration 0:   log likelihood = -6595.9125
Iteration 1:   log likelihood = -6589.5645
Iteration 2:   log likelihood = -6589.5637
Iteration 3:   log likelihood = -6589.5637
Logistic regression                               Number of obs     =     10,000
                                                  LR chi2(9)        =      12.70
                                                  Prob > chi2       =     0.1768
Log likelihood = -6589.5637                       Pseudo R2         =     0.0010

------------------------------------------------------------------------------
         trt |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0015863   .0026061    -0.61   0.543    -.0066941    .0035215
      female |   -.046352   .0424815    -1.09   0.275    -.1296142    .0369102
         ses |   .0228782   .0389346     0.59   0.557    -.0534322    .0991886
       smoke |  -.0659094   .0694756    -0.95   0.343    -.2020791    .0702602
         alc |   .0887163   .0663556     1.34   0.181    -.0413384    .2187709
      bmicat |   .0049882    .046912     0.11   0.915    -.0869576    .0969341
    nsaid_rx |  -.0454822   .0451712    -1.01   0.314    -.1340161    .0430518
      cancer |   .0544823   .0518726     1.05   0.294    -.0471861    .1561507
       hyper |  -.0875655   .0421824    -2.08   0.038    -.1702414   -.0048895
       _cons |   .6215302   .1473132     4.22   0.000     .3328016    .9102589
------------------------------------------------------------------------------

. predict pscore, pr
. gen wts = 1/ps if trt == 1
(3,712 missing values generated)
. replace wts = 1/(1-ps) if trt == 0
(3,712 real changes made)
```

for a time-to-event outcome. The binary outcome indicator would be used throughout the HDPS procedure to select the HDPS covariates. In the PS analysis, the outcome model would be the appropriate survival model.

```
. logistic outcome i.trt [pw=wts], vce(robust)
Logistic regression                               Number of obs     =     10,000
                                                  Wald chi2(1)      =       5.26
                                                  Prob > chi2       =     0.0219
Log pseudolikelihood = -13720.103                 Pseudo R2         =     0.0004

------------------------------------------------------------------------------
             |               Robust
     outcome | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       1.trt |   1.100317   .0458824     2.29   0.022     1.013966    1.194022
       _cons |   1.206206   .0398105     5.68   0.000     1.130649    1.286813
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

For the investigator analysis, we obtain some evidence supporting an increased risk of the outcome in those receiving the study drug compared to the those receiving the comparator drug (OR 1.10; 95% CI: 1.01 to 1.19).

5.6.4 High-dimensional propensity scores analysis

We now illustrate how to incorporate the selected HDPS covariates into a PS analysis.

Ensure the cohort dataset is still loaded into memory. The first step is to either drop or rename the previous `pscore` and `wts` variables as we will now re-estimate the PS. We need to `merge` the generated set of 100 HDPS covariates to the cohort datatset using the patient identifier (`patid`). As before, we fit a logistic regression model to estimate the propensity score and now additionally include the HDPS covariates in this model (the prefixes "d1" and "d2" represent covariates derived from the clinical and prescription dimensions, respectively). For brevity, we suppress the output from the logistic regression model containing 109 covariates. However it is important to inspect large models, especially in small samples, where covariates might perfectly predict treatment allocation. Furthermore, note that when adjusting for several hundred HDPS covariates it may be necessary to increase the maximum matrix size in Stata, for more details see

```
. drop pscore wts
. merge 1:1 patid using "../output/example_hdps_covariates_top_100.dta", assert(match) nogen
    Result                           # of obs.
    -----------------------------------------
    not matched                             0
    matched                            10,000
    -----------------------------------------
. logit trt age female ses smoke alc bmicat nsaid_rx cancer hyper d1* d2*
  (output omitted)
```

We now estimate the propensity score and generate new inverse probability of treatment weights, before estimating the treatment effect using a weighted logistic regression model.

For the HDPS analysis, we observe that the inclusion of the HDPS covariates has led to a result closer to the expected null association (OR 1.03; 95% CI: 0.94 to 1.11).

As previously mentioned, the number of covariates selected is a key decision in the HDPS procedure and we recommend testing the sensitivity of results to this decision. The analysis outlined above can easily be repeated for a different set of covariates by updating the `merge` file.

```
. predict pscore, pr
. gen wts = 1/ps if trt == 1
(3,712 missing values generated)
. replace wts = 1/(1-ps) if trt == 0
(3,712 real changes made)
. logistic outcome i.trt [pw=wts], vce(robust)
Logistic regression                               Number of obs     =     10,000
                                                  Wald chi2(1)      =       0.35
                                                  Prob > chi2       =     0.5513
Log pseudolikelihood = -13712.248                 Pseudo R2         =     0.0000

------------------------------------------------------------------------------
             |               Robust
     outcome | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       1.trt |   1.025467   .0432824     0.60   0.551     .9440489    1.113907
       _cons |   1.261751    .042279     6.94   0.000     1.181548    1.347398
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

5.7 Discussion

In this article we have introduced the `hdps` suite of commands for applying the HDPS algorithm in Stata. This suite consists of 5 commands for generating, prioritizing, and visualizing the properties of HDPS covariates. We have illustrated these commands using simulated data and demonstrated how to incorporate the resulting HDPS covariates within a propensity score analysis.

For illustrative purposes, the analysis presented is based on data simulated with a relatively simple structure. In practice, there will be complex relationships between the codes identified and investigators will often specify many more data dimensions. The plasmode framework has become a popular method for simulating data more reflective of large healthcare databases and is often used to evaluate the performance of methods in this setting (*Franklin et al.*, 2014).

The main benefit of HDPS methods can be seen in settings where information recorded within the healthcare databsase is likely to be strongly correlated to key confounders that are hard to measure. However, in settings with a well-established or basic confounding structure, the HDPS is not likely to outperform traditional PS or outcome regression methods. Furthermore, it is important to acknowledge that unmeasured confounding may remain an issue even after adjustment for HDPS covariates.

Methodological work surrounding HDPS methods continues to develop rapidly and any new features in the `hdps` suite will aim to reflect best practices, as they becomes apparent. A recent review summarises key areas of development (*Schneeweiss*, 2018). One topic of growing interest surrounds the possible benefits of combining HDPS and machine learning approaches (*Franklin et al.*, 2015; *Karim et al.*, 2018; *Schneeweiss et al.*, 2017; *Tian et al.*, 2018).

The `hdps` suite will be updated and developed, and we would welcome suggestions for improvements and new features. We are also interested in how the data management commands presented might be used to create data-driven covariates in alternative contexts, e.g. prediction modelling (*Franklin et al.*, 2016).

5.8 Acknowledgments

We thank Tim P. Morris for his suggestions and help surrounding the design of this package.

5.9 Supporting information

Help files for the developed Stata commands

Title

hdps Suite of commands for 1) performing data manipulation and variable selection steps of the high-dimensional propensity score (HDPS) algorithm, 2) graphically assess the properties of selected covariates.

Syntax

hdps setup	Identify data dimensions and key patient variables
hdps prevalence	Applies a prevalence filter to features within each dimension
hdps recurrence	Creates binary covariates based on feature recording frequency
hdps prioritize	Prioritizes covariates and selects a subset(s) for analysis
hdps graphics	Standalone command for investigating the properties of selected covariates

Description

The HDPS algorithm is a multi-step procedure for confounder generation and selection in large healthcare databases.

Step 1: Declare data dimensions and key variables (**hdps setup**)

Step 2: Apply a feature prevalence filter within each dimension (**hdps prevalence**)

Step 3: Assess the recurrence of features based on data driven cut-offs (creating a large pool of binary HDPS covariates) (**hdps recurrence**)

Steps 4 & 5: Prioritize the set of binary HDPS covariates and select a subset to incorporate into a propensity score analysis (**hdps prioritize**)

Step 6: Graphically investigate the properties of selected covariates (**hdps graphics**)

Step 7: Apply a traditional propensity score analysis

The **hdps** suite of programs conducts steps 1 - 6 of the high-dimensional propensity (HDPS) algorithm.

Data formats

There are two data formats to be discussed, both of which are needed to run **hdps setup** and the subsequent commands.

The first relates to the study dataset. This is expected to be a cohort, formatted to 1 observation per patient. Additionally, it must contain a patient identifier and binary exposure and outcome variables.

Cohort format

patid	*exposure*	*outcome*
1001A	1	0

1002A	0	0
1003A	1	1
1004A	0	1

The second relates to the data dimensions. This is expected to be a *long* format dataset with many observations per patient and feature (e.g. code). To incorporate 'ever' data, separate dimensions should be specified containing these data.

Dimension format (using International Classification of Disease Edition 10 (ICD-10) codes as an example)

patid	*icd10*
1001A	W22
1001A	W22
1001A	X52
1003A	V97
1004A	W61
1004A	Y92

Examples

Code and data for the examples below are available at https://github.com/johntaz/HDPS-Stata-Demo.

Change the current directory to the folder with HDPS data dimensions and load the cohort data

```
. use "cohort.dta", clear
```

Declare data dimensions, output folder and key patient variables

```
. hdps setup (clinical_dim, icd10 ever) (therapy_dim, bnf), patid(patid) exp(trt) out(outcome) study(example)
save(../output/)
```

Apply the prevalence filter, selecting the top 100 most prevalent codes from each dimension

```
. hdps prevalence, top(100)
```

Assess recurrence of the selected codes for each patient in the pre-exposure covariate window

```
. hdps recurrence
```

Prioritize covariates using the Bross formula and select the top 100 covariates

```
. hdps prioritize, method(bross) top(100)
```

Methodological considerations

The HDPS was developed in claims data by Schneeweiss et al (2009). Whilst typically the Bross formula is used to prioritize covariates in the HDPS algorithm, ranking covariates by the strength of covariate-exposure relationship has been suggested in small samples with rare outcomes (Rassen et al, 2011).

Developments to the cut-offs described by Schneeweiss et al (applied in **hdps recurrence**), incorporating pre-exposure information recorded across a patient's entire medical history (so-called 'ever' information) have been proposed by Tazare et al (2020) and are implemented in this suite of commands.

hdps prevalence applies a prevalence filter restricting further steps of the HDPS to only the most prevalent features within each dimension. There is ongoing debate in the literature surrounding the use of a prevalence filter (Schuster et al, 2015) and users of **hdps** have the option to consider all features throughout the procedure. However, it should be noted that selecting all features will, in some cases, substantially add to the computational burden.

hdps prioritize allows the user to select a number of covariates based on covariate prioritization. Whilst convention often leads to the selection of 200 and 500 covariates, it is unclear what the optimal number is for a given setting (Patorno et al, 2014). We recommend assessing the sensitivity of results to the number of covariates selected.

Limitations

With large dimension files and numbers of patients, you may run into memory problems. Where possible, ensure you have reduced the size of data dimensions prior to running the **hdps** commands.

If your version of Stata allows you to do so, you may need to increase matsize in order to fit propensity score models containing a large number of covariates.

Please report any other problems to john.tazare1@lshtm.ac.uk or submit an issue to the GitHub page for this project http://www.http://github.com/johntaz/hdps.

References

Schneeweiss S, Rassen JA, Glynn RJ, Avorn J, Mogun H, Brookhart MA. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data [published correction appears in Epidemiology. 2018 Nov;29(6):e63-e64]. Epidemiology, 20(4):512-522, 2009

Rassen JA, Glynn RJ, Brookhart MA, Schneeweiss S. Covariate selection in high-dimensional propensity score analyses of treatment effects in small samples. Am J Epidemiol, 173(12):1404-1413, 2011

Tazare J, Smeeth L, Evans SJW, Williamson EJ, Douglas IJ. Implementing high-dimensional propensity score principles to improve confounder adjustment in UK electronic health records. Pharmacoepidemiol Drug Saf 29:1373-1381, 2020

Schuster T, Pang M, Platt RW. On the role of marginal confounder prevalence - implications for the high-dimensional propensity score algorithm. Pharmacoepidemiol Drug Saf 24(9):1004-1007, 2015

Patorno E, Glynn RJ, Hernández-Díaz S, Liu J, Schneeweiss S. Studies with many covariates and few outcomes: selecting covariates and implementing propensity-score-based confounding adjustments. Epidemiology 25(2):268-78, 2014

Authors

John Tazare, Liam Smeeth, Stephen JW Evans, Ian J Douglas and Elizabeth J Williamson; London School of Hygiene & Tropical Medicine, UK.

Contact Email: john.tazare1@lshtm.ac.uk.

You can see the latest updates and discussions surrounding the **hdps** suite on GitHub and Twitter

Title

hdps setup — Specifies data dimensions and key patient variables for the HDPS procedure

Syntax

hdps setup *dimension(s)* [, *options*]

Dimensions are specified using the filename:

(*filename*, *varname* [*ever*])

Dimension syntax

filename specifices file name for data dimension

varname specifies variable in the data dimension containing the codes. This option is required and must be the first option specified

ever optionally specifies that the recurrence assessment should incorporate `Ever' information

options	Description
Required	
save(*string*)	Output directory
study(*string*)	Study name and output prefix
patid(*varname*)	Patient identifier variable
exposure(*varname*)	Exposure variable
outcome(*varname*)	Outcome variable

Description

The hdps setup command declares the data dimensions and key variables used through- out the HDPS procedure, further specifying the directory for outputted datasets. Set the current directory to a folder containing all necessary data and load the cohort dataset into memory before running this command.

Options

Main

save(*string*) specifies a study name that serves as a prefix on all output files. This option is required.

study(*string*) specifies a directory where output files will be saved. This option is required.

patid(*varname*) specifies variable containing the patient identifiers in cohort data set and data dimensions. This option is required

exposure(*varname*) specifies the binary treatment or exposure variable. THis option is required.

outcome(*varname*) specifies the binary outcome variable. This option is required.

Output

"study_cohort_info.dta" contains the patient identifier, treatment and outcome variables.

Examples

Change current directory to folder containing cohort and data dimension datasets

Load cohort dataset

```
. use "cohort.dta", replace
```

Specify two data dimensions (clinical and therapy) with Ever incorporated for clinical dimensions.

```
. hdps setup (clinical_dim, icd10 ever) (therapy_dim, bnf), patid(patid) exp(trt) out(outcome) study(example)
  save(../output/)
```

Return to main help page for hdps

Title

hdps prevalence — Applies a code prevalence filter within each dimension (Step 2)

Syntax

hdps prevalence [, *options*]

options	Description
Optional - one of the following must be specified	
top(*integer***)**	Number of codes to be selected from each dimension
nofilter	No filter applied (all codes assessed)

You must run **hdps setup** before using **hdps prevalence.**

Description

hdps prevalence performs Step 2 of the HDPS procedure, identifying the most prevalent codes within each dimension and calculating distribution cut-offs used to assess code recurrence. The command also, for each patient, assesses the total frequency of each of the selected codes.

Options

Main

top(*integer***)** specifies the number of codes to be selected from each dimension.

nofilter calculates distribution cut-offs and patient frequencies for all available codes.

Output

The number of codes successfully selected from each dimension is reported in the Results Window.

"study_feature_prevalence.dta" contains a summary of the codes selected.

"study_patient_totals.dta" reports the per patient code totals for each of the codes selected.

Examples

Select top 100 most prevalent codes from each dimension

```
. hdps prevalence, top(100)
```

Select all codes from each dimension

```
. hdps prevalence, nofilter
```

Return to main help page for hdps

Title

hdps recurrence — Creates a pool of HDPS covariates and assesses the frequency of recording (Step 3)

Syntax

hdps recurrence

You must run **hdps setup** and **hdps prevalence** before using **hdps recurrence.**

Description

Using information about the distributions of features across the data dimensions, **hdps recurrence** generates HDPS covariates. For each of the features, up to three binary covariates are generated: '*code*_once', '*code*_spor' and '*code*_freq' as described by Schneeweiss et al (2009).

If during **hdps setup** the 'ever' option is specified for a particular data dimension, the bottom cut-off generated will be '*code*_ever'. For more details, see Tazare et al (2020).

Based on the frequency a feature is recorded during an individual's pre-exposure period, **hdps recurrence** then assigns either 1 or 0 for each HDPS covariate generated.

Output

The total number of binary HDPS covariates generated is return in the Results Window.

"study_hdps_covariates.dta" is returned in the specified output folder.

Return to main help page for hdps

Title

hdps prioritize — Prioritizes covariates and selects a subset(s) for analysis (Steps 4 & 5)

Syntax

hdps prioritize [, *options*]

options	Description
Required	
method(*string***)**	Covariate prioritization method
top(*numlist***)**	Number of covariates to select
Optional	
zerocell	Applies zero cell correction in calculation of Bross formula

Description

hdps prioritize prioritizes and performs variable selection on the set of covariates created by the **hdps recurrence command.**

Options

Main

method(*string***)** specifies the method of covariate prioritization. Available methods are 'bross' or 'exposure'.

top(*numlist***)** specifies the number of covariates to be selected. To obtain multiple datasets varying the number of covariates selected, a list of integers can be provided, e.g. top(200 500).

zerocell applies a correction of 0.1 to cells used in the calculation of the Bross formula. This is useful in settings with few outcomes, where computation of these values can be challenging.

Output

The hdps prioritize command outputs a dataset containing the data used to calculate the ranking information for each of the HDPS covariates ("study_bias_info.dta").

"study_hdps_covariates_top_k.dta" is return in the specified output folder containing the selected number of covariates (k) for each scenario specified.

Examples

Prioritize covariates using the Bross formula and select the top 100 covariates

```
. hdps prioritize, method(bross) top(100)
```

Prioritize covariates using the strength of exposure association and select the top 100 covariates

```
. hdps prioritize, method(exposure) top(100)
```

Select sets including the top 50 and 100 covariates

```
. hdps prioritize, method(bross) top(50 100)
```

Return to main help page for hdps

Title

hdps graphics — Graphically investigates the properties of selected covariates (Step 6)

Syntax

hdps graphics varlist(min = 2) [if] [, *options*]

options	Description
Required	
type(*string*)	Plot type (bross, prevalence or strength)
dimension(*varname*)	Dimension identifier (only required for prevalence or strength types)
Optional	
pr(#)	Plot prevalence ratios ('prevalence' plots only)
graph_options	Twoway graph options **twoway_options**

Description

The hdps graphics command is a standalone command for graphically assessing the properties of covariates generated and selected by the HDPS procedure. There are three graphical diagnostic tools available:

Bross: inspects the distribution of ranked Bross values.

Prevalence: compares the prevalence of selected codes in the two treatment groups.

Strength: compares the relationship between the covariate-exposure and covarite-outcome association strengths.

Options

Main

type(*string*) specifies one of three plot types: 'bross', 'prevalence' or 'strength'. Only one type can be specified at a time. This option is required.

dimension(*varname*) specifies a numeric variable identifying the dimension a covariate is derived from. Note this option is only required for 'prevalence' and 'strength' plot types.

pr(#) optionally specifies prevalence ratios. The prevalence ratio and its reciprocal will be plotted as dashed lines. The default is prevalence ratios of 2 and 0.5 Note pr() is option an option for the `prevalence' plot type.

graph_options are any of the options in **twoway_options**.

Variables

hdps prioritize returns a data set called "study_bias_info.dta" that stores variables which can be used to generate these visualizations. The following variables are available:

abs_log_bias stores the Bross bias values used to rank HDPS covariates.

rank stores the rank of each HDPS covariate.

pc1/pc0 stores the prevalence of codes in the exposure and unexposred groups.

ce_strength/cd_strength stores the strength of association between HDPS covariate and a) exposure (ce), b) outcome (cd).

Examples

Load bias information dataset and generate a dimension identifier

```
. use "../output/example_bias_info.dta", clear
. gen dimension=substr(code,1,2)
. encode dimension, gen(dim)
```

Inspect Bross values for top 100 selected covariates

```
. hdps graphics abs_log_bias rank if rank<=100, type(bross)
```

Compare prevalence in treated and untreated groups

```
. hdps graphics pc1 pc0, type(prevalence) dim(dim)
. hdps graphics pc1 pc0 if rank<=100, type(prevalence) dim(dim)
```

Compare covariate-exposure and covariate-outcome association strengths

```
. hdps graphics ce_strength cd_strength, type(strength) dim(dim)
. hdps graphics ce_strength cd_strength if rank<=100, type(strength) dim(dim)
```

Return to main help page for hdps

Chapter 6

High-dimensional propensity score analysis of upper GI bleed risk in NSAID and COX-2I users

John Tazare[1], Daniel C Gibbons[2], Liam Smeeth[1,3], M Sanni Ali[1], Elizabeth Williamson[1,3], Ian J Douglas[1,3], John Logie[2]

1. London School of Hygiene and Tropical Medicine, London, UK.
2. GlaxoSmithKline, London UK.
3. Health Data Research (HDR) UK, London, UK.

6.1 Overview

Summary

In Chapter 3, I introduced modifications for translating high-dimensional propensity score (HDPS) principles to UK electronic health records (EHRs). These modifications performed well in a case-study that explored the interaction between clopidogrel and proton pump inhibitors (PPIs), obtaining results closer to the expected null association. In this chapter, I further assess the modified-HDPS by applying it to a question extensively investigated in both randomised controlled trials and observational studies. The association between non-steroidal anti-inflammatory drug and cyclo-oxygenase-2 inhibitor use on the risk of upper gastrointestinal bleeding (UGIB) is often used as a case-study for investigating new methodology in pharmacoepidemiology and this example has been applied in a number of settings to assess the performance of the HDPS compared to investigator-led propensity score models. Furthermore, the majority of studies applying the HDPS in this context concluded that the HDPS successfully captured subtle risk factors for UGIB (thought to be the primary mechanism for residual confounding). This is therefore an important case study testing the ability of the proposed modifications to replicate this widely studied association.

Thesis objective addressed

This chapter addresses the following objective of the overall thesis (Section 1.3):

3. Apply and discuss the use of these modification when applying the HDPS to UK EHRs.

Role of candidate

All authors were involved in the study design. Daniel Gibbons and I extracted the datasets. I performed the data management to create analysis-ready datasets and lead

the statistical analysis. I conducted the HDPS analysis, developing `R` code for implementing the modifications described in Chapter 3. All authors interpreted the results and contributed to the write up of this chapter. This is part of a larger project comparing the active-comparator new user and prevalent new user designs. In this chapter we focus on the results obtained from applying the HDPS in the active-comparator new user design.

6.2 Introduction

Non-steroidal anti-inflammatory drugs (NSAIDs) and cyclo-oxygenase-2 (COX-2) inhibitors are commonly used for the long-term treatment of persistent pain and inflammation, especially in patients with chronic conditions such as osteoarthritis and other musculoskeletal disorders. However, historically there have been safety concerns surrounding the use of these drugs, including surrounding the risk of gastrointestinal complications (*Pham and Hirschberg*, 2005). Evidence arising from randomised controlled trials comparing NSAID and COX-2 inhibitor use on the risk of upper gastrointestinal bleeding (simplified to GI bleed throughout) suggest that COX-2 inhibitors reduce the risk of UGIB by between 10 and 25% (*Bombardier et al.*, 2000; *Silverstein et al.*, 2000).

Despite these results suggesting that COX-2 inhibitor use can result in less gastrointestinal toxicity compared to traditional NSAIDs, in practice there appeared to be a higher incidence of upper GI disorders amongst users of COX-2 inhibitors (*Martin et al.*, 2000). It was hypothesised that COX-2 inhibitors were being prescribed to patients at high risk of GI complications, leading to channelling bias (*MacDonald et al.*, 2003). Furthermore, after accounting for GI risk factors the results obtained were much closer to those obtained in the randomised controlled trials (*MacDonald et al.*, 2003).

In large healthcare databases, it can be hard to accurately measure some of these risk factors and this has motivated the extensive use of this example for testing the performance of the high-dimensional propensity score (HDPS) in a diverse range of data sources, including: the US, Germany, Denmark and UK (*Garbe et al.*, 2013; *Hallas and Pottegard*, 2017; *Schneeweiss et al.*, 2009; *Toh et al.*, 2011). In all case studies, the HDPS performed similarly to the investigator model or obtained results closer to those from randomised controlled trials (*Schneeweiss*, 2018).

In this chapter, we study this association in the Clinical Practice Research Datalink (CPRD) using an active-comparator new user design to investigate the ability of the modified-HDPS (introduced in Chapter 3) to obtain comparable results to trials and other pharmacoepidemiological literature applying the HDPS.

6.3 Methods

6.3.1 Data source

The Clinical Practice Research Datalink (CPRD) GOLD is a de-identified primary care database broadly representative of patients registered with General Practitioners in the United Kingdom. This database includes data pertaining to prescribing, diagnosis, referrals and some lifestyle factors for approximately 9% of the UK population (*Herrett et al.*, 2015). The CPRD was used to identify patients with prescriptions for either of the study drugs and establish relevant osteoarthritis diagnoses, baseline characteristics, HDPS dimensions and follow-up. Hospital Episode Statistics (HES) is a records-based system including information on admissions, outpatient appointments and Accident & Emergency attendances per period of care at NHS hospitals in England (*NHS Digital*, 2020). Linked HES data was used for defining the outcome of interest (GI bleed leading to hospitalisation or death), covariates and a HDPS dimension. Office for National Statistics (ONS) mortality data was used to accurately ascertain date and cause of death. Additionally, we linked to Patient Level Index of Multiple Deprivation and Rural-Urban Classification at LSOA level to identify additional baseline covariates.

6.3.2 Study population

The study population consisted of patients with osteoarthritis aged 18 years or older, who initiated NSAIDs or COX-2 inhibitors between 1st January 2000 and the 31st December 2004 (the date of this prescription was considered the index date and date of cohort entry). We chose to conduct our study within a population of patients with osteoarthritis since these are patients more likely to use these treatments chronically. Patients were required to have at least 12 months of up-to-standard data available prior to cohort entry; allowing us to adequately assess baseline confounder information. Furthermore, a washout window excluded patients with a prescription for either NSAIDs or COX-2 inhibitors in the previous 365 days. Patients with a diagnosis for cirrhotic liver disease in their medical history prior to cohort entry were excluded (since cirrhosis

is a known cause of UGIB).

Patients were followed and censored at the earliest of outcome occurrence, study end date, death, incident cirrhotic liver disease, transfer out of practice, treatment switching or treatment discontinuation. Treatment discontinuation was defined as absence of a refill prescription 30 days after the end of the previous prescription.

Codelists defining osteoarthritis, NSAID and COX-2 inhibitor use, cirrhotic liver disease and upper GI bleed are included in Supporting Information Tables S1-S4.

The study population is summarised in Figure 6.1.

6.3.3 Exposure

All patients receiving first prescription for NSAIDs or COX-2 inhibitors between 1st January 2000 and 31st December 2004. NSAID and COX-2 inhibitor was defined using product codes (Supporting Information, Tables S1 and S2). We focused on product codes referring only to oral use of either NSAIDs or COX-2 inhibitors.

Continuous exposure was defined as: 1st Prescription Date + Duration of Prescription + Duration of any successive overlapping prescriptions (of same drug) + 30 days

6.3.4 Covariates

We adjusted for the following variables in the traditional propensity score (PS) model:

- Demographics: Age, sex, index of multiple deprivation score rank decile, body mass index (BMI), smoking status, alcohol consumption
- Comorbidities/ behaviours: Hypertension, chronic renal failure, inflammatory bowel disease, gastrointestinal tract tumours, coagulopathies, gastro-oesophageal reflux disease, diabetes, heart failure, previous upper GI bleed (defined in Read and ICD-10), number of hospital admissions in previous 6 months.

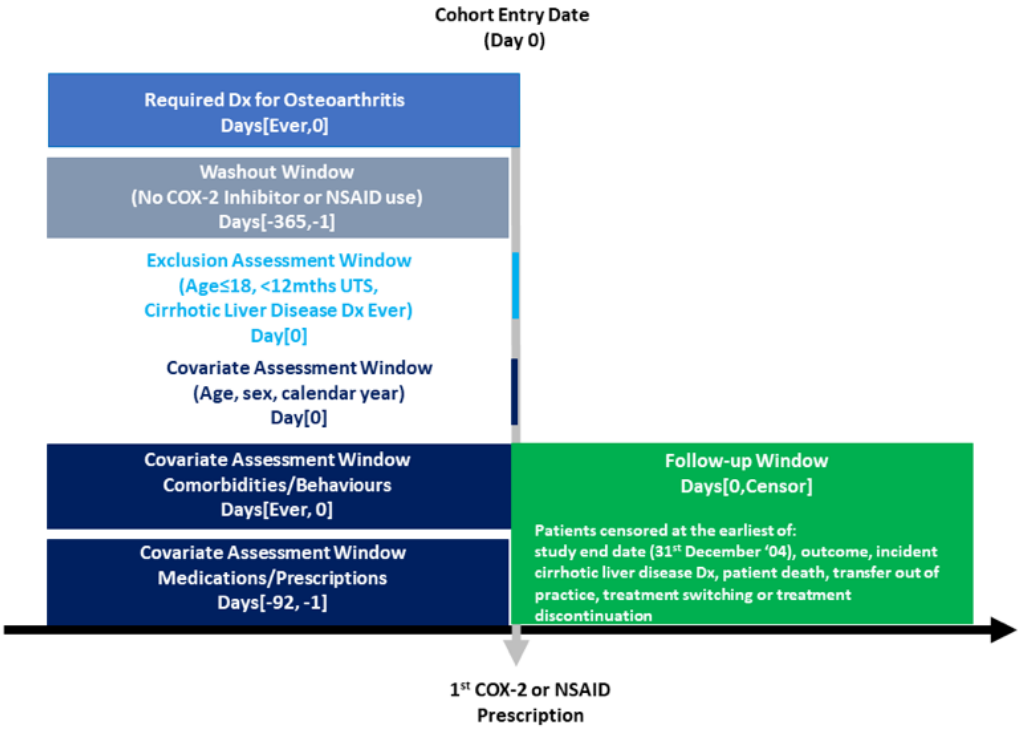


Figure 6.1: *Schematic showing active comparator study design for a study of NSAID and COX-2 inhibitor use on upper GI bleeding risk. Covariate assessment window only refers to covariates in the traditional propensity score model.*

- Medications/therapies (any recording in the 3 months prior to cohort entry): anticoagulants, systemic corticosteroids, proton pump inhibitors, H2 antagonists, coronary angioplasty, selective serotonin reuptake inhibitors, statins and clopidogrel
- Other: Calendar year

6.3.5 Outcomes

The study outcome was the first occurrence of an upper GI bleed leading to hospitalisation or death, as recorded in HES and ONS Mortality Data. This was defined as a binary variable and analysed in a time-to-event framework.

6.3.6 Statistical analysis

We analysed the hazard ratio (HR) for the association between COX-2 inhibitor and NSAID use on upper GI bleeding risk using Cox models, adjusting for confounders using propensity scores (*Williamson et al.*, 2012). The propensity score will be estimated using multivariable logistic regression to model the relationship between treatment and potential confounders. A propensity score-matched cohort was created by matching each COX-2 inhibitor user to an NSAID user using a nearest-neighbour matching algorithm (with no replacement and a caliper of 0.05); this approach estimated the average treatment effect in the treated (ATT) (*Stuart*, 2010).

Given previous research using the CPRD, we anticipated missing covariate data for categorised BMI, smoking status and alcohol consumption. A missing indicator approach was used for missing covariate information (*Blake et al.*, 2020; *Groenwold et al.*, 2012). The missing indicator approach adds a 'missing' category to these variables, allowing for the inclusion of patients with missing values (*Blake et al.*, 2020; *Groenwold et al.*, 2012). Despite historically being considered an unprincipled approach (and likely to yield biased results) (*Greenland and Finkle*, 1995; *Groenwold et al.*, 2012), recent work has highlighted that the missing indicator approach can be applied in a principled way

in the context of PS analysis (*Blake et al.*, 2020). Specifically, the missing indicator method is valid under the key assumption that a variable is only a confounder when observed (*Blake et al.*, 2020). In this study, this implies that the value for these variables only contribute to the treatment decision (i.e. the decision to initiate COX-2 inhibitors or NSAIDs) if they are measured. In the context of EHR studies, where we likely have access to (mostly) the same information the clinician had when deciding whether or not to prescribe a medication, the assumption that missing values did not affect this decision may well be reasonable (*Blake et al.*, 2020).

Alternative approaches often applied for handling missing data in the context of EHRs include complete case (or complete record) analysis and multiple imputation (*Farmer et al.*, 2018). A complete case analysis only considers patients with fully observed information on all necessary covariates (*Farmer et al.*, 2018). Despite being easy to implement, this approach leads to a loss of efficiency (since patients with missing data are discarded from the analysis) and can lead to biased treatment effect estimates when missingness depends on both the treatment and outcome (*Bartlett et al.*, 2015; *Blake et al.*, 2020). Multiple imputation involves filling in missing covariate information with plausible values (obtained by drawing from the predictive distribution based on the observed data (*Sterne et al.*, 2009a)) a number of times to create multiple 'complete' datasets (*Carpenter and Kenward*, 2013). The full PS analysis (i.e. including estimation of the treatment effect) is performed within each imputed data and an overall estimate of the treatment effect is obtained via Rubin's rules (*Carpenter and Kenward*, 2013; *Leyrat et al.*, 2019; *Rubin*, 1976). The incorporation of multiple imputation within a PS analysis is complex (*Granger et al.*, 2019b; *Leyrat et al.*, 2019) and furthermore relies on the missing at random assumption, i.e. that the missingness can be explained by the observed data (*Bhaskaran and Smeeth*, 2014; *Carpenter and Kenward*, 2013; *Sterne et al.*, 2009a). This assumption is often unlikely to be plausible in the context of EHR studies (*Farmer et al.*, 2018).

In relation to these alternatives, the missing indicator approach has several advantages. Firstly, it is easy to implement. Secondly, unlike the complete case analysis, the missing indicator approach does not discard any patients from the analysis. Finally, the missing

indicator approach does not require the missing at random assumption to hold and instead relies on an assumption we believe is likely to be at least approximately true in the context of EHR studies (*Blake et al.*, 2020).

We additionally used the HDPS to investigate residual confounding. Whilst details of the HDPS procedure are given in Chapters 2 and 3, the investigator decisions are described in the following paragraphs. As in previous work (*Tazare et al.*, 2020) (Chapter 3), we identified clinical, referral and therapy data dimensions capturing relevant information from primary care records in the 12-months prior to cohort entry. We then applied our modifications: mapping the clinical and referral dimensions to ICD-10 and extending the frequency assessment. Prescriptions were classified at the BNF paragraph level.

Given the availability of linked HES data, we additionally incorporated a data dimension capturing HES diagnoses in the year prior to cohort entry. These were classified using ICD-10 codes. Whilst in primary care data, we extended the frequency assessment to capture information recorded 'Ever' during a patient's medical history, this was not relevant for the HES dimension. In primary care data, a patient may consult for a reason (for example, a diagnosis) that has not been recently recorded in the GP records. However, HES data will capture all relevant diagnoses pertaining to a specific hospitalisation. Therefore, frequency assessment for the HES dimension was conducted using the traditional cut-offs (*Schneeweiss et al.*, 2009) (see Chapters 2 and 3 for details).

Having defined 4 data dimensions (clinical, referral, therapy and hospitalisations), we selected the top 200 most prevalent codes in the each dimension. Additionally, for the primary HDPS analysis we selected the top 500 covariates as ranked by the Bross formula (*Schneeweiss et al.*, 2009).

For the HDPS analysis, we varied the covariate assessment period used to identify codes in all dimensions to 6 and 24 months. Since the inclusion of a HES data dimension creates a larger pool of potential HDPS covariates, it was hypothesised that accounting for more than the typical 500 covariates might improve confounding control. Therefore, we investigated the robustness of results to the number of covariates selected (250, 750,

900 and 1,000).

All analyses were conducted using the R Statistical Software package (*R Core Team*, 2020).

6.4 Results

We identified a cohort of 74,443 new users of NSAIDs and 25,742 new users of COX-2 inhibitors. During the follow-up period, there were 113 cases of upper GI bleeds in the NSAID users and 78 in the COX-2 inhibitor users. Furthermore, the average duration of treatment use was 61 days for COX-2 inhibitor users and 59 days for NSAID users, consistent with a study by *Toh* (2017) conducted using the The Health Improvement Network. As expected, COX-2 inhibitor users were on average older, had more hospitalisations in the previous 6 months and had consistently higher prevalence of comorbidities and medications Table 6.1. After propensity score matching using investigator and HDPS models, good covariate balance was achieved between the two treatment groups (Table 6.1).

Investigation of the prescribing patterns for NSAID and COX-2 inhibitors across the study period highlighted the expected increase use of COX-2 inihibitors over time Figure 6.2.

6.4.1 Investigator-led traditional PS analysis

Results obtained in the unmatched sample indicated a greater risk of upper GI bleed in COX-2 inhibitor users compared to NSAID users (HR 1.28; 95% CI: 0.95 - 1.72).

The investigator analysis included all covariates described in Section 6.3.4 in the propensity score model. Nearest neighbour 1:1 propensity score matching resulted in successful matches for 97% of COX-2 inhibitor users. Furthermore, the estimated propensity score distributions by treatment group are presented in Figure 6.3. After adjustment

for factors identified by the investigators based on clinical knowledge, results indicated a slight reduction in the risk of upper GI bleeding (compared to the unmatched sample), however, the 95% confidence interval suggested these data were still consistent with a substantial increased risk (HR 1.08; 95% CI: 0.73 - 1.61) (Table 6.2).

6.4.2 HDPS analysis

In the primary HDPS analysis, the set of investigator covariates was supplemented by the top 500 Bross-prioritised HDPS covariates. Propensity score matching resulted in successful matches for 93% of COX-2 inhibitor users. Augmenting the investigator covariates with a set of HDPS-derived covariates reduced the HR for the association between COX-2 inihibitor and NSAID use further towards the expected result (HR 0.86; 95% 0.58 - 1.26).

The top-500 HDPS covariates included covariates derived from codes originating in each of the dimensions, as follows: 112 (22%) Clinical, 39 (8%) Referral, 220 (44%) Therapy and 129 (26%) Hospitalisation. This highlights the potential importance of including hospital discharge data in the HDPS procedure. The characteristics of these covariates is summarised in Figures 6.4 and 6.5. Figure 6.4 shows that COX-2 inihibitor users have a higher prevalence of HDPS derived covariates compared to NSAID users. Furthermore, Figure 6.6 highlights that differences in covariate balance between the two groups improves after propensity score matching. Figure 6.5 highlights a number of covariates with strong association with the outcome but mild to weak association with treatment allocation.

Compared to the investigator-matched sample, the distribution of estimated propensity scores between the treatment groups is similar in the HDPS-matched sample (Figure 6.7). As seen in Chapter 3, when very strong indicators of treatment are included by the HDPS this can lead to bi-modal propensity score distributions; suggesting that these types of variables were not identified in this example. Given the number of covariates with a strong outcome association (Figure 6.5) and the improvement in balance of all covariates after matching (Figure 6.6), this tends to support the hypothesis that in this

study confounding bias is being driven by previously unmeasured risk factors for UGIB.

In sensitivity analyses, the covariate assessment period and number of HDPS covariates selected was varied 6.3. Overall, the interpretation of results remained unchanged by these decisions. However, assessing HDPS covariate information in the 24-months prior to cohort entry resulted in further reductions in the risk of UGIB for COX-2 inhibitor users compared to NSAID users, suggesting this could lead to the inclusion of additional relevant information in some settings.

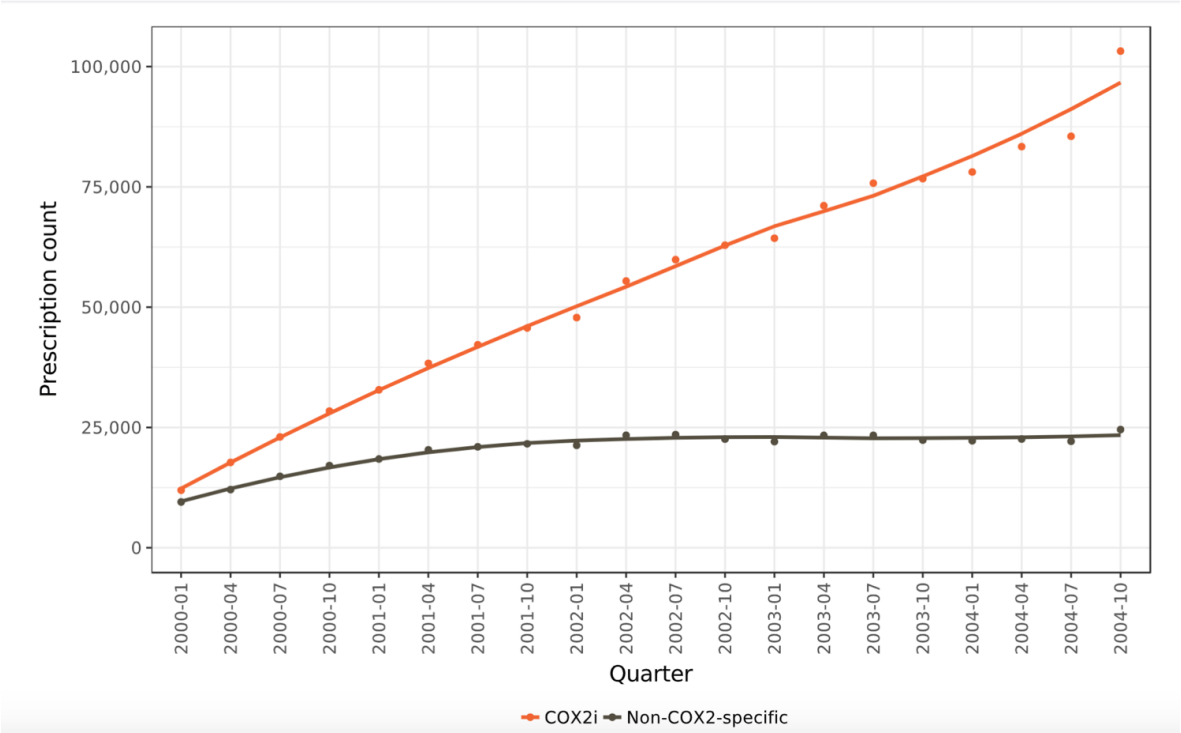


Figure 6.2: *Prescribing trends for NSAIDs (Non-COX2-specific) and COX-2 inhibitors across the study period.*

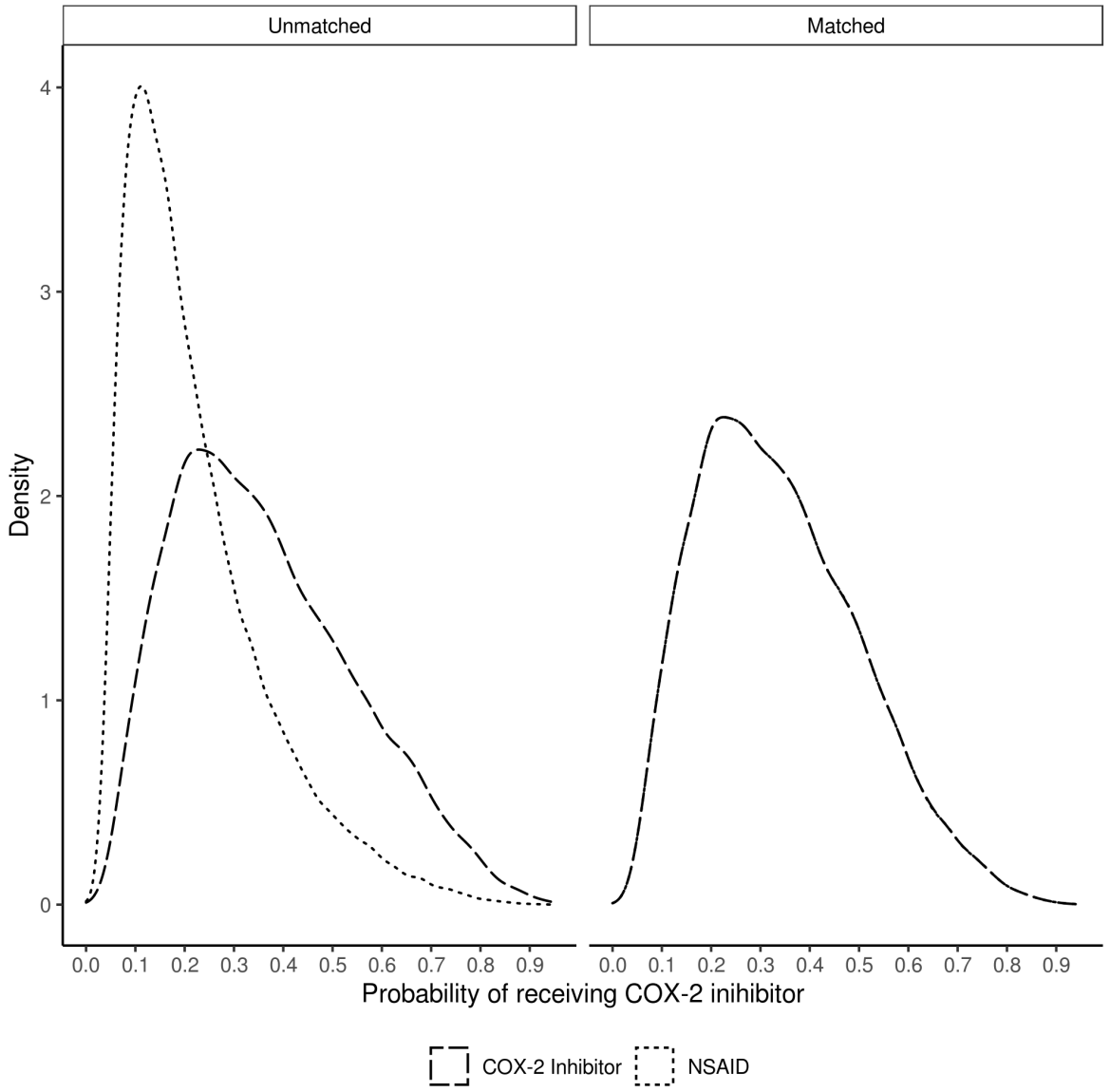


Figure 6.3: *Comparison of estimated propensity score distributions by treatment group in the investigator-matched sample.*

Table 6.1: *Characteristics of NSAID and COX-2 inhibitor users in unmatched, investigator-matched and HDPS-matched samples* ***Abbreviations:*** *BMI, body mass index; CKD, chronic kidney disease; GERD, gastroesophageal reflux disease; H2RA, h2 receptor antagonists; IBD, inflammatory bowel disease; IMD, index of multiple deprivation; OCS, oral corticosteroids; PPI, proton pump inhibitor; SD, standard deviation; SMD, standardised mean difference; SSRI, Selective serotonin reuptake inhibitors; UGIB, upper gastrointestinal bleeding*

;

	Unmatched			Investigator-Matched			HDPS-Matched		
	NSAID	COX2-I	SMD	NSAID	COX2-I	SMD	NSAID	COX2-I	SMD
N	74443	25742		24996	24996		24064	24064	
Age (Mean (SD))	64.95 (13.09)	70.19 (12.03)	0.416	69.99 (12.18)	69.93 (12.02)	0.005	69.76 (12.11)	69.72 (12.07)	0.004
Male	30208 (40.6)	7993 (31.1)	0.2	7840 (31.4)	7884 (31.5)	0.004	7713 (32.1)	7722 (32.1)	0.001
Urban	11739 (15.8)	4570 (17.8)	0.053	4412 (17.7)	4410 (17.6)	<0.001	4221 (17.5)	4210 (17.5)	0.001
IMD			0.038			0.031			0.043
1	8599 (11.6)	2910 (11.3)		2849 (11.4)	2842 (11.4)		2679 (11.1)	2746 (11.4)	
2	8331 (11.2)	2843 (11.0)		2760 (11.0)	2776 (11.1)		2676 (11.1)	2680 (11.1)	
3	8618 (11.6)	2815 (10.9)		2821 (11.3)	2740 (11.0)		2797 (11.6)	2643 (11.0)	
4	8410 (11.3)	2837 (11.0)		2714 (10.9)	2742 (11.0)		2630 (10.9)	2632 (10.9)	
5	8205 (11.0)	3011 (11.7)		2752 (11.0)	2921 (11.7)		2650 (11.0)	2818 (11.7)	
6	7518 (10.1)	2645 (10.3)		2688 (10.8)	2567 (10.3)		2498 (10.4)	2462 (10.2)	
7	6985 (9.4)	2503 (9.7)		2449 (9.8)	2434 (9.7)		2269 (9.4)	2341 (9.7)	
8	6827 (9.2)	2491 (9.7)		2322 (9.3)	2411 (9.6)		2209 (9.2)	2318 (9.6)	
9	5403 (7.3)	1819 (7.1)		1797 (7.2)	1759 (7.0)		1796 (7.5)	1686 (7.0)	
10	5547 (7.5)	1868 (7.3)		1844 (7.4)	1804 (7.2)		1860 (7.7)	1738 (7.2)	
Hospital Admissions			0.089			0.02			0.007
0	65033 (87.4)	21688 (84.3)		21143 (84.6)	21111 (84.5)		20449 (85.0)	20398 (84.8)	
1	7124 (9.6)	3076 (11.9)		2840 (11.4)	2946 (11.8)		2759 (11.5)	2787 (11.6)	
2	1526 (2.0)	650 (2.5)		675 (2.7)	624 (2.5)		573 (2.4)	584 (2.4)	
>2	760 (1.0)	328 (1.3)		338 (1.4)	315 (1.3)		283 (1.2)	295 (1.2)	
Alcohol Status			0.081			0.008			0.005
High	1205 (1.6)	356 (1.4)		359 (1.4)	345 (1.4)		331 (1.4)	343 (1.4)	
Low	36333 (48.8)	13595 (52.8)		13206 (52.8)	13134 (52.5)		12557 (52.2)	12585 (52.3)	
Missing	36905 (49.6)	11791 (45.8)		11431 (45.7)	11517 (46.1)		11176 (46.4)	11136 (46.3)	
Smoking Status			0.095			0.011			0.007
Current	11063 (14.9)	3457 (13.4)		3324 (13.3)	3382 (13.5)		3332 (13.8)	3284 (13.6)	
Ex	13072 (17.6)	5285 (20.5)		5117 (20.5)	5087 (20.4)		4886 (20.3)	4875 (20.3)	
Non-smoker	36701 (49.3)	12855 (49.9)		12418 (49.7)	12472 (49.9)		11935 (49.6)	11950 (49.7)	
Missing	13607 (18.3)	4145 (16.1)		4137 (16.6)	4055 (16.2)		3911 (16.3)	3955 (16.4)	

Continued on next page

	Unmatched			Investigator-Matched			HDPS-Matched		
	NSAID	COX2-I	SMD	NSAID	COX2-I	SMD	NSAID	COX2-I	SMD
BMI			0.028			0.004			0.011
<18.5	564 (0.8)	238 (0.9)		228 (0.9)	230 (0.9)		211 (0.9)	215 (0.9)	
18.5-25	20523 (27.6)	6892 (26.8)		6724 (26.9)	6702 (26.8)		6370 (26.5)	6457 (26.8)	
25-30	24847 (33.4)	8555 (33.2)		8275 (33.1)	8311 (33.2)		8065 (33.5)	8019 (33.3)	
30+	13989 (18.8)	4991 (19.4)		4859 (19.4)	4841 (19.4)		4729 (19.7)	4650 (19.3)	
Missing	14520 (19.5)	5066 (19.7)		4910 (19.6)	4912 (19.7)		4689 (19.5)	4723 (19.6)	
Comorbidities									
IBD	572 (0.8)	303 (1.2)	0.042	289 (1.2)	282 (1.1)	0.003	268 (1.1)	265 (1.1)	0.001
Heart Failure	2519 (3.4)	1492 (5.8)	0.115	1396 (5.6)	1389 (5.6)	0.001	1294 (5.4)	1320 (5.5)	0.005
Hypertension	23771 (31.9)	10613 (41.2)	0.194	10247 (41.0)	10181 (40.7)	0.005	9781 (40.6)	9708 (40.3)	0.006
GI Cancer	758 (1.0)	297 (1.2)	0.013	308 (1.2)	289 (1.2)	0.007	285 (1.2)	279 (1.2)	0.002
CKD	289 (0.4)	141 (0.5)	0.023	140 (0.6)	130 (0.5)	0.005	117 (0.5)	121 (0.5)	0.002
Diabetes	5428 (7.3)	2291 (8.9)	0.059	2248 (9.0)	2215 (8.9)	0.005	2170 (9.0)	2124 (8.8)	0.007
Coronary Angioplasty	384 (0.5)	185 (0.7)	0.026	172 (0.7)	178 (0.7)	0.003	157 (0.7)	170 (0.7)	0.007
Coagulopathy	339 (0.5)	181 (0.7)	0.033	170 (0.7)	169 (0.7)	<0.001	149 (0.6)	156 (0.6)	0.004
Previous UGIB	1520 (2.0)	1057 (4.1)	0.12	889 (3.6)	938 (3.8)	0.01	845 (3.5)	861 (3.6)	0.004
GERD	2617 (3.5)	1627 (6.3)	0.13	1477 (5.9)	1482 (5.9)	0.001	1453 (6.0)	1415 (5.9)	0.007
Medications/Therapies									
Statin	7533 (10.1)	3645 (14.2)	0.124	3427 (13.7)	3485 (13.9)	0.007	3351 (13.9)	3316 (13.8)	0.004
PPI/H2RA	6681 (9.0)	5823 (22.6)	0.381	5021 (20.1)	5174 (20.7)	0.015	4668 (19.4)	4763 (19.8)	0.01
SSRI	3316 (4.5)	1539 (6.0)	0.069	1436 (5.7)	1452 (5.8)	0.003	1399 (5.8)	1381 (5.7)	0.003
Anticoagulant	644 (0.9)	524 (2.0)	0.098	445 (1.8)	464 (1.9)	0.006	386 (1.6)	417 (1.7)	0.01
Antiplatetes	11249 (15.1)	5575 (21.7)	0.17	5343 (21.4)	5326 (21.3)	0.002	5133 (21.3)	5083 (21.1)	0.005
OCS	1826 (2.5)	1189 (4.6)	0.117	1098 (4.4)	1085 (4.3)	0.003	985 (4.1)	1002 (4.2)	0.004
Other Respiratory	6293 (8.5)	3201 (12.4)	0.13	3014 (12.1)	3020 (12.1)	0.001	2837 (11.8)	2860 (11.9)	0.003
Other									
Calendar			0.457			0.139			0.125
2000	16561 (22.2)	2429 (9.4)		2987 (11.9)	2428 (9.7)		2893 (12.0)	2405 (10.0)	
2001	18103 (24.3)	4479 (17.4)		4415 (17.7)	4453 (17.8)		4253 (17.7)	4371 (18.2)	
2002	15169 (20.4)	6025 (23.4)		5011 (20.0)	5922 (23.7)		5003 (20.8)	5747 (23.9)	
2003	13077 (17.6)	6622 (25.7)		5817 (23.3)	6411 (25.6)		5595 (23.3)	6106 (25.4)	
2004	11533 (15.5)	6187 (24.0)		6766 (27.1)	5782 (23.1)		6320 (26.3)	5435 (22.6)	

Table 6.2: *Results from primary analysis comparing investigator and HDPS models*

Analysis	Hazard Ratio (95% CI)
Unmatched	1.28 (0.95 - 1.72)
Investigator-Matched	1.08 (0.73 - 1.61)
HDPS-Matched*	0.86 (0.58 - 1.26)

* Based on data dimensions capturing clinical, referral, therapy and hospitalisation information, selecting the top 200 most prevalent codes per dimension and selecting the top 500 covariates as ranked by the Bross formula

Table 6.3: *Sensitivity analyses for the HDPS analysis extending the covariate assessment period and number of covariates selected*

Covariate assessment period	Number of HDPS covariates	Hazard Ratio (95% CI)
12-months	250	0.84 (0.58 - 1.22)
12-months	500	0.86 (0.58 - 1.26)
12-months	750	0.81 (0.55 - 1.18)
12-months	900	0.87 (0.59 - 1.28)
12-months	1000	0.83 (0.56 - 1.22)
24-months	250	0.89 (0.61 - 1.28)
24-months	500	0.85 (0.59 - 1.23)
24-months	750	0.73 (0.51 - 1.06)
24-months	900	0.69 (0.48 - 1.00)
24-months	1000	0.77 (0.53 - 1.13)
6-month	250	0.83 (0.57 - 1.21)
6-month	500	0.90 (0.61 - 1.33)
6-month	750	0.87 (0.59 - 1.28)
6-month	900	0.76 (0.52 - 1.12)
6-month	1000	0.78 (0.53 - 1.15)

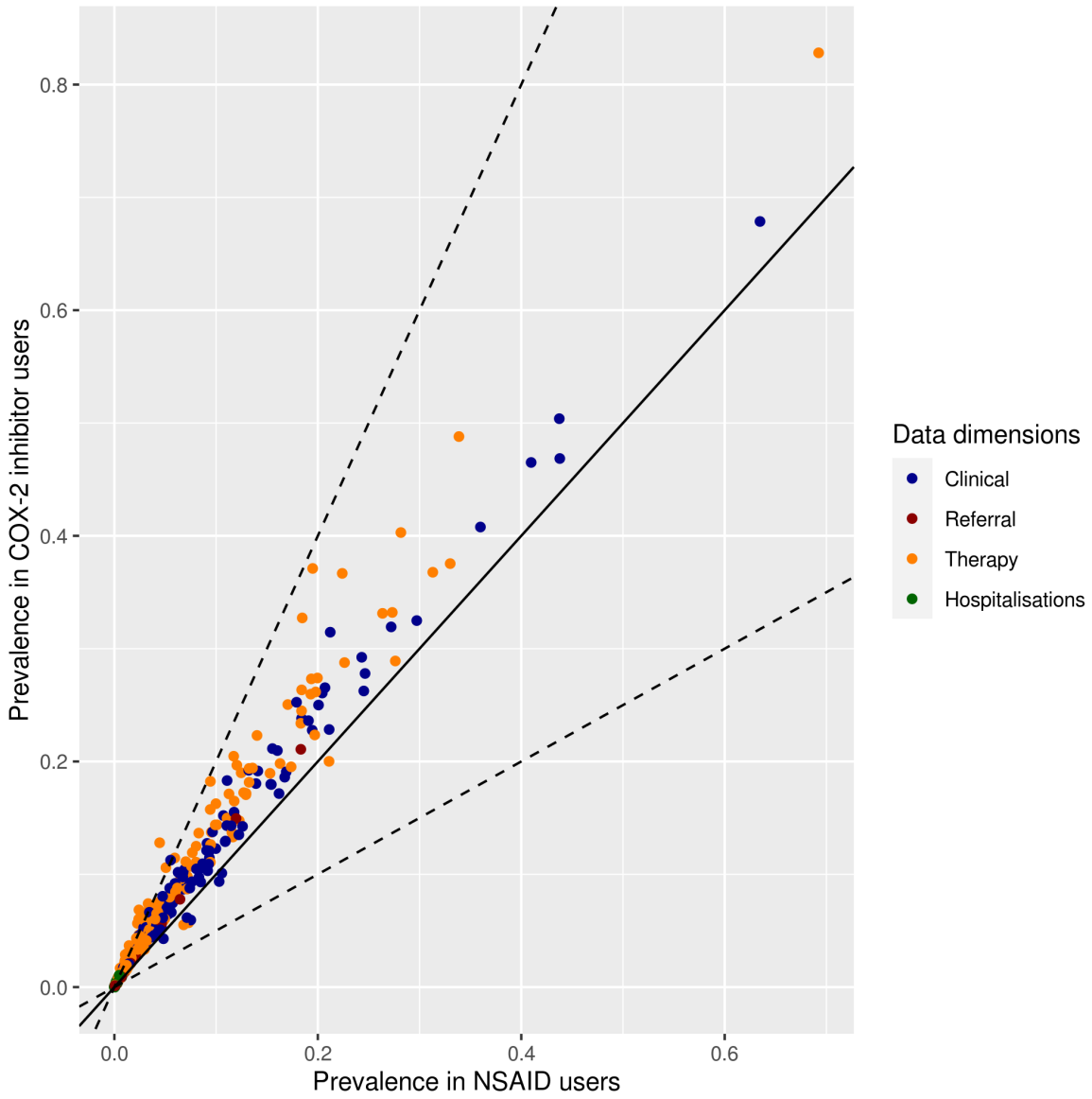


Figure 6.4: *Prevalence of the top 500 Bross-prioritised covariates by treatment group and data dimension. The diagonal line indicate equal prevalence in both groups and the dashed lines show prevalence ratios (PR) of 0.5 and 2.0. The colour coding highlights which dimension the covariate originated from.*

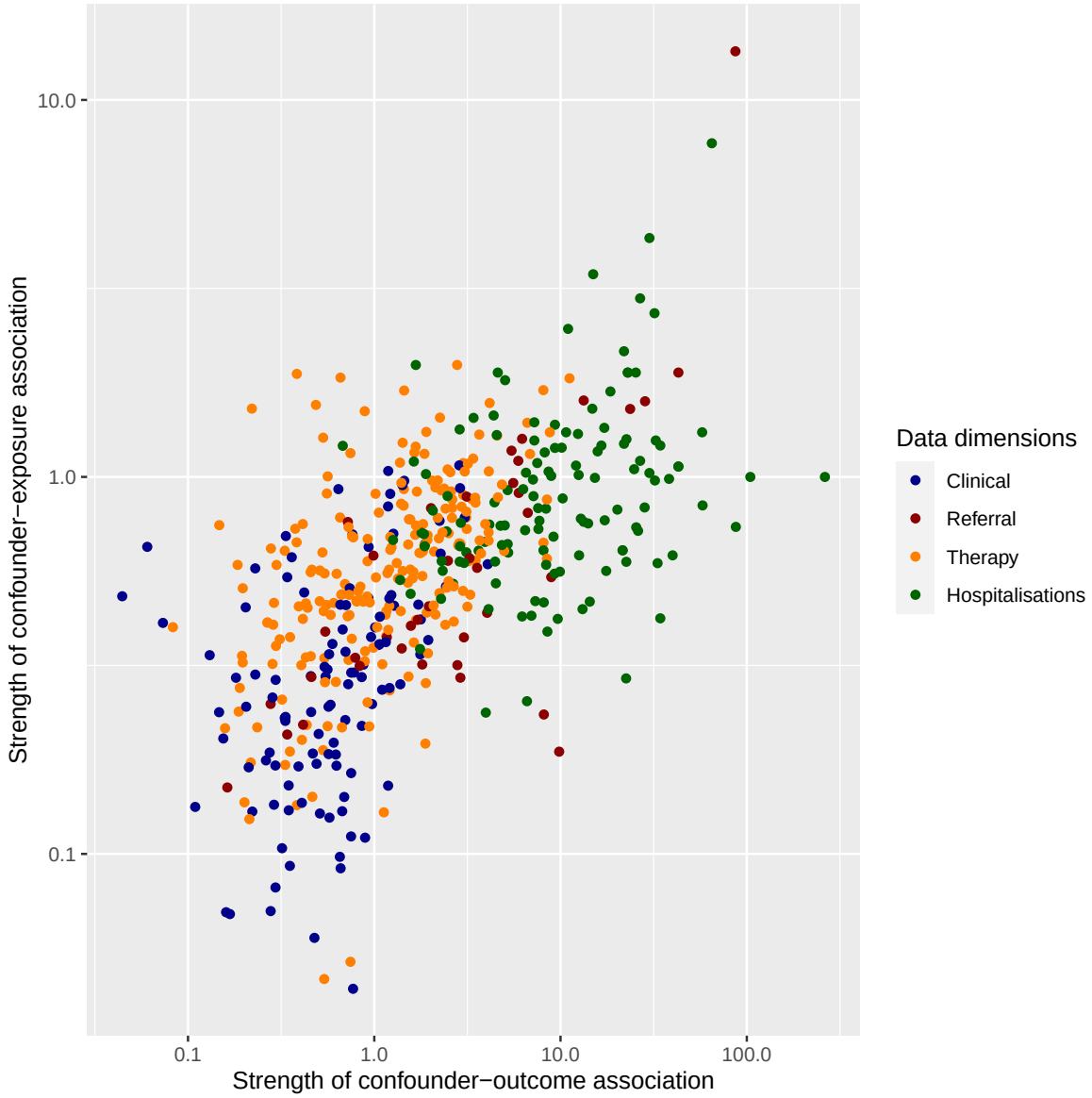


Figure 6.5: *Strength of covariate-exposure and covariate-outcome associations for the top 500 Bross-prioritised HDPS covariates. The colour coding highlights which dimension the covariate originated from.*

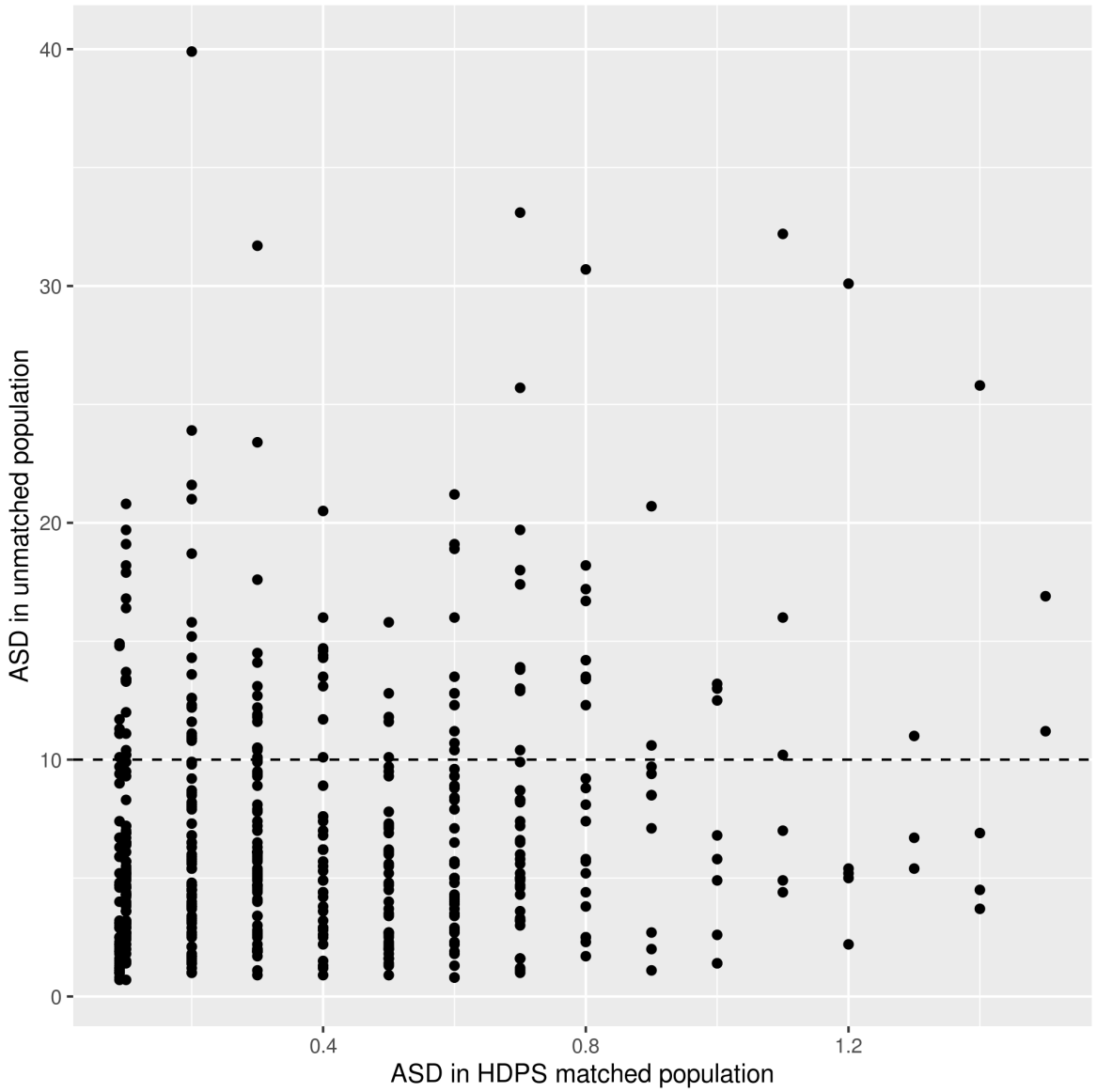


Figure 6.6: *Comparison of absolute standardised differences (ASDs) between unmatched and HDPS matched samples, selecting the top 500 HDPS covariates. The dashed line indicates ASDs of 10%.*

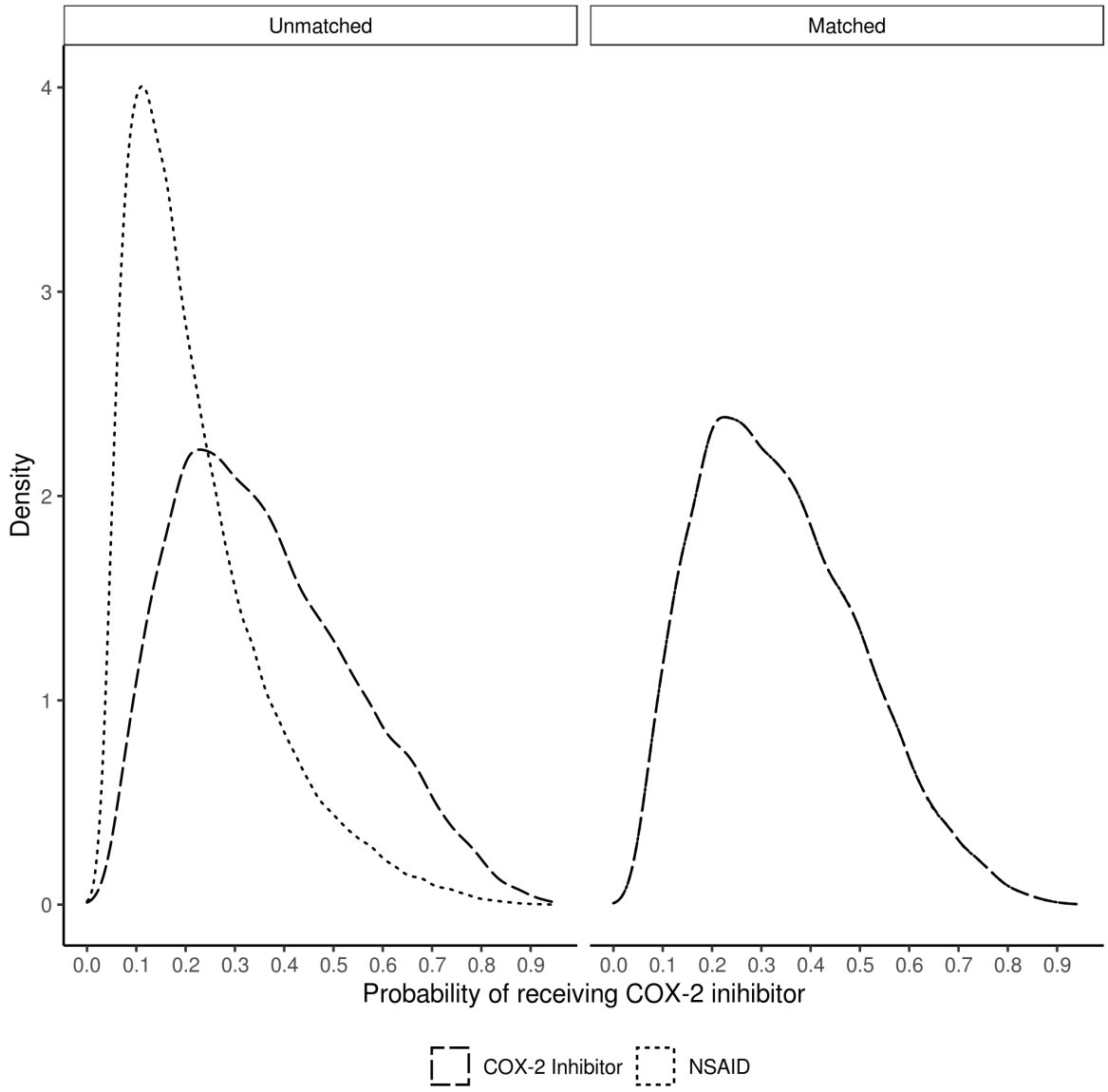


Figure 6.7: *Comparison of estimated propensity score distributions by treatment group in the HDPS-matched sample.*

6.5 Discussion

In this chapter, we applied the modified HDPS (proposed in Chapter 3) to a study question where HDPS had been extensively applied (*Schneeweiss*, 2018). The aim was to compare the results obtained to existing evidence generated by randomised controlled trials and replicated in pharmacoepidemiological studies (*Bombardier et al.*, 2000; *Schneeweiss*, 2018; *Silverstein et al.*, 2000). Compared to results from an investigator-led propensity score model (HR 1.08; 95% CI: 0.73 - 1.61), the modified-HDPS obtained results closer to expected association (HR 0.86; 95% 0.58 - 1.26); although the 95% confidence intervals suggest these data are also compatible with an increased risk. Furthermore, the pattern of results were remarkably similar to those obtained by 4 separate pharmacoepidemiological studies applying the HDPS to the same question in a range of healthcare databases (*Garbe et al.*, 2013; *Hallas and Pottegard*, 2017; *Schneeweiss et al.*, 2009; *Toh et al.*, 2011). Upon graphical inspection of the characertistics of the HDPS covariates and estimated propensity score distributions, it appeared that the pattern of results was largely driven by the ability of the HDPS to identify and measure risk factors for upper GI bleeding that were not captured by the set of investigator covariates. This is consistent with previous hypotheses surrounding the likely confounding structure when comparing COX-2 inhibitor and NSAID use on the risk of upper GI bleeding (*Schneeweiss*, 2018).

This work contributes to evidence surrounding the performance of the modified HDPS presented in Chapter 3. Firstly, sensitivity analyses surrounding the number of covariates selected highlighted that conclusions were robust to this decision. Furthermore, slight improvements in the expected effect estimates when assessing HDPS covariates in the 24-months prior to cohort entry (compared to 12-months and 6-months) suggests that extending the covariate assessment period for all dimensions might be useful when applying the HDPS in UK EHRs. In particular, this may highlight important differences between EHR and insurance claims databases in terms of patterns of recording practice. Finally, we incorporated hospital discharge information, specifying a separate data dimension capturing ICD-10 codes in HES discharge data. In the primary analysis, these codes accounted for 26% of covariates selected, highlighting the potential

importance of these data for successful confounder capture and control.

This study has several limitations. Firstly, since NSAIDs are an over the counter (OTC) medication we cannot exclude the possibility that patients may have been chronically self-medicating prior to study entry. It is unknown whether such patients would have a different likelihood of receiving one treatment over another but we have no evidence to support a differential lead time bias between groups of patients as defined by initiation of COX-2 inihibtors and NSAIDs. Furthermore, we consider it likely that individuals who have a need for chronic NSAID use would be likely to engage with primary care and receive prescriptions for treatment, rather than obtaining via OTC routes. Whilst these issues would be of legitimate concern for an observational study that seeks to explore differentials in risk between two compounds, such limitations are likely to have applied to a number of extant studies in the considerable body of work describing NSAIDs and COX-2 inhibitors. Finally, our study period spans 2000-2004 which covers the introduction of the Quality and Outcomes Framework (QOF) to UK primary care in 2004 (*Roland and Guthrie*, 2016). This may introduce a bias insofar as the recording of certain incentivised comorbidities or other directly or indirectly QOF-induced changes in provider behaviour during the latter part of our observation period. Furthermore, several COX-2s were withdrawn from the market in 2003.

This study has shown an example where the modified-HDPS was able to obtain results similar to those from randomised controlled trials and other pharmacoepidemiological studies. Importantly, these results were only replicated after the HDPS was used to augment the investigator set of covariates, further highlighting the potential benefits of these approaches for successful confounder control in UK EHRs.

6.6 Supporting information

S1 - Code list for COX-2 inhibitors

Prodcode	Gemscript code	Product name
474	79023020	Celecoxib 100mg capsules
5080	81598020	Celebrex 200mg capsules (Pfizer Ltd)
5175	81597020	Celebrex 100mg capsules (Pfizer Ltd)
5254	79024020	Celecoxib 200mg capsules
43616	98663020	Celecoxib 400mg capsules
50059	8520020	Celebrex 100mg capsules (Necessity Supplies Ltd)
55582	8512020	Celebrex 200mg capsules (Lexon (UK) Ltd)
66757	8508020	Celebrex 200mg capsules (Waymade Healthcare Plc)
3311	60932020	Etodolac 200mg capsules
4368	58036020	Lodine 200mg Capsule (Shire Pharmaceuticals Ltd)
8969	58038020	Lodine 300mg Capsule (Shire Pharmaceuticals Ltd)
10033	60934020	Etodolac 300mg capsules
24356	83390020	Eccoxolac 300mg capsules (Meda Pharmaceuticals Ltd)
66323	86577020	Ebretin 300mg capsules (Ranbaxy (UK) Ltd)
5266	75111020	Lodine sr 600mg Modified-release tablet (Shire Pharmaceuticals Ltd)
5455	75114020	Etodolac 600mg modified-release tablets
35653	91953020	Etopan XL 600mg tablets (Sun Pharmaceuticals UK Ltd)
38770	95557020	Lodine SR 600mg tablets (Almirall Ltd)
52714	8117020	Etodolac 600mg modified-release tablets (Alliance Healthcare (Distribution) Ltd)
71908	70658021	Etolyn 600mg modified-release tablets (Mylan)
76419	75667020	Etodolac sr 600mg Tablet (Winthrop Pharmaceuticals Ltd)
8451	60933020	Etodolac 200mg Tablet
16194	58037020	Lodine 200mg Tablet (Shire Pharmaceuticals Ltd)
20386	60937020	Ramodar 200mg Tablet (Wyeth Pharmaceuticals)
650	77851020	Etoricoxib 60mg tablets
5812	77854020	Etoricoxib 90mg tablets
5938	84667020	Etoricoxib 120mg tablets
6464	84678020	Arcoxia 60mg tablets (Grunenthal Ltd)
6498	77845020	Arcoxia 90mg tablets (Grunenthal Ltd)
9822	77848020	Arcoxia 120mg tablets (Grunenthal Ltd)
37562	94524020	Arcoxia 30mg tablets (Grunenthal Ltd)
37587	94522020	Etoricoxib 30mg tablets
51284	10595020	Arcoxia 60mg tablets (Sigma Pharmaceuticals Plc)
51874	39400020	Arcoxia 30mg tablets (Lexon (UK) Ltd)
53576	10607020	Arcoxia 120mg tablets (DE Pharmaceuticals)
62658	10606020	Arcoxia 120mg tablets (Waymade Healthcare Plc)

62843	10603020	Arcoxia 90mg tablets (Lexon (UK) Ltd)
74952	10592020	Arcoxia 60mg tablets (Waymade Healthcare Plc)
75549	77465021	Etoricoxib 30mg tablets (Accord Healthcare Ltd)
7118	88358020	Prexige 100mg tablets (Novartis Pharmaceuticals UK Ltd)
10212	88352020	Lumiracoxib 100mg tablets
28171	88356020	Lumiracoxib 400mg tablets
28383	88362020	Prexige 400mg tablets (Novartis Pharmaceuticals UK Ltd)
76595	36645020	Meloxicam 7.5mg/5ml oral suspension
57370	17125021	Meloxicam 15mg orodispersible tablets sugar free
57475	17127021	Meloxicam 7.5mg orodispersible tablets sugar free
850	81638020	Mobic 7.5mg tablets (Boehringer Ingelheim Ltd)
1469	81615020	Meloxicam 15mg tablets
1470	81639020	Mobic 15mg tablets (Boehringer Ingelheim Ltd)
2243	81614020	Meloxicam 7.5mg tablets
35935	72118020	Meloxicam 7.5mg tablets (Somex Pharma)
56275	72217020	Meloxicam 7.5mg tablets (Teva UK Ltd)
66364	72791020	Meloxicam 15mg tablets (Actavis UK Ltd)
76191	72175020	Meloxicam 7.5mg tablets (A A H Pharmaceuticals Ltd)
77260	72221020	Meloxicam 15mg tablets (Teva UK Ltd)
28190	!8505504	VIOXX
28193	!8505505	VIOXX
32362	!8505522	ROFECOXIB
36669	!8505521	ROFECOXIB
613	76748020	Vioxx 12.5mg/5ml oral suspension (Merck Sharp & Dohme Ltd)
637	83963020	Rofecoxib 25mg/5ml oral suspension sugar free
640	79862020	Rofecoxib 12.5mg/5ml oral suspension sugar free
5739	59878020	Vioxx 25mg/5ml oral suspension (Merck Sharp & Dohme Ltd)
518	79860020	Rofecoxib 12.5mg tablets
538	76746020	Vioxx 12.5mg tablets (Merck Sharp & Dohme Ltd)
666	76747020	Vioxx 25mg tablets (Merck Sharp & Dohme Ltd)
706	79861020	Rofecoxib 25mg tablets
5695	80118020	VioxxAcute 50mg tablets (Merck Sharp & Dohme Ltd)
5841	81743020	Rofecoxib 50mg tablets
6460	78899020	VioxxAcute 25mg tablets (Merck Sharp & Dohme Ltd)
723	77194020	Valdecoxib 10mg tablets
6663	79996020	Valdecoxib 20mg tablets
9899	85809020	Bextra 10mg tablets (Pfizer Ltd)
9912	85182020	Bextra 20mg tablets (Pfizer Ltd)
9978	49090020	Bextra 40mg tablets (Pfizer Ltd)
18066	83422020	Valdecoxib 40mg tablets

S2 - Code list for NSAIDs

Prodcode	Gemscript code	Product name
526	81068020	Aceclofenac 100mg tablets
344	73771020	Acemetacin 60mg capsules
55099	86649020	Acoflam 100mg Retard tablets (Mercury Pharma Group Ltd)
40086	86647020	Acoflam 50mg gastro-resistant tablets (Mercury Pharma Group Ltd)
75442	86651020	Acoflam 75mg SR tablets (Mercury Pharma Group Ltd)
25257	83095020	Advil 200mg tablets (Wyeth Consumer Healthcare)
40394	83096020	Advil 400mg Tablet (Wyeth Consumer Healthcare)
32704	86005020	Advil cold and sinus 200mg+30mg Tablet (Wyeth Consumer Healthcare)
13347	48091020	Alrheumat 50mg Capsule (Bayer Plc)
32509	73886020	Anadin Ibuprofen 200mg tablets (Pfizer Consumer Healthcare Ltd)
38493	94461020	Anadin Joint Pain 200mg tablets (Pfizer Consumer Healthcare Ltd)
46860	99382020	Anadin LiquiFast 200mg effervescent tablets (Pfizer Consumer Healthcare Ltd)
43456	96457020	Anadin LiquiFast 400mg capsules (Pfizer Consumer Healthcare Ltd)
40516	96455020	Anadin Ultra 200mg capsules (Pfizer Consumer Healthcare Ltd)
37253	91330020	Anadin ultra double strength 400mg Capsule (Wyeth Consumer Healthcare)
20978	77321020	Anadin Ultra liquid capsules (Wyeth Consumer Healthcare)
31482	57638020	Apsifen 200mg Tablet (Approved Prescription Services Ltd)
27968	57639020	Apsifen 400mg Tablet (Approved Prescription Services Ltd)
31469	57645020	Apsifen -f 600mg Tablet (Approved Prescription Services Ltd)
19036	62435020	Arthrofen 200 tablets (Ashbourne Pharmaceuticals Ltd)
15068	62436020	Arthrofen 400 tablets (Ashbourne Pharmaceuticals Ltd)
21815	62437020	Arthrofen 600 tablets (Ashbourne Pharmaceuticals Ltd)
21840	62238020	Arthrosin 250 tablets (Ashbourne Pharmaceuticals Ltd)
20385	62239020	Arthrosin 500 tablets (Ashbourne Pharmaceuticals Ltd)
25341	86551020	Arthrosin EC 250 tablets (Ashbourne Pharmaceuticals Ltd)
25342	86553020	Arthrosin EC 500 tablets (Ashbourne Pharmaceuticals Ltd)
162	74346020	Arthrotec 50 gastro-resistant tablets (Pfizer Ltd)
50269	8107020	Arthrotec 75 gastro-resistant tablets (Mawdsley-Brooks & Company Ltd)
2387	81620020	Arthrotec 75 gastro-resistant tablets (Pfizer Ltd)
30168	62234020	Arthroxen 250mg Tablet (C P Pharmaceuticals Ltd)
23121	62235020	Arthroxen 500mg Tablet (C P Pharmaceuticals Ltd)

41366	97646020	Axorid 100mg/20mg modified-release capsules (Meda Pharmaceuticals Ltd)
41365	97648020	Axorid 200mg/20mg modified-release capsules (Meda Pharmaceuticals Ltd)
4049	60168020	Azapropazone 300mg capsules
3262	60169020	Azapropazone 600mg tablets
71584	60687021	Boots Ibuprofen 3 Months Plus 100mg/5ml oral suspension strawberry (The Boots Company Plc)
76093	8186020	Boots Ibuprofen 6 Months Plus 100mg/5ml oral suspension strawberry (The Boots Company Plc)
69285	60686021	Boots Ibuprofen and Codeine 200mg/12.8mg tablets (The Boots Company Plc)
71779	60688021	Boots Ibuprofen Long Lasting 200mg capsules (The Boots Company Plc)
48568	14108020	Boots Rapid Ibuprofen lysine 342mg tablets (The Boots Company Plc)
10169	77261020	Brexidol 20mg tablets (Chiesi Ltd)
19537	!0838201	BRUFEN
19538	!0838101	BRUFEN
50117	39329020	Brufen 100mg/5ml syrup (Lexon (UK) Ltd)
53397	38583020	Brufen 100mg/5ml syrup (Mawdsley-Brooks & Company Ltd)
360	48494020	Brufen 100mg/5ml syrup (Mylan)
1621	48493020	Brufen 200mg tablets (Abbott Laboratories Ltd)
1739	53998020	Brufen 400mg tablets (Mylan)
50314	8169020	Brufen 600mg effervescent granules sachets (DE Pharmaceuticals)
407	68366020	Brufen 600mg effervescent granules sachets (Mylan)
4216	54001020	Brufen 600mg tablets (Mylan)
74806	8166020	Brufen Retard 800mg tablets (Dowelhurst Ltd)
39019	95831020	Brufen Retard 800mg tablets (Mylan)
2129	68365020	Brufen retard tabs 800mg Modified-release tablet (Abbott Laboratories Ltd)
167	65704020	Butacote 100mg gastro-resistant tablets (Novartis Pharmaceuticals UK Ltd)
7483	65708020	Butazolidin 100mg Tablet (Novartis Pharmaceuticals UK Ltd)
29674	65709020	Butazolidin 200mg Tablet (Novartis Pharmaceuticals UK Ltd)
7058	83707020	Calprofen 100mg/5ml Oral suspension (McNeil Products Ltd)
49432	8177020	Calprofen 100mg/5ml oral suspension (McNeil Products Ltd)
56441	14184020	Calprofen 100mg/5ml oral suspension 5ml sachets (McNeil Products Ltd)
29316	86537020	Care ibuprofen 400mg Tablet (Thornton & Ross Ltd)
66194	49091021	Care Ibuprofen for Children 100mg/5ml oral suspension (Thornton & Ross Ltd)

7434	48703020	Clinoril 100mg tablets (Merck Sharp & Dohme Ltd)
13380	48704020	Clinoril 200mg tablets (Merck Sharp & Dohme Ltd)
28764	84951020	Closteril 100mg Modified-release tablet (Pharmalife Healthcare Services Ltd)
20036	83580020	Clotam 200mg Capsule (Thames Laboratories Ltd)
14994	85000020	Clotam Rapid 200mg tablets (Galen Ltd)
1708	54315020	Codafen Continus tablets (Napp Pharmaceuticals Ltd)
10519	4147007	CODEINE PHOS/IBUPROFEN SR (20MG/300MG) TAB
17733	86585020	Condrotec 500mg+200microgram Tablet (Pharmacia Ltd)
30389	80085020	Contraflam 250mg Capsule (Berk Pharmaceuticals Ltd)
30391	80086020	Contraflam 500mg Tablet (Berk Pharmaceuticals Ltd)
14385	50220020	Cuprofen 200mg Tablet (SSL International Plc)
37094	93812020	Cuprofen 200mg tablets (SSL International Plc)
11980	50221020	Cuprofen 400mg Tablet (SSL International Plc)
24469	73716020	Cuprofen for Children 100mg/5ml oral suspension (SSL International Plc)
39873	95907020	Cuprofen Maximum Strength 400mg tablets (SSL International Plc)
37816	87860020	Cuprofen PLUS tablets (SSL International Plc)
25362	57275020	Defanac 25mg gastro-resistant tablets (Ranbaxy (UK) Ltd)
25358	78548020	Defanac 50mg gastro-resistant tablets (Ranbaxy (UK) Ltd)
14672	75883020	Defanac 75mg SR tablets (Ranbaxy (UK) Ltd)
14707	79162020	Defanac Retard 100mg tablets (Ranbaxy (UK) Ltd)
14678	75884020	Defanac sr 100mg Modified-release tablet (Ranbaxy (UK) Ltd)
10325	89560020	Dexibuprofen 300mg tablets
11907	89572020	Dexibuprofen 400mg tablets
5173	67445020	Dexketoprofen 25mg tablets
31383	86166020	Dexomon 75mg SR tablets (Hillcross Pharmaceuticals Ltd)
16225	84262020	Dexomon retard 100mg Modified-release tablet (Hillcross Pharmaceuticals Ltd)
34744	68904020	Diclofenac 100mg Modified-release capsule (Sandoz Ltd)
27362	56927020	Diclofenac 100mg Modified-release tablet (Actavis UK Ltd)
42793	61462020	Diclofenac 100mg Modified-release tablet (IVAX Pharmaceuticals UK Ltd)
45213	92541020	Diclofenac 10mg dispersible tablets
60368	19739020	Diclofenac 10mg/5ml oral solution
61762	19741020	Diclofenac 10mg/5ml oral suspension
51808	29485020	Diclofenac 12.5mg/5ml oral solution
68849	29487020	Diclofenac 12.5mg/5ml oral suspension
73131	73411020	Diclofenac 25mg Gastro-resistant tablet (Almus Pharmaceuticals Ltd)
34362	61922020	Diclofenac 25mg Gastro-resistant tablet (Genus Pharmaceuticals Ltd)
34218	59232020	Diclofenac 25mg Gastro-resistant tablet (Pharmacia Ltd)
32536	49461020	Diclofenac 25mg Tablet (Berk Pharmaceuticals Ltd)

75136	51184020	Diclofenac 25mg Tablet (C P Pharmaceuticals Ltd)
417	72775020	Diclofenac 50mg dispersible tablets sugar free
59595	8087020	Diclofenac 50mg dispersible tablets sugar free (Sigma Pharmaceuticals Plc)
42406	73414020	Diclofenac 50mg Gastro-resistant tablet (Almus Pharmaceuticals Ltd)
33669	61923020	Diclofenac 50mg Gastro-resistant tablet (Genus Pharmaceuticals Ltd)
30297	59231020	Diclofenac 50mg Gastro-resistant tablet (Pharmacia Ltd)
54463	54642020	Diclofenac 50mg Tablet (Approved Prescription Services Ltd)
28256	49462020	Diclofenac 50mg Tablet (Berk Pharmaceuticals Ltd)
33559	51185020	Diclofenac 50mg Tablet (C P Pharmaceuticals Ltd)
30942	59698020	Diclofenac 50mg Tablet (Regent Laboratories Ltd)
64759	29493020	Diclofenac 50mg/5ml oral solution
54906	29442020	Diclofenac 50mg/5ml oral suspension
32916	68901020	Diclofenac 75mg Modified-release capsule (Sandoz Ltd)
42905	59298020	Diclofenac 75mg Modified-release tablet (Actavis UK Ltd)
30282	59264020	Diclofenac 75mg Modified-release tablet (Galen Ltd)
34212	61924020	Diclofenac 75mg Modified-release tablet (Genus Pharmaceuticals Ltd)
33645	53793020	Diclofenac 75mg Modified-release tablet (IVAX Pharmaceuticals UK Ltd)
38817	95565020	Diclofenac potassium 12.5mg tablets
628	79361020	Diclofenac potassium 25mg tablets
58572	8370020	Diclofenac potassium 25mg tablets (A A H Pharmaceuticals Ltd)
597	79362020	Diclofenac potassium 50mg tablets
43045	77025020	Diclofenac potassium 50mg tablets (Accord Healthcare Ltd)
52338	8378020	Diclofenac potassium 50mg tablets (Focus Pharmaceuticals Ltd)
1115	73894020	Diclofenac sodium 100mg modified-release capsules
72546	69323020	Diclofenac sodium 100mg modified-release capsules (A A H Pharmaceuticals Ltd)
1984	62457020	Diclofenac sodium 100mg modified-release tablets
3416	83823020	Diclofenac sodium 100mg modified-release tablets
34271	60405020	Diclofenac sodium 100mg modified-release tablets (A A H Pharmaceuticals Ltd)
649	83871020	Diclofenac sodium 25mg gastro-resistant tablets
1096	73892020	Diclofenac sodium 25mg gastro-resistant tablets
24128	51193020	Diclofenac sodium 25mg gastro-resistant tablets (A A H Pharmaceuticals Ltd)
24121	56925020	Diclofenac sodium 25mg gastro-resistant tablets (Actavis UK Ltd)
33994	53791020	Diclofenac sodium 25mg gastro-resistant tablets (IVAX Pharmaceuticals UK Ltd)

53164	68114020	Diclofenac sodium 25mg gastro-resistant tablets (Kent Pharmaceuticals Ltd)
31944	60370020	Diclofenac sodium 25mg gastro-resistant tablets (Mylan)
34091	51270020	Diclofenac sodium 25mg gastro-resistant tablets (Sandoz Ltd)
62636	60043020	Diclofenac sodium 25mg gastro-resistant tablets (Sterwin Medicines)
32108	59804020	Diclofenac sodium 25mg gastro-resistant tablets (Teva UK Ltd)
928	62455020	Diclofenac sodium 25mg tablets
1692	74349020	Diclofenac sodium 50mg gastro-resistant / Misoprostol 200microgram tablets
40	83872020	Diclofenac sodium 50mg gastro-resistant tablets
1075	73893020	Diclofenac sodium 50mg gastro-resistant tablets
26165	51194020	Diclofenac sodium 50mg gastro-resistant tablets (A A H Pharmaceuticals Ltd)
24122	56926020	Diclofenac sodium 50mg gastro-resistant tablets (Actavis UK Ltd)
34487	53792020	Diclofenac sodium 50mg gastro-resistant tablets (IVAX Pharmaceuticals UK Ltd)
27055	63521020	Diclofenac sodium 50mg gastro-resistant tablets (Kent Pharmaceuticals Ltd)
21387	60371020	Diclofenac sodium 50mg gastro-resistant tablets (Mylan)
29330	51271020	Diclofenac sodium 50mg gastro-resistant tablets (Sandoz Ltd)
31950	60044020	Diclofenac sodium 50mg gastro-resistant tablets (Sterwin Medicines)
28553	59805020	Diclofenac sodium 50mg gastro-resistant tablets (Teva UK Ltd)
917	62456020	Diclofenac sodium 50mg tablets
4880	74350020	Diclofenac sodium 75mg gastro-resistant / Misoprostol 200microgram tablets
2904	75877020	Diclofenac sodium 75mg gastro-resistant modified-release capsules
447	52256020	Diclofenac sodium 75mg modified-release capsules
32854	69320020	Diclofenac sodium 75mg modified-release capsules (A A H Pharmaceuticals Ltd)
580	83822020	Diclofenac sodium 75mg modified-release tablets
1233	62462020	Diclofenac sodium 75mg modified-release tablets
31589	51195020	Diclofenac sodium 75mg modified-release tablets (A A H Pharmaceuticals Ltd)
20653	!1857105	DICLOFENAC SODIUM S/R
20105	81306020	Dicloflex 25mg Gastro-resistant tablet (Ratiopharm UK Ltd)
40756	96608020	Dicloflex 25mg gastro-resistant tablets (Almus Pharmaceuticals Ltd)
612	50464020	Dicloflex 25mg gastro-resistant tablets (Dexcel-Pharma Ltd)

35711	92493020	Dicloflex 25mg gastro-resistant tablets (Teva UK Ltd)
9886	82072020	Dicloflex 50mg Gastro-resistant tablet (Ratiopharm UK Ltd)
39823	96610020	Dicloflex 50mg gastro-resistant tablets (Almus Pharmaceuticals Ltd)
4692	50465020	Dicloflex 50mg gastro-resistant tablets (Dexcel-Pharma Ltd)
46844	99460020	Dicloflex 75mg SR tablets (Actavis UK Ltd)
29181	91117020	Dicloflex 75mg SR tablets (Almus Pharmaceuticals Ltd)
9222	79017020	Dicloflex 75mg SR tablets (Dexcel-Pharma Ltd)
20621	84866020	Dicloflex 75mg SR tablets (Kent Pharmaceuticals Ltd)
20805	85934020	Dicloflex 75mg SR tablets (Teva UK Ltd)
35893	91115020	Dicloflex Retard 100mg tablets (Almus Pharmaceuticals Ltd)
39264	95833020	Dicloflex Retard 100mg tablets (Dexcel-Pharma Ltd)
17532	84868020	Dicloflex Retard 100mg tablets (Kent Pharmaceuticals Ltd)
42455	97700020	Dicloflex Retard 100mg tablets (Teva UK Ltd)
8789	50466020	Dicloflex retard tabs 100 100mg Modified-release tablet (Dexcel-Pharma Ltd)
17124	85935020	Dicloflex sr 100mg Tablet (IVAX Pharmaceuticals UK Ltd)
48218	629021	Dicloflex sr 100mg Tablet (Teva UK Ltd)
30790	84507020	Dicloflex sr 75mg Tablet (Genus Pharmaceuticals Ltd)
17491	76698020	Dicloflex sr 75mg Tablet (Ratiopharm UK Ltd)
3852	74400020	Diclomax 100mg Modified-release capsule (Provalis Healthcare Ltd)
74835	8102020	Diclomax Retard 100mg capsules (DE Pharmaceuticals)
38948	95891020	Diclomax Retard 100mg capsules (Galen Ltd)
74048	8101020	Diclomax Retard 100mg capsules (Mawdsley-Brooks & Company Ltd)
71362	8100020	Diclomax Retard 100mg capsules (Waymade Healthcare Plc)
71117	8046020	Diclomax SR 75mg capsules (DE Pharmaceuticals)
38881	95549020	Diclomax SR 75mg capsules (Galen Ltd)
74028	8047020	Diclomax SR 75mg capsules (Waymade Healthcare Plc)
3421	74401020	Diclomax sr 75mg Modified-release capsule (Provalis Healthcare Ltd)
9465	84244020	Diclotard 100 100mg Modified-release tablet (Galen Ltd)
9500	84230020	Diclotard 75mg modified-release tablets (Galen Ltd)
25361	86198020	Diclovol 25mg gastro-resistant tablets (Arun Pharmaceuticals Ltd)
15732	86199020	Diclovol 50mg gastro-resistant tablets (Arun Pharmaceuticals Ltd)
14084	86201020	Diclovol 75mg SR tablets (Arun Pharmaceuticals Ltd)
9688	86796020	Diclovol 75mg SR tablets (Mylan)
14085	86203020	Diclovol Retard 100mg tablets (Arun Pharmaceuticals Ltd)
27200	86798020	Diclovol Retard 100mg tablets (Mylan)

16221	57190020	Diclozip 25mg gastro-resistant tablets (Ashbourne Pharmaceuticals Ltd)
16222	57191020	Diclozip 50mg gastro-resistant tablets (Ashbourne Pharmaceuticals Ltd)
26888	85826020	Difenor xl 100mg Modified-release tablet (IVAX Pharmaceuticals UK Ltd)
18371	83511020	Digenac xl 100mg Modified-release tablet (Genus Pharmaceuticals Ltd)
48810	8229020	Dysman 250 capsules (Ashbourne Pharmaceuticals Ltd)
21831	56666020	Dysman 250mg Capsule (Ashbourne Pharmaceuticals Ltd)
13459	56667020	Dysman 500 tablets (Ashbourne Pharmaceuticals Ltd)
29587	59451020	Ebufac 400mg Tablet (DDSA Pharmaceuticals Ltd)
31787	87660020	Econac SR 75mg tablets (AMCo)
36486	87664020	Econac XL 100mg tablets (AMCo)
2258	53285020	Emflex 60mg capsules (Merck Serono Ltd)
23468	!2511301	FELDENE
341	49419020	Feldene 10mg capsules (Pfizer Ltd)
2827	55305020	Feldene 10mg dispersible tablets (Pfizer Ltd)
3935	49420020	Feldene 20 capsules (Pfizer Ltd)
7524	55306020	Feldene 20mg dispersible tablets (Pfizer Ltd)
3409	55307020	Feldene 20mg Orodispersible tablet (Pfizer Ltd)
19560	!2511302	FELDENE DISPERSIBLE
19788	!2511103	FELDENE DISPERSIBLE
39109	95883020	Feldene Melt 20mg tablets (Pfizer Ltd)
73981	8287020	Feldene Melt 20mg tablets (Sigma Pharmaceuticals Plc)
67815	8286020	Feldene Melt 20mg tablets (Waymade Healthcare Plc)
43904	98003020	Feminax Express 342mg tablets (Bayer Plc)
38511	94811020	Feminax Ultra 250mg gastro-resistant tablets (Bayer Plc)
18921	77276020	Fenactol 25mg gastro-resistant tablets (Discovery Pharmaceuticals)
17128	77273020	Fenactol 50mg gastro-resistant tablets (Discovery Pharmaceuticals)
17525	78919020	Fenactol Retard 100mg tablets (Discovery Pharmaceuticals)
17126	55714020	Fenactol SR 75mg tablets (Discovery Pharmaceuticals)
10785	53757020	Fenbid 300mg Spansules (Mercury Pharma Group Ltd)
24687	!2517102	FENBUFEN
7424	61799020	Fenbufen 300mg capsules
8145	61800020	Fenbufen 300mg tablets
14422	61806020	Fenbufen 450mg Effervescent tablet
8544	61801020	Fenbufen 450mg tablets
74641	53555020	Fenbufen 450mg tablets (A A H Pharmaceuticals Ltd)
26205	74759020	Fenbuzip 300mg Capsule (Ashbourne Pharmaceuticals Ltd)
26994	74757020	Fenbuzip 300mg Tablet (Ashbourne Pharmaceuticals Ltd)
26214	74758020	Fenbuzip 450mg Tablet (Ashbourne Pharmaceuticals Ltd)
18647	78164020	Fenoket 200mg modified-release capsules (Opus Pharmaceuticals Ltd)

4564	61812020	Fenoprofen 200mg Tablet
4469	61813020	Fenoprofen 300mg tablets
4565	61814020	Fenoprofen 600mg tablets
22158	4985007	FENOPROFEN disp 300 MG TAB
15477	2579007	FENOPRON 200 MG TAB
10678	49424020	Fenopron 300 tablets (Typharm Ltd)
10589	49425020	Fenopron 600 tablets (Typharm Ltd)
25760	5834007	FENOPRON D 300 MG TAB
18820	84269020	Fenpaed 100mg/5ml Oral suspension (Pinewood Healthcare)
65121	8176020	Fenpaed 100mg/5ml oral suspension (Pinewood Healthcare)
25800	55616020	Feverfen 100mg/5ml oral suspension (Wise Pharmaceuticals Ltd)
45814	99265020	First Resort Double Action Pain Relief 12.5mg tablets (Actavis UK Ltd)
20384	85672020	Flamatak MR 100mg tablets (Actavis UK Ltd)
20395	85673020	Flamatak MR 75mg tablets (Actavis UK Ltd)
26234	75358020	Flamatrol 10mg Capsule (Berk Pharmaceuticals Ltd)
21807	74801020	Flamrase 25 EC tablets (Teva UK Ltd)
21824	74802020	Flamrase 50 EC tablets (Teva UK Ltd)
38992	95555020	Flamrase 75mg SR tablets (Teva UK Ltd)
26212	!8504179	FLAMRASE SR
10917	74803020	Flamrase SR 100mg tablets (Teva UK Ltd)
11322	83762020	Flamrase sr 75mg Modified-release tablet (APS Berk)
71949	78255021	Flarin 200mg capsules (infirst Healthcare Ltd)
29455	84251020	Flexotard MR 100mg tablets (Pfizer Ltd)
20161	4975007	FLUFENAMIC ACID 100 MG CAP
6249	49560020	Froben 100mg tablets (Abbott Laboratories Ltd)
3182	49559020	Froben 50mg tablets (Abbott Laboratories Ltd)
38944	95551020	Froben SR 200mg capsules (Abbott Laboratories Ltd)
4043	68387020	Froben sr 200mg Modified-release capsule (Abbott Laboratories Ltd)
39354	90397020	Galpharm ibuprofen for children 100mg/5ml Oral suspension (Galpharm International Ltd)
71374	14183020	Galpharm Ibuprofen For Children 100mg/5ml oral suspension 5ml sachets (Galpharm International Ltd)
30724	86821020	Galprofen 100mg/5ml oral suspension (Galpharm International Ltd)
33785	83372020	Galprofen 200mg tablets (Galpharm International Ltd)
28888	84592020	Galprofen Long Lasting 200mg capsules (Galpharm International Ltd)
75305	93283020	Galprofen Long Lasting 300mg capsules (Galpharm International Ltd)
36597	82132020	Hedex Ibuprofen 200mg tablets (Omega Pharma Ltd)
38332	94471020	Ibucalm 200mg tablets (Aspar Pharmaceuticals Ltd)
37553	94475020	Ibucalm 400mg tablets (Aspar Pharmaceuticals Ltd)

24305	80560020	Ibufac 400mg Tablet (DDSA Pharmaceuticals Ltd)
10209	80061020	Ibufem 200mg tablets (Galpharm International Ltd)
32136	58455020	Ibular 200mg Tablet (Lagap)
76284	82706021	Ibular 200mg tablets (Ennogen Pharma Ltd)
18364	58456020	Ibular 400mg Tablet (Lagap)
76041	82708021	Ibular 400mg tablets (Ennogen Pharma Ltd)
849	74754020	Ibumed 400mg Tablet (Medipharma Ltd)
21045	59456020	Ibumetin 400mg Tablet (Alfred Benzon (UK) Ltd)
29524	63506020	Ibumetin 600mg Tablet (Alfred Benzon (UK) Ltd)
66247	65964021	Ibuprofen 100mg chewable capsules
37235	94243020	Ibuprofen 100mg/5ml / Pseudoephedrine 15mg/5ml oral suspension sugar free
647	63502020	Ibuprofen 100mg/5ml oral suspension
2938	63503020	Ibuprofen 100mg/5ml Oral suspension
29345	65371020	Ibuprofen 100mg/5ml Oral suspension (Hillcross Pharmaceuticals Ltd)
34663	68074020	Ibuprofen 100mg/5ml Oral suspension (Neo Laboratories Ltd)
48562	14177020	Ibuprofen 100mg/5ml oral suspension 5ml sachets sugar free
25205	89068020	Ibuprofen 100mg/5ml oral suspension 5ml sachets sugar free (Thornton & Ross Ltd)
48326	8170020	Ibuprofen 100mg/5ml oral suspension sugar free
33704	63505020	Ibuprofen 100mg/5ml oral suspension sugar free (A A H Pharmaceuticals Ltd)
53331	8171020	Ibuprofen 100mg/5ml oral suspension sugar free (Alliance Healthcare (Distribution) Ltd)
51828	8175020	Ibuprofen 100mg/5ml oral suspension sugar free (Kent Pharmaceuticals Ltd)
29332	62712020	Ibuprofen 100mg/5ml oral suspension sugar free (Sandoz Ltd)
52617	8184020	Ibuprofen 100mg/5ml oral suspension sugar free (Sigma Pharmaceuticals Plc)
26970	68245020	Ibuprofen 100mg/5ml oral suspension sugar free (Teva UK Ltd)
32862	71142020	Ibuprofen 100mg/5ml oral suspension sugar free (Thornton & Ross Ltd)
29352	68023020	Ibuprofen 100mg/5ml oral suspension sugar free (Vantage)
215	4856007	IBUPROFEN 200 MG CAP
11554	77460020	Ibuprofen 200mg / Codeine 12.8mg tablets
45988	97483020	Ibuprofen 200mg / Phenylephrine 5mg tablets
28522	75899020	Ibuprofen 200mg / Pseudoephedrine hydrochloride 30mg tablets
49277	8143020	Ibuprofen 200mg caplets (Bristol Laboratories Ltd)
40083	88937020	Ibuprofen 200mg caplets (Galpharm International Ltd)
51614	8135020	Ibuprofen 200mg caplets (Lloyds Pharmacy Ltd)
50266	8136020	Ibuprofen 200mg caplets (The Boots Company Plc)

61953	29732021	Ibuprofen 200mg caplets (Wockhardt UK Ltd)
586	80144020	Ibuprofen 200mg Capsule
10149	86472020	Ibuprofen 200mg capsules
59067	17816021	Ibuprofen 200mg capsules (AM Distributions (Yorkshire) Ltd)
30243	50930020	Ibuprofen 200mg effervescent tablets
75338	82223021	Ibuprofen 200mg medicated plasters
392	53037020	Ibuprofen 200mg modified-release capsules
5648	80145020	Ibuprofen 200mg orodispersible tablets sugar free
1468	71645020	Ibuprofen 200mg Soluble tablet
30382	53953020	Ibuprofen 200mg Tablet (C P Pharmaceuticals Ltd)
34911	49827020	Ibuprofen 200mg Tablet (Celltech Pharma Europe Ltd)
45331	62584020	Ibuprofen 200mg Tablet (Co-Pharma Ltd)
34621	66405020	Ibuprofen 200mg Tablet (Nucare Plc)
34931	62843020	Ibuprofen 200mg Tablet (Regent Laboratories Ltd)
416	59354020	Ibuprofen 200mg tablets
16001	49832020	Ibuprofen 200mg tablets (A A H Pharmaceuticals Ltd)
65471	69610020	Ibuprofen 200mg tablets (Almus Pharmaceuticals Ltd)
52154	67026020	Ibuprofen 200mg tablets (Galpharm International Ltd)
41513	57811020	Ibuprofen 200mg tablets (IVAX Pharmaceuticals UK Ltd)
42108	62385020	Ibuprofen 200mg tablets (OBG Pharmaceuticals Ltd)
29749	60159020	Ibuprofen 200mg tablets (Ranbaxy (UK) Ltd)
45320	59470020	Ibuprofen 200mg tablets (Sandoz Ltd)
28348	53366020	Ibuprofen 200mg tablets (Teva UK Ltd)
34447	57304020	Ibuprofen 200mg tablets (Thornton & Ross Ltd)
34354	56266020	Ibuprofen 200mg tablets (Vantage)
34527	59965020	Ibuprofen 200mg tablets (Zentiva)
60035	69002020	Ibuprofen 200mg tablets film coated (Actavis UK Ltd)
34980	53948020	Ibuprofen 200mg tablets sugar coated (Actavis UK Ltd)
48084	94572020	Ibuprofen 200mg/5ml oral suspension
75893	85432021	Ibuprofen 200mg/5ml oral suspension sugar free
28172	75901020	Ibuprofen 300mg / Pseudoephedrine 45mg modified-release capsules
11461	59563020	Ibuprofen 300mg modified-release / Codeine 20mg tablets
784	63501020	Ibuprofen 300mg modified-release capsules
48546	8156020	Ibuprofen 400mg caplets (Bristol Laboratories Ltd)
48644	8148020	Ibuprofen 400mg caplets (Lloyds Pharmacy Ltd)
50628	8149020	Ibuprofen 400mg caplets (The Boots Company Plc)
14333	91245020	Ibuprofen 400mg capsules
4911	80143020	Ibuprofen 400mg Granules
45216	53954020	Ibuprofen 400mg Tablet (C P Pharmaceuticals Ltd)
34889	49829020	Ibuprofen 400mg Tablet (Celltech Pharma Europe Ltd)
34425	53852020	Ibuprofen 400mg Tablet (Family Health)
34757	59877020	Ibuprofen 400mg Tablet (Unichem)
15	59355020	Ibuprofen 400mg tablets

19046	49833020	Ibuprofen 400mg tablets (A A H Pharmaceuticals Ltd)
57112	8146020	Ibuprofen 400mg tablets (Alliance Healthcare (Distribution) Ltd)
34536	57812020	Ibuprofen 400mg tablets (IVAX Pharmaceuticals UK Ltd)
34729	62386020	Ibuprofen 400mg tablets (OBG Pharmaceuticals Ltd)
75677	8155020	Ibuprofen 400mg tablets (Phoenix Healthcare Distribution Ltd)
46921	60160020	Ibuprofen 400mg tablets (Ranbaxy (UK) Ltd)
32875	59471020	Ibuprofen 400mg tablets (Sandoz Ltd)
27782	53365020	Ibuprofen 400mg tablets (Teva UK Ltd)
33589	57305020	Ibuprofen 400mg tablets (Thornton & Ross Ltd)
34359	56114020	Ibuprofen 400mg tablets (Vantage)
34550	69005020	Ibuprofen 400mg tablets film coated (Actavis UK Ltd)
27783	53949020	Ibuprofen 400mg tablets sugar coated (Actavis UK Ltd)
56213	8145020	Ibuprofen 400mg tablets sugar coated (Kent Pharmaceuticals Ltd)
3599	50929020	Ibuprofen 600mg effervescent granules sachets
43911	53955020	Ibuprofen 600mg Tablet (C P Pharmaceuticals Ltd)
45842	49828020	Ibuprofen 600mg Tablet (Celltech Pharma Europe Ltd)
40253	62183020	Ibuprofen 600mg Tablet (Sovereign Medical Ltd)
1086	59356020	Ibuprofen 600mg tablets
32100	49834020	Ibuprofen 600mg tablets (A A H Pharmaceuticals Ltd)
41701	53950020	Ibuprofen 600mg tablets (Actavis UK Ltd)
67740	73278020	Ibuprofen 600mg tablets (Fannin UK Ltd)
46942	55366020	Ibuprofen 600mg tablets (IVAX Pharmaceuticals UK Ltd)
34961	59472020	Ibuprofen 600mg tablets (Sandoz Ltd)
58652	8163020	Ibuprofen 600mg tablets (Sigma Pharmaceuticals Plc)
34850	53364020	Ibuprofen 600mg tablets (Teva UK Ltd)
1392	50928020	Ibuprofen 800mg modified-release tablets
2622	68109020	Ibuprofen 800mg tablets
12709	77459020	Ibuprofen and codeine 200mg + 12.5mg Tablet
49266	8183020	Ibuprofen for Children 100mg/5ml oral suspension (Galpharm International Ltd)
4309	86342020	Ibuprofen lysine 200mg tablets
54514	46754020	Ibuprofen lysine 400mg oral powder sachets
26095	90473020	Ibuprofen lysine 400mg tablets
345	3721007	IBUPROFEN S/R 300 MG CAP
39502	95773020	Ibuprofen sodium dihydrate 200mg tablets
66567	95775020	Ibuprofen sodium dihydrate 400mg tablets
76234	85433021	Ibuprofen Twelve Plus Pain Relief 200mg/5ml oral suspension (Aspire Pharma Ltd)
28822	75900020	Ibuprofen with pseudoephedrine hc 400mg + 60mg Liquid
30821	6285007	INDOPROFEN 200 MG TAB
43032	84356020	Inoven 200mg Tablet (Janssen-Cilag Ltd)
33457	68415020	Isclofen 50mg Gastro-resistant tablet (Isis Products Ltd)

25794	56702020	Isisfen 400mg Tablet (Isis Products Ltd)
30327	84738020	Jomethid XL 200mg capsules (Actavis UK Ltd)
1030	72737020	Junifen 100mg/5ml Oral suspension (Crookes Healthcare Ltd)
50652	8173020	Junior Ibuprofen 100mg/5ml oral suspension (Numark Ltd)
9637	79359020	Keral 25mg tablets (A. Menarini Farmaceutica Internazionale SRL)
15286	77485020	Ketocid 200 modified-release capsules (Chiesi Ltd)
21050	74798020	Ketonal 100mg Capsule (Lagap)
41364	97642020	Ketoprofen 100mg / Omeprazole 20mg modified-release capsules
1231	63884020	Ketoprofen 100mg capsules
40141	55793020	Ketoprofen 100mg capsules (A A H Pharmaceuticals Ltd)
46940	59918020	Ketoprofen 100mg capsules (Mylan)
1571	69511020	Ketoprofen 100mg modified-release capsules
75573	55791020	Ketoprofen 100mg modified-release capsules (A A H Pharmaceuticals Ltd)
8385	69513020	Ketoprofen 150mg modified-release capsules
41367	97644020	Ketoprofen 200mg / Omeprazole 20mg modified-release capsules
33568	57010020	Ketoprofen 200mg Modified-release capsule (Actavis UK Ltd)
46920	56131020	Ketoprofen 200mg Modified-release capsule (Generics (UK) Ltd)
3043	69512020	Ketoprofen 200mg modified-release capsules
77293	55792020	Ketoprofen 200mg modified-release capsules (A A H Pharmaceuticals Ltd)
389	63883020	Ketoprofen 50mg capsules
75581	53673020	Ketoprofen cr 100mg Capsule (Bristol-Myers Squibb Pharmaceuticals Ltd)
33180	53674020	Ketoprofen cr 200mg Capsule (Bristol-Myers Squibb Pharmaceuticals Ltd)
42500	53591020	Ketoprofen sr 100mg Capsule (Approved Prescription Services Ltd)
46919	53592020	Ketoprofen sr 200mg Capsule (Approved Prescription Services Ltd)
16637	69066020	Ketorolac 10mg tablets
29772	82614020	Ketotard XL 200mg capsules (Galen Ltd)
17818	74336020	Ketovail 100mg modified-release capsules (Teva UK Ltd)
25701	74337020	Ketovail 200mg modified-release capsules (Teva UK Ltd)
21955	53056020	Ketozip 200 XL capsules (Ashbourne Pharmaceuticals Ltd)
31962	84318020	Ketpron XL 200mg capsules (Mercury Pharma Group Ltd)
75771	69043020	Larafen 100mg Capsule (Sandoz Ltd)
32227	80610020	Larafen CR 200mg capsules (Ennogen Pharma Ltd)
7426	50135020	Lederfen 300mg Capsule (Wyeth Pharmaceuticals)
14380	84600020	Lederfen 300mg capsules (Mercury Pharma Group Ltd)
7522	50134020	Lederfen 300mg Tablet (Wyeth Pharmaceuticals)

17131	57844020	Lederfen 300mg tablets (Mercury Pharma Group Ltd)
7481	50136020	Lederfen 450mg Tablet (Wyeth Pharmaceuticals)
16176	77247020	Lederfen 450mg tablets (Mercury Pharma Group Ltd)
10481	54744020	Lederfen f 450mg Tablet (Wyeth Pharmaceuticals)
30164	88340020	Lemsip Cold and Flu Sinus 12 Hr Ibuprofen + Pseudoephedrine modified-release capsules (Reckitt Benckiser Healthcare (UK) Ltd)
22283	85582020	Lemsip flu 12 hr Modified-release capsule (Reckitt Benckiser Healthcare (UK) Ltd)
21811	59661020	Lidifen 200mg Tablet (Berk Pharmaceuticals Ltd)
21813	59662020	Lidifen 400mg Tablet (Berk Pharmaceuticals Ltd)
21821	59663020	Lidifen f 600mg Tablet (Berk Pharmaceuticals Ltd)
25329	79838020	Lofensaid 25mg gastro-resistant tablets (Opus Pharmaceuticals Ltd)
18798	79839020	Lofensaid 50mg gastro-resistant tablets (Opus Pharmaceuticals Ltd)
16272	82238020	Lofensaid Retard 100 tablets (Opus Pharmaceuticals Ltd)
16286	82237020	Lofensaid Retard 75 tablets (Opus Pharmaceuticals Ltd)
29110	83611020	Lornoxicam 4mg tablets
30122	83613020	Lornoxicam 8mg tablets
18527	86224020	Mandafen 400mg tablets (M & A Pharmachem Ltd)
30892	84841020	Mandafen for Children 100mg/5ml oral suspension sugar free (M & A Pharmachem Ltd)
36606	83343020	Manorfen 400mg tablets (The Manor Drug Company (Nottingham) Ltd)
46342	99120020	Medifen 3with months 100mg/5ml Oral suspension (SSL International Plc)
4710	49999020	Mefenamic acid 250mg Capsule (Actavis UK Ltd)
34898	49994020	Mefenamic acid 250mg Capsule (Berk Pharmaceuticals Ltd)
41677	54078020	Mefenamic acid 250mg Capsule (IVAX Pharmaceuticals UK Ltd)
46967	59436020	Mefenamic acid 250mg Capsule (Sandoz Ltd)
34924	55454020	Mefenamic acid 250mg Capsule (Teva UK Ltd)
259	59793020	Mefenamic acid 250mg capsules
34438	50007020	Mefenamic acid 250mg capsules (A A H Pharmaceuticals Ltd)
70221	8228020	Mefenamic acid 250mg capsules (Alliance Healthcare (Distribution) Ltd)
57007	76223020	Mefenamic acid 250mg capsules (Essential Generics Ltd)
46968	60540020	Mefenamic acid 250mg capsules (Mylan)
34793	60065020	Mefenamic acid 250mg capsules (Zentiva)
1983	64363020	Mefenamic acid 250mg Dispersible tablet
75154	20463020	Mefenamic acid 250mg/5ml oral suspension
34910	49995020	Mefenamic acid 500mg Tablet (Berk Pharmaceuticals Ltd)
1073	64364020	Mefenamic acid 500mg tablets
32105	50008020	Mefenamic acid 500mg tablets (A A H Pharmaceuticals Ltd)

32090	50000020	Mefenamic acid 500mg tablets (Actavis UK Ltd)
57297	8234020	Mefenamic acid 500mg tablets (Alliance Healthcare (Distribution) Ltd)
64103	78544020	Mefenamic acid 500mg tablets (Almus Pharmaceuticals Ltd)
61581	76226020	Mefenamic acid 500mg tablets (Essential Generics Ltd)
32234	54079020	Mefenamic acid 500mg tablets (IVAX Pharmaceuticals UK Ltd)
51827	8236020	Mefenamic acid 500mg tablets (Sigma Pharmaceuticals Plc)
41524	55455020	Mefenamic acid 500mg tablets (Teva UK Ltd)
34595	60066020	Mefenamic acid 500mg tablets (Zentiva)
76310	20465020	Mefenamic acid 500mg/5ml oral suspension
9736	64365020	Mefenamic acid 50mg/5ml oral suspension
20709	!4405104	MEFENAMIC ACID DISPERSIBLE
22230	78062020	Meflam 250mg Capsule (Trinity Pharmaceuticals Ltd)
26522	78063020	Meflam 500mg Tablet (Trinity Pharmaceuticals Ltd)
36260	85863020	Mendys 250mg Capsule (Kent Pharmaceuticals Ltd)
37053	82155020	Migrafen 200mg tablets (Chatfield Laboratories)
64595	16443021	Misofen 50mg/200microgram gastro-resistant tablets (Morningside Healthcare Ltd)
58842	16444021	Misofen 75mg/200microgram gastro-resistant tablets (Morningside Healthcare Ltd)
24531	69422020	Mobiflex 20mg Effervescent tablet (Roche Products Ltd)
31064	69421020	Mobiflex 20mg Granules (Roche Products Ltd)
12075	69420020	Mobiflex 20mg Tablet (Roche Products Ltd)
71152	8317020	Mobiflex 20mg tablets (Dowelhurst Ltd)
42604	98118020	Mobiflex 20mg tablets (Meda Pharmaceuticals Ltd)
8062	75820020	Motifene 75mg modified-release capsules (Daiichi Sankyo UK Ltd)
16192	56328020	Motrin 200mg Tablet (Pharmacia Ltd)
8401	56329020	Motrin 400mg tablets (Pfizer Ltd)
17201	56330020	Motrin 600mg tablets (Pfizer Ltd)
16193	68112020	Motrin 800mg tablets (Pfizer Ltd)
16474	73719020	Nabumetone 500mg dispersible tablets sugar free
2234	68000020	Nabumetone 500mg tablets
42821	61088020	Nabumetone 500mg tablets (A A H Pharmaceuticals Ltd)
13818	60823020	Nabumetone 500mg tablets (Actavis UK Ltd)
64297	62443020	Nabumetone 500mg tablets (Mylan)
11466	68001020	Nabumetone 500mg/5ml oral-suspension
19559	!4802101	NAPROSYN
23268	!4802102	NAPROSYN
4320	57758020	Naprosyn 125mg/5ml oral suspension (Roche Products Ltd)
2288	57757020	Naprosyn 250mg tablets (Atnahs Pharma UK Ltd)
34143	73699020	Naprosyn 375 Tablet (Roche Products Ltd)
19007	52497020	Naprosyn 500mg Granules (Roche Products Ltd)
1866	57763020	Naprosyn 500mg tablets (Atnahs Pharma UK Ltd)

3972	67330020	Naprosyn EC 250mg tablets (Atnahs Pharma UK Ltd)
4045	67331020	Naprosyn EC 375mg tablets (Atnahs Pharma UK Ltd)
3901	67332020	Naprosyn EC 500mg tablets (Atnahs Pharma UK Ltd)
8663	74549020	Naprosyn S/R 500mg tablets (Roche Products Ltd)
28313	!4804105	NAPROXEN
56762	31715020	Naproxen 100mg/5ml oral suspension
5407	64942020	Naproxen 125mg/5ml oral suspension
66993	68881021	Naproxen 125mg/5ml oral suspension sugar free
39693	96624020	Naproxen 200mg/5ml oral suspension
2391	4699007	NAPROXEN 250 MG CAP
65862	61301021	Naproxen 250mg effervescent tablets sugar free
34670	59135020	Naproxen 250mg Gastro-resistant tablet (Galen Ltd)
3431	74064020	Naproxen 250mg gastro-resistant tablets
34738	60413020	Naproxen 250mg gastro-resistant tablets (A A H Pharmaceuticals Ltd)
65348	40593020	Naproxen 250mg gastro-resistant tablets (Genesis Pharmaceuticals Ltd)
40401	57817020	Naproxen 250mg gastro-resistant tablets (IVAX Pharmaceuticals UK Ltd)
34289	59490020	Naproxen 250mg gastro-resistant tablets (Mylan)
34290	53361020	Naproxen 250mg gastro-resistant tablets (Teva UK Ltd)
34923	49074020	Naproxen 250mg Tablet (Berk Pharmaceuticals Ltd)
661	58922020	Naproxen 250mg tablets
39085	49087020	Naproxen 250mg tablets (A A H Pharmaceuticals Ltd)
51829	8242020	Naproxen 250mg tablets (Kent Pharmaceuticals Ltd)
53980	8247020	Naproxen 250mg tablets (Phoenix Healthcare Distribution Ltd)
54783	53359020	Naproxen 250mg tablets (Teva UK Ltd)
68685	47562020	Naproxen 250mg tablets (Waymade Healthcare Plc)
28255	49079020	Naproxen 250mg tablets (Wockhardt UK Ltd)
56554	20548020	Naproxen 250mg/5ml oral suspension
68470	68882021	Naproxen 25mg/ml oral suspension sugar free (Orion Pharma (UK) Ltd)
3432	74066020	Naproxen 375mg gastro-resistant tablets
15023	74183020	Naproxen 375mg Modified-release tablet
2197	58924020	Naproxen 375mg Tablet
44800	99218020	Naproxen 500mg / Esomeprazole 20mg modified-release tablets
46848	75583020	Naproxen 500mg Gastro-resistant tablet (Almus Pharmaceuticals Ltd)
34977	59136020	Naproxen 500mg Gastro-resistant tablet (Galen Ltd)
31945	59110020	Naproxen 500mg Gastro-resistant tablet (Sterwin Medicines)
3053	74065020	Naproxen 500mg gastro-resistant tablets
34743	49089020	Naproxen 500mg gastro-resistant tablets (A A H Pharmaceuticals Ltd)

30982	61530020	Naproxen 500mg gastro-resistant tablets (Actavis UK Ltd)
54476	40594020	Naproxen 500mg gastro-resistant tablets (Genesis Pharmaceuticals Ltd)
34610	59491020	Naproxen 500mg gastro-resistant tablets (Mylan)
27366	57207020	Naproxen 500mg gastro-resistant tablets (Teva UK Ltd)
15104	74182020	Naproxen 500mg Granules
5268	74184020	Naproxen 500mg modified-release tablets
48161	75930020	Naproxen 500mg Tablet (Almus Pharmaceuticals Ltd)
34922	49075020	Naproxen 500mg Tablet (Berk Pharmaceuticals Ltd)
46440	63631020	Naproxen 500mg Tablet (M & A Pharmachem Ltd)
807	58923020	Naproxen 500mg tablets
34769	49088020	Naproxen 500mg tablets (A A H Pharmaceuticals Ltd)
54304	57065020	Naproxen 500mg tablets (Actavis UK Ltd)
55486	53360020	Naproxen 500mg tablets (Teva UK Ltd)
39317	49080020	Naproxen 500mg tablets (Wockhardt UK Ltd)
4984	66530020	Naproxen 500mg tablets and Misoprostol 200microgram tablets
56106	20550020	Naproxen 500mg/5ml oral suspension
71709	75871021	Naproxen 50mg/ml oral suspension (A A H Pharmaceuticals Ltd)
69828	75196021	Naproxen 50mg/ml oral suspension (Alliance Healthcare (Distribution) Ltd)
69645	73768021	Naproxen 50mg/ml oral suspension (Thornton & Ross Ltd)
76955	31720020	Naproxen 75mg/5ml oral suspension
15180	86566020	Naproxen and misoprostol 500mgwith200microgram combined Tablet
45262	92261020	Naproxen Oral solution
20704	!4805101	NAPROXEN SODIUM
1043	64945020	Naproxen sodium 275mg tablets
35890	93688020	Nurofen 200mg caplets (Reckitt Benckiser Healthcare (UK) Ltd)
7535	85652020	Nurofen 200mg Capsule (Crookes Healthcare Ltd)
35292	93690020	Nurofen 200mg liquid capsules (Reckitt Benckiser Healthcare (UK) Ltd)
3597	73682020	Nurofen 200mg Soluble tablet (Crookes Healthcare Ltd)
402	85651020	Nurofen 200mg Tablet (Crookes Healthcare Ltd)
4298	73681020	Nurofen 200mg Tablet (Crookes Healthcare Ltd)
36650	93686020	Nurofen 200mg tablets (Reckitt Benckiser Healthcare (UK) Ltd)
25619	73683020	Nurofen 400mg Tablet (Crookes Healthcare Ltd)
24887	86353020	Nurofen Advance 200mg tablets (Crookes Healthcare Ltd)
28479	88706020	Nurofen Back Pain SR 300mg capsules (Reckitt Benckiser Healthcare (UK) Ltd)
72156	21548021	Nurofen Cold & Flu Relief 200mg/5mg tablets (Reckitt Benckiser Healthcare (UK) Ltd)
15363	75904020	Nurofen Cold and Flu tablets (Reckitt Benckiser Healthcare (UK) Ltd)

37002	93924020	Nurofen Express 200mg liquid capsules (Reckitt Benckiser Healthcare (UK) Ltd)
39758	95779020	Nurofen Express 256mg caplets (Reckitt Benckiser Healthcare (UK) Ltd)
42397	95785020	Nurofen Express 256mg tablets (Reckitt Benckiser Healthcare (UK) Ltd)
37731	93916020	Nurofen Express 342mg caplets (Reckitt Benckiser Healthcare (UK) Ltd)
37648	93920020	Nurofen Express 400mg liquid capsules (Reckitt Benckiser Healthcare (UK) Ltd)
44483	95783020	Nurofen Express 512mg tablets (Reckitt Benckiser Healthcare (UK) Ltd)
36787	93922020	Nurofen Express 684mg caplets (Reckitt Benckiser Healthcare (UK) Ltd)
61878	45498020	Nurofen Express Period Pain 200mg capsules (Reckitt Benckiser Healthcare (UK) Ltd)
55153	46755020	Nurofen Express Soluble 400mg oral powder sachets (Reckitt Benckiser Healthcare (UK) Ltd)
29068	91247020	Nurofen Extra Strength 400mg capsules (Reckitt Benckiser Healthcare (UK) Ltd)
73040	65965021	Nurofen for Children 100mg chewable capsules (Reckitt Benckiser Healthcare (UK) Ltd)
4731	86753020	Nurofen for children 100mg/5ml Oral suspension (Reckitt Benckiser Healthcare (UK) Ltd)
48738	8174020	Nurofen for Children 100mg/5ml oral suspension orange (Reckitt Benckiser Healthcare (UK) Ltd)
49133	8180020	Nurofen for Children 100mg/5ml oral suspension strawberry (Reckitt Benckiser Healthcare (UK) Ltd)
35265	93625020	Nurofen for children 3 months to 9 years 100mg/5ml Oral suspension (Reckitt Benckiser Healthcare (UK) Ltd)
44233	98805020	Nurofen for children baby 100mg/5ml Oral suspension (Reckitt Benckiser Healthcare (UK) Ltd)
60510	21431021	Nurofen for Children Cold, Pain and Fever Orange Flavour 100mg/5ml oral suspension (Reckitt Benckiser Healthcare (UK) Ltd)
59502	21430021	Nurofen for Children Cold, Pain and Fever Strawberry Flavour 100mg/5ml oral suspension (Reckitt Benckiser Healthcare (UK) Ltd)
51769	14178020	Nurofen for Children Singles 100mg/5ml oral suspension 5ml sachets orange (Reckitt Benckiser Healthcare (UK) Ltd)
50363	14181020	Nurofen for Children Singles 100mg/5ml oral suspension 5ml sachets strawberry (Reckitt Benckiser Healthcare (UK) Ltd)
70878	60950021	Nurofen Joint & Back Pain Relief 200mg capsules (Reckitt Benckiser Healthcare (UK) Ltd)
69018	61136021	Nurofen Joint & Back Pain Relief 256mg tablets (Reckitt Benckiser Healthcare (UK) Ltd)

22206	82951020	Nurofen Long Lasting 300mg capsules (Crookes Healthcare Ltd)
33935	90475020	Nurofen Maximum Strength Migraine Pain 684mg caplets (Reckitt Benckiser Healthcare (UK) Ltd)
11550	68069020	Nurofen Meltlets 200mg tablets (Reckitt Benckiser Healthcare (UK) Ltd)
18812	76985020	Nurofen meltlets lemon 200mg Orodispersible tablet (Reckitt Benckiser Healthcare (UK) Ltd)
23425	82652020	Nurofen Migraine Pain 342mg tablets (Reckitt Benckiser Healthcare (UK) Ltd)
13893	77456020	Nurofen Plus tablets (Reckitt Benckiser Healthcare (UK) Ltd)
28168	83270020	Nurofen Recovery 200mg orodispersible tablets (Reckitt Benckiser Healthcare (UK) Ltd)
46141	89638020	Nurofen Tension Headache 342mg caplets (Reckitt Benckiser Healthcare (UK) Ltd)
46904	130021	Nuromol 200mg/500mg tablets (Reckitt Benckiser Healthcare (UK) Ltd)
3496	74060020	Nycopren 250mg gastro-resistant tablets (Ardern Healthcare Ltd)
17165	74061020	Nycopren 500mg gastro-resistant tablets (Ardern Healthcare Ltd)
33801	78352020	Opustan 250mg Capsule (Opus Pharmaceuticals Ltd)
26247	78353020	Opustan 500mg Tablet (Opus Pharmaceuticals Ltd)
38182	94245020	Orbifen Cold & Flu oral suspension (Orbis Consumer Products Ltd)
18196	81617020	Orbifen for children 100mg/5ml Oral suspension (Orbis Consumer Products Ltd)
51943	8172020	Orbifen For Children 100mg/5ml oral suspension (Orbis Consumer Products Ltd)
11999	50920020	Orudis 100mg Capsule (Hawgreen Ltd)
40484	96899020	Orudis 100mg capsules (Sanofi)
12122	50919020	Orudis 50mg Capsule (Hawgreen Ltd)
40336	96897020	Orudis 50mg capsules (Sanofi)
71376	8218020	Oruvail 100 modified-release capsules (Mawdsley-Brooks & Company Ltd)
40215	96903020	Oruvail 100 modified-release capsules (Sanofi)
71127	8217020	Oruvail 100 modified-release capsules (Waymade Healthcare Plc)
3326	50923020	Oruvail 100mg Modified-release capsule (Hawgreen Ltd)
40664	96907020	Oruvail 150 modified-release capsules (Sanofi)
7840	50925020	Oruvail 150mg Modified-release capsule (Hawgreen Ltd)
74005	8222020	Oruvail 200 modified-release capsules (Dowelhurst Ltd)
67803	8225020	Oruvail 200 modified-release capsules (Lexon (UK) Ltd)
40185	96905020	Oruvail 200 modified-release capsules (Sanofi)
71104	8221020	Oruvail 200 modified-release capsules (Waymade Healthcare Plc)
838	50924020	Oruvail 200mg Modified-release capsule (Hawgreen Ltd)

21814	!5214001	ORUVAIL S/R
75720	71032021	Paracetamol 500mg / Ibuprofen 150mg tablets
46638	124021	Paracetamol 500mg / Ibuprofen 200mg tablets
29704	59460020	Paxofen 200mg Tablet (M A Steinhard Ltd)
11952	65700020	Phenylbutazone 100mg gastro-resistant tablets
29010	90907020	Phenylbutazone 100mg tablets
27723	65701020	Phenylbutazone 200mg tablets
28695	79273020	Piroflam 10mg Capsule (Opus Pharmaceuticals Ltd)
19320	79274020	Piroflam 20mg Capsule (Opus Pharmaceuticals Ltd)
20663	!5677102	PIROXICAM
44703	50314020	Piroxicam 10mg Capsule (Berk Pharmaceuticals Ltd)
141	65858020	Piroxicam 10mg capsules
41622	50328020	Piroxicam 10mg capsules (A A H Pharmaceuticals Ltd)
43541	50320020	Piroxicam 10mg capsules (Actavis UK Ltd)
41624	54331020	Piroxicam 10mg capsules (IVAX Pharmaceuticals UK Ltd)
2463	68349020	Piroxicam 10mg dispersible tablets
77185	69505020	Piroxicam 20mg Capsule (Ashbourne Pharmaceuticals Ltd)
21123	50315020	Piroxicam 20mg Capsule (Berk Pharmaceuticals Ltd)
1755	65859020	Piroxicam 20mg capsules
41621	50329020	Piroxicam 20mg capsules (A A H Pharmaceuticals Ltd)
29465	50321020	Piroxicam 20mg capsules (Actavis UK Ltd)
74659	54593020	Piroxicam 20mg capsules (Approved Prescription Services Ltd)
41623	54332020	Piroxicam 20mg capsules (IVAX Pharmaceuticals UK Ltd)
37750	60589020	Piroxicam 20mg capsules (Mylan)
3710	68350020	Piroxicam 20mg dispersible tablets
67608	56350020	Piroxicam 20mg dispersible tablets (A A H Pharmaceuticals Ltd)
31777	60595020	Piroxicam 20mg dispersible tablets (Mylan)
4965	65860020	Piroxicam 20mg orodispersible tablets sugar free
11495	50157020	Piroxicam betadex 20mg tablets
20699	!5677103	PIROXICAM DISPERSIBLE
20742	!5677104	PIROXICAM DISPERSIBLE
21864	56769020	Pirozip 10 capsules (Ashbourne Pharmaceuticals Ltd)
21846	56770020	Pirozip 20 capsules (Ashbourne Pharmaceuticals Ltd)
126	51186020	Ponstan 250mg capsules (Chemidex Pharma Ltd)
1246	51187020	Ponstan 250mg Dispersible tablet (Chemidex Pharma Ltd)
14541	53445020	Ponstan 50mg/5ml paediatric Liquid (Chemidex Pharma Ltd)
296	51182020	Ponstan Forte 500mg tablets (Chemidex Pharma Ltd)
21843	57369020	Pranoxen continus 375mg Tablet (Napp Pharmaceuticals Ltd)
21816	57370020	Pranoxen continus 500mg Tablet (Napp Pharmaceuticals Ltd)
9474	81066020	Preservex 100mg tablets (Almirall Ltd)

19575	63156020	Proflex 200mg Tablet (Novartis Consumer Health UK Ltd)
30811	63157020	Proflex 300mg Modified-release capsule (Novartis Consumer Health UK Ltd)
17754	51288020	Progesic 200mg Tablet (Eli Lilly and Company Ltd)
33111	56630020	Prosaid 250mg Tablet (BHR Pharmaceuticals Ltd)
23323	56631020	Prosaid 500mg Tablet (BHR Pharmaceuticals Ltd)
28519	75925020	Pseudoephedrine 30mg with ibuprofen 200mg tablet
27438	75926020	Pseudoephedrine 45mg with ibuprofen 300mg modified-release capsule
32366	82148020	Relcofen 200mg Tablet (Actavis UK Ltd)
32365	82149020	Relcofen 400mg tablets (Actavis UK Ltd)
16473	60242020	Relifex 500mg dispersible tablets (Meda Pharmaceuticals Ltd)
2235	67974020	Relifex 500mg tablets (Meda Pharmaceuticals Ltd)
10295	67975020	Relifex 500mg/5ml oral suspension (Meda Pharmaceuticals Ltd)
25750	57373020	Rheuflex 250mg Tablet (Goldshield Pharmaceuticals Ltd)
28816	57374020	Rheuflex 500mg Tablet (Goldshield Pharmaceuticals Ltd)
26351	83497020	Rheumatac Retard 75 tablets (AMCo)
10296	4453007	RHEUMOX 100 MG CAP
3739	53354020	Rheumox 300mg capsules (Mercury Pharma Group Ltd)
7688	51376020	Rheumox 600mg tablets (Mercury Pharma Group Ltd)
25790	68882020	Rhumalgan 25mg Tablet (Lagap)
30806	68883020	Rhumalgan 50mg Tablet (Lagap)
21610	80211020	Rhumalgan CR 100 tablets (Sandoz Ltd)
17029	80210020	Rhumalgan CR 75 tablets (Sandoz Ltd)
56898	15708021	Rhumalgan SR 75mg capsules (Actavis UK Ltd)
47501	59021	Rhumalgan SR 75mg capsules (Almus Pharmaceuticals Ltd)
17030	88558020	Rhumalgan SR 75mg capsules (Sandoz Ltd)
56078	61021	Rhumalgan XL 100mg capsules (Almus Pharmaceuticals Ltd)
26631	88560020	Rhumalgan XL 100mg capsules (Sandoz Ltd)
21419	89574020	Seractil 300mg tablets (Thornton & Ross Ltd)
21421	89582020	Seractil 400mg tablets (Thornton & Ross Ltd)
24201	6994007	SLOFENAC 100 MG TAB
24236	83889020	Slofenac 100mg Modified-release tablet (Sterwin Medicines)
21620	6993007	SLOFENAC 75 MG TAB
19382	83888020	Slofenac 75mg SR tablets (Sterwin Medicines)
18922	87111020	Solpadeine Headache soluble tablets (Omega Pharma Ltd)
10196	87109020	Solpadeine Headache tablets (GlaxoSmithKline Consumer Healthcare)
39461	90563020	Solpadeine Migraine Ibuprofen & Codeine tablets (Omega Pharma Ltd)
10178	89794020	Solpadeine Plus capsules (Omega Pharma Ltd)
10226	89796020	Solpadeine Plus tablets (Omega Pharma Ltd)

25330	82982020	Solpaflex tablets (GlaxoSmithKline Consumer Healthcare)
67117	61302021	Stirlescent 250mg effervescent tablets (Stirling Anglian Pharmaceuticals Ltd)
44892	99013020	Sudafed sinus pressure & pain Tablet (McNeil Products Ltd)
20907	82037020	Sudafed Sinus Pressure & Pain tablets (McNeil Products Ltd)
3897	66832020	Sulindac 100mg tablets
5482	66833020	Sulindac 200mg tablets
27916	!6853201	SURGAM
387	51745020	Surgam 200mg tablets (Sanofi)
25643	51747020	Surgam 300mg Sachets (Sanofi)
1778	51746020	Surgam 300mg Tablet (Sanofi)
14776	77350020	Surgam 300mg tablets (Sanofi)
2257	67923020	Surgam SA 300mg capsules (Sanofi)
3817	57767020	Synflex 275mg tablets (Roche Products Ltd)
24682	69417020	Tenoxicam 20mg effervescent tablets
47816	66615020	Tenoxicam 20mg Tablet (Sovereign Medical Ltd)
3974	69415020	Tenoxicam 20mg tablets
27013	86742020	Tiloket 200mg Modified-release capsule (Tillomed Laboratories Ltd)
31916	86741020	Tiloket CR 100mg capsules (Tillomed Laboratories Ltd)
31429	86902020	Timpron 250mg Gastro-resistant tablet (Berk Pharmaceuticals Ltd)
26242	74810020	Timpron 250mg Tablet (Berk Pharmaceuticals Ltd)
26231	74811020	Timpron 500mg Gastro-resistant tablet (Berk Pharmaceuticals Ltd)
26216	74809020	Timpron 500mg Tablet (Berk Pharmaceuticals Ltd)
18640	51959020	Tolectin 200mg Capsule (Cilag Pharmaceuticals Ltd)
10711	51960020	Tolectin 400mg Capsule (Cilag Pharmaceuticals Ltd)
15159	83568020	Tolfenamic acid 200mg Capsule
7222	83569020	Tolfenamic acid 200mg tablets
22410	5629007	TOLMETIN 200 MG TAB
26404	67137020	Tolmetin 200mg Capsule
20016	67138020	Tolmetin 400mg Capsule
3336	69061020	Toradol 10mg tablets (Roche Products Ltd)
29037	81085020	Valdic 100 Retard tablets (Fannin UK Ltd)
30849	81084020	Valdic 75 Retard tablets (Fannin UK Ltd)
28390	73897020	Valenac ec 25mg Gastro-resistant tablet (Shire Pharmaceuticals Ltd)
25283	73898020	Valenac ec 50mg Gastro-resistant tablet (Shire Pharmaceuticals Ltd)
57943	98801020	Valket 200 Retard capsules (Tillomed Laboratories Ltd)
24020	57154020	Valrox 250mg Tablet (Shire Pharmaceuticals Ltd)
24007	57155020	Valrox 500mg Tablet (Shire Pharmaceuticals Ltd)

44986	99220020	Vimovo 500mg/20mg modified-release tablets (AstraZeneca UK Ltd)
21444	74266020	Volraman 25mg gastro-resistant tablets (LPC Medical (UK) Ltd)
15201	74267020	Volraman 50mg gastro-resistant tablets (LPC Medical (UK) Ltd)
11168	80723020	Volsaid Retard 100 tablets (Chiesi Ltd)
4506	80722020	Volsaid Retard 75 tablets (Chiesi Ltd)
497	83709020	Voltarol 25mg gastro-resistant tablets (Novartis Pharmaceuticals UK Ltd)
1139	53288020	Voltarol 25mg Tablet (Novartis Pharmaceuticals UK Ltd)
50058	8086020	Voltarol 50mg dispersible tablets (DE Pharmaceuticals)
49059	8090020	Voltarol 50mg dispersible tablets (Lexon (UK) Ltd)
589	72769020	Voltarol 50mg dispersible tablets (Novartis Pharmaceuticals UK Ltd)
4631	83710020	Voltarol 50mg gastro-resistant tablets (Novartis Pharmaceuticals UK Ltd)
1446	53289020	Voltarol 50mg Tablet (Novartis Pharmaceuticals UK Ltd)
4625	83774020	Voltarol 75mg SR tablets (Novartis Pharmaceuticals UK Ltd)
44112	98747020	Voltarol Joint Pain 12.5mg tablets (Novartis Consumer Health UK Ltd)
39722	95573020	Voltarol Pain-eze 12.5mg tablets (Novartis Consumer Health UK Ltd)
47820	164021	Voltarol Pain-eze Extra Strength 25mg tablets (Novartis Consumer Health UK Ltd)
5401	82738020	Voltarol Rapid 25mg tablets (Novartis Pharmaceuticals UK Ltd)
53345	8375020	Voltarol Rapid 50mg tablets (Lexon (UK) Ltd)
51099	8377020	Voltarol Rapid 50mg tablets (Mawdsley-Brooks & Company Ltd)
5085	82739020	Voltarol Rapid 50mg tablets (Novartis Pharmaceuticals UK Ltd)
70145	8376020	Voltarol Rapid 50mg tablets (Stephar (U.K.) Ltd)
27901	!7713301	VOLTAROL RETARD
2386	53298020	Voltarol Retard 100mg tablets (Novartis Pharmaceuticals UK Ltd)
1766	53294020	Voltarol sr 75mg Modified-release tablet (Novartis Pharmaceuticals UK Ltd)

S3: Code list for upper GI bleed

ICD-10	Description
K25.0	Gastric ulcer Acute with haemorrhage
K25.2	Gastric ulcer Acute with both haemorrhage and perforation
K25.4	Gastric ulcer Chronic or unspecified with haemorrhage
K25.6	Gastric ulcer Chronic or unspecified with both haemorrhage and perforation
K26.0	Duodenal ulcer Acute with haemorrhage
K26.2	Duodenal ulcer Acute with both haemorrhage and perforation
K26.4	Duodenal ulcer Chronic or unspecified with haemorrhage
K26.6	Duodenal ulcer Chronic or unspecified with both haemorrhage and perforation
K27.0	Peptic ulcer, site unspecified Acute with haemorrhage
K27.2	Peptic ulcer, site unspecified Acute with both haemorrhage and perforation
K27.4	Peptic ulcer, site unspecified Chronic or unspecified with haemorrhage
K27.6	Peptic ulcer, site unspecified Chronic or unspecified with both haemorrhage and perforation
K28.0	Gastrojejunal ulcer Acute with haemorrhage
K28.2	Gastrojejunal ulcer Acute with both haemorrhage and perforation
K28.4	Gastrojejunal ulcer Chronic or unspecified with haemorrhage
K28.6	Gastrojejunal ulcer Chronic or unspecified with both haemorrhage and perforation
K29.0	Acute haemorrhagic gastritis
K92.0	Haematemesis
K92.1	Melaena
K92.2	Gastrointestinal haemorrhage, unspecified

S4: Code list for Osteoarthritis

Read Code	Medcode	Description
N05..00	3057	Osteoarthritis and allied disorders
N05..11	396	Osteoarthritis
N050.00	4353	Generalised osteoarthritis - OA
N050000	38631	Generalised osteoarthritis of unspecified site
N050100	36327	Generalised osteoarthritis of the hand
N050111	4015	Heberdens' nodes
N050112	35919	Bouchards' nodes
N050200	23676	Generalised osteoarthritis of multiple sites
N050300	38018	Bouchard's nodes with arthropathy
N050400	23646	Primary generalized osteoarthrosis
N050500	11256	Secondary multiple arthrosis
N050600	38019	Erosive osteoarthrosis
N050700	24432	Heberden's nodes with arthropathy
N050z00	34867	Generalised osteoarthritis NOS
N051.00	32839	Localised, primary osteoarthritis
N051000	54224	Localised, primary osteoarthritis of unspecified site
N051100	24022	Localised, primary osteoarthritis of the shoulder region
N051200	24217	Localised, primary osteoarthritis of the upper arm
N051300	34806	Localised, primary osteoarthritis of the forearm
N051400	21350	Localised, primary osteoarthritis of the hand
N051500	15839	Localised, primary osteoarthritis of the pelvic region/thigh
N051600	21159	Localised, primary osteoarthritis of the lower leg
N051700	25793	Localised, primary osteoarthritis of the ankle and foot
N051800	20472	Localised, primary osteoarthritis of other specified site
N051900	24287	Primary coxarthrosis, bilateral
N051A00	25812	Coxarthrosis resulting from dysplasia, bilateral
N051B00	24146	Primary gonarthrosis, bilateral
N051C00	36182	Primary arthrosis of first carpometacarpal joints, bilateral
N051D00	24958	Localised, primary osteoarthritis of the wrist
N051E00	28908	Localised, primary osteoarthritis of toe
N051F00	18602	Localised, primary osteoarthritis of elbow
N051G00	106678	Osteoarthritis of spinal facet joint
N051z00	20660	Localised, primary osteoarthritis NOS
N052.00	42045	Localised, secondary osteoarthritis
N052000	68712	Localised, secondary osteoarthritis of unspecified site
N052100	33574	Localised, secondary osteoarthritis of the shoulder region

N052200	41088	Localised, secondary osteoarthritis of the upper arm
N052300	45815	Localised, secondary osteoarthritis of the forearm
N052400	23638	Localised, secondary osteoarthritis of the hand
N052500	44041	Localised, secondary osteoarthritis of pelvic region/thigh
N052511	101479	Coxae malum senilis
N052600	33479	Localised, secondary osteoarthritis of the lower leg
N052700	34035	Localised, secondary osteoarthritis of the ankle and foot
N052800	32891	Localised, secondary osteoarthritis of other specified site
N052900	64503	Post-traumatic coxarthrosis, bilateral
N052A00	24392	Post-traumatic gonarthrosis, bilateral
N052B00	60183	Post-traumatic arthrosis of first carpometacarpal jt bilat
N052C00	50470	Post-traumatic gonarthrosis, unilateral
N052z00	57912	Localised, secondary osteoarthritis NOS
N053.00	34122	Localised osteoarthritis, unspecified
N053000	49545	Localised osteoarthritis, unspecified, of unspecified site
N053100	15441	Localised osteoarthritis, unspecified, of shoulder region
N053200	59637	Localised osteoarthritis, unspecified, of the upper arm
N053300	60537	Localised osteoarthritis, unspecified, of the forearm
N053400	16242	Localised osteoarthritis, unspecified, of the hand
N053500	20626	Localised osteoarthritis, unspecified, pelvic region/thigh
N053511	52925	Otto's pelvis
N053512	1104	Hip osteoarthitis NOS
N053600	34804	Localised osteoarthritis, unspecified, of the lower leg
N053611	1296	Patellofemoral osteoarthritis
N053700	4461	Localised osteoarthritis, unspecified, of the ankle and foot
N053800	18112	Localised osteoarthritis, unspecified, of other spec site
N053900	21177	Arthrosis of first carpometacarpal joint, unspecified
N053z00	31200	Localised osteoarthritis, unspecified, NOS
N054.00	21528	Oligoarticular osteoarthritis, unspecified
N054000	48214	Oligoarticular osteoarthritis, unspec, of unspecified sites
N054100	52095	Oligoarticular osteoarthritis, unspecified, of shoulder
N054200	97073	Oligoarticular osteoarthritis, unspecified, of upper arm
N054300	112556	Oligoarticular osteoarthritis, unspecified, of forearm
N054400	59616	Oligoarticular osteoarthritis, unspecified, of hand

N054500	68648	Oligoarticular osteoarthritis, unspecified, of pelvis/thigh
N054600	41090	Oligoarticular osteoarthritis, unspecified, of lower leg
N054700	72109	Oligoarticular osteoarthritis, unspecified, of ankle/foot
N054800	41985	Oligoarticular osteoarthritis, unspecified, other spec sites
N054900	57267	Oligoarticular osteoarthritis, unspecified, multiple sites
N054z00	53858	Osteoarthritis of more than one site, unspecified, NOS
N05z.00	5776	Osteoarthritis NOS
N05z.11	1509	Joint degeneration
N05z000	35527	Osteoarthritis NOS, of unspecified site
N05z100	3147	Osteoarthritis NOS, of shoulder region
N05z200	50848	Osteoarthritis NOS, of the upper arm
N05z211	639	Elbow osteoarthritis NOS
N05z300	24152	Osteoarthritis NOS, of the forearm
N05z311	15206	Wrist osteoarthritis NOS
N05z400	658	Osteoarthritis NOS, of the hand
N05z411	4490	Finger osteoarthritis NOS
N05z412	1959	Thumb osteoarthritis NOS
N05z500	4967	Osteoarthritis NOS, pelvic region/thigh
N05z511	2209	Hip osteoarthritis NOS
N05z600	15144	Osteoarthritis NOS, of the lower leg
N05z611	665	Knee osteoarthritis NOS
N05z700	15447	Osteoarthritis NOS, of ankle and foot
N05z711	52897	Ankle osteoarthritis NOS
N05z712	1312	Foot osteoarthritis NOS
N05z713	4878	Toe osteoarthritis NOS
N05z800	15052	Osteoarthritis NOS, other specified site
N05z900	5802	Osteoarthritis NOS, of shoulder
N05zA00	3814	Osteoarthritis NOS, of sternoclavicular joint
N05zB00	2229	Osteoarthritis NOS, of acromioclavicular joint
N05zC00	19713	Osteoarthritis NOS, of elbow
N05zD00	65748	Osteoarthritis NOS, of distal radio-ulnar joint
N05zE00	9649	Osteoarthritis NOS, of wrist
N05zF00	7866	Osteoarthritis NOS, of MCP joint
N05zG00	11032	Osteoarthritis NOS, of PIP joint of finger
N05zH00	9681	Osteoarthritis NOS, of DIP joint of finger
N05zJ00	6812	Osteoarthritis NOS, of hip
N05zK00	34023	Osteoarthritis NOS, of sacro-iliac joint
N05zL00	2487	Osteoarthritis NOS, of knee
N05zM00	70425	Osteoarthritis NOS, of tibio-fibular joint
N05zN00	8202	Osteoarthritis NOS, of ankle

N05zP00	40972	Osteoarthritis NOS, of subtalar joint
N05zQ00	55388	Osteoarthritis NOS, of talonavicular joint
N05zR00	54350	Osteoarthritis NOS, of other tarsal joint
N05zS00	6887	Osteoarthritis NOS, of 1st MTP joint
N05zT00	9010	Osteoarthritis NOS, of lesser MTP joint
N05zU00	27834	Osteoarthritis NOS, of IP joint of toe
N05zz00	27972	Osteoarthritis NOS
N11..00	2001	Spondylosis and allied disorders
N11..11	2294	Arthritis of spine
N11..12	7429	Osteoarthritis of spine
N110.00	2881	Cervical spondylosis without myelopathy
N110.11	771	Cervical spondylosis
N110.12	17092	Osteoarthritis cervical spine
N110000	38501	Single-level cervical spondylosis without myelopathy
N110100	51531	Two-level cervical spondylosis without myelopathy
N110200	15744	Multiple-level cervical spondylosis without myelopathy
N111.00	8208	Cervical spondylosis with myelopathy
N111000	27583	Single-level cervical spondylosis with myelopathy
N111100	63192	Two-level cervical spondylosis with myelopathy
N111200	58865	Multiple-level cervical spondylosis with myelopathy
N112.00	18217	Thoracic spondylosis without myelopathy
N112.11	5183	Thoracic spondylosis
N112000	69912	Single-level thoracic spondylosis without myelopathy
N112100	62914	Two-level thoracic spondylosis without myelopathy
N112200	50448	Multiple-level thoracic spondylosis without myelopathy
N112300	18205	Dorsal spondylosis without myelopathy
N113.00	55628	Thoracic spondylosis with myelopathy
N113000	64854	Single-level thoracic spondylosis with myelopathy
N113200	96103	Multiple-level thoracic spondylosis with myelopathy
N114.00	15015	Lumbosacral spondylosis without myelopathy
N114.11	1565	Degeneration of lumbar spine
N114.12	1100	Lumbar spondylosis
N114000	20791	Single-level lumbosacral spondylosis without myelopathy
N114100	52991	Two-level lumbosacral spondylosis without myelopathy
N114200	37097	Multiple-level lumbosacral spondylosis without myelopathy
N115.00	11688	Lumbosacral spondylosis with myelopathy
N115000	41516	Single-level lumbosacral spondylosis with myelopathy
N115100	45730	Two-level lumbosacral spondylosis with myelopathy

N115200	63578	Multiple-level lumbosacral spondylosis with myelopathy
N119.00	10121	Cervical spondylosis with radiculopathy
N119000	55810	Single-level cervical spondylosis with radiculopathy
N119100	51318	Two-level cervical spondylosis with radiculopathy
N119200	56212	Multiple-level cervical spondylosis with radiculopathy
N11A.00	35851	Cervical spondylosis with vascular compression
N11B.00	19386	Thoracic spondylosis with radiculopathy
N11B000	54852	Single-level thoracic spondylosis with radiculopathy
N11B100	103137	Two-level thoracic spondylosis with radiculopathy
N11B200	93977	Multiple-level thoracic spondylosis with radiculopathy
N11C.00	9834	Lumbosacral spondylosis with radiculopathy
N11C000	54843	Single-level lumbosacral spondylosis with radiculopathy
N11C100	65641	Two-level lumbosacral spondylosis with radiculopathy
N11C200	48810	Multiple-level lumbosacral spondylosis with radiculopathy
N11D.00	18826	Osteoarthritis of spine
N11D000	41378	Osteoarthritis of cervical spine
N11D100	47024	Osteoarthritis of thoracic spine
N11D200	22452	Osteoarthritis of lumbar spine
N11D300	53184	Osteoarthritis of spine NOS
N11E.00	96948	Cervical spondylosis
N11F.00	109023	Axial spondyloarthritis
N11z.00	3447	Spondylosis NOS
N11z.11	829	Osteoarthritis spine
N11z000	56594	Spondylosis without myelopathy, NOS
N11z100	35838	Spondylosis with myelopathy, NOS
N11zz00	17766	Spondylosis NOS

S5: Code list for Cirrhotic Liver Disease

Read Code	Medcode	Description
C310400	19512	Glycogenosis with hepatic cirrhosis
C350012	8206	Pigmentary cirrhosis of liver
C370800	102922	Cystic fibrosis related cirrhosis
G852200	26319	Oesophageal varices in cirrhosis of the liver
G852300	8363	Oesophageal varices in alcoholic cirrhosis of the liver
J612.00	4743	Alcoholic cirrhosis of liver
J612.11	68376	Florid cirrhosis
J612.12	100474	Laennec's cirrhosis
J612000	21713	Alcoholic fibrosis and sclerosis of liver
J615.00	16725	Cirrhosis - non alcoholic
J615.11	47257	Portal cirrhosis
J615100	69204	Multilobular portal cirrhosis
J615300	3450	Diffuse nodular cirrhosis
J615400	44676	Fatty portal cirrhosis
J615500	92909	Hypertrophic portal cirrhosis
J615600	40567	Capsular portal cirrhosis
J615700	27438	Cardiac portal cirrhosis
J615711	108819	Congestive cirrhosis
J615800	96664	Juvenile portal cirrhosis
J615811	112867	Childhood function cirrhosis
J615812	58184	Indian childhood cirrhosis
J615C00	100253	Xanthomatous portal cirrhosis
J615D00	73482	Bacterial portal cirrhosis
J615F00	112044	Syphilitic portal cirrhosis
J615G00	109540	Zooparasitic portal cirrhosis
J615H00	48928	Infectious cirrhosis NOS
J615y00	55454	Portal cirrhosis unspecified
J615z00	16455	Non-alcoholic cirrhosis NOS
J615z11	22841	Macronodular cirrhosis of liver
J615z12	18739	Cryptogenic cirrhosis of liver
J615z13	1638	Cirrhosis of liver NOS
J616.00	9494	Biliary cirrhosis
J616000	5638	Primary biliary cirrhosis
J616100	15424	Secondary biliary cirrhosis
J616200	91591	Biliary cirrhosis of children
J616z00	58630	Biliary cirrhosis NOS
J61y500	60104	Hepatic sclerosis
J61y600	100592	Hepatic fibrosis with hepatic sclerosis
J635600	44120	Toxic liver disease with fibrosis and cirrhosis of liver
Jyu7100	6015	[X]Other and unspecified cirrhosis of liver

C310411	85188	Glycogenosis, type 4
C310412	31944	Andersen's disease

Chapter 7

Paper F: Proton pump inhibitors and risk of all-cause and cause-specific mortality: a cohort study

Jeremy P Brown[1], John Tazare[1], Elizabeth Williamson[1], Kathryn E. Mansfield[1], Stephen JW Evans[1], Laurie A Tomlinson[1], Krishnan Bhaskaran[1], Liam Smeeth[1], Kevin Wing[1], Ian J Douglas[1]

1. London School of Hygiene and Tropical Medicine, London, UK.

7.1 Overview

Summary

This chapter introduces a study comparing the risk of all-cause and cause-specific mortality in users of proton pump inhibitors (PPI) and H2 receptor antagonists (H2RAS). A number of non-interventional studies have obtained results suggesting that PPIs are associated with a range of adverse health outcomes, including increased risk of mortality. However, since patients taking PPIs typically have poorer health compared to those who do not, residual confounding is a key concern across these studies. In this study, we applied the HDPS (without the modifications proposed in Chapter 3) with the aim of obtaining better capture and control for potentially important, but hard to measure, confounding factors relating to disease severity, healthcare utilisation and frailty. Results from this study highlighted the poorer health of PPI users compared to individuals prescribed alternative acid suppression therapy. Furthermore, whilst results indicated an association between PPI prescription and both all-cause and cause-specific mortality, the pattern of results indicate that residual confounding is still likely despite HDPS adjustment. In the next chapter, this study is used as a case-study for exploratory work investigating whether incorporation of laboratory test data within the HDPS framework can potentially lead to improved confounder capture and control in UK electronic health records Initially, the work was presented as an oral presentation at the *35th International Conference on Pharmacoepidemiology & Therapeutic Risk Management (2019)*. This paper was published in January 2021 in the *British Journal of Clinical Pharmacology*.

Thesis objective addressed

This chapter addresses the following objective of the overall thesis (Section 1.3):

3. Apply the HDPS and proposed modifications in the context of UK EHRs.

Role of candidate

Ian Douglas (ID), Krishnan Bhaskaran and Laurie Tomlinson conceived the study. All authors were involved in the study design. Jeremy Brown (JB) extracted and performed the data management to create analysis-ready datasets. I co-lead the propensity score modelling and analysis with JB, with input from ID and Elizabeth Williamson. I implemented and conducted the HDPS analysis. JB lead the writing of the initial draft. I contributed to the methods, results and discussion sections of the initial draft. All authors interpreted the results, contributed to revisions and approved the final manuscript.

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

London School of Hygiene & Tropical Medicine
Keppel Street, London WC1E 7HT

T: +44 (0)20 7299 4646
F: +44 (0)20 7299 4656
www.lshtm.ac.uk

RESEARCH PAPER COVER SHEET

Please note that a cover sheet must be completed for each research paper included within a thesis.

SECTION A – Student Details

Student ID Number	LSH1401926	**Title**	Mr
First Name(s)	John		
Surname/Family Name	Tazare		
Thesis Title	High-dimensional propensity scores for data-driven confounder adjustment in UK electronic health records		
Primary Supervisor	Elizabeth Williamson & Ian Douglas		

If the Research Paper has previously been published please complete Section B, if not please move to Section C.

SECTION B – Paper already published

Where was the work published?	British Journal of Clinical Pharmacology		
When was the work published?	January 2021		
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion	N/A		
Have you retained the copyright for the work?*	**Yes**	Was the work subject to academic peer review?	**Yes**

*If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	-
Please list the paper's authors in the intended authorship order:	-
Stage of publication	Choose an item.

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	I co-lead the propensity score modelling and analysis. Furthermore, I implemented and conducted the HDPS analysis. I contributed to the initial draft and, along with all authors, contributed to revisions and the final manuscript.

SECTION E

Student Signature	John Tazare
Date	14/05/2021

Supervisor Signature	Ian Douglas
Date	14/05/2021

7.2 Abstract

Aim

To investigate the association between proton pump inhibitors (PPIs) and both all-cause and cause-specific mortality.

Methods

We conducted a cohort study using the UK Clinical Practice Research Datalink GOLD database. We compared 733,885 new users of PPIs to 124,410 new users of H2 receptor antagonists (H2RAs). In a secondary analysis we compared 689,602 PPI new users to 1,361,245 non-users of acid suppression therapy matched on age, sex, and calendar year. Hazard ratios for all-cause and cause-specific mortality were estimated using propensity score (PS) weighted Cox models.

Results

PPI prescription was associated with increased risk of all-cause mortality, with hazard ratios decreasing considerably by increasing adjustment (unadjusted hazard ratio [HR] 1.65, 95% confidence interval (CI) 1.62-1.69; PS-weighted HR 1.38, 95% CI 1.33-1.44; high-dimensional PS-weighted HR 1.31, 95% CI 1.26-1.37). Short-term associations were observed with mortality from causes where a causal short-term association is unexpected (e.g. lung cancer mortality: PS-weighted HR at 6 months 1.77; 95% CI 1.39-2.25). Adjusted hazard ratios were substantially higher when comparing to non-users (PS-weighted HR all-cause mortality 1.96, 95% CI 1.94-1.99) rather than H2RA users.

Conclusions

PPI prescription was strongly associated with all-cause and cause-specific mortality. However, the change in hazard ratios by (1) increasing adjustment and (2) between comparator groups indicate that residual confounding is likely to explain the association between poor health outcomes and PPI use, and fully accounting for this using observational data may not be possible.

7.3 Introduction

Proton pump inhibitors (PPIs) are a group of commonly prescribed drugs used to suppress gastric acid production. They are prescribed for a variety of indications including the treatment of dyspepsia, peptic ulcers, and gastro-oesophageal reflux disease, the eradication of H. pylori, and prophylaxis to prevent drug-induced gastrointestinal damage (e.g. from non-steroidal anti-inflammatory drugs (NSAID)).

Concern over the safety of PPIs has grown, given associations observed in non-interventional studies between PPI use and a range of outcomes including pneumonia, chronic kidney disease, cancer, and alcoholic liver disease (*Batchelor et al.*, 2018; *Benson et al.*, 2015; *Cheema*, 2019; *Dial et al.*, 2005; *Laheij et al.*, 2004; *Li et al.*, 2019; *Llorente et al.*, 2017; *Tvingsholm et al.*, 2018; *Yang et al.*, 2017). Furthermore, recent non-interventional studies identified associations between PPI prescription and increased all-cause and cause-specific mortality (*Dultz et al.*, 2015; *Tvingsholm et al.*, 2018; *Xie et al.*, 2017, 2019).

Previous safety concerns about PPIs have highlighted important limitations of statistical techniques used to account for differences between PPI users and non-users in non-interventional studies; several studies identified a harmful association between combined clopidogrel and PPI use, whilst randomised controlled trials (RCTs) found no clinically relevant interaction (*Demcsák et al.*, 2018). Given that PPIs are globally one of the most frequently used classes of drugs, it is vital that we are able to reliably

evaluate their potential risks and benefits when making treatment decisions.

In this study we aimed to examine the association between PPIs and all-cause and cause-specific mortality, and to investigate the robustness of results to confounding by (1) applying different methods to adjust for confounding, (2) using different comparator groups, (3) examining the pattern of the associations across different time periods, 4) and including control outcomes not previously associated with PPI use.

7.4 Methods

We conducted a cohort study comparing mortality among new users of PPIs to, in the first instance, new users of an alternative acid suppression drug, H2 receptor antagonists (H2RAs), and as a secondary analysis to non-users of either H2RAs or PPIs.

7.4.1 Data source

The Clinical Practice Research Datalink (CPRD) GOLD database consists of primary care electronic medical records of people registered at one of over 700 general practices in the United Kingdom (UK). The dataset is widely validated for epidemiological research and broadly representative of the UK population in terms of age, sex and ethnicity (*Herrett et al.*, 2015). Our study included the subset of CPRD GOLD practices that have consented to linkage with other datasets.

We incorporated linked data from the Office of National Statistics (ONS) death registration data, Hospital Episode Statistics Admitted Patient Care (HES APC) data, and Index of Multiple Deprivation (IMD) data. Date and cause of death were ascertained from ONS death registration data. In the UK all deaths are registered and cause of death is certified by a clinician. The number of hospital admissions in the 6 months prior to study entry, a covariate, was calculated from HES APC (*Herbert et al.*, 2017). Socioeconomic deprivation, another covariate, was ascertained from postcode-based IMD data. The IMD is an index of relative socioeconomic deprivation based upon seven do-

mains, which include income, employment, education and health (*Sooriakumaran et al.*, 2014).

7.4.2 Study population

We included all adults in CPRD GOLD who were eligible for person-level linkage to HES APC and ONS, had acceptable research standard data, and who were prescribed a PPI or H2RA for the first time on or after the latest of: their 18th birthday, date of registration at current practice plus 1 year, first appointment with clinician after registration at current practice, date practice began contributing research quality data plus 1 year, or 02/01/1998 (start of ONS data coverage).

In a secondary analysis, to identify the extent to which confounding by indication may be an issue, we compared PPI users to matched non-users. We would expect similar results from both comparisons (PPI/H2RA and PPI/non-user) if our statistical models control for all confounding, and assuming no causal effect of H2RAs on mortality.

In calendar date order PPI users were matched to non-users of either acid-suppression medication (PPI or H2RA), who met the same date-based eligibility criteria as PPI users, on year of birth (+/- 2 years), sex, calendar year, and clinical practice. Up to two non-users meeting the matching criteria, and with the closest year of birth, were randomly matched (without replacement) to each PPI user. PPI and H2RA users were eligible as potential non-users prior to first PPI/H2RA prescription.

Cohort entry was defined as date of prescription for H2RA and PPI users, and for non-users as cohort entry date of matched PPI user. We followed individuals up until the earliest of death date, date the individual was no longer registered with the practice, date of last practice data collection, 17/04/2017 (end of coverage period of included ONS mortality data), date of first PPI prescription (H2RA users and non-users), or date of first H2RA prescription (non-users only).

7.4.3 Exposure

Prescription of a PPI (omeprazole, lansoprazole, pantoprazole, rabeprazole or esomeprazole) was the main exposure of interest. The choice of comparator group is an important consideration in observational studies of drug effects, with an active comparator generally considered the best approach to mitigate confounding. H2RA prescription (cimetidine, ranitidine, famotidine, nizatidine) was therefore chosen as the main comparator given that H2RAs are a gastric-acid suppressing medication used for similar indications to PPIs. PPIs are predated by H2RAs, but are now the most commonly prescribed acid-suppression therapy in the UK with superior efficacy observed for many indications in RCTs (*Alhazzani et al.*, 2018; *Alshamsi et al.*, 2016; *Gisbert et al.*, 2003; *Van Pinxteren et al.*, 2003). Key protein targets and ligands in this article are hyperlinked to corresponding entries in `http://www.guidetopharmacology.org`, and are permanently archived in the Concise Guide to Pharmacology 2019/20 (*Alexander et al.*, 2019a,b).

7.4.4 Covariates

We adjusted for demographic and lifestyle variables, potential indications for PPI treatment, indicators of frailty, previous comorbidities and calendar year in our statistical models (Table 7.1 - further detail provided in Supporting Information - Supplementary Methods).

Table 7.1: *Covariates adjusted for in statistical models*

Type	Covariates
Demographic and lifestyle variables at baseline	Age, sex, index of multiple deprivation score (IMD), body mass index (BMI), smoking status, alcohol consumption
Potential indications for PPI treatment in 6 months prior to baseline:	Prescription for NSAID, aspirin, clopidogrel, oral anticoagulant or corticosteroid, upper gastrointestinal endoscopy, gastric cancer, gastro-oesophageal reflux disease, peptic ulcers, upper gastrointestinal (GI) bleeding, pancreatitis, cirrhosis, oesophagitis, Barrett's oesophagus, and H. pylori infection
Indicators of frailty in 6 months prior to baseline	Number of hospital admissions, number of general practitioner (GP) appointments, number of different drug types prescribed (based on distinct British National Formulary (BNF) chapters)
Ever recorded previous comorbidities	hypertension, cardiovascular disease, peripheral artery disease, cerebrovascular disease, chronic obstructive pulmonary disease (COPD), cancer, non-viral liver disease, human immunodeficiency virus (HIV), chronic kidney disease (CKD), dementia, and diabetes mellitus
Other	Calendar year at cohort entry

7.4.5 Outcomes

All-cause mortality was the primary outcome. Cause of death was ascertained from the International Classification of Diseases (ICD) 9 or 10 code recorded for the underlying cause of death on the death certificate. Secondary outcomes included cause-specific mortality: 1) categorised into groupings used in the Global Burden of Diseases Study (*Abubakar et al.*, 2015; *Bhaskaran et al.*, 2018); 2) a priori causes that have previously been associated with PPIs; and 3) control outcomes we would not expect to be associated with PPIs.

Global Burden of Diseases Study groupings included the high-level categories of cause-specific mortality: communicable disease, non-communicable disease, and injury/external cause. Global Burden of Diseases groupings also included the lower-level categories:

neoplasms; cardiovascular/circulatory; chronic respiratory diseases; liver cirrhosis; digestive other than cirrhosis; neurological; mental and behavioural; diabetes, urogenital, blood and endocrine; and musculoskeletal.

We included pre-specified individual causes of death where the cause was:

- Previously associated with PPIs and a short term causal association was considered plausible: pneumonia, acute kidney injury, C. difficile enterocolitis, atrial fibrillation/flutter, heart failure, and aortic aneurysm
- Previously associated with PPIs but where a short term causal association was considered to be unexpected based on disease pathogenesis: dementia and Alzheimer's, chronic kidney disease, hypertensive heart disease, ischaemic heart disease, lung cancer, mesothelioma, breast cancer, liver cancer, prostate cancer, gastric cancer, alcoholic liver disease, and chronic obstructive pulmonary disease (COPD)

We also included, as control outcomes, individual causes of mortality that had not been previously associated with PPIs: accidental trauma (excluding falls), and pulmonary embolism. We did not expect an association between PPI use and accidental trauma, which is unlikely to be confounded by underlying health status, whereas the association with pulmonary embolism may be affected by unmeasured differences between PPI exposed and unexposed individuals. ICD codes for all outcomes are included in Supporting Information Table S1.

7.4.6 Statistical analysis

Propensity scores were used to adjust for differences in baseline covariates (*Williamson et al.*, 2012). We generated propensity scores for PPI prescription using logistic regression, or conditional logistic regression (in the case of the matched non-user cohort (*Smeeth et al.*, 2009)). In the PPI/non-user matched analysis the matching factors age, sex, and calendar year were excluded from the conditional logistic regression model. In the PPI/H2RA analysis propensity scores were estimated separately within each

category of calendar year (1998-2003, 2004-2009, 2010-2015) given strong trends in prescribing of the two drugs over time.

A missing indicator approach was used for missing covariate information (for BMI, smoking status, and alcohol consumption). The missing indicator method has been found to be unbiased for propensity score analysis under assumptions that may be more plausible in the context of electronic health records than the complete records approach (*Blake et al.*, 2020).

Estimated propensity scores were incorporated using average effect of treatment in the treated (ATT) weights. These weights estimate the average effect of treatment among individuals similar to the treated (PPI users) rather than in the overall study population (*Austin and Stuart*, 2015). ATT weights were chosen to increase comparability between the PPI versus H2RA, and PPI versus non-user analyses. By using ATT weights our effect estimates in both the PPI/H2RA and PPI/non-user comparisons pertain to the same population, PPI users.

ATT weighted Cox regression models, or ATT weighted stratified Cox regression models (in the case of the matched non-user cohort), were used to estimate the relative risk of each mortality outcome with PPI exposure over 0 to 6 months (censoring follow-up at 6 months), 0 to 1 year, 0 to 10 years, and over all follow-up (*Hernan*, 2010). An early increase in risk for associations where a short-term association with outcome incidence is unexpected causally (based on disease pathogenesis) may indicate residual confounding.

As a secondary analysis, high-dimensional propensity scores (HDPS) were used to investigate residual confounding of the primary analysis. The HDPS approach selects a large number of covariates (500 in our study), prioritising for inclusion those with the greatest potential to confound the association of interest (*Schneeweiss et al.*, 2009). It has been suggested that the HDPS may control for additional confounding by adjusting for proxies of unmeasured covariates (see Supporting Information - Supplementary Methods for further detail).

Sensitivity analyses included: 1) direct adjustment for covariates in the Cox model

rather than propensity score weighting, 2) defining cause of death based on any listed cause rather than restricting to primary cause of death , 3) censoring follow-up at first PPI/H2RA treatment break (further detail in Supporting Information - Supplementary Methods), 4) censoring follow-up at first prescription of an H2RA among PPI users, 5) censoring follow-up on 31st December 2014 in order to only include follow-up when PPIs were solely available through pharmacy or prescription in the UK, rather than more generally in shops, 6) a post-hoc analysis excluding gastric cancer deaths from the definition of neoplasms deaths, and 7) propensity score trimming excluding individuals with propensity scores outside the range [0.1, 0.9] to assess sensitivity of findings to extreme weights (*Crump et al.*, 2009; *Ding and VanderWeele*, 2016; *Fedeli et al.*, 2015). Additionally, to quantify sensitivity to unmeasured confounding we calculated, using e-value formulae, the strength of association that an unmeasured confounder would need to have with exposure or outcome to fully explain the observed association (*Ding and VanderWeele*, 2016).

All analyses were conducted using Stata MP Version 15.

7.5 Results

The primary cohort consisted of 733,885 new users of PPIs and 124,410 new users of H2RAs (Figure 7.1). PPI users were on average older, more often male, and had a higher baseline prevalence of comorbidities and co-medication use (Table 7.2). Covariate balance improved after propensity score weighting with absolute standardised differences below 0.1 for all measured covariates.

7.5.1 Risk of mortality relative to H2RA users

There were 95,489 (26.5 per 1,000 person-years [PY]) deaths observed among PPI users and 8,800 (16.1 per 1,000 PY) among H2RA users. Median follow-up was 4.1 years (interquartile range [IQR] 1.8-7.2) among PPI users and 3.0 years (IQR 0.8-7.0) years

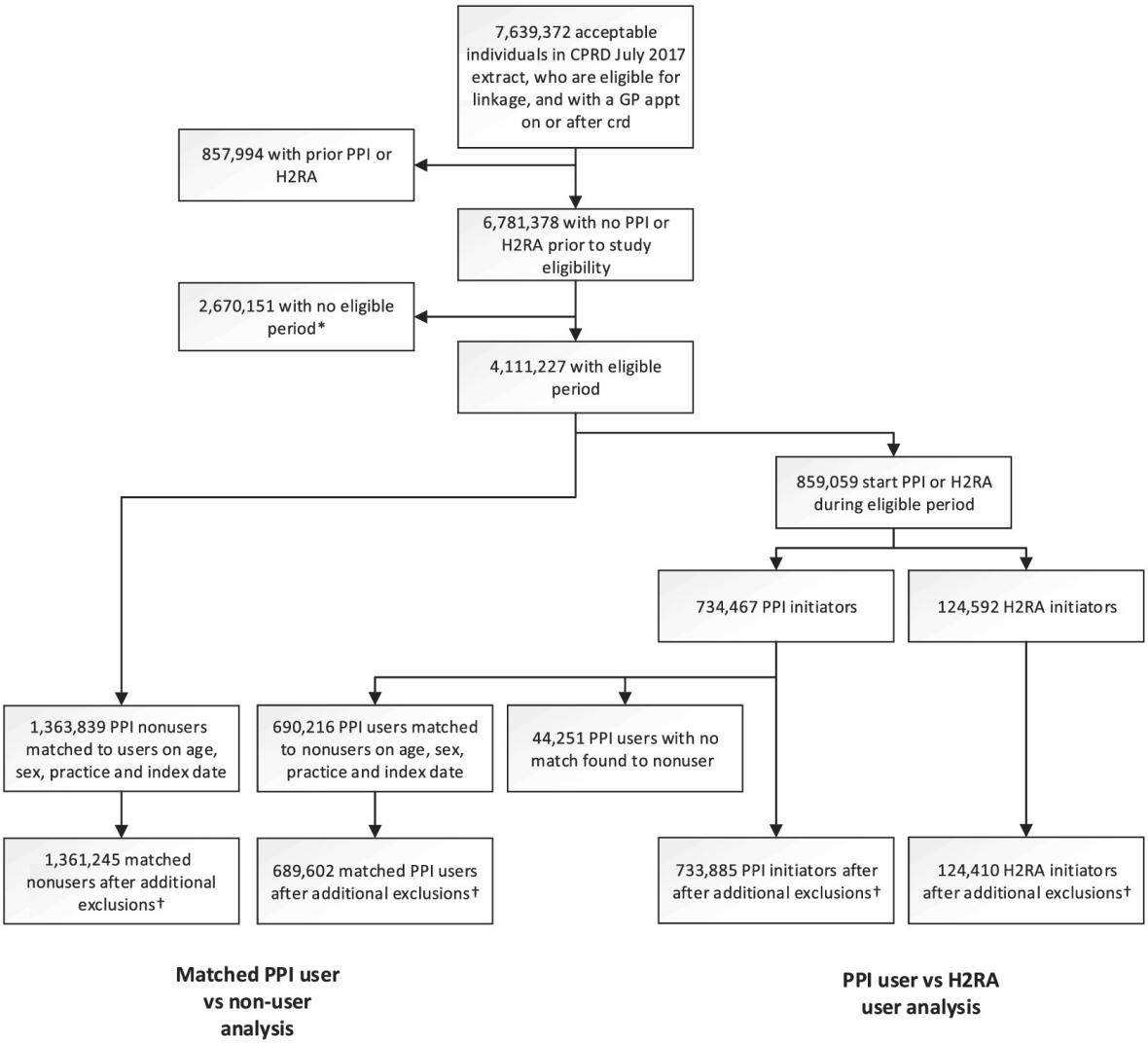


Figure 7.1: *Study flow chart.* ***Abbreviations:*** *CPRD, Clinical Practice Research Datalink; GP, general practitioner; crd, current registration date i.e. date of registration at current practice; PPI, proton pump inhibitor; H2RA, H2 receptor antagonist. * Beginning of eligibility is latest of on one year after current registration date, one year after up-to-standard date, study start (02/01/1998), 18th birthday, and first GP appointment after current registration date. End of eligibility is earliest of transfer out date, last collection date, study end date (31/12/2015) and death date. † Additional exclusions occurred to remove individuals with missing Index of Multiple Deprivation data, and when ONS death date obtained.*

Table 7.2: *Absolute standardised differences between PPI and H2RA users before and after weighting.* ***Abbreviations:*** *ASD, absolute standardised difference; H2RA, H2 receptor antagonist; PPI, proton pump inhibitor; BMI, body mass index; IMD, Index of Multiple Deprivation; GP, General Practitioner; BNF, British National Formulary; GI, gastrointestinal; GORD, gastro-oesophageal reflux disease; COPD, chronic obstructive pulmonary disease; CKD, chronic kidney disease.*

Characteristic*	Unweighted		Weighted		Unweighted ASD	Weighted ASD
	H2RA	PPI	H2RA	PPI		
Effective sample size	124,410	733,885	732,547.6	733,885		
Mean age in years	51.2	54.9	55	54.9	0.204	0.006
Mean BMI	26.5	27.2	27.2	27.2	0.118	<0.001
Calendar year						
1998-2003	56.7%	14.2%	14.2%	14.2%	1.135	0.001
2004-2009	31.3%	41.6%	41.6%	41.6%	0.210	<0.001
2010-2015	12.0%	44.2%	44.2%	44.2%	0.678	0.001
Female	57.3%	54.7%	54.1%	54.7%	0.052	0.014
Current smoker	24.3%	19.6%	19.7%	19.6%	0.119	0.002
Ex-smoker	24.3%	33.3%	33.2%	33.4%	0.195	0.005
High alcohol intake	2.7%	3.4%	3.4%	3.4%	0.039	<0.001
Below national median IMD	49.4%	51.8%	51.7%	51.8%	0.048	0.002
In 6 months prior to PPI/H2RA treatment initiation						
Mean no. of hospital admissions	0.3	0.4	0.4	0.4	0.052	0.001
Mean no. of GP appointments	4.8	5.9	6	5.9	0.165	0.003
Mean no. of BNF drug chapters	2.3	2.5	2.5	2.5	0.078	0.016
NSAID	21.3%	32.1%	31.5%	32.1%	0.236	0.012
Aspirin	11.7%	15.0%	14.9%	15.0%	0.093	0.004
Clopidogrel	1.6%	2.1%	2.2%	2.1%	0.036	0.009
Oral anticoagulant	2.0%	2.4%	2.5%	2.4%	0.022	0.01
Inhaled steroid	11.4%	12.8%	13.1%	12.8%	0.045	0.009
Systemic steroid	6.7%	7.2%	7.2%	7.2%	0.018	0.002
GORD	7.0%	8.4%	8.7%	8.4%	0.052	0.009
Oesophagitis	2.7%	3.5%	3.5%	3.5%	0.048	<0.001
Ever previous						
Hypertension	19.7%	26.1%	25.9%	26.1%	0.149	0.005
Coronary heart disease	8.0%	8.2%	8.3%	8.2%	0.01	<0.001
Peripheral artery disease	2.0%	2.2%	2.2%	2.2%	0.013	0.001
Cerebrovascular disease	3.7%	4.6%	4.8%	4.6%	0.043	0.011
COPD	2.6%	3.6%	3.7%	3.6%	0.055	0.003
Cancer	7.4%	10.1%	10.5%	10.1%	0.093	0.012
CKD	8.8%	13.9%	13.9%	13.9%	0.15	0.002
Diabetes	5.2%	7.7%	7.7%	7.7%	0.097	<0.001

* Only covariates with a frequency greater than 2% among PPI users or H2RA users are in this table. Standardised differences for all measured covariates including those with frequency less than 2% are provided in Supporting Information Table S2.

among H2RA users.

The risk of all-cause mortality was greater among PPI users relative to H2RA users (ATT weighted hazard ratio [wHR] 1.38; 95% confidence interval [CI] 1.33-1.44; Figure 7.2). At the broadest level, cause-specific mortality was elevated from communicable (wHR 1.40; 95% CI 1.22-1.60), and non-communicable (wHR 1.39; 95% CI 1.34-1.45) diseases but not from injuries/external causes (wHR 1.00; 95% CI 0.78-1.26).

By more specific cause-of-death category, mortality was higher in PPI users compared to H2RA users from neoplasms (wHR 1.74; 95% CI 1.63-1.86), cardiovascular/circulatory causes (wHR 1.17; 95% CI 1.10-1.25), chronic respiratory diseases (wHR 1.40; 95% CI 1.22-1.62), liver cirrhosis (wHR 1.95; 95% CI 1.10-3.46), digestive causes other than cirrhosis (wHR 1.43; 95% 1.20-1.69), and diabetes, urogenital, blood and endocrine causes (wHR 1.27; 95% CI 1.06-1.51). Excluding gastric cancer deaths from neoplasms made little difference to the effect estimate for neoplasms mortality (wHR 1.72; 95% CI 1.61-1.83). There was no evidence of an increased risk of mortality from neurological, mental and behavioural, or musculoskeletal causes.

There was strong evidence of an association with mortality from a number of individual causes previously associated with PPI use including pneumonia, cardiovascular events, cancer, alcoholic liver disease, and chronic obstructive pulmonary disease. There was no evidence for an association with the control outcome of mortality due to accidental trauma excluding falls (wHR 1.05; 95% CI 0.69-1.59), and the hazard ratio for the second control outcome, mortality from pulmonary embolism, was raised but had wide confidence intervals (wHR 1.33; 95% CI 0.85-2.09).

Adjustment via weighting reduced all hazard ratios (Figures 7.2 and 7.3). For most outcomes, further adjustment using the HDPS, reduced hazard ratios further towards the null (compared to a propensity score based on investigator chosen covariates).

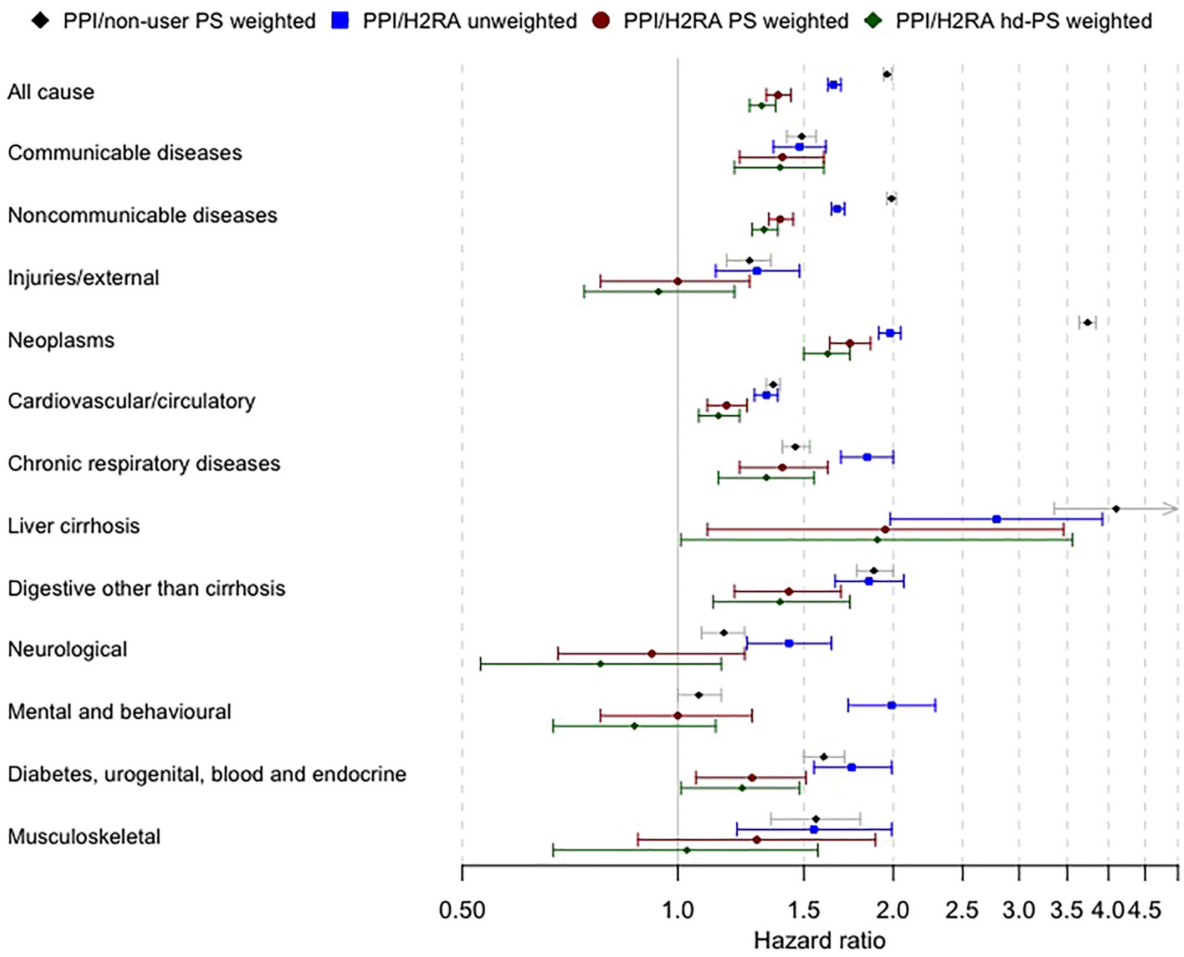


Figure 7.2: *Forest plot of the associations between PPI prescription and both all-cause and broad-level cause-specific mortality. Hazard ratios and 95% CI are plotted here and listed in Supporting Information Tables S3 and S4.* ***Abbreviations:*** *PPI, proton pump inhibitor; PS, propensity score; H2RA, H2 receptor antagonist; HDPS, high-dimensional propensity score.*

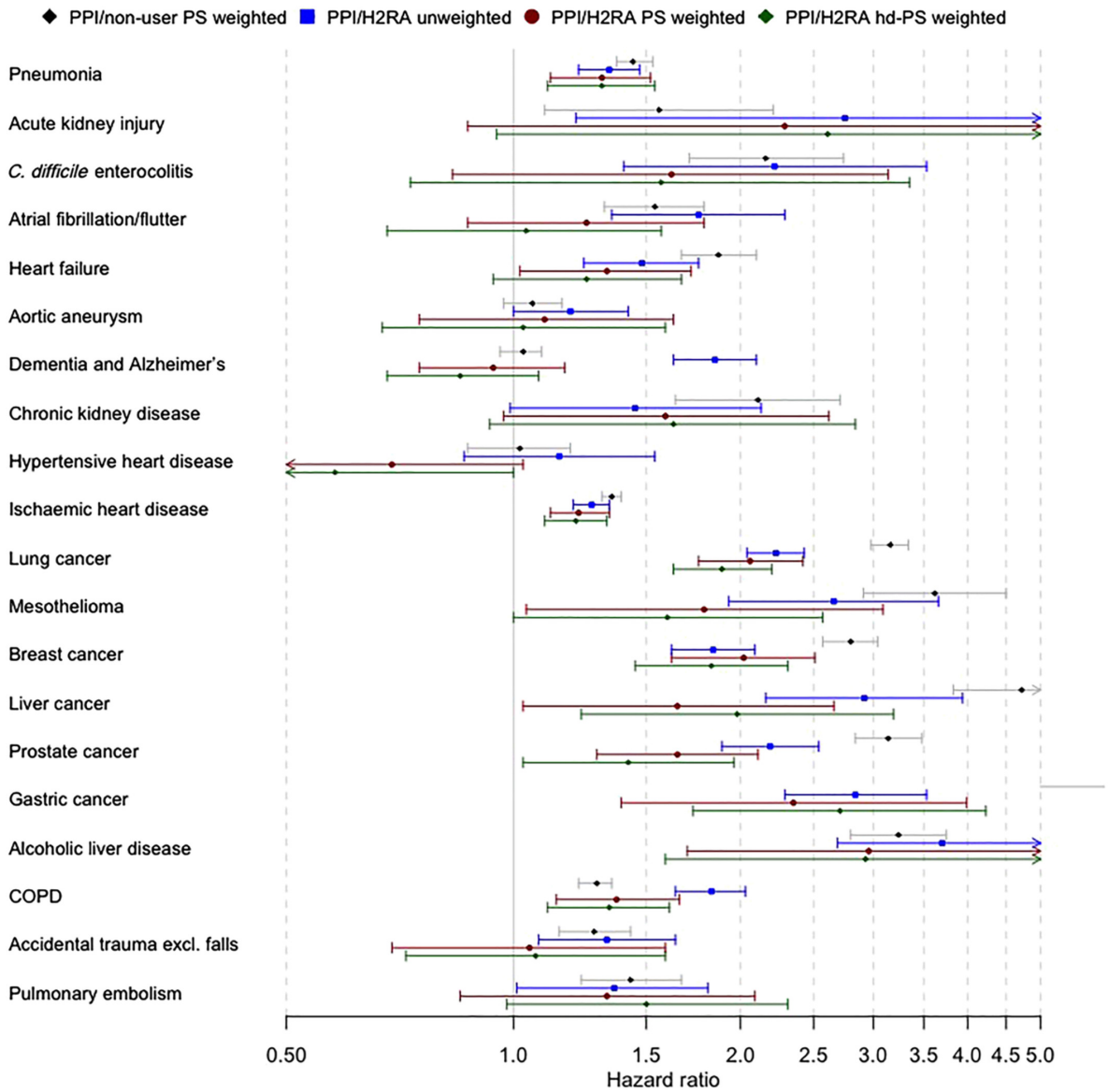


Figure 7.3: *Forest plot of the associations between PPI prescription and mortality from individual specified causes. Hazard ratios and 95% CI are plotted here and listed in Supporting Information Tables S3 and S4. For gastric cancer both the unweighted hazard ratio and 95% CI are outside of the x-axis.* ***Abbreviations:*** *PPI, proton pump inhibitor; PS, propensity score; H2RA, H2 receptor antagonist; HDPS, high-dimensional propensity score.*

7.5.2 Risk over different time periods

Examining hazard ratios comparing PPI and H2RA users at different time points revealed that, for many of the outcomes, including outcomes (lung, liver and breast cancer) where a short-term causal association was unexpected, an association was apparent within 6 months of treatment initiation (Figures 7.4 and 7.5, Supporting Information Figure S1). For all-cause mortality the weighted hazard ratio was 1.34 (95% CI 1.25-1.43) over the first 6 months.

7.5.3 Non-user comparison

For the secondary non-user comparison, 689,602 PPI users were matched (on age, sex, calendar year and clinical practice) to 1,361,245 non-users of acid suppression therapy (Figure 7.1). No suitable match could be found for 44,283 (6%) of PPI users (characteristics of matched/non-matched patients in Supporting Information Table S6). Matched non-users, relative to both PPI users and H2RA users, had a lower baseline prevalence of several comorbidities, and a lower mean number of GP appointments in the 6 months prior to cohort entry date (Supporting Information Table S7).

7.5.4 Risk of mortality relative to non-users

There were 86,825 (24.8 per 1,000 PY) deaths observed among matched PPI users and 69,402 (11.5 per 1,000 PY) deaths among non-users. Median follow-up was 4.3 (IQR 1.9-7.5) years among matched PPI users and 3.6 years (IQR 1.6-6.5) among non-users.

Weighted hazard ratios for all outcomes (with the exception of acute kidney injury, aortic aneurysm and COPD) were greater for PPI users compared to non-users, than for PPI users compared to H2RA users (Figures 7.2 and 7.3). For PPI use, relative to non-use, the weighted hazard ratio for all-cause mortality was 1.96 (95% CI 1.94-1.99) which was substantially higher than the comparison with H2RA users (wHR 1.38; 95% CI 1.33-1.44). Similarly, cause-specific mortality was substantially higher for a number

of outcomes such as mortality from neoplasms (3.74, 95% CI 3.63-3.84 vs. 1.74, 95% CI 1.63-1.86), liver cirrhosis (4.10, 95% CI 3.36-5.01 vs. 1.95, 95% CI 1.10-3.46), and gastric cancer (14.59, 95% CI 11.16-19.08 vs. 2.35, 95% CI 1.39-3.99).

7.5.5 Sensitivity analysis

To fully explain the lower bound of the observed association (HR 1.33) with all-cause mortality an unmeasured confounder would need to be associated with either exposure or outcome by at least RR 1.99 (risk ratio) and associated with both exposure and outcome by at least RR 1.33 (*Ding and VanderWeele*, 2016). Differences between estimates obtained from direct adjustment for covariates in the Cox model (adjusted HR all-cause mortality 1.39, 95% CI 1.35-1.42) relative to propensity score weighting (wHR 1.38, 95% CI 1.33-1.44) were minor (Supporting Information Tables S8 and S9). Censoring follow-up among PPI users at first prescription of a H2RA similarly had minimal impact on effect estimates (wHR all-cause mortality 1.36, 95% CI 1.31-1.41; Supporting Information Table S10). Censoring follow-up at treatment discontinuation consistently reduced effect estimates (wHR all-cause mortality 1.12, 95% CI 1.04-1.20; Supporting Information Table 11) which may reflect both reduced follow-up and informative censoring whereby treatment is discontinued prior to death. Censoring follow-up at 31st December 2014 before PPIs became more widely available had little impact on effect estimates (wHR all-cause mortality 1.41, 95% CI 1.36-1.47; Supporting Information Table S12). The differences between estimates of cause-specific mortality when defining cause of death based on any recorded, rather than primary recorded cause, were small (Supporting Information Table S13). Propensity score trimming had minor effect on estimated associations (Supporting Information Table S14 and S15).

7.6 Discussion

In this cohort study we found associations between prescription of PPIs and both all-cause and cause-specific mortality. However, our findings also clearly indicated there are important differences between PPI users and comparator groups on characteristics predictive of death. PPI users were sicker and in order to draw any causal conclusions from these findings we must first decide whether these baseline differences were fully captured by measured covariates. In line with previous non-interventional studies, at baseline PPI users had a higher prevalence of measures of comorbidity and indicators of frailty, both when compared to H2RA users and even more so when compared to non-users (*Charlot et al.*, 2010; *Xie et al.*, 2017). We would therefore expect the PPI users to have a higher risk of mortality than either comparator group, which may bias a causal assessment of the observed association with PPIs.

With both comparator groups (H2RAs and non-users), hazard ratios decreased towards the null with increasing adjustment, indicative of increasing control of confounding. The unweighted hazard ratio for all-cause mortality was 1.65, which decreased to 1.38 after adjustment for covariates chosen by the study investigator, and to 1.31 after adjustment for the HDPS (a methodology that has been suggested to control for additional confounding in studies using electronic health record data) (*Schneeweiss et al.*, 2009). However, it is not clear whether all confounding was fully controlled by any of these approaches. The HDPS, as with any covariate adjustment method, requires confounders (or proxies of those confounders) to be measured to eliminate confounding.

Success in adjusting for confounding in all non-interventional studies hinges on the quality and completeness of data recording for all relevant variables. If we had accounted for all confounding, and the associations we reported were causal, we would expect the adjusted effect size to be very similar for both the non-user and H2RA comparator groups. However, the adjusted effect estimates were substantially higher when PPI users were compared to non-users, rather than H2RA users. This suggests residual confounding in one or both of these comparisons. Our estimates are consistent with, though slightly higher than those observed in a cohort of United States veterans in a

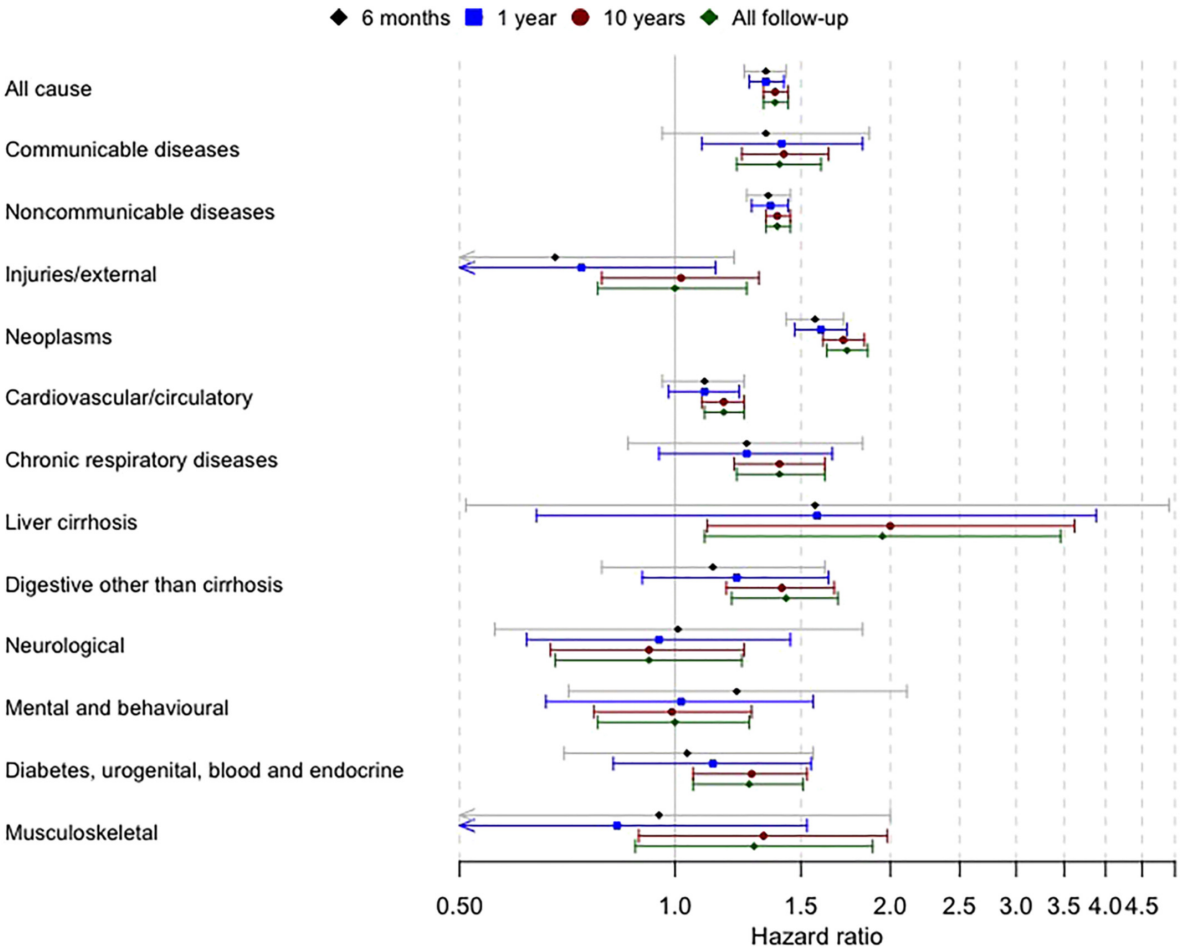


Figure 7.4: *Forest plot of the associations between PPI prescription, relative to H2RA prescription, and both all-cause and broad-level cause-specific mortality over up to 6 months, 1 year, 10 years, and all follow-up. Hazard ratios and 95% confidence intervals are plotted here and listed in Supporting Information Table S5. All figures represent propensity score (based on investigator chosen covariates) weighted hazard ratios.*

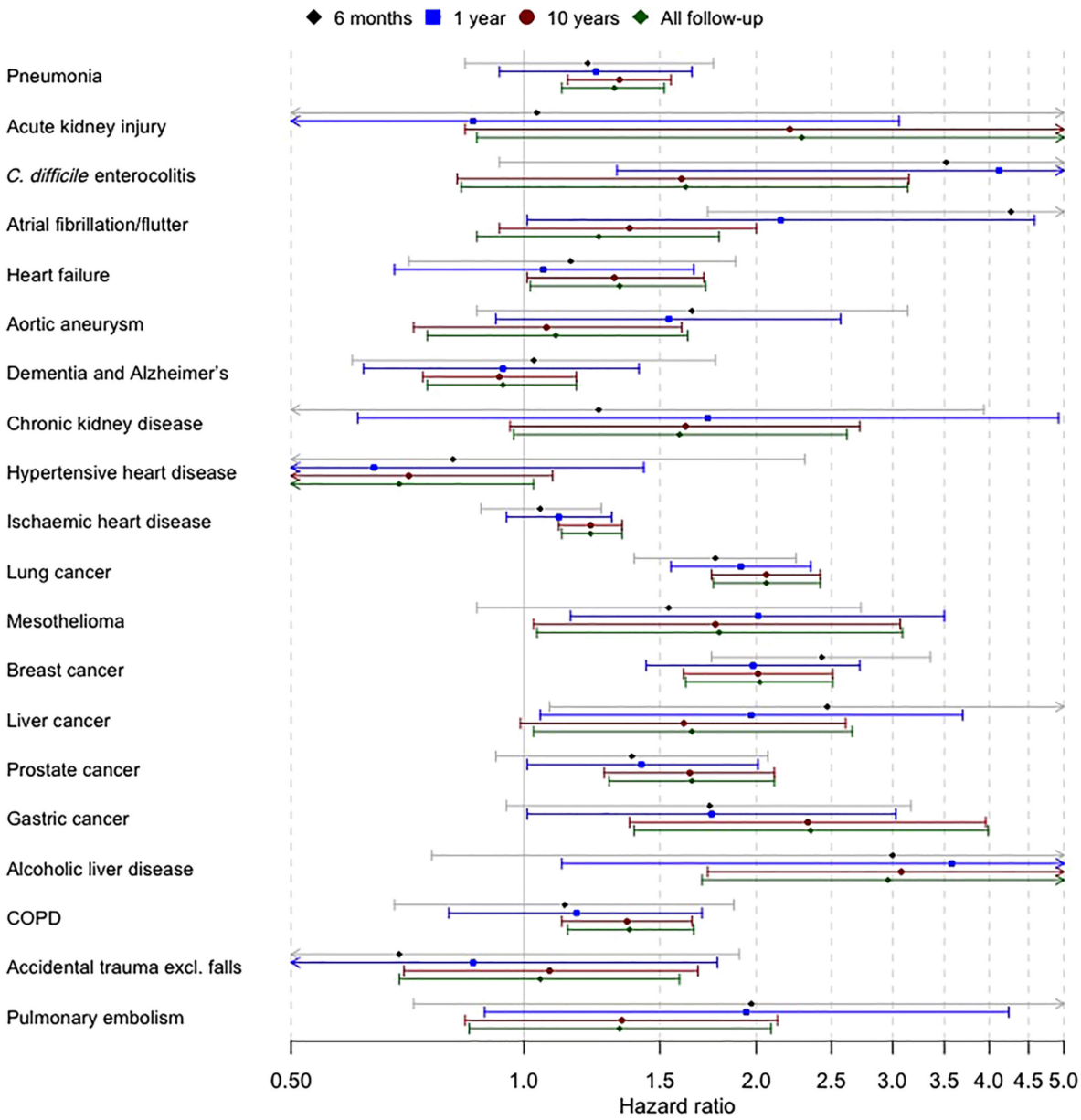


Figure 7.5: *Forest plot of the associations between PPI prescription, relative to H2RA prescription, and mortality from individual specified causes over up to 6 months, 1 year, 10 years, and all follow-up. Hazard ratios and 95% confidence intervals are plotted here and listed in Supporting Information Table S5. All figures represent propensity score (based on investigator chosen covariates) weighted hazard ratios.*

non-interventional study examining the association between PPIs and all-cause mortality (*Xie et al.*, 2017). In this previous cohort, the unadjusted and adjusted hazard ratios for all-cause mortality were 1.46 (95% CI 1.43-1.49) and 1.25 (95% CI 1.23-1.28), compared to 1.65, 95% CI 1.62-1.69 and 1.38 (1.33-1.44) in our study.

We did not find an association with the control outcome, mortality from accidental trauma excluding falls, which is expected given that this is less likely than other causes of death to be strongly related to health status. There was weak evidence for an association with the control outcome of mortality from pulmonary embolism, an outcome which might be affected by differences in underlying frailty between comparator groups, though confidence intervals were wide as this outcome was relatively rare in our cohort.

We found associations within six months of commencing PPI therapy for a number of very varied diseases that typically have a prolonged course from initial development to diagnosis (e.g. lung cancer). If causal, they would represent the actions of PPI on prevalent disease which could only be explained by a wide range of distinct biological mechanisms since the diseases themselves have different aetiologies and patterns of progression. Alternatively, such short-term associations could be explained by confounding, whereby PPIs are prescribed for symptoms in the early stages of a serious progressive illness. Notably, short-term associations are generally not reported in non-interventional studies of drugs as they are judged as unlikely to be causal, but we believe that reporting them is informative in showing a more rounded picture of the general problem of confounding.

Randomised controlled trials have not replicated the findings from non-interventional studies, providing further evidence that non-interventional studies are likely confounded. A recent randomised placebo-controlled trial of 17,598 patients with stable cardiovascular disease (median follow-up of 3.01 years) found no association between PPI use and all-cause mortality (HR 1.03, 95% CI 0.92–1.15), mortality from cardiovascular causes (HR 1.03, 95% CI 0.89–1.20), or mortality from non-cardiovascular causes (1.02, 95% CI 0.87–1.21) (*Moayyedi et al.*, 2019). Whilst it could be argued that any causal association may require a longer duration of exposure, these results at least mitigate against a short-to-medium-term effect of PPIs on undiagnosed disease. No association

was observed in the RCT with incidence of cause-specific mortality outcomes previously associated with PPI use in non-interventional studies including cancer, chronic kidney disease, dementia, pneumonia and COPD. The one exception to these negative findings was an increased incidence of enteric infections.

Previous non-interventional studies found differences in patient baseline characteristics similar to those observed in our study (*Charlot et al.*, 2010; *Lee et al.*, 2015; *Xie et al.*, 2017). The range of comorbidities that are more prevalent among PPI users reflects the multiple indications for, and broad patient population prescribed, PPIs. No observational study can deal with unmeasured confounding, and in the case of prescribing of PPIs the data suggest that they are given at a greater rate to people who are frail, but we cannot fully assess how frail they may be. An unmeasured confounder associated with both exposure and outcome by a risk ratio of at least 1.33, and with either by at least 1.99, could potentially fully explain the observed association (*Ding and VanderWeele*, 2016). Given strong associations previously observed between frailty and mortality (RR $>$ 2) and the possibility that more than one relevant variable may be under- or un-recorded, such unmeasured confounding is plausible (*Puts et al.*, 2005). This could be related to either the recording of presence or absence of a disease, but possibly more importantly, could also be related to the severity of a disease. For example, PPI users may have not only a higher prevalence of diseases such as hypertension and diabetes; they may also have more severe disease, which is less readily captured through routine health records.

Residual confounding may explain the wide ranging associations with PPI use observed both in the literature, where PPIs have been associated with over a dozen conditions, and in this study with cause-specific mortality from a number of causes (*Batchelor et al.*, 2018; *Benson et al.*, 2015; *Cheema*, 2019; *Cheung et al.*, 2018; *Dial et al.*, 2005; *Laheij et al.*, 2004; *Li et al.*, 2019; *Llorente et al.*, 2017; *Targownik et al.*, 2008; *Tvingsholm et al.*, 2018; *Yang et al.*, 2017; *Yuan et al.*, 2020). Notably, non-interventional research on the interaction between clopidogrel and PPIs similarly suffered from hard to account for confounding, and ultimately randomised trials suggested the harmful associations detected in many studies were not causal (*Demcsák et al.*, 2018; *Douglas et al.*, 2012).

Our study has several strengths. It is the largest study to date to examine the association of PPI prescription with all-cause and cause-specific mortality. Furthermore, our population was broadly representative of patients taking PPIs in the general population, given that the database used, CPRD GOLD, is similar to the UK population on age, sex, and ethnicity (*Herrett et al.*, 2015). The validity of health data recording in CPRD GOLD has been found to be very high (*Herrett et al.*, 2010b).

There were limitations to our study. We expect some misclassification of acid-suppression drug usage as the data capture primary care prescriptions, but not over-the-counter or pharmacy medications sold without a prescription. However, sensitivity analysis limiting the study period to when PPIs were solely available through prescription or pharmacy (before January 2015), had little effect on results. Given the large number of cause-specific mortality associations estimated, which increases the risk of observing some statistically significant associations that are purely due to chance, caution is warranted in the interpretation of any one individual association.

There will have been some misclassification due to non-adherence to prescribed medication, which is not recorded in these electronic health records. Assuming such misclassification was non-differential with respect to the outcome, this would tend to bias any causal association towards the null. There may be some misclassification of cause of death due to incorrect attribution of cause by the clinician certifying the death certificate. However, we expect misclassification to be non-differential with respect to PPI prescribing. Propensity score trimming did not lead to a systematic or major change in the hazard ratios, which we might have anticipated had it led to more valid estimates.

We have demonstrated that PPIs are associated with an increased risk of mortality from a wide range of illnesses. However, PPIs are preferentially given to people at increased risk of death. The change in hazard ratios with increasing adjustment and between comparison groups is indicative of residual confounding, and as such, we believe causality is unclear. Randomised trials are generally the ideal source of evidence to answer important questions about drug safety, but are not always available in sufficient size. Whilst non-interventional studies can often be helpful in assessing drug safety, we have presented an example where extra caution is needed in their design and reporting

due to intractable confounding. We recommend a strong emphasis on informative sensitivity analyses, such as negative controls and quantitative bias analyses, to assess this problem in order to inform appropriate interpretation and application to clinical practice.

As with all medications, care should be taken to ensure PPIs are prescribed appropriately and for the correct duration. What is clear is that PPIs have a well-defined clinical benefit, and that uncertainty over their safety can lead to adverse unintended consequences (*Platt et al.*, 2019).

7.7 Acknowledgments

This study evolved from an earlier project conceived and developed by our much missed colleague Adrian Root.

7.8 Ethics statement

This study was approved by the London School of Hygiene and Tropical Medicine Research Ethics Committee (reference no.15655) and by the CPRD Independent Scientific Advisory Committee (ISAC reference 17_252) (see Appendices E & F for details).

7.9 Supporting information

Supplementary information

Contents

Supplementary Methods ---------- 2

Summary of methods for examining robustness of results to confounding ---------- 2

Covariate definition and parameterisation ---------- 2

High-dimensional propensity score estimation ---------- 5

Supplementary Results ---------- 6

Table S1: International Classification of Disease (ICD) chapters and codes used for ---------- 6

Table S2: Absolute standardised differences between PPI and H2RA users before and after weighting ---------- 7

Table S3 - Association between PPI prescription and mortality among PPI and H2RA users without weighting, with propensity score weighting, and with high-dimensional propensity score weighting ---------- 9

Figure S1: Weighted* cumulative hazard curve for all-cause mortality among PPI and H2RA users ---------- 10

Table S4 - Association between PPI prescription and mortality among PPI and non-users with and without propensity score weighting ---------- 11

Table S5: Association between PPI prescription and mortality among PPI and H2RA users over up to 6 months, 1 year and 10 year follow-up after treatment initiation ---------- 12

Table S6: Characteristics of matched and unmatched PPI users ---------- 13

Table S7: Absolute standardised differences between matched PPI and non-users before and after weighting -- 14

Table S8 - Association between PPI prescription and mortality among PPI and H2RA users with Cox model adjustment of covariates ---------- 16

Table S9 - Associations between covariates and all-cause mortality among PPI and H2RA users with Cox model adjustment for covariates ---------- 17

Table S10 - Association between PPI prescription and mortality among PPI and H2RA users censoring follow-up at prescription of a H2RA among PPI users ---------- 19

Table S11 - Association between PPI prescription and mortality among PPI and H2RA users censoring follow-up at first treatment break ---------- 20

Table S12 - Association between PPI prescription and mortality among PPI and H2RA users censoring follow-up at 31st December 2014 ---------- 21

Table S13 - Association between PPI prescription and mortality among PPI and H2RA users with cause of death defined based on any rather than primary cause recorded ---------- 22

Table S14 - Weighted association between PPI prescription and mortality among PPI and H2RA users with and without propensity score trimming ---------- 23

Table S15 - Weighted association between PPI prescription and mortality among PPI users and non-users with and without propensity score trimming ---------- 24

Supplementary Methods

Summary of methods for examining robustness of results to confounding

Method	Description
Multiple comparator groups	We included two comparator groups (H2 receptor antagonist [H2RA] users and non-users) to compare with PPI users. Using average effect of treatment in the treated (ATT) weights, we estimated the same effect in both analyses: the effect of proton pump inhibitor (PPI) prescription amongst PPI users. We would expect to observe similar results using both comparators in the absence of uncontrolled confounding and assuming no causal effect of H2RAs on mortality.
Negative control outcomes	We included outcomes that have not previously been associated with PPI use. One of these variables, accidental trauma, we did not expect to be associated with the exposure or confounders. An association observed with this variable may indicate selection bias or other biases. We also included an outcome that we did not expect to be associated causally (pulmonary embolism), but that we expected may be associated with PPIs through a confounder (patient frailty). An association with this control outcome may indicate residual confounding by this confounder.
Weighting using the propensity score derived from investigator-chosen covariates	A reduction in effect estimates after adjusting for covariates provides an indication that there is confounding which is being adjusted for. The extent of the reduction in effect estimates can provide an indication on the strength of confounding.
Weighting using the high-dimensional propensity score (hd-PS)	The hd-PS is a method for selecting a large number of covariates into a propensity score model that has been developed for use in insurance claims and electronic health record data. It has been suggested that using a high-dimensional propensity score may control for additional confounding relative to a propensity score based on investigator-chosen covariates. A reduction in effect estimates using the hd-PS, relative to a propensity score based on investigator-chosen covariates, provides an indication that the analysis using the investigator-chosen covariate propensity score may have residual confounding.
Hazard ratios over different periods of follow-up	We estimated hazard ratios over different periods of follow-up (6 months, 1 year, up to ten years, all follow-up). For many outcomes we would not expect to observe an association with outcome incidence shortly after treatment initiation given disease pathogenesis. An association observed at 6 months for these outcomes may be explained by confounding.

Covariate definition and parameterisation

Covariate	Definition and parameterisation
Demographic and lifestyle variables	
Age	Continuous covariate calculated at baseline based on date of birth. Included in propensity score (PS) model as a restricted cubic spline

	with five knots. Interaction between age and sex included in PS model.
Sex	Binary covariate. Interaction with age and BMI included in PS model.
Index of multiple deprivation (IMD) score	Categorical covariate defined by IMD score twentile. IMD score was calculated by the data provider based on patient postcode.
Body mass index	Continuous covariate calculated based on recorded height and weight measurements. Interaction with sex included in PS model. Missing indicator approach used for missing values.
Smoking status	Categorical covariate (non-smoker, smoker, ex-smoker). Missing indicator approach used for missing values.
Alcohol status and level	Categorical covariates (non-drinker, ex-drinker, current drinker; light, moderate, heavy consumption). Missing indicator approach used for missing values.
Potential indications for PPI treatment in 6 months prior to baseline	
NSAID	Binary covariate based on presence of a relevant recorded prescription within 6 months prior to index date.
Aspirin	
Oral anticoagulant	
Clopidogrel	
Inhaled steroid	
Systemic steroid	
Upper gastrointestinal endoscopy	Binary covariate based on presence of a relevant Read code within 6 months prior to index date.
Gastric cancer	
Gastro-oesophageal reflux disease	
Peptic ulcers	
Upper gastrointestinal bleeding	
Pancreatitis	
Cirrhosis	
Oesophagitis	
Barrett's oesophagus	
H. pylori infection	
Indicators of frailty in 6 months prior to baseline	
Number of hospital admissions	Categorical covariate (0, 1, 2, 3, 4+) derived from number of hospital admissions in 6 months prior to index date within linked Hospital Episode Statistics Admitted Patient Care data.
Number of general practitioners (GP) appointments	Categorical covariate (0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11-12, 13-14, 15-19, 20-29, 30+) derived from number of GP appointments in 6 months prior to index date.
Number of different drug types prescribed	Categorical covariate (0, 1, 2, 3, 4, 5, 6, 7, 8+) derived from number of different BNF chapters prescribed in 6 months prior to index date.
Ever recorded previous comorbidities	
Hypertension	Binary covariate based on presence of a Read code ever prior to index date.
Coronary heart disease	
Heart failure	
Peripheral artery disease	
Cerebrovascular disease	
Other atheroma	
Chronic obstructive pulmonary disease	
Cancer	
Non-viral liver disease	
HIV	
Dementia	

Diabetes	
Chronic kidney disease	Binary covariate derived from calculated eGFR.
Other	
Calendar year	Categorical covariate derived from calendar year at index date (1998-2003, 2004-2009, 2010-2015). Given strong association between exposure and calendar year, propensity scores were estimated within each category of calendar year.

High-dimensional propensity score estimation

The high-dimensional propensity score (hd-PS) is a semi-automated approach to confounder selection in large healthcare databases. There are a number of investigator decisions to be made when using this approach:

1. **Identify dimensions which reflect aspects of care.**
   We identified the following dimensions:
   Clinical (Read codes): Signs, symptoms and diagnoses
   Referral (Read codes): Indicate a possible escalation in care
   Prescriptions (BNF codes): Patterns of drug usage

2. **Sort codes by prevalence within each dimension.**
   We selected the top 200 most prevalent codes from each dimension.

3. **Prioritise result hd-PS covariates**
   Covariates were prioritised using the Bross Formula, which prioritises covariates with most potential to bias the treatment-outcome relationship of interest.

4. **Select the top k covariates**
   Separately for each outcome we selected the top 500 covariates to be included in the propensity score alongside the investigator-chosen covariates.

We added hd-PS covariates to the propensity score model to supplement investigator-chosen covariates.

Supplementary Results

Table S1: International Classification of Disease (ICD) chapters and codes used for outcome definitions

Cause of death outcomes	**Relevant ICD-10 chapters/codes**	**Relevant ICD-9 codes**
Top-level outcomes		
Communicable diseases	A, B, J00-22	1-139, 460-469, 480-488
Non-communicable diseases	C through R	140-459, 470-479, 490-799
Injuries/external	S through Y	800-999, E001-E999
Second-level outcomes		
Neoplasms	C	140-239
Cardiovascular/circulatory	I	390-459
Chronic respiratory diseases	J23-99	470-478, 490-519
Liver cirrhosis	K70.3, K71.7, K74.3-6	571.2, 571.5, 571.6
Digestive other than cirrhosis	K except codes above	520-579 (except 571.2, 571.5 & 571.6)
Neurological	G	320-359, 290
Mental and behavioural	F	291-319
Diabetes, urogenital, blood and endocrine	D50-89, E, N	240-289, 580-629
Musculoskeletal	M	710-739
Individual causes		
Pneumonia	J10.0, J11.0, J12-J18	480-486, 487.0, 514
Acute kidney injury	N17	584
Enterocolitis due to *Clostridium difficile*	A04.7	008.45
Atrial fibrillation/flutter	I48	427.3
Heart failure	I50	428
Aortic Aneurysm	I71	441
Dementia and Alzheimer's disease	F00, F01, F03, G30	290, 294.2, 331
Chronic kidney disease	N18	585
Hypertensive heart disease	I11	402
Ischaemic heart disease	I20-I25	410-414
Accidental trauma (excluding falls)	V01-X59 (excluding W00-W19), Y86, Y86	E800-E928 (excluding E870-E888)
Pulmonary embolism	I26	415.1-415.19
Lung cancer	C34	162 (except 162.0 & 162.2)
Mesothelioma	C45	163
Breast cancer	C50	174
Liver cancer	C22	155
Prostate cancer	C61	185
COPD	J40-J44	490-492, 496
Alcoholic liver disease	K70	571.0-571.3
Gastric cancer	C16	151

Table S2: Absolute standardised differences between PPI and H2RA users before and after weighting

Characteristic	Unweighted		Weighted		Unweighted ASD	Weighted ASD
	H2RA user	PPI user	H2RA user	PPI user		
Effective sample size	124,410	733,885	732,547.6	733,885		
Mean age in years	51.2	54.9	55.0	54.9	0.204	0.006
Mean BMI	26.5	27.2	27.2	27.2	0.118	<0.001
Calendar year						
1998-2003	56.7%	14.2%	14.2%	14.2%	1.135	0.001
2004-2009	31.3%	41.6%	41.6%	41.6%	0.210	<0.001
2010-2015	12.0%	44.2%	44.2%	44.2%	0.678	0.001
Female	57.3%	54.7%	54.1%	54.7%	0.052	0.014
Current smoker	24.3%	19.6%	19.7%	19.6%	0.119	0.002
Ex-smoker	24.3%	33.3%	33.2%	33.4%	0.195	0.005
High alcohol intake	2.7%	3.4%	3.4%	3.4%	0.039	<0.001
Below national median IMD	49.4%	51.8%	51.7%	51.8%	0.048	0.002
In 6 months prior to treatment initiation						
Number of hospital admissions	0.3	0.4	0.4	0.4	0.052	0.001
Number of GP appointments	4.8	5.9	6	5.9	0.165	0.003
Number of BNF drug chapters	2.3	2.5	2.5	2.5	0.078	0.016
NSAID	21.3%	32.1%	31.5%	32.1%	0.236	0.012
Aspirin	11.7%	15.0%	14.9%	15.0%	0.093	0.004
Clopidogrel	1.6%	2.1%	2.2%	2.1%	0.036	0.009
Oral anticoagulant	2.0%	2.4%	2.5%	2.4%	0.022	0.010
Inhaled steroid	11.4%	12.8%	13.1%	12.8%	0.045	0.009
Systemic steroid	6.7%	7.2%	7.2%	7.2%	0.018	0.002
Upper GI bleed	0.8%	1.3%	1.3%	1.3%	0.045	0.001
Gastric cancer	0.0%	0.1%	0.0%	0.1%	0.014	0.005
GORD	7.0%	8.4%	8.7%	8.4%	0.052	0.009
Peptic ulcer	0.3%	0.9%	0.9%	0.9%	0.067	0.001
Upper GI endoscopy	0.0%	0.1%	0.1%	0.1%	0.020	0.002
Pancreatitis	0.1%	0.1%	0.2%	0.1%	0.002	0.007
Cirrhosis	0.0%	0.1%	0.1%	0.1%	0.022	0.006
Oesophagitis	2.7%	3.5%	3.5%	3.5%	0.048	<0.001
Barrett's oesophagus	0.0%	0.2%	0.2%	0.2%	0.048	0.003
H pylori infection	0.5%	1.5%	1.6%	1.5%	0.088	0.005
Ever previous						
Hypertension	19.7%	26.1%	25.9%	26.1%	0.149	0.005
Coronary Heart Disease	8.0%	8.2%	8.3%	8.2%	0.010	<0.001
Heart failure	2.0%	2.0%	2.0%	2.0%	0.002	0.003
Peripheral artery disease	2.0%	2.2%	2.2%	2.2%	0.013	0.001
Cerebrovascular disease	3.7%	4.6%	4.8%	4.6%	0.043	0.011
Other atherosclerosis	0.1%	0.1%	0.1%	0.1%	0.006	0.006
COPD	2.6%	3.6%	3.7%	3.6%	0.055	0.003
Cancer	7.4%	10.1%	10.5%	10.1%	0.093	0.012

Non-viral liver disease	0.6%	1.0%	0.9%	1.0%	0.045	0.007
HIV	0.1%	0.1%	0.1%	0.1%	0.004	0.004
CKD	8.8%	13.9%	13.9%	13.9%	0.150	0.002
Dementia	0.5%	0.9%	1.0%	0.9%	0.047	0.011
Diabetes	5.2%	7.7%	7.7%	7.7%	0.097	<0.001

Abbreviations: ASD, absolute standardised difference; H2RA, H2 receptor antagonist; PPI, proton pump inhibitor; BMI, body mass index; IMD, Index of Multiple Deprivation; GP, General Practitioner; BNF, British National Formulary; GI, gastrointestinal; GORD, gastro-oesophageal reflux disease; COPD, chronic obstructive pulmonary disease; CKD, chronic kidney disease.

Table S3 - Association between PPI prescription and mortality among PPI and H2RA users without weighting, with propensity score weighting, and with high-dimensional propensity score weighting

	PPI users/H2RA users					
Cause of death	Unweighted HR		Weighted HR		HDPS Weighted HR	
	HR	95% CI	HR	95% CI	HR	95% CI
Top-level causes						
All cause	1.65	1.62-1.69	1.38	1.33-1.44	1.31	1.26-1.37
Communicable diseases	1.48	1.36-1.61	1.40	1.22-1.60	1.39	1.20-1.60
Non-communicable diseases	1.67	1.64-1.71	1.39	1.34-1.45	1.32	1.27-1.38
Injuries/external	1.29	1.13-1.48	1.00	0.78-1.26	0.94	0.74-1.20
Second-level outcomes						
Neoplasms	1.98	1.91-2.05	1.74	1.63-1.86	1.62	1.50-1.74
Cardiovascular/circulatory	1.33	1.28-1.38	1.17	1.10-1.25	1.14	1.07-1.22
Chronic respiratory diseases	1.84	1.69-2.00	1.40	1.22-1.62	1.33	1.14-1.55
Liver cirrhosis	2.79	1.98-3.92	1.95	1.10-3.46	1.90	1.01-3.56
Digestive other than cirrhosis	1.85	1.66-2.07	1.43	1.20-1.69	1.39	1.12-1.74
Neurological	1.43	1.25-1.64	0.92	0.68-1.24	0.78	0.53-1.15
Mental and behavioural	1.99	1.73-2.29	1.00	0.78-1.27	0.87	0.67-1.13
Diabetes, urogenital, blood and endocrine	1.75	1.55-1.99	1.27	1.06-1.51	1.23	1.01-1.48
Musculoskeletal	1.55	1.21-1.99	1.29	0.88-1.89	1.03	0.67-1.57
Individual causes that been associated with PPIs						
Pneumonia	1.34	1.22-1.47	1.31	1.12-1.52	1.31	1.11-1.54
Acute kidney injury	2.75	1.21-6.26	2.29	0.87-6.02	2.61	0.95-7.17
C. difficile enterocolitis	2.22	1.40-3.53	1.62	0.83-3.14	1.57	0.73-3.35
Atrial fibrillation/flutter	1.76	1.35-2.29	1.25	0.87-1.79	1.04	0.68-1.57
Heart failure	1.48	1.24-1.76	1.33	1.02-1.72	1.25	0.94-1.67
Aortic aneurysm	1.19	1.00-1.42	1.10	0.75-1.63	1.03	0.67-1.59
Dementia and Alzheimer's	1.85	1.63-2.10	0.94	0.75-1.17	0.85	0.68-1.08
Chronic kidney disease	1.45	0.99-2.13	1.59	0.97-2.62	1.63	0.93-2.84
Hypertensive heart disease	1.15	0.86-1.54	0.69	0.46-1.03	0.58	0.34-1.00
Ischaemic heart disease	1.27	1.20-1.34	1.22	1.12-1.34	1.21	1.10-1.33
Lung cancer	2.23	2.04-2.43	2.06	1.76-2.42	1.89	1.63-2.20
Mesothelioma	2.66	1.93-3.66	1.79	1.04-3.09	1.60	1.00-2.57
Breast cancer	1.84	1.62-2.09	2.02	1.62-2.51	1.83	1.45-2.31
Liver cancer	2.92	2.16-3.94	1.65	1.03-2.66	1.98	1.23-3.19
Prostate cancer	2.19	1.89-2.54	1.65	1.29-2.11	1.42	1.03-1.96
Gastric cancer	2.84	2.29-3.53	2.35	1.39-3.99	2.71	1.73-4.23
Alcoholic liver disease	3.70	2.69-5.10	2.96	1.70-5.14	2.93	1.59-5.39
COPD	1.83	1.64-2.03	1.37	1.14-1.66	1.34	1.11-1.61
Control outcomes						
Accidental trauma excl. falls	1.33	1.08-1.64	1.05	0.69-1.59	1.07	0.72-1.59
Pulmonary embolism	1.36	1.01-1.81	1.33	0.85-2.09	1.50	0.98-2.31

Abbreviations: HR, hazard ratio; HDPS, high-dimensional propensity score; COPD, Chronic Obstructive Pulmonary Disease.

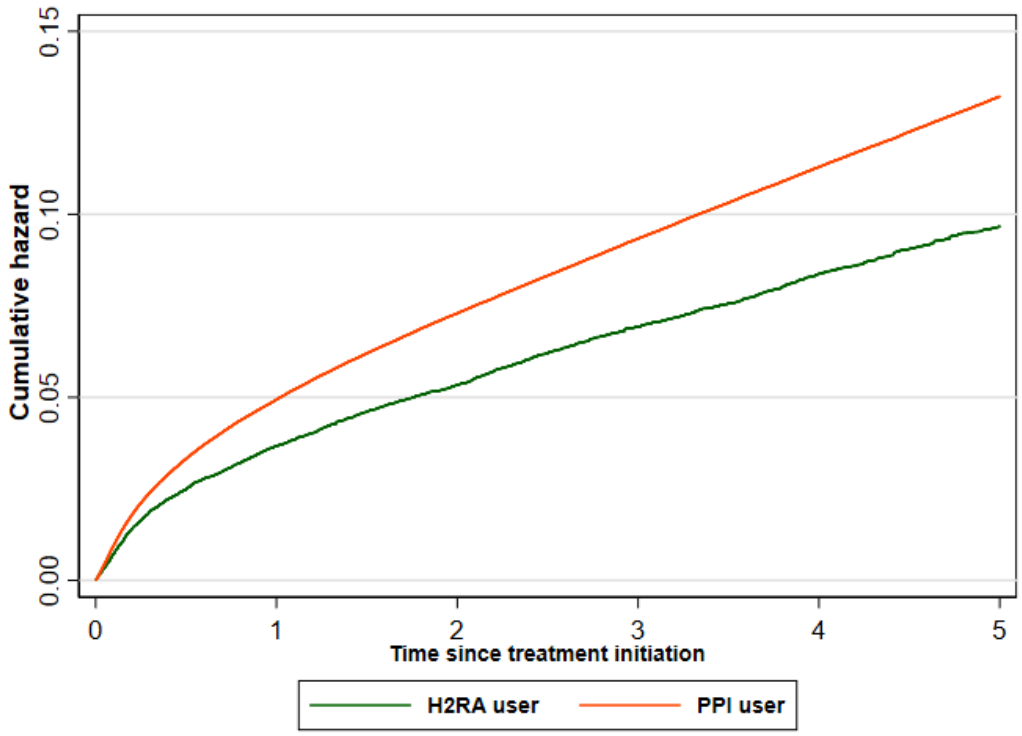


Figure S1: Weighted* cumulative hazard curve for all-cause mortality among PPI and H2RA users

* Average effect of treatment in the treated (ATT) weights were used. These weights were calculated using a propensity score estimated with investigator-chosen covariates.

Table S4 - Association between PPI prescription and mortality among PPI and non-users with and without propensity score weighting

Cause of death	Unweighted HR		Weighted HR	
	HR	95% CI	HR	95% CI
Top-level causes				
All cause	2.55	2.52-2.58	1.96	1.94-1.99
Communicable diseases	1.70	1.62-1.79	1.49	1.42-1.56
Non-communicable diseases	2.59	2.56-2.62	1.99	1.96-2.02
Injuries/external	1.58	1.47-1.70	1.26	1.17-1.35
Second-level outcomes				
Neoplasms	5.74	5.61-5.88	3.74	3.64-3.84
Cardiovascular/Circulatory	1.63	1.59-1.66	1.36	1.33-1.39
Chronic respiratory diseases	2.29	2.19-2.39	1.46	1.40-1.53
Liver cirrhosis	5.35	4.51-6.34	4.10	3.36-5.01
Digestive other than cirrhosis	2.58	2.44-2.73	1.88	1.78-2.00
Neurological	1.13	1.05-1.20	1.16	1.08-1.24
Mental and behavioural	0.83	0.78-0.89	1.07	1.00-1.15
Diabetes, urogenital, blood and endocrine	2.23	2.08-2.38	1.60	1.50-1.71
Musculoskeletal	2.32	2.01-2.68	1.56	1.35-1.80
Individual causes that been associated with PPIs				
Pneumonia	1.58	1.49-1.66	1.44	1.37-1.53
Acute kidney injury	2.48	1.70-3.63	1.56	1.10-2.21
C. difficile enterocolitis	2.76	2.19-3.48	2.16	1.71-2.74
Atrial fibrillation/flutter	1.57	1.35-1.82	1.54	1.32-1.79
Heart failure	2.20	1.98-2.45	1.87	1.67-2.10
Aortic aneurysm	1.24	1.13-1.36	1.06	0.97-1.16
Dementia and Alzheimer's	0.74	0.70-0.79	1.03	0.96-1.09
Chronic kidney disease	2.84	2.22-3.64	2.11	1.64-2.71
Hypertensive heart disease	1.27	1.08-1.51	1.02	0.87-1.19
Ischaemic heart disease	1.73	1.68-1.79	1.35	1.31-1.39
Lung cancer	5.28	5.02-5.55	3.16	2.98-3.34
Mesothelioma	6.98	5.80-8.40	3.62	2.91-4.50
Breast cancer	5.51	5.09-5.97	2.80	2.57-3.04
Liver cancer	6.08	5.15-7.18	4.72	3.83-5.82
Prostate cancer	6.05	5.51-6.64	3.14	2.84-3.48
Gastric cancer	15.10	12.79-17.82	14.59	11.16-19.08
Alcoholic liver disease	5.32	4.65-6.10	3.24	2.80-3.75
COPD	2.19	2.07-2.31	1.29	1.22-1.35
Control outcomes				
Accidental trauma excl. falls	1.61	1.44-1.80	1.28	1.15-1.43
Pulmonary embolism	1.54	1.32-1.80	1.43	1.23-1.67

Abbreviations: HR, hazard ratio; COPD, Chronic Obstructive Pulmonary Disease.

Table S5: Association between PPI prescription and mortality among PPI and H2RA users over up to 6 months, 1 year and 10 year follow-up after treatment initiation

Cause of death	Weighted HR over 6 months		Weighted HR over 1 year		Weighted HR over 10 years	
	HR	95% CI	HR	95% CI	HR	95% CI
Top-level causes						
All cause	1.34	1.25-1.43	1.34	1.27-1.42	1.38	1.33-1.44
Communicable diseases	1.34	0.96-1.87	1.41	1.09-1.83	1.42	1.24-1.64
Non-communicable diseases	1.35	1.26-1.45	1.36	1.28-1.44	1.39	1.34-1.45
Injuries/external	0.68	0.38-1.21	0.74	0.48-1.14	1.02	0.79-1.31
Second-level outcomes						
Neoplasms	1.57	1.43-1.72	1.60	1.47-1.74	1.72	1.61-1.84
Cardiovascular/Circulatory	1.10	0.96-1.25	1.10	0.98-1.23	1.17	1.09-1.25
Chronic respiratory diseases	1.26	0.86-1.83	1.26	0.95-1.66	1.4	1.21-1.62
Liver cirrhosis	1.57	0.51-4.91	1.58	0.64-3.88	2.00	1.11-3.62
Digestive other than cirrhosis	1.13	0.79-1.62	1.22	0.90-1.64	1.41	1.18-1.67
Neurological	1.01	0.56-1.83	0.95	0.62-1.45	0.92	0.67-1.25
Mental and behavioural	1.22	0.71-2.11	1.02	0.66-1.56	0.99	0.77-1.28
Diabetes, urogenital, blood and endocrine	1.04	0.70-1.56	1.13	0.82-1.55	1.28	1.06-1.53
Musculoskeletal	0.95	0.45-2.00	0.83	0.46-1.53	1.33	0.89-1.98
Individual causes that been associated with PPIs						
Short term association plausible						
Pneumonia	1.21	0.84-1.76	1.24	0.93-1.65	1.33	1.14-1.55
Acute kidney injury	1.04	0.21-5.12	0.86	0.24-3.06	2.21	0.84-5.83
C. difficile enterocolitis	3.52	0.93-13.37	4.12	1.32-12.88	1.60	0.82-3.15
Atrial fibrillation/flutter	4.27	1.73-10.49	2.15	1.01-4.58	1.37	0.93-2.00
Heart failure	1.15	0.71-1.88	1.06	0.68-1.66	1.31	1.01-1.71
Aortic aneurysm	1.65	0.87-3.14	1.54	0.92-2.57	1.07	0.72-1.60
COPD	1.13	0.68-1.87	1.17	0.80-1.70	1.36	1.12-1.65
Short term association not expected						
Dementia and Alzheimer's	1.03	0.60-1.77	0.94	0.62-1.41	0.93	0.74-1.17
Chronic kidney disease	1.25	0.39-3.94	1.73	0.61-4.92	1.62	0.96-2.72
Hypertensive heart disease	0.81	0.29-2.31	0.64	0.29-1.43	0.71	0.46-1.09
Ischaemic heart disease	1.05	0.88-1.26	1.11	0.95-1.30	1.22	1.11-1.34
Lung cancer	1.77	1.39-2.25	1.91	1.55-2.35	2.06	1.75-2.42
Mesothelioma	1.54	0.87-2.73	2.01	1.15-3.50	1.77	1.03-3.07
Breast cancer	2.43	1.75-3.36	1.98	1.44-2.72	2.01	1.61-2.51
Liver cancer	2.47	1.08-5.63	1.97	1.05-3.70	1.61	0.99-2.61
Prostate cancer	1.38	0.92-2.07	1.42	1.01-2.01	1.64	1.27-2.11
Gastric cancer	1.74	0.95-3.17	1.75	1.01-3.03	2.33	1.37-3.96
Alcoholic liver disease	3.00	0.76-11.92	3.58	1.12-11.43	3.08	1.73-5.50
Control outcomes						
Accidental trauma excl. falls	0.69	0.25-1.90	0.86	0.42-1.78	1.08	0.70-1.68
Pulmonary embolism	1.97	0.72-5.40	1.94	0.89-4.24	1.34	0.84-2.13

Abbreviations: HR, hazard ratio; COPD, Chronic Obstructive Pulmonary Disease.

Table S6: Characteristics of matched and unmatched PPI users

Characteristic	Unmatched PPI users	Matched PPI users
N	44,283	689,602
Mean age in years	76.3	53.6
Mean BMI	26.8	27.2
Mean calendar year	2012.2	2008.2
Female	58.59%	54.49%
Current smoker	8.85%	20.27%
High alcohol intake	2.19%	3.45%
Below national median IMD	55.82%	51.56%
In 6 months prior to treatment initiation		
Mean number of hospital admissions	0.5	0.4
Mean number of GP appointments	9.0	5.7
Mean number of BNF drug chapters	3.5	2.4
NSAID	40.6%	31.5%
Aspirin	30.3%	14.0%
Clopidogrel	4.7%	1.9%
Oral anticoagulant	6.1%	2.1%
Inhaled steroid	15.6%	12.7%
Systemic steroid	10.7%	6.9%
Upper GI bleed	0.1%	0.1%
Gastric cancer	0.1%	0.1%
GORD	5.0%	8.6%
Peptic ulcer	1.0%	0.9%
Upper GI endoscopy	1.0%	1.3%
Pancreatitis	0.2%	0.1%
Cirrhosis	0.1%	0.1%
Oesophagitis	1.9%	3.6%
Barrett's oesophagus	0.4%	0.2%
H pylori infection	0.7%	1.6%
Ever previous		
Hypertension	53.8%	24.4%
Coronary Heart Disease	15.2%	7.8%
Heart failure	4.3%	1.8%
Peripheral artery disease	4.7%	2.0%
Cerebrovascular disease	11.0%	4.2%
Other atherosclerosis	0.2%	0.1%
COPD	8.2%	3.3%
Cancer	22.8%	9.3%
Non-viral liver disease	1.1%	1.0%
HIV	0.0%	0.1%
CKD	39.5%	12.2%
Dementia	3.8%	0.7%
Diabetes	14.0%	7.3%

Abbreviations: PPI, proton pump inhibitor; BMI, body mass index; IMD, Index of Multiple Deprivation; GP, General Practitioner; BNF, British National Formulary; GI, gastrointestinal; GORD, gastro-oesophageal reflux disease; COPD, chronic obstructive pulmonary disease; CKD, chronic kidney disease.

Table S7: Absolute standardised differences between matched PPI and non-users before and after weighting

Characteristic	Unweighted		Weighted		Unweighted ASD	Weighted ASD
	Non-user	PPI user	Non-user	PPI user		
Effective sample size	1,361,245	689,602	744,078.80	689,602		
Mean age in years	53.3	53.6	53.3	53.6	0.017	0.017
Mean BMI	26.6	27.2	27.4	27.2	0.112	0.033
Mean calendar year	2008.1	2008.2	2007.9	2008.2	0.011	0.051
Female	54.4%	54.5%	56.1%	54.5%	0.002	0.033
Current smoker	18.3%	20.3%	21.1%	20.3%	0.049	0.019
High alcohol intake	2.3%	3.5%	3.7%	3.5%	0.061	0.012
Below national median IMD	53.3%	51.6%	50.2%	51.6%	0.034	0.027
In 6 months prior to treatment initiation						
Mean number of hospital admissions	0.1	0.4	0.5	0.4	0.194	0.048
Mean number of GP appointments	3.2	5.7	6.3	5.7	0.393	0.093
Mean number of BNF drug chapters	1.5	2.4	2.7	2.4	0.434	0.142
NSAID	7.6%	31.5%	37.5%	31.5%	0.515	0.127
Aspirin	8.3%	14.0%	14.7%	14.0%	0.166	0.019
Clopidogrel	0.5%	1.9%	2.1%	1.9%	0.103	0.018
Oral anticoagulant	1.5%	2.1%	2.0%	2.1%	0.043	0.008
Inhaled steroid	7.3%	12.7%	14.3%	12.7%	0.162	0.049
Systemic steroid	2.0%	6.9%	8.3%	6.9%	0.195	0.054
Upper GI bleed	1.0%	1.3%	1.5%	1.3%	0.031	0.014
Gastric cancer	0.0%	0.1%	0.1%	0.1%	0.016	0.006
GORD	1.6%	8.6%	12.2%	8.6%	0.251	0.127
Peptic ulcer	0.8%	0.9%	0.9%	0.9%	0.014	0.001
Upper GI endoscopy	0.0%	0.1%	0.1%	0.1%	0.017	0.017
Pancreatitis	0.2%	0.1%	0.1%	0.1%	0.018	0.001
Cirrhosis	0.1%	0.1%	0.1%	0.1%	0.004	0.002
Oesophagitis	0.7%	3.6%	5.0%	3.6%	0.154	0.072
Barrett's oesophagus	0.0%	0.2%	0.2%	0.2%	0.043	0.007
H pylori infection	0.0%	1.6%	0.1%	1.6%	0.125	0.122
Ever previous						
Hypertension	20.7%	24.4%	24.4%	24.4%	0.086	0.002
Coronary Heart Disease	4.6%	7.8%	8.4%	7.8%	0.121	0.021
Heart failure	1.1%	1.8%	1.9%	1.8%	0.056	0.005
Peripheral artery disease	1.2%	2.0%	2.1%	2.0%	0.057	0.007
Cerebrovascular disease	2.8%	4.2%	4.4%	4.2%	0.068	0.012
Other atherosclerosis	0.1%	0.1%	0.1%	0.1%	0.01	<0.001
COPD	2.0%	3.3%	3.6%	3.3%	0.074	0.017
Cancer	6.2%	9.3%	9.2%	9.3%	0.108	0.003
Non-viral liver disease	0.6%	1.0%	1.1%	1.0%	0.046	0.005
HIV	0.1%	0.1%	0.1%	0.1%	0.002	0.001
CKD	8.8%	12.2%	12.5%	12.2%	0.103	0.008
Dementia	1.1%	0.7%	0.7%	0.7%	0.043	<0.001
Diabetes	5.5%	7.3%	7.4%	7.3%	0.069	0.003

Abbreviations: ASD, absolute standardised difference; H2RA, H2 receptor antagonist; PPI, proton pump inhibitor; BMI, body mass index; IMD, Index of Multiple Deprivation; GP, General Practitioner; BNF, British National Formulary; GI, gastrointestinal; GORD, gastro-oesophageal reflux disease; COPD, chronic obstructive pulmonary disease; CKD, chronic kidney disease.

Table S8 - Association between PPI prescription and mortality among PPI and H2RA users with Cox model adjustment of covariates

Cause of death	Unadjusted HR		Adjusted HR	
	HR	95% CI	HR	95% CI
Top-level causes				
All cause	1.65	1.62-1.69	1.39	1.35-1.42
Communicable diseases	1.48	1.36-1.61	1.18	1.08-1.29
Non-communicable diseases	1.67	1.64-1.71	1.40	1.37-1.43
Injuries/external	1.29	1.13-1.48	1.10	0.96-1.27
Second-level outcomes				
Neoplasms	1.98	1.91-2.05	1.77	1.70-1.84
Cardiovascular/circulatory	1.33	1.28-1.38	1.13	1.09-1.18
Chronic respiratory diseases	1.84	1.69-2.00	1.36	1.25-1.49
Liver cirrhosis	2.79	1.98-3.92	1.79	1.26-2.55
Digestive other than cirrhosis	1.85	1.66-2.07	1.48	1.32-1.66
Neurological	1.43	1.25-1.64	1.04	0.91-1.20
Mental and behavioural	1.99	1.73-2.29	1.16	1.00-1.34
Diabetes, urogenital, blood and endocrine	1.75	1.55-1.99	1.34	1.18-1.53
Musculoskeletal	1.55	1.21-1.99	1.29	0.99-1.67
Individual causes that been associated with PPIs				
Pneumonia	1.34	1.22-1.47	1.11	1.00-1.22
Acute kidney injury	2.75	1.21-6.26	1.84	0.79-4.28
C. difficile enterocolitis	2.22	1.40-3.53	1.90	1.17-3.06
Atrial fibrillation/flutter	1.76	1.35-2.29	1.17	0.89-1.55
Heart failure	1.48	1.24-1.76	1.28	1.06-1.54
Aortic aneurysm	1.19	1.00-1.42	1.07	0.89-1.29
Dementia and Alzheimer's	1.85	1.63-2.10	1.06	0.94-1.21
Chronic kidney disease	1.45	0.99-2.13	1.10	0.74-1.64
Hypertensive heart disease	1.15	0.86-1.54	0.76	0.56-1.03
Ischaemic heart disease	1.27	1.20-1.34	1.12	1.06-1.19
Lung cancer	2.23	2.04-2.43	1.89	1.72-2.06
Mesothelioma	2.66	1.93-3.66	2.16	1.55-3.00
Breast cancer	1.84	1.62-2.09	1.78	1.56-2.03
Liver cancer	2.92	2.16-3.94	2.09	1.53-2.84
Prostate cancer	2.19	1.89-2.54	1.88	1.62-2.19
Gastric cancer	2.84	2.29-3.53	2.74	2.19-3.42
Alcoholic liver disease	3.70	2.69-5.10	2.49	1.79-3.46
COPD	1.83	1.64-2.03	1.29	1.16-1.44
Control outcomes				
Accidental trauma excl. falls	1.33	1.08-1.64	1.16	0.93-1.44
Pulmonary embolism	1.36	1.01-1.81	1.20	0.88-1.62

Abbreviations: HR, hazard ratio; COPD, Chronic Obstructive Pulmonary Disease.

Table S9 - Associations between covariates and all-cause mortality among PPI and H2RA users with Cox model adjustment for covariates

Characteristic	Adjusted HR*	95% CI
PPI exposure	1.46	(1.42-1.49)
Age in years	1.07	(1.07-1.07)
BMI		
18.5 ≥ X < 25	1 (REF)	
< 18.5	1.71	(1.66-1.76)
≥ 25	0.82	(0.81-0.83)
Missing	1.65	(1.61-1.69)
Calendar year		
1998-2003	1 (REF)	
2004-2009	0.83	(0.82-0.85)
2010-2015	0.64	(0.63-0.66)
Female	0.73	(0.72-0.74)
Smoking		
Non-smoker	1 (REF)	
Current smoker	1.77	(1.73-1.80)
Current or ex-smoker	1.53	(1.47-1.59)
Ex-smoker	1.14	(1.12-1.15)
Missing	2.30	(2.21-2.39)
Alcohol consumption		
None	1 (REF)	
Low	0.93	(0.90-0.96)
Medium	1.01	(0.97-1.05)
High	1.58	(1.51-1.65)
Missing	0.95	(0.92-0.98)
Below national median IMD	1.18	(1.16-1.19)
In 6 months prior to treatment initiation		
Number of hospital admissions	1.04	(1.04-1.04)
Number of GP appointments	1.02	(1.02-1.02)
Number of BNF drug chapters	1.08	(1.08-1.09)
NSAID	0.84	(0.83-0.86)
Aspirin	0.98	(0.96-0.99)
Clopidogrel	0.96	(0.93-0.99)
Oral anticoagulant	1.14	(1.11-1.17)
Inhaled steroid	0.82	(0.80-0.84)
Systemic steroid	1.66	(1.64-1.69)
Upper GI bleed	1.07	(1.03-1.12)
Gastric cancer	2.78	(2.48-3.11)
GORD	0.71	(0.68-0.74)
Peptic ulcer	1.09	(1.04-1.14)
Upper GI endoscopy	0.95	(0.81-1.12)
Pancreatitis	0.96	(0.84-1.10)
Cirrhosis	2.35	(2.08-2.65)
Oesophagitis	1.10	(1.05-1.16)
Barrett's oesophagus	1.22	(1.10-1.34)

H pylori infection	0.78	(0.73-0.82)
Ever previous		
Hypertension	0.99	(0.98-1.00)
Coronary Heart Disease	1.05	(1.03-1.06)
Heart failure	1.41	(1.38-1.45)
Peripheral artery disease	1.26	(1.23-1.29)
Cerebrovascular disease	1.21	(1.18-1.23)
Other atherosclerosis	0.94	(0.83-1.05)
COPD	1.42	(1.39-1.46)
Cancer	2.22	(2.19-2.25)
Non-viral liver disease	1.58	(1.50-1.66)
HIV	1.18	(0.78-1.77)
CKD	1.10	(1.08-1.12)
Dementia	1.64	(1.58-1.69)
Diabetes	1.22	(1.20-1.25)

* For ease of interpretation these hazard ratios were generated from a simplified Cox models without splines or interactions and with categorical variables with a large number of categories (IMD, no. of GP appointments, no. of BNF chapters, no. of hospital admissions) replaced by continuous or binary variables. Abbreviations: PPI, proton pump inhibitor; BMI, body mass index; IMD, Index of Multiple Deprivation; GP, General Practitioner; BNF, British National Formulary; GI, gastrointestinal; GORD, gastro-oesophageal reflux disease; COPD, chronic obstructive pulmonary disease; CKD, chronic kidney disease.

Table S10 - Association between PPI prescription and mortality among PPI and H2RA users censoring follow-up at prescription of a H2RA among PPI users

Cause of death	Unweighted HR		Weighted HR	
	HR	95% CI	HR	95% CI
Top-level causes				
All cause	1.62	1.58-1.65	1.36	1.31-1.41
Communicable diseases	1.46	1.35-1.59	1.39	1.21-1.59
Non-communicable diseases	1.64	1.60-1.68	1.37	1.32-1.42
Injuries/external	1.28	1.12-1.46	0.99	0.78-1.25
Second-level outcomes				
Neoplasms	1.92	1.85-1.99	1.69	1.58-1.80
Cardiovascular/circulatory	1.31	1.26-1.36	1.16	1.08-1.23
Chronic respiratory diseases	1.80	1.65-1.96	1.38	1.19-1.59
Liver cirrhosis	2.76	1.96-3.88	1.94	1.09-3.44
Digestive other than cirrhosis	1.81	1.62-2.03	1.40	1.18-1.67
Neurological	1.43	1.25-1.63	0.91	0.68-1.24
Mental and behavioural	1.98	1.72-2.28	1.00	0.78-1.27
Diabetes, urogenital, blood and endocrine	1.73	1.53-1.96	1.25	1.05-1.50
Musculoskeletal	1.53	1.19-1.96	1.27	0.87-1.87
Individual causes that been associated with PPIs				
Pneumonia	1.33	1.22-1.46	1.30	1.12-1.51
Acute kidney injury	2.68	1.18-6.12	2.24	0.85-5.94
C. difficile enterocolitis	2.20	1.38-3.50	1.60	0.82-3.13
Atrial fibrillation/flutter	1.74	1.33-2.28	1.24	0.87-1.78
Heart failure	1.45	1.22-1.74	1.31	1.01-1.70
Aortic aneurysm	1.19	0.99-1.42	1.11	0.75-1.64
Dementia and Alzheimer's	1.85	1.63-2.10	0.94	0.76-1.17
Chronic kidney disease	1.40	0.95-2.06	1.54	0.93-2.54
Hypertensive heart disease	1.16	0.86-1.55	0.69	0.46-1.04
Ischaemic heart disease	1.25	1.18-1.32	1.20	1.10-1.32
Lung cancer	2.19	2.01-2.39	2.03	1.73-2.38
Mesothelioma	2.58	1.87-3.55	1.74	1.01-3.00
Breast cancer	1.81	1.60-2.06	1.99	1.60-2.48
Liver cancer	2.81	2.08-3.80	1.60	0.99-2.58
Prostate cancer	2.15	1.86-2.50	1.62	1.27-2.08
Gastric cancer	2.70	2.17-3.35	2.24	1.32-3.79
Alcoholic liver disease	3.73	2.71-5.14	2.97	1.71-5.17
COPD	1.79	1.61-1.99	1.35	1.11-1.63
Control outcomes				
Accidental trauma excl. falls	1.31	1.06-1.61	1.04	0.68-1.58
Pulmonary embolism	1.33	0.99-1.77	1.32	0.84-2.07

Abbreviations: HR, hazard ratio; COPD, Chronic Obstructive Pulmonary Disease.

Table S11 - Association between PPI prescription and mortality among PPI and H2RA users censoring follow-up at first treatment break

Cause of death	Unweighted HR		Weighted HR	
	HR	95% CI	HR	95% CI
Top-level causes				
All cause	1.28	1.22-1.33	1.12	1.04-1.20
Communicable diseases	1.01	0.85-1.20	1.19	0.91-1.55
Non-communicable diseases	1.29	1.24-1.35	1.12	1.04-1.21
Injuries/external	1.31	0.85-2.01	0.68	0.37-1.26
Second-level outcomes				
Neoplasms	1.47	1.38-1.56	1.35	1.21-1.52
Cardiovascular/Circulatory	1.05	0.96-1.14	0.97	0.86-1.10
Chronic respiratory diseases	1.36	1.12-1.65	1.08	0.81-1.43
Liver cirrhosis	2.22	1.04-4.76	1.62	0.35-7.44
Digestive other than cirrhosis	1.31	1.04-1.65	1.07	0.76-1.51
Neurological	0.93	0.72-1.20	0.51	0.27-0.95
Mental and behavioural	1.55	1.16-2.07	0.79	0.52-1.22
Diabetes, urogenital, blood and endocrine	1.35	1.03-1.77	1.08	0.75-1.58
Musculoskeletal	1.65	0.84-3.25	1.39	0.60-3.25
Individual causes that been associated with PPIs				
Pneumonia	0.89	0.75-1.07	1.10	0.83-1.46
Acute kidney injury	1.53	0.36-6.44	1.34	0.31-5.74
C. difficile enterocolitis	3.31	0.81-13.56	8.26	1.98-34.44
Atrial fibrillation/flutter	1.14	0.64-2.02	1.05	0.47-2.35
Heart failure	1.05	0.73-1.50	0.97	0.58-1.62
Aortic aneurysm	1.19	0.74-1.91	1.19	0.63-2.26
Dementia and Alzheimer's	1.30	1.02-1.65	0.70	0.48-1.03
Chronic kidney disease	1.08	0.47-2.52	1.11	0.38-3.21
Hypertensive heart disease	0.75	0.41-1.39	0.31	0.14-0.68
Ischaemic heart disease	1.00	0.89-1.13	1.01	0.85-1.21
Lung cancer	1.84	1.59-2.12	1.57	1.20-2.04
Mesothelioma	1.62	1.02-2.59	1.19	0.66-2.13
Breast cancer	1.68	1.33-2.13	2.17	1.43-3.28
Liver cancer	3.11	1.65-5.87	1.87	0.79-4.39
Prostate cancer	1.40	1.09-1.78	1.26	0.84-1.90
Gastric cancer	1.61	1.15-2.25	1.59	0.84-3.01
Alcoholic liver disease	3.82	1.57-9.32	1.29	0.37-4.48
COPD	1.41	1.09-1.82	1.07	0.71-1.60
Control outcomes				
Accidental trauma excl. falls	1.20	0.63-2.29	0.79	0.29-2.20
Pulmonary embolism	1.47	0.74-2.90	2.41	0.80-7.27

Abbreviations: HR, hazard ratio; COPD, Chronic Obstructive Pulmonary Disease.

Table S12 - Association between PPI prescription and mortality among PPI and H2RA users censoring follow-up at 31st December 2014

Cause of death	Unweighted HR		Weighted HR	
	HR	95% CI	HR	95% CI
Top-level causes				
All cause	1.70	1.66-1.74	1.41	1.36-1.47
Communicable diseases	1.50	1.38-1.64	1.45	1.26-1.66
Non-communicable diseases	1.73	1.69-1.77	1.42	1.37-1.48
Injuries/external	1.31	1.14-1.50	1.05	0.83-1.32
Second-level outcomes				
Neoplasms	2.02	1.95-2.10	1.76	1.65-1.88
Cardiovascular/Circulatory	1.39	1.34-1.44	1.19	1.11-1.27
Chronic respiratory diseases	1.90	1.74-2.07	1.43	1.24-1.65
Liver cirrhosis	2.85	2.02-4.03	1.93	1.07-3.48
Digestive other than cirrhosis	1.90	1.70-2.13	1.42	1.19-1.68
Neurological	1.46	1.27-1.68	0.93	0.67-1.29
Mental and behavioural	1.91	1.65-2.21	0.97	0.75-1.24
Diabetes, urogenital, blood and endocrine	1.87	1.64-2.12	1.32	1.10-1.59
Musculoskeletal	1.70	1.31-2.20	1.36	0.90-2.03
Individual causes that been associated with PPIs				
Pneumonia	1.37	1.25-1.51	1.37	1.18-1.60
Acute kidney injury	2.78	1.22-6.34	2.14	0.81-5.65
C. difficile enterocolitis	2.27	1.42-3.62	1.55	0.80-3.02
Atrial fibrillation/flutter	1.84	1.39-2.44	1.37	0.94-1.99
Heart failure	1.51	1.26-1.80	1.26	0.97-1.64
Aortic aneurysm	1.24	1.03-1.48	1.12	0.74-1.69
Dementia and Alzheimer's	1.79	1.57-2.04	0.91	0.72-1.15
Chronic kidney disease	1.48	0.99-2.20	1.68	0.99-2.85
Hypertensive heart disease	1.17	0.87-1.58	0.66	0.43-1.02
Ischaemic heart disease	1.33	1.26-1.41	1.22	1.11-1.34
Lung cancer	2.27	2.08-2.48	2.10	1.79-2.46
Mesothelioma	2.66	1.92-3.67	1.80	1.02-3.19
Breast cancer	1.87	1.65-2.13	2.02	1.61-2.53
Liver cancer	3.02	2.21-4.12	1.84	1.14-2.98
Prostate cancer	2.24	1.93-2.61	1.62	1.25-2.09
Gastric cancer	2.98	2.39-3.72	2.96	1.74-5.02
Alcoholic liver disease	3.78	2.73-5.23	2.90	1.63-5.13
COPD	1.87	1.68-2.09	1.43	1.18-1.73
Control outcomes				
Accidental trauma excl. falls	1.40	1.13-1.74	1.24	0.87-1.77
Pulmonary embolism	1.39	1.03-1.86	1.26	0.80-1.98

Abbreviations: HR, hazard ratio; COPD, Chronic Obstructive Pulmonary Disease.

Table S13 - Association between PPI prescription and mortality among PPI and H2RA users with cause of death defined based on any rather than primary cause recorded

Cause of death	Unweighted HR		Weighted HR	
	HR	95% CI	HR	95% CI
Top-level causes				
All cause	1.65	1.62-1.69	1.38	1.33-1.44
Communicable diseases	1.68	1.61-1.76	1.30	1.21-1.41
Non-communicable diseases	1.67	1.63-1.70	1.39	1.34-1.44
Injuries/external	1.48	1.36-1.60	1.15	1.01-1.31
Second-level outcomes				
Neoplasms	1.96	1.89-2.03	1.70	1.60-1.81
Cardiovascular/Circulatory	1.54	1.49-1.59	1.24	1.17-1.32
Chronic respiratory diseases	1.86	1.75-1.97	1.42	1.29-1.56
Liver cirrhosis	2.77	2.13-3.59	1.84	1.24-2.73
Digestive other than cirrhosis	1.90	1.75-2.05	1.49	1.30-1.71
Neurological	1.50	1.36-1.66	0.98	0.81-1.20
Mental and behavioural	1.88	1.71-2.06	1.12	0.96-1.32
Diabetes, urogenital, blood and endocrine	1.96	1.85-2.07	1.27	1.16-1.39
Musculoskeletal	1.79	1.54-2.08	1.30	1.02-1.65
Individual causes that been associated with PPIs				
Pneumonia	1.59	1.51-1.67	1.30	1.20-1.41
Acute kidney injury	2.13	1.80-2.52	1.24	0.98-1.56
C. difficile enterocolitis	1.97	1.45-2.67	1.94	1.23-3.07
Atrial fibrillation/flutter	1.98	1.76-2.22	1.14	0.92-1.42
Heart failure	1.59	1.49-1.70	1.38	1.25-1.52
Aortic aneurysm	1.25	1.07-1.48	1.16	0.82-1.62
Dementia and Alzheimer's	1.79	1.64-1.96	0.97	0.83-1.14
Chronic kidney disease	1.93	1.70-2.19	1.32	1.11-1.57
Hypertensive heart disease	1.31	1.09-1.57	0.90	0.69-1.18
Ischaemic heart disease	1.45	1.39-1.52	1.24	1.14-1.34
Lung cancer	2.22	2.04-2.42	2.03	1.75-2.37
Mesothelioma	2.61	1.92-3.57	1.83	1.07-3.13
Breast cancer	1.81	1.61-2.04	1.65	1.29-2.11
Liver cancer	2.84	2.14-3.78	1.71	1.09-2.71
Prostate cancer	2.11	1.86-2.40	1.58	1.29-1.95
Gastric cancer	2.92	2.36-3.60	2.33	1.41-3.83
Alcoholic liver disease	3.70	2.77-4.93	2.94	1.83-4.73
COPD	1.89	1.75-2.05	1.41	1.23-1.63
Control outcomes				
Accidental trauma excl. falls	1.50	1.30-1.73	1.22	0.94-1.57
Pulmonary embolism	1.57	1.36-1.80	1.21	0.95-1.55

Abbreviations: HR, hazard ratio; COPD, Chronic Obstructive Pulmonary Disease.

Table S14 - Weighted association between PPI prescription and mortality among PPI and H2RA users with and without propensity score trimming

Cause of death	Without PS trimming		With PS trimming	
	HR	95% CI	HR	95% CI
Top-level causes				
All cause	1.38	1.33-1.44	1.36	1.33-1.39
Communicable diseases	1.40	1.22-1.60	1.22	1.11-1.34
Non-communicable diseases	1.39	1.34-1.45	1.37	1.34-1.41
Injuries/external	1.00	0.78-1.26	1.13	0.96-1.32
Second-level outcomes				
Neoplasms	1.74	1.63-1.86	1.61	1.55-1.68
Cardiovascular/Circulatory	1.17	1.10-1.25	1.12	1.07-1.17
Chronic respiratory diseases	1.40	1.22-1.62	1.51	1.38-1.66
Liver cirrhosis	1.95	1.10-3.46	2.07	1.38-3.11
Digestive other than cirrhosis	1.43	1.20-1.69	1.47	1.30-1.67
Neurological	0.92	0.68-1.24	1.21	1.03-1.41
Mental and behavioural	1.00	0.78-1.27	1.38	1.17-1.62
Diabetes, urogenital, blood and endocrine	1.27	1.06-1.51	1.30	1.13-1.50
Musculoskeletal	1.29	0.88-1.89	1.14	0.87-1.50
Individual causes that been associated with PPIs				
Pneumonia	1.31	1.12-1.52	1.12	1.01-1.24
Acute kidney injury	2.29	0.87-6.02	1.96	0.85-4.55
C. difficile enterocolitis	1.62	0.83-3.14	1.69	1.02-2.80
Atrial fibrillation/flutter	1.25	0.87-1.79	1.38	1.02-1.87
Heart failure	1.33	1.02-1.72	1.23	1.01-1.50
Aortic aneurysm	1.10	0.75-1.63	1.04	0.85-1.26
Dementia and Alzheimer's	0.94	0.75-1.17	1.26	1.09-1.46
Chronic kidney disease	1.59	0.97-2.62	1.23	0.82-1.85
Hypertensive heart disease	0.69	0.46-1.03	0.76	0.55-1.05
Ischaemic heart disease	1.22	1.12-1.34	1.09	1.02-1.16
Lung cancer	2.06	1.76-2.42	1.85	1.68-2.04
Mesothelioma	1.79	1.04-3.09	2.22	1.57-3.15
Breast cancer	2.02	1.62-2.51	1.47	1.28-1.69
Liver cancer	1.65	1.03-2.66	1.91	1.32-2.74
Prostate cancer	1.65	1.29-2.11	1.52	1.29-1.80
Gastric cancer	2.35	1.39-3.99	2.43	1.87-3.15
Alcoholic liver disease	2.96	1.70-5.14	3.26	2.21-4.81
COPD	1.37	1.14-1.66	1.48	1.32-1.67
Control outcomes				
Accidental trauma excl. falls	1.05	0.69-1.59	1.14	0.89-1.45
Pulmonary embolism	1.33	0.85-2.09	1.11	0.79-1.56

Abbreviations: HR, hazard ratio; PS, propensity score; COPD, Chronic Obstructive Pulmonary Disease.

Table S15 - Weighted association between PPI prescription and mortality among PPI users and non-users with and without propensity score trimming

Cause of death	Without PS trimming		With PS trimming	
	HR	95% CI	HR	95% CI
Top-level causes				
All cause	1.96	1.94-1.99	1.78	1.74-1.82
Communicable diseases	1.49	1.42-1.56	1.40	1.33-1.48
Non-communicable diseases	1.99	1.96-2.02	1.80	1.75-1.84
Injuries/external	1.26	1.17-1.35	1.24	1.13-1.35
Second-level outcomes				
Neoplasms	3.74	3.64-3.84	3.48	3.36-3.60
Cardiovascular/Circulatory	1.36	1.33-1.39	1.30	1.24-1.35
Chronic respiratory diseases	1.46	1.40-1.53	1.45	1.38-1.54
Liver cirrhosis	4.10	3.36-5.01	4.52	3.58-5.71
Digestive other than cirrhosis	1.88	1.78-2.00	1.79	1.64-1.95
Neurological	1.16	1.08-1.24	0.93	0.66-1.32
Mental and behavioural	1.07	1.00-1.15	1.01	0.93-1.09
Diabetes, urogenital, blood and endocrine	1.60	1.50-1.71	1.45	1.32-1.60
Musculoskeletal	1.56	1.35-1.80	1.32	1.11-1.58
Individual causes that been associated with PPIs				
Pneumonia	1.44	1.37-1.53	1.35	1.27-1.43
Acute kidney injury	1.56	1.10-2.21	1.59	1.03-2.46
C. difficile enterocolitis	2.16	1.71-2.74	1.94	1.47-2.56
Atrial fibrillation/flutter	1.54	1.32-1.79	1.52	1.29-1.79
Heart failure	1.87	1.67-2.10	1.61	1.42-1.83
Aortic aneurysm	1.06	0.97-1.16	1.05	0.94-1.18
Dementia and Alzheimer's	1.03	0.96-1.09	0.98	0.92-1.05
Chronic kidney disease	2.11	1.64-2.71	1.40	0.96-2.04
Hypertensive heart disease	1.02	0.87-1.19	1.14	0.95-1.39
Ischaemic heart disease	1.35	1.31-1.39	1.24	1.15-1.34
Lung cancer	3.16	2.98-3.34	3.31	3.07-3.56
Mesothelioma	3.62	2.91-4.50	3.25	2.43-4.36
Breast cancer	2.80	2.57-3.04	2.91	2.55-3.32
Liver cancer	4.72	3.83-5.82	3.25	2.58-4.11
Prostate cancer	3.14	2.84-3.48	2.81	2.48-3.18
Gastric cancer	14.59	11.16-19.08	12.71	10.05-16.08
Alcoholic liver disease	3.24	2.80-3.75	3.53	2.75-4.53
COPD	1.29	1.22-1.35	1.31	1.22-1.40
Control outcomes				
Accidental trauma excl. falls	1.28	1.15-1.43	1.29	1.13-1.47
Pulmonary embolism	1.43	1.23-1.67	1.45	1.20-1.76

Abbreviations: HR, hazard ratio; PS, propensity score; COPD, Chronic Obstructive Pulmonary Disease.

Chapter 8

Exploration of methods for incorporating test result information within the high-dimensional propensity score framework

John Tazare[1], Jeremy P Brown[1], Daniel R Morales[2], Liam Smeeth[1,3], Stephen JW Evans[1], Elizabeth Williamson[1,3], Ian J Douglas[1,3]

1. London School of Hygiene and Tropical Medicine, London, UK.
2. University of Dundee, Dundee, UK.
3. Health Data Research (HDR) UK, London, UK.

8.1 Overview

Summary

Chapter 7 introduced a study comparing the risk of all-cause and cause-specific mortality in users of proton pump inhibitors (PPI) and H2 receptor antagonists (H2RAS) in the UK Clinical Practice Research Datalink. Despite multiple comparisons and attempts to minimise confounding bias using the high-dimensional propensity score (HDPS), the results obtained suggested that confounding was not fully controlled by the approaches implemented. Furthermore, it was hypothesised that the residual confounding was likely to be driven by factors relating to the frailty and disease severity of PPI users but not captured by the specified covariates and HDPS data dimensions. In this chapter, I explore whether the use of laboratory test information can help to further mitigate confounding bias in this setting. I focus on the chronic-obstructive pulmonary disease mortality outcome and present initial proposals surrounding the semi-automated use of laboratory test information to improve confounding capture and control within the HDPS framework. I apply the HDPS modifications proposed in Chapter 3 before extending this framework to include information relating to requested laboratory tests and continuous test result values. The work was presented as an oral presentation at the *ICPE All Access 2020* online conference.

Thesis objectives addressed

This chapter addresses the following objectives of the overall thesis (Section 1.3):

3. Apply the HDPS and proposed modifications in the context of UK EHRs.

6. Investigate extensions to the HDPS framework that allow for the incorporation of laboratory test information.

Role of candidate

I planned and conducted the statistical analysis and wrote the initial chapter draft. Daniel Morales provided clinical input surrounding the cleaning of laboratory blood test results and selection of suitable 'normal' therapeutic ranges. I transcoded the test information data cleaning from SPPS to Stata and applied the data preparation steps to the PPI-Mortality study presented in Chapter 7. All authors were involved in the study design, interpreted the results and contributed to revisions of this chapter.

8.2 Introduction

The high dimensional propensity score (HDPS) procedure is guided by a set of underlying principles that create a scalable framework for data-driven confounder generation and selection in large healthcare databases (*Schneeweiss*, 2018; *Schneeweiss et al.*, 2009; *Tazare et al.*, 2020). Whilst data-driven covariates are usually derived from the presence of codes recorded in the healthcare database, this does not preclude the incorporation of other types of data in the HDPS procedure (*Schneeweiss*, 2018). For example, *Rassen et al.* (2013) investigated the use of free-text information within the HDPS framework.

Laboratory test results are commonly available in electronic health records (EHRs) and, in comparison to administrative claims databases, can often be available for a large proportion of patients (*Schneeweiss and Avorn*, 2005). Despite increases in the availability of linked laboratory test result data in healthcare databases (*Platt et al.*, 2012), they are rarely considered for confounder adjustment; often amid concerns surrounding the completeness and quality of data available (*Schneeweiss et al.*, 2012). Furthermore, since the HDPS was not originally developed to handle continuous values, it is unclear how this framework extends to incorporate these data (*Schneeweiss et al.*, 2009).

In this chapter, a cohort study from the UK Clinical Practice Research Datalink, investigating the association between proton pump inhibitor (PPI) use and both all-cause and cause-specific mortality (presented in Chapter 7) is used to propose and illustrate methods for incorporating laboratory test result information when applying HDPS approaches in EHRs.

8.3 PPI-Mortality study

In this section, I summarise relevant information from the study by *Brown et al.* (2021), including the use of the chronic obstructive pulmonary disease (COPD) mortality outcome and details surrounding the original HDPS analysis (which did not apply the modifications proposed in Chapter 3). Furthermore, I re-analyse the COPD-specific

mortality outcome using the modifications to the HDPS presented in Chapter 3 (*Tazare et al.*, 2020).

8.3.1 Data summary

Throughout this chapter, I focus on the primary analysis comparing PPI users to H2 recepter antagonist (H2RA) users. Findings from the study highlighted a higher prevalence of comorbidities and indicators of frailty in the PPI users compared to groups of non-users and H2RA users. *Brown et al.* (2021) concluded that it was not clear whether baseline differences between the comparator groups were fully accounted for by the measured covariates and suggested residual confounding remained a concern despite adjustment for investigator-specified and HDPS-derived covariates. Since the HDPS is constrained by the availability and completeness of data on key confounders (or proxies thereof) to successfully mitigate confounding bias, this motivates investigation of the ability of information relating to laboratory tests to further improve the capture and control of confounding in this setting.

To illustrate the proposed methods in this chapter, I focus on the COPD mortality outcome where a causal association was considered unexpected based on disease pathogenesis and evidence from trials (*Brown et al.*, 2021; *Moayyedi et al.*, 2019).

8.3.2 Results summary

All analyses estimated hazard ratios (HR) using average treatment effect in the treated (ATT) weighted Cox regression models (*Brown et al.*, 2021). The primary analysis estimated propensity scores using logistic regression based on 35 investigator covariates. A secondary analysis used the HDPS to adjust for an additional 500 covariates, as ranked by the Bross formula. For the HDPS analysis, we defined clinical, referral and prescription dimensions 1 year prior to cohort entry to identify relevant codes. The prescription dimension captured British National Formularly codes and the other dimensions captured Read codes truncated to the first 3-digits. Additionally, we applied

the prevalence filter and selected the top 200 most prevalent codes from each dimension (Step 2 of the HDPS procedure, see Chapters 2 and 3 for a summary of the steps of the HDPS). Note that this study was planned and conducted concurrently with the development of the work presented in Chapter 3. Therefore, the proposed modifications to the HDPS were not originally considered for this study.

The cohort identified 733,885 new users of PPIs and 124,410 new users of H2RAS. Furthermore, 7,846 (1.1%) PPI users and 538 (0.4%) of H2RA users had COPD-specific mortality. The association between PPI prescription and COPD-mortality among PPI and H2RA users without weighting was 1.83 (95% CI: 1.65-2.03). After adjustment for the investigator covariates, the HR for COPD-mortality reduced to 1.37 (95% CI: 1.14-1.66). Further adjustment for the HDPS covariates gave similar results (HR 1.34; 95% CI: 1.11-1.61).

8.3.3 Re-analysis using HDPS modifications

I re-analysed the COPD-specific mortality outcome using the HDPS modifications described in Chapter 3 (*Tazare et al.*, 2020): mapping the clinical and referral dimensions to ICD-10 and extending the frequency assessment to incorporate information recorded in patient's entire medical history. All other HDPS tuning parameters from the original primary analysis remained the same (*Brown et al.*, 2021): the top 200 most prevalent codes from each dimension were selected, ranking was performed using the Bross formula and the top 500 covariates were selected. A summary of the results are presented in Table 8.1.

In the re-analysis, results were similar to those obtained by the investigator and original HDPS models (HR 1.36; 95% CI: 1.14 - 1.64). Figure 8.1 shows the similarities in the PS distribution between the investigator and modified-HDPS models. Furthermore, in a sensitivity analysis, these results were robust to the number of covariates selected (Table 8.1).

Given similarities in the effect estimates and PS distributions under the investigator and

modified-HDPS models, it appears that the HDPS does not meaningfully contribute additional confounder information when incorporating data dimensions capturing clinical diagnoses, referrals and prescriptions. In subsequent sections, we will focus on the ability of test result information to contribute additional confounder information. Therefore, comparisons will focus on the primary modified-HDPS model (adjusting for the top 500 covariates) as a benchmark.

Table 8.1: *Association between PPI prescription and COPD-mortality among PPI and H2RA users applying the HDPS modifications proposed in Chapter 3.* ***Abbreviations:*** *HDPS; high-dimensional propensity score, HR; hazard ratio, CI; confidence interval*

Weighting	Number of variables included	HR (95% CI)
Unweighted	-	1.83 (1.65 - 2.03)
Investigator	35	1.37 (1.14 - 1.66)
Original HDPS (*Brown et al.*, 2021)	+500	1.34 (1.11 - 1.61)
Modified HDPS	+100	1.37 (1.14 - 1.65)
	+250	1.35 (1.13 - 1.63)
	+500	1.36 (1.14 - 1.64)
	+750	1.39 (1.16 - 1.67)

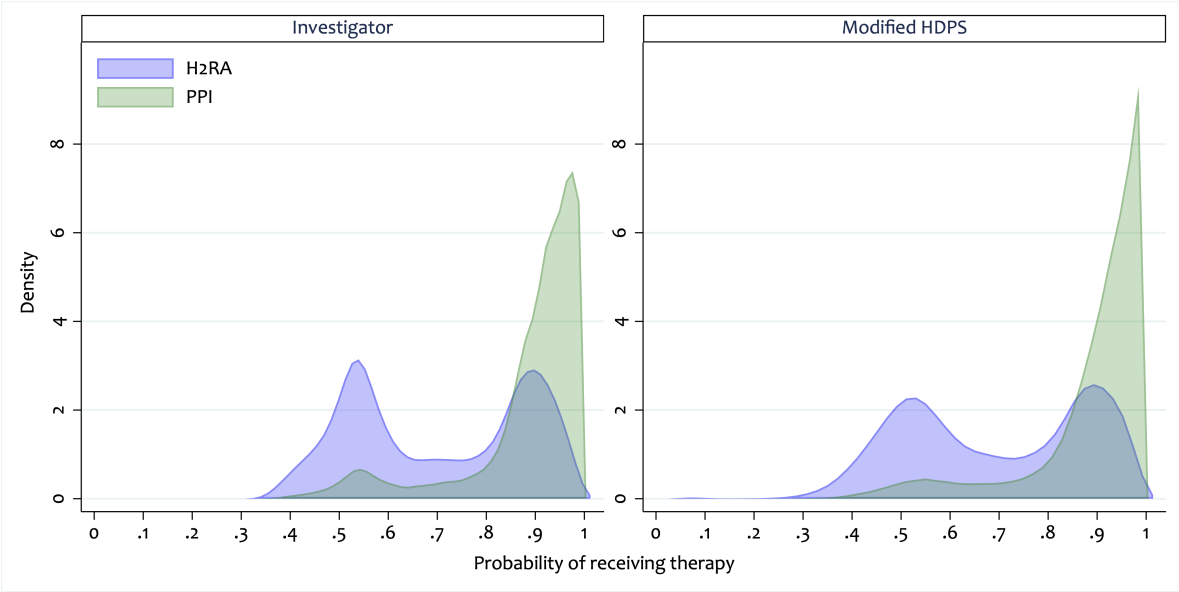


Figure 8.1: *Overlap plot comparing the propensity score distributions including 35 investigator pre-exposure covariates and additionally including the 500 from the modified-HDPS model.* ***Abbreviations:*** *H2RA, H2-receptor antagonist; PPI proton pump inhibitor.*

8.4 Types of test result information

In this section we consider the types of test result result data available in the CPRD, focusing on data that is potentially relevant for inclusion within the HDPS framework.

8.4.1 Overview in UK EHRs

Test information recorded in UK primary care is available within the CPRD GOLD and Aurum databases (*Herrett et al.*, 2015; *Wolf et al.*, 2019).

The test information available across these databases can be crudely grouped as laboratory, imaging, or other. The other category includes tests such as spirometry tests, cervical smears and colonoscopies (*O'Sullivan et al.*, 2018). Since 2000 there has been a substantial increase in the number of tests recorded even after adjustment for population growth (*O'Sullivan et al.*, 2018). This is likely due to many reasons, including services being diverted from secondary to primary care and the increased incentivisation of monitoring of chronic diseases across this period (through the introduction of the quality and outcomes framework, see *Lester* (2008) for an overview) (*O'Sullivan et al.*, 2018). Whilst the rates of tests have increased, the size of this increase may be slightly exaggerated due to greater integration of electronic blood test requesting systems within UK primary care EHRs (*Morales*, 2018a,b).

Given the variety of test information available, tests have a wide range of uses in studies using the CPRD. One common use is to validate specific concepts, for example, diabetes mellitus (*Mathur et al.*, 2020), chronic kidney disease (*Iwagami et al.*, 2017) and COPD (*Rothnie et al.*, 2017). Furthermore, work assessing the quality of blood and spirometry tests suggests that these data are of a high quality in UK primary care (*Rothnie et al.*, 2017; *Virdee et al.*, 2020).

In the context of confounder adjustment, data relating to testing are rarely directly adjusted for in studies using UK EHRs, often amid concerns of missing data (*Petersen et al.*, 2019). However, as highlighted above, they may be indirectly used to define a

specific concept considered for adjustment. We now consider several aspects of testing data that may be useful for confounder adjustment. In particular, we focus on information relating to whether a test was requested (yes or no) and the actual continuous test result value.

8.4.2 Test requested

The first type of test data indicates whether a test was requested and conducted in primary care. Similarly to other information used within the HDPS framework, this relates to capturing the presence of a code; in this case, relating to a specific test.

The reason for a test being requested is likely to be related to a number of factors and these will vary depending on the specific test. More generally, tests being requested signal increased contact and engagement with the healthcare system which is an important marker of healthcare utilisation and potentially reflective of underlying health status. Furthermore, certain tests have a specific indication and will relate to either the monitoring of a current diagnosis or the potential discovery of a new diagnosis. For example, testing the level of creatinine in the blood might be requested due to a potential or confirmed diagnosis relating to decreased kidney function. Conversely, blood pressure (BP) will often be routinely tested but not necessarily for a specific indication.

8.4.3 Continuous test results

The second type of test data considered in this chapter are continuous test result values. Whilst test result values are likely to more directly signal the underlying health status of an individual, the use of these data can be complex. Two issues with important consequences for the analysis of these data are data cleaning and missing data:

- **Data cleaning:** In the CPRD (and in EHR data more widely), these data are not automatically checked for implausible or impossible values. Furthermore, a given test is often recorded using a range of measurement units. This necessitates

an initial data cleaning step to remove unlikely values and harmonise the units before the test results can be incorporated into an analysis (*Virdee et al.*, 2020).

- **Missing data:** We often define a variable based on whether a patient has evidence of a condition (i.e. presence of a code, 'yes' or 'no'). Despite the possibility of misclassification, this allows the investigator to define the variable for the whole study population (*Farmer et al.*, 2018). In the case of continuous test results, data are typically missing when the test has not been requested, however, this raises important questions surrounding the missing data mechanism (*Farmer et al.*, 2018; *Schneeweiss et al.*, 2012). Therefore, in response to these missing data, it important to consider possible missing data mechanisms and the potential impact on an analysis (*Carpenter and Kenward*, 2013; *Sterne et al.*, 2009a). For example, if analysis is conducted only on the subset of individuals with complete data for a set of continuous test result values, the sample size might be dramatically reduced.

.

The HDPS procedure does not readily support the inclusion of continuous variables (*Schneeweiss et al.*, 2009). In the subsequent sections we describe simple methods for incorporating these data within the HDPS framework. Finally, in this pilot work we focus only on continuous blood test results, which are likely to be some of the most prevalent test results recorded for our cohort. This allows us to draw on specific clinical and operational knowledge available in our research team (*Morales*, 2018b).

8.5 Data analysis

In this section, we describe the data preparation and analysis steps for incorporating information relating to tests within the HDPS framework. As per the original study, all analyses estimate HRs using ATT weighted Cox regression models (*Brown et al.*, 2021).

Analyses were conducted using Stata 15 (*StataCorp*, 2017).

8.5.1 Tests requested

Proposed methods

The simplest method for incorporating test result information in the HDPS framework is to focus on tests requested since these data are comparable to data commonly included in the HDPS procedure (i.e. they focus on the presence of a code in a patient's medical history).

General practitioner experience within our research team suggested that recording of tests requested was likely to be complete and most relevant in the period directly preceding cohort entry. Therefore, we propose defining an additional data dimension identifying all tests requested (referred to as the test-requested dimension). This dimension captures a combination of 1) concern about possible illnesses that would lead to an abnormality in the test parameter and 2) genuine illness that would be shown through the test parameter. Similar to the other dimensions, the covariate assessment window will be defined prior to cohort entry. Since the frequency of tests is likely to be complete, we propose assessing frequency of code recurrence using the original cut-offs for generating binary HDPS covariates (*Schneeweiss et al.*, 2009):

- Once: Test is requested $\geq$ once.
- Sporadic: Test is requested $\geq$ the median
- Frequent: Test is requested $\geq$ the 75th percentile

Finally, in CPRD GOLD, laboratory test results are coded using the Read coding system. In this context, Read codes successfully capture distinct tests and therefore mapping or truncation of codes was not required.

Analysis

All tests requested in the 1 year prior to cohort entry were included as an additional data dimension alongside the existing clinical, referral and prescription dimensions.

Covariates from all dimensions were ranked by the Bross formula (*Schneeweiss et al.*, 2009; *Wyss et al.*, 2018a) and the top 500 covariates were used to augment the set of investigator covariates.

In a sensitivity analysis, we varied the number of covariates selected (100, 250 and 750).

8.5.2 Cleaning of continuous blood test results

Before incorporating the continuous blood test result values it was necessary to clean these data to remove implausible values and harmonise the units of measurement. We transcoded blood test extraction and cleaning rules (previously developed by *Morales* (2018b) using clinical knowledge) from SPSS (a statistical software package developed by International Business Machines (IBM)) to Stata 15.

We identified all test values for 35 blood tests (selected based on being the most prevalent blood tests) in a baseline covariate assessment window prior to cohort entry. For a given individual and blood test, we then selected the cleaned blood test result value closest to cohort entry since this was likely to be most relevant to the decision to initiate treatment and most reflective of current health status. Finally, for duplicated cleaned blood test results (i.e. the presence of more than one blood test result for a specific test on the same date) we selected the lowest test result. Further work could investigate the sensitivity of results to this decision. For example, by instead selecting the test result closest to the therapeutic normal range.

We defined two sets of test result values, defined 1 year and 2 years prior to cohort entry. The rationale was to investigate potential gains in the proportion of patients with a test result. For example, extending the covariate assessment window to 2 years allows us to account for any annual check-ups not strictly 1 year prior to cohort entry. Whilst there is likely a trade-off between relevance of the test values obtained and ensuring a high proportion of patients have a usable test result value, in many cases older test results are likely to still be relevant. For example, if a patient had a HbA1c value within the normal therapeutic range recorded in the last few years, their HbA1c level today is

likely to be similar.

8.5.3 Cut-offs

Proposed methods

One approach for incorporating continuous test result values is to generate binary covariates based on whether an individual's most recent test result falls within a therapeutic range. This approach, based on cut-offs, provides a simple method for incorporating continuous test result values and is conceptually consistent with the HDPS approach more generally.

In the context of blood tests, therapeutic ranges are either bounded in one direction or can plausibly take values either side of the specified interval.

- **Bounded in one direction:** When the therapeutic range is bounded, we propose generating a single binary covariate representing:

$$\text{Normal} = \begin{cases} 1 & \text{Test result value is within therapeutic range} \\ 0 & \text{otherwise} \end{cases}$$

- **Plausible values either side of interval:** When the therapeutic range is not bounded in one direction and the plausible values can lie either side, we propose generating three binary covariates representing:.

$$\text{Low} = \begin{cases} 1 & \text{Test result value is below the therapeutic range} \\ 0 & \text{otherwise} \end{cases}$$

$$\text{Normal} = \begin{cases} 1 & \text{Test result value is within therapeutic range} \\ 0 & \text{otherwise} \end{cases}$$

$$\text{High} = \begin{cases} 1 & \text{Test result value is above the therapeutic range} \\ 0 & \text{otherwise} \end{cases}$$

As with binary covariates in the HDPS paradigm more generally, we propose prioritising and selecting these covariates using usual standard HDPS methods (e.g. via the Bross formula (*Schneeweiss et al.*, 2009)). However, the proposed cut-off method differs from traditional HDPS covariates since a patient can only be represented by one of the variables (since we only focus on the most recent test result value). In comparison, under the original HDPS framework, a patient could have 'Yes' for each of the 'Once', 'Sporadic' and 'Frequent' variables.

Since missing data are coded as '0', the selection of all three covariates for a specific test is equivalent to a four level categorical variable (i.e. with categories 'Low', 'Normal', 'High' and 'Missing'). This approach is widely used for incorporating continuous values in the context of UK EHR analyses, e.g. a common example is categorised body mass index (*Blake et al.*, 2020). In the proposed approach, instead of adjusting for all categories automatically, we are allowing for the inclusion of one or many of the indicator variables included when modelling a categorical variable. For example, if the 'High' HbA1c covariate was selected, we would not include the 'Low' and 'Normal' categories, unless they were independently selected. This might be a reasonable approach in the context of a specific test since certain test result values are likely to have varying clinical importance. For example, high blood pressure is clinically more important to capture than blood pressure in the normal therapeutic range. However, in the context of glucose blood tests both high and low results are clinically important.

Whilst cut-offs provide a simple solution, the dichotomisation of continuous variables reduces statistical power and can result in residual confounding (*Brenner and Blettner*, 1997; *Groenwold et al.*, 2013; *Royston et al.*, 2006; *Steryerberg*, 2009).

The proposed cut-offs for the available blood tests are summarised in Table 8.4. These were specified based on reference ranges published by NHS Scotland and additionally checked by Daniel Morales (*NHS*, 2017). Note that some ranges are stratified by age or sex.

Table 8.2: *Proposed cut-offs for generating binary test result HDPS covariates.* ***Abbreviations:*** *HDPS; high-dimensional propensity score.*

Test	Normal Range Definition	Number of binary covariates
	Blood Pressure (BP)	
Systolic Blood Pressure	Age $<$ 80: $<$140 mmHg	1
	Age $\geq$ 80: $<$150 mmHg	1
Diastolic Blood Pressure	$<$90 mmHg	1
	Calcium	
Calcium	2.15 - 2.65 mmol/L	3
Calcium Adjusted	2.20 - 2.60 mmol/L	3
	Full blood count (FBC)	
Basophil	0 - 0.1 x10^9/L	1
Eosinophil	0 - 0.4 x10^9/L	1
Haemoglobin	Men: 135 - 180 g/L	3
	Women: 115 - 160 g/L	3
Lymphocyte	1.0 - 4.8 x10^9/L	3
Monocytes	0.2 - 0.8 x10^9/L	3
MCV	80 - 100 fL	3
MCH	27 - 34 pg	3
Platelets	130 - 400 K/μL	3
RBC	Men: 4.5 - 6.0	3
	Women: 4.0 - 5.6	3
WBC	4 - 11 x10^9/L	3
	Glucose	
Glucose	3.3 - 6.1 mmol/L	3
Glucose Fasting	3.3 - 6.1 mmol/L	3
	Lipids	
Cholesterol	$\leq$ 5 mmol/L	1

Continued on next page

Test	Normal Range Definition	Number of binary covariates
HDL Cholesterol	$\geq$ 1 mmol/L	1
LDL Cholesterol	$\leq$ 3 mmol/L	1
Triglycerides	0.85 - 2.0 mmol/L	3
	Liver Function Tests (LFTs)	
Albumin	38 - 50 g/L	3
AST	15 - 42 IU/L	3
ALT	0 - 40 U/L	1
ALP	Age 17-60: 35 - 115 U/L	3
	Age $\geq$ 60: 35 - 150 U/L	3
AKP	Age 17-60: 35 - 115 U/L	3
	Age $\geq$ 60: 35 - 150 U/L	3
Bilirubin	2 - 20 umol/L	3
	Urea & electrolytes	
Creatinine	Age $<$ 60: 20 - 120 μmol/L	3
	Age $\geq$ 60: 70 - 140 μmol/L	3
Potassium	3.6 - 5.4 mmol/L	3
Sodium	134 - 144 mmol/L	3
Urea	Age $<$ 60: 3.0 - 8.5 mmol/L	3
	Age $\geq$ 60: 3.0 - 10.0 mmol/L	3
GFR	$\geq$60 mls/min	2
	Other	
C-reactive Protein Test	$<$ 5 mg/L	2
HbA1c	$<$ 60 mmol/mol	2
Total Protein	60 - 80 g/L	3
Urate	Men: 0.12 - 0.42 mmol/L	3
	Women 0.12 - 0.38 mmol/L	3

Notes: Glomerular filtration rate (GFR), Alkaline Phosphatase (AKP), Alkaline

Continued on next page

Test	Normal Range Definition	Number of binary covariates
Phosphatase Level (ALP), Alanine Aminotransferase (ALT), Aspartate Aminotransferase (AST), Mean Corpsucular Volume (MCV), Mean Corpsucular Haemoglobin (MCH), Red Blood Count (RBC), White Blood Count (WBC)		

Analysis

The resulting pool of binary cut-off covariates was added as a fifth data dimension (referred to as the cut-offs dimension). These covariates were considered by the HDPS procedure alongside those derived from other data dimensions, covariates were prioritised by the Bross formula and the top 500 were selected for inclusion alongside the investigator covariates.

In sensitivity analyses, we varied the number of covariates selected (100, 250 and 750) and investigated extending the baseline covariate assessment window for the cut-offs dimension to 2 years.

8.5.4 Continuous modelling

Proposed methods

Whilst the cut-offs approach provides a simple solution, given the limitations discussed, we propose the following to minimise information loss and make better use of these continuous data.

For each of the test cut-off variables selected in the top 100 HDPS covariates, we propose additionally including the corresponding continuous test variable (in a linear form) and a missing indicator in the HDPS model. For continuous variables, the missing indicator approach involves setting missing values to a fixed value, for example 0, and including

both the variable and its missing indicator in the HDPS model (*Blake et al.*, 2020; *Groenwold et al.*, 2012).

One concern surrounding the use of continuous values is missing data. Even in this example, where patients are expected to have a high level of engagement with health services, we expect missing test result data for a potentially large proportion (often over 50%). Since we are attempting to incorporate many of these variables, a complete case approach (only including those with test results present for all included tests) would likely significantly reduce the study sample size (*Carpenter and Kenward*, 2013).

The proposed approach is based on recent work by *Blake et al.* (2020) showing that the missing indicator approach is unbiased under assumptions that can be summarised as follows. In this setting, the missing indicator method is appropriate if the continuous test value only contributes to the treatment decision if it is measured, i.e. the missing value did not inform the decision to initiate treatment. When this is true, or approximately true, the missing indicator approach gives unbiased estimates (*Blake et al.*, 2020).

Analysis

For each of the test cut-off variables selected in the top 100 HDPS covariates, we additionally incorporated the continuous test result variable (linearly) and a missing indicator in the HDPS model.

In sensitivity analyses, we varied the number of covariates selected (100, 250 and 750) and investigated extending the baseline covariate assessment window to 2 years to investigate potential gains in the proportion of patients with complete test data.

8.6 Results

This section presents the results of incorporating laboratory test information in the HDPS framework. Results for all the methods considered are summarised in Table 8.3

Table 8.3: *Comparison of methods for incorporating laboratory test information in the HDPS framework. All analyses contain the clinical, referral and prescription dimensions, plus the corresponding test-based information*

Number of HDPS covariates selected	+ Test-requested Dimension	+ Test-requested & Cut-off Dimensions (1-year)	+ Test-requested & Cut-off Dimensions (2-years)	+ Test-requested & Cut-off dimensions & Continuous test variables (1-year)	+ Test-requested & Cut-off dimensions & Continuous test variables (2-years)
100	1.36 (1.12 - 1.66)	1.36 (1.11 – 1.68)	1.34 (1.08 – 1.65)	1.35 (1.08 - 1.67)	1.36 (1.10 - 1.67)
250	1.34 (1.09 - 1.63)	1.30 (1.05 – 1.62)	1.30 (1.05 – 1.61)	1.32 (1.06 – 1.64)	1.32 (1.08 – 1.64)
500	1.25 (1.01 - 1.54)	1.24 (1.00 - 1.54)	1.25 (1.01 – 1.54)	1.24 (1.00 – 1.54)	1.26 (1.02 – 1.55)
750	1.25 (1.00 - 1.55)	1.21 (0.97 – 1.52)	1.22 (0.98 – 1.52)	1.22 (0.97 – 1.53)	1.23 (0.98 – 1.53)

Note: These dimensions are added to the benchmark HDPS model presented in Section 8.3.3

8.6.1 Tests requested

Initially, we incorporated test result information by adding an additional data dimension capturing tests requested in the year prior to cohort entry.

A description of the top 50 test requested and the respective Read codes is presented in Table 8.4. This table highlights that a high proportion of patients have test information recorded in the 1-year prior to cohort entry. Furthermore, Table 8.4 illustrates that Read codes accurately capture distinct tests requested, at least for the most prevalent tests.

In the primary analysis selecting the top 500 HDPS covariates, 38% of the covariates selected originated from the test-requested dimension. Furthermore, 53 out of the top 100 covariates were from this dimension.

Compared to the modified HDPS presented in Section 8.3.3 (HR 1.36; 95% CI: 1.14 - 1.64), incorporation of the test requested information appeared to further contribute to confounding control and obtained results closer to the expected null association (HR 1.25; 95% CI: 1.01 - 1.54). Figure 8.2 highlights similarity in the PS distributions compared to the model presented in Section 8.3.3. Figure 8.3 highlights that the covariates based on tests requested were typically more prevalent in PPI users compared to the H2RA users. However, these differences could be partly due to calendar differences in the start of therapy. Figure 8.4 highlights several covariates from the test-requested dimension with strong outcome associations and these all relate to respiratory tests, for example, forced expiratory volume tests.

Despite the larger pool of covariates, in sensitivity analyses, selecting 750 covariates did not appear to meaningfully alter the conclusions (Table 8.3).

Table 8.4: *Read codes for the top 50 tests requested in the PPI-Mortality cohort*

Read code	Description	Total*
423..00	Haemoglobin estimation	585,841
42P..00	Platelet count	582,052

Continued on next page

Read code	Description	Total*
42A..00	Mean corpuscular volume (MCV)	575,489
44J3.00	Serum creatinine	559,046
42H..00	Total white cell count	558,474
426..00	Red blood cell (RBC) count	555,576
42M..00	Lymphocyte count	553,596
42J..00	Neutrophil count	553,413
42N..00	Monocyte count	551,326
44I5.00	Serum sodium	550,262
44I4.00	Serum potassium	550,101
42K..00	Eosinophil count	549,551
428..00	Mean corpusc. haemoglobin(MCH)	546,025
42L..00	Basophil count	523,588
44M4.00	Serum albumin	508,079
44F..00	Serum alkaline phosphatase	502,823
424..00	Full blood count - FBC	467,489
429..00	Mean corpusc. Hb. conc. (MCHC)	441,299
44P..00	Serum cholesterol	433,481
4258.00	Haematocrit	432,365
44J9.00	Serum urea level	424,306
442W.00	Serum TSH level	387,705
44G3.00	ALT/SGPT serum level	369,175
44Q..00	Serum triglycerides	355,280
44P5.00	Serum HDL cholesterol level	355,196
44M3.00	Serum total protein	337,372
44D6.00	Liver function test	329,489
44EC.00	Serum total bilirubin level	327,003
42B6.00	Erythrocyte sedimentation rate	319,506
42Z7.00	Red blood cell distribution width	275,417
44JB.00	Urea and electrolytes	273,437

Continued on next page

Read code	Description	Total*
466..00	Urine test for glucose	269,767
467..00	Urine protein test	264,306
44g..00	Plasma glucose level	261,667
44P6.00	Serum LDL cholesterol level	252,907
44I8.00	Serum calcium	251,460
44O..00	Serum lipids	248,893
451E.00	GFR calculated abbreviated MDRD	241,175
44E..00	Serum bilirubin level	226,428
44IC.00	Corrected serum calcium level	220,891
44M5.00	Serum globulin	205,394
44G9.00	Serum gamma-glutamyl transferase level	183,860
442V.00	Serum free T4 level	177,675
44I9.00	Serum inorganic phosphate	174,968
535..00	Standard chest X-ray	156,372
44U..00	Blood glucose result	150,787
44PF.00	Total cholesterol:HDL ratio	148,586
442J.00	Thyroid function test	148,298
442A.00	TSH - thyroid stim. hormone	147,971
44CS.00	Serum C reactive protein level	138,351

* Total represents the number of patients with at least one test requested in the 1-year prior to cohort entry

8.6.2 Cleaning

To incorporate continuous blood test data it was first necessary to remove implausible values and harmonise the units of measurement.

The proportion of patients with an eligible continuous test result in the 1-year and 2-years prior to cohort entry is presented in Table 8.5. Furthermore, we highlight that extending the assessment period from 1-year to 2-years, typically increases the

proportion of patients with a valid test by approximately 10%. In the Supporting Information A, we provide the distributions before and after cleaning of the continuous test results when incorporating information in the 1-year prior to cohort entry. After the cleaning steps have been applied we observe sensible distributions of test result values for majority of blood tests included. However, we note that for some blood tests applying sensible cleaning rules can be difficult. For example, a patient might record a one-off and clinically valid high total protein value which makes identifying recording errors or implausibly large values for these tests particularly challenging.

8.6.3 Cut-offs

We first incorporated continuous blood test result data using cut-offs described in Section 8.5.3. These were included as an additional data dimensions alongside the clinical, referral, prescription and tests requested dimensions.

In the primary analysis selecting the top 500 HDPS covariates, 46% related to test-related covariates (35% from the test-requested dimension and 11% from the cut-offs dimension).

Compared to the HDPS model incorporating the tests requested dimension, additionally incorporating the blood test cut-offs obtained similar results (HR 1.24; 95% CI: 1.00 – 1.54). Figure 8.3 highlights that the covariates based on the cut-offs were typically more prevalent in PPI users compared to the H2RA users. Furthermore, Figure 8.4 highlights two covariates from the cut-offs dimension with strong outcome associations and these relate to patients with normal recordings for glomerular filtration rate and high-density lipoproteins. In sensitivity analyses, results were similar when defining the cut-offs in a 2-year covariate assessment period. Furthermore, the pattern of results from increasing adjustment was similar compared to the model incorporating tests-requested information Table 8.3.

8.6.4 Continuous modelling

Finally, for the blood test cut-offs selected in the top 100 HDPS covariates, we additionally incorporated the continuous variable and a missing indicator.

This resulted in the inclusion of 19 continuous blood test variables (listed in the Supporting Information B).

For the primary analysis selecting the top 500 HDPS covariates, we obtained similar results compared to the cut-offs analysis (HR 1.24; 95% CI: 1.00 - 1.54). The pattern of results obtained when varying the number of HDPS covariates selected was also similar between the two approaches. Finally, Figure 8.2 highlights similarity in the PS distributions compared to the other models incorporating test information.

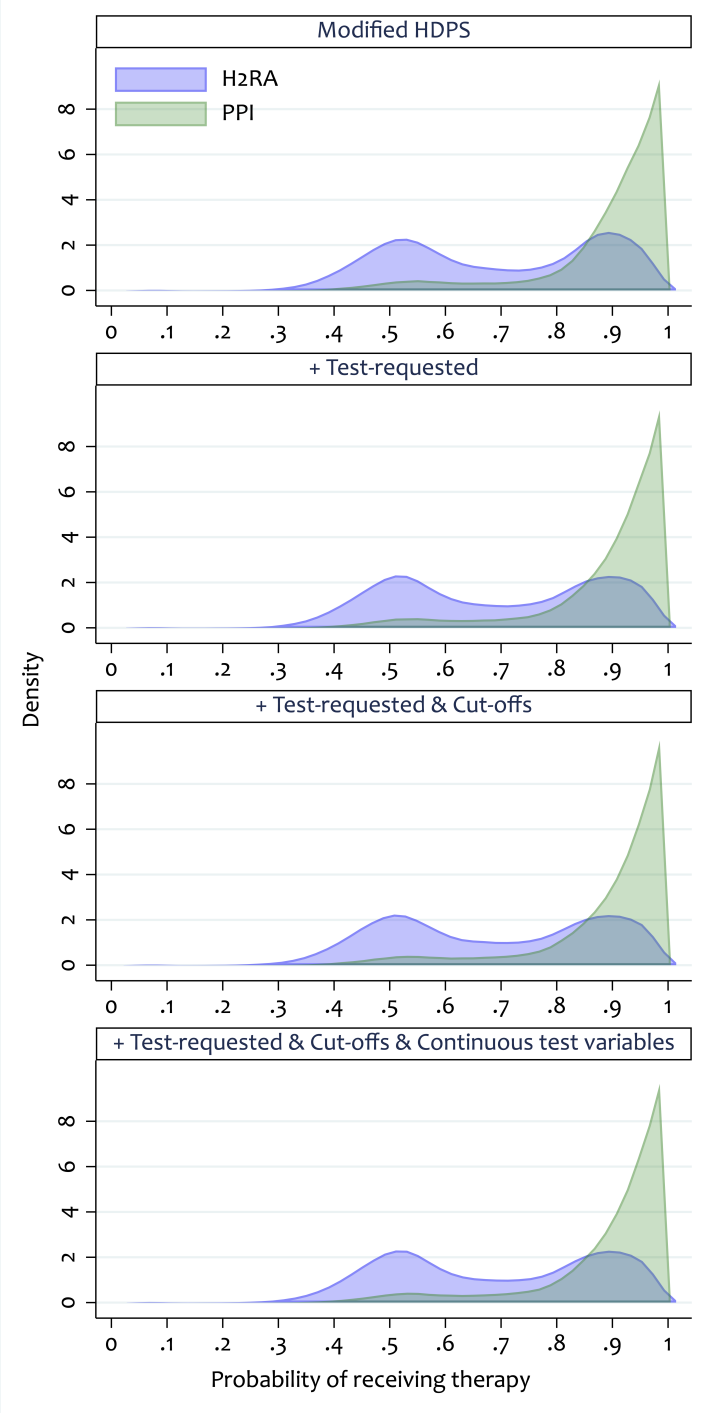


Figure 8.2: *Overlap plot comparing the propensity score distributions between the modified-HDPS model (containing clinical, referral and prescription dimensions) and HDPS models additionally incorporating test result information.* ***Abbreviations:*** *H2RA, H2-receptor antagonist; PPI proton pump inhibitor.*

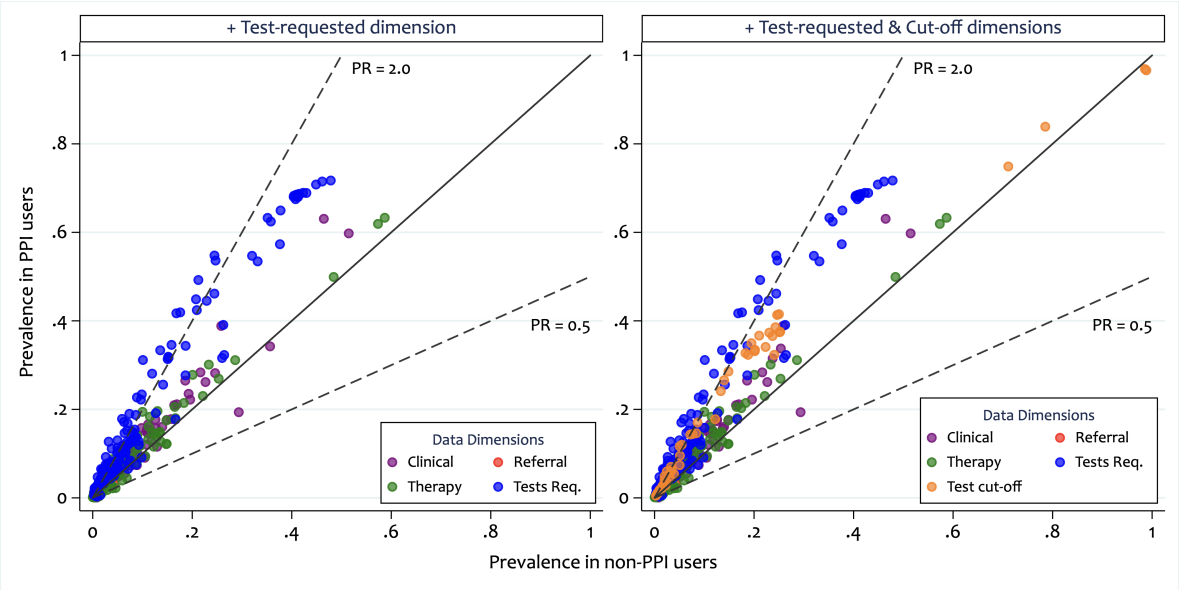


Figure 8.3: *Prevalence of the top 500 Bross-prioritised HDPS covariates by treatment group. All HDPS models contain the clinical, referral and prescription dimensions, plus the corresponding test-based dimensions (with test data assessed in a 1-year pre-exposure covariate window).* ***Abbreviations:*** *PR, prevalence ratio.*

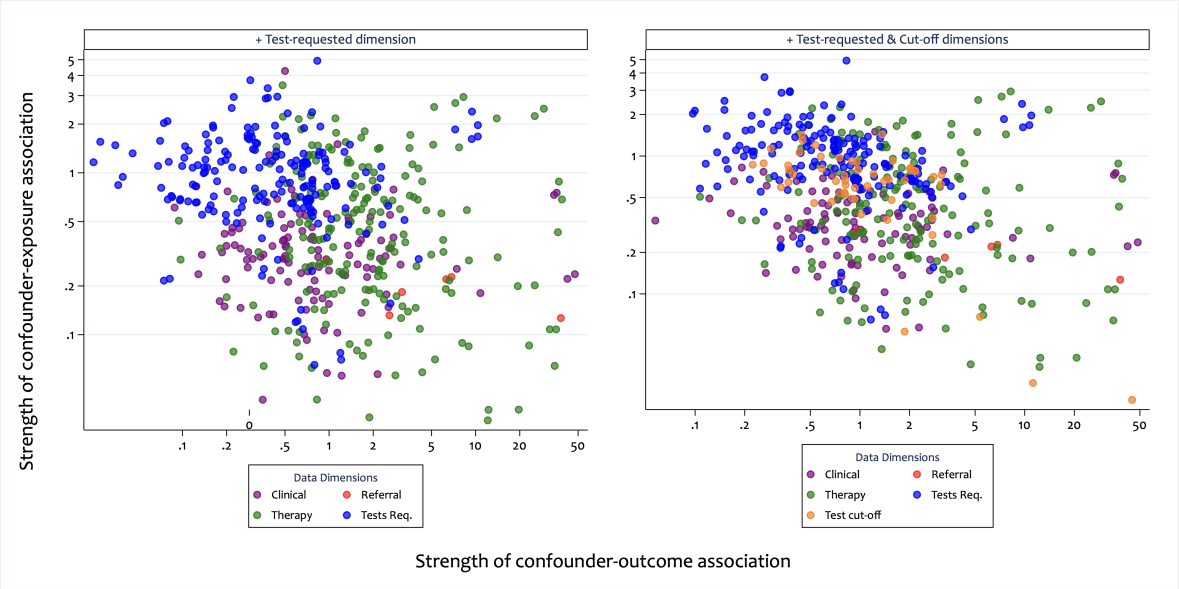


Figure 8.4: *Comparison of the covariate-exposure and covariate-outcome associations for the top 500 bias-based HDPS covariates. All HDPS models contain the clinical, referral and prescription dimensions, plus the corresponding test-based dimensions (with test data assessed in a 1-year pre-exposure covariate window).*

Table 8.5: *Summary of the 35 clean blood test results highlighting the proportion of PPI and H2RA users with a cleaned and eligible continuous value*

Test Result	Cleaning rules applied	Percentage with result within 1 year of cohort entry (%)	Percentage with result within 2 years of cohort entry (%)	Percentage increase (%)
AKP	Yes	36	47	11
ALP	No	31	40	9
ALT	Yes	31	39	8
AST	Yes	8	11	3
Albumin	Yes	37	47	10
Basophil	Yes	22	30	8
Bilirubin	Yes	37	47	10
CRP	Yes	11	15	4
Calcium	Yes	14	30	16
Calcium Adjusted	Yes	13	17	4
Cholesterol	Yes	29	39	10
Creatinine	Yes	43	53	10
Eosinophil	Yes	36	47	11
GFR	Yes	19	24	5
Glucose	Yes	22	32	10
Glucose Fasting	Yes	8	13	5
HDL	Yes	23	32	9
Haemoglobin	Yes	40	52	12
HbA1c	Yes	8	10	2
LDL	Yes	19	26	7
Lymphocyte	Yes	37	48	11
MCH	No	36	47	11
MCV	Yes	39	51	12
Monocytes	Yes	37	48	11
Platelets	Yes	39	51	12
Potassium	Yes	42	52	10
RBC	Yes	1	1	0
Sodium	Yes	42	53	11
Total Protein	No	24	31	7
Triglycerides	Yes	22	30	8
Urate	Yes	1	2	1
Urea	Yes	34	43	9
WBC	Yes	38	50	12

8.7 Discussion

In this chapter, we investigated the incorporation of laboratory test information to improve confounding control within the developed HDPS framework. We used the study introduced in Chapter 7 to illustrate the proposed methods and focussed on the COPD-specific mortality outcome. Compared to a HDPS model incorporating clinical, referral and prescription information (HR 1.36; 95%: CI 1.14 - 1.64), the final model incorporating test information (relating to tests requested, test cut-offs and continuous test variables) appeared to improve confounder adjustment and obtained results closer to the expected null association (HR 1.24; 95% CI: 1.00 - 1.54). Furthermore, 46% of the top 500 covariates in the final model incorporating test information were derived from test-related data dimensions, highlighting the potential importance of these data for mitigating confounding bias in UK EHRs. Despite this, in this example, residual confounding is likely to remain surrounding unmeasured factors relating to the frailty and disease severity of PPI users (*Brown et al.*, 2021).

The work presented contributes to evidence surrounding the incorporation of test results for mitigating confounding bias in pharmacoepidemiological studies (*Schneeweiss et al.*, 2012). Furthermore, we have demonstrated how relatively simple methods, such as therapeutically-led cut-offs, can be used to incorporate continuous laboratory test data within the HDPS framework in a way that is consistent with the semi-automated nature of this approach. Finally, we have highlighted how continuous variables can be incorporated in these models using a missing indicator approach valid under assumptions likely to be approximately true in UK EHRs.

This pilot work has highlighted several issues which could result in further exploration and expansion of the proposed methods. These are briefly outlined below.

Firstly, when incorporating cut-offs, we have focused on the inclusion of data from 35 blood tests. This had key advantages since we were able to benefit from clinical knowledge in the research team and develop a framework on a set of test results likely to be recorded for a relatively high proportion of patients in our cohort. In the future, the developed framework could be expanded to incorporate additional test data, for

example, spirometry results (*Rothnie et al.*, 2017).

Secondly, when using the HDPS, increasing the number of data dimensions leads to a larger pool of covariates for prioritisation and selection. Whilst, in this example, results were robust to the number of covariates selected, in other settings it might be necessary to adjust for more than the typical 500 HDPS covariates to optimise confounding capture and control (*Schneeweiss*, 2018).

Thirdly, when selecting the cut-off covariates, we only incorporated the categories selected by the Bross formula. For example, if the 'High' variable was selected the 'Low' or 'Normal' variables were not necessarily included (unless independently selected). Future work could explore the selection of all categories if one of the categories for a certain test is selected.

The final issues surround the inclusion of continuous test variables and the missing indicator approach. In the work presented, we have focussed on including continuous variables in a linear form in the PS model. Whilst this approach can lead to residual confounding, it is likely to perform well in most settings (*Groenwold et al.*, 2013). Future work could investigate the use of fractional polynomials and splines for more precise modelling of the relationship between treatment initiation and these continuous test variables (*Binder et al.*, 2013; *Steryerberg*, 2009). However, in feasibility work, the use of cubic splines in this example led to issues surrounding fitting of the PS models. Furthermore, these approaches are less automatable and therefore require more investigator input in the context of the HDPS.

Missing indicators were used to handle missing data arising from the incorporation of the continuous test results and work by *Blake et al.* (2020) suggests this approach is valid under a set of assumptions likely to hold in UK EHRs (discussed in relation to alternative methods in Section 6.3.6). Therefore, despite previous criticism of this method in the context of incorporating test data for confounding control (*Schneeweiss et al.*, 2012), it may prove a useful approach in this context. Whilst multiple imputation (MI) has gained popularity as a method for handling missing data (*Carpenter and Kenward*, 2013), it would add several complexities in this setting. For example, combining

MI and PS models is not straightforward (*Granger et al.*, 2019a; *Leyrat et al.*, 2019) and MI would also lead to specific issues surrounding increased computational burden and the need to specify imputation models (*Sterne et al.*, 2009b). Finally, despite a lot of methodological work in the context of HDPS (*Schneeweiss*, 2018), implications surrounding the use of missing data methods (e.g. MI or the missing indicator approach) when applying the HDPS have not been fully explored and this highlights an important area for future research.

Our study period covers the introduction of the Quality and Outcomes Framework (from 2000). Given the recording of certain incentivised information (including relating to laboratory test information), this framework might have resulted in changes to recording practice over the early period of this study. These changes may have impacted differently on PPI users and H2RA users. Future work could investigate the incorporation of test-related data in the HDPS using an example with less potential for temporal variation.

The focus of this work was to empirically investigate several methods for incorporating test results in the HDPS framework. This study illustrates the potential for laboratory test result information to improve the ability of HDPS approaches to reduce residual confounding in UK EHRs. In any application of the HDPS, the performance will be constrained by the data available within the defined data dimensions. Therefore, whilst the incorporation of laboratory test data will help to minimise residual confounding arising from these factors (and associated proxies), residual confounding may still remain from important unmeasured factors.

8.8 Ethics statement

This study was approved by the London School of Hygiene and Tropical Medicine Research Ethics Committee (reference no.15655) and by the CPRD Independent Scientific Advisory Committee (ISAC reference 17_252) (see Appendices E & F for details).

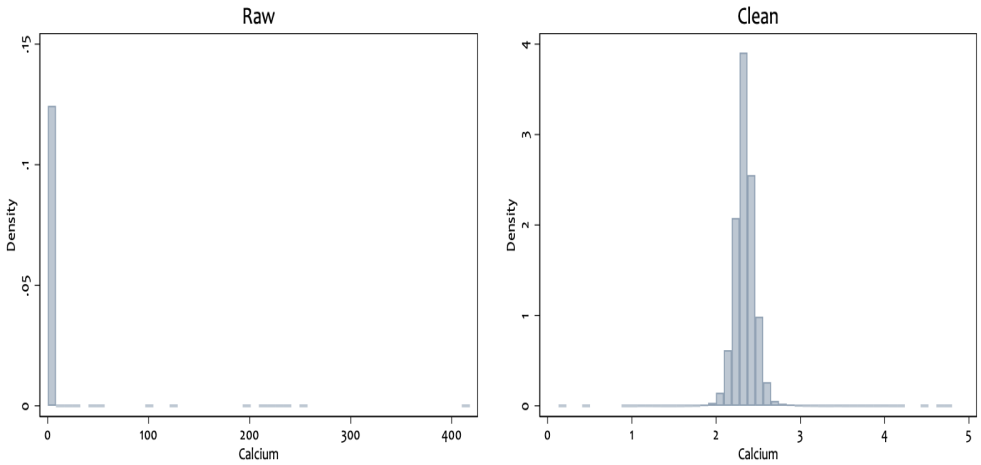
8.9 Supporting information

8.9.1 A: Cleaned test results

Test results in the year prior to index

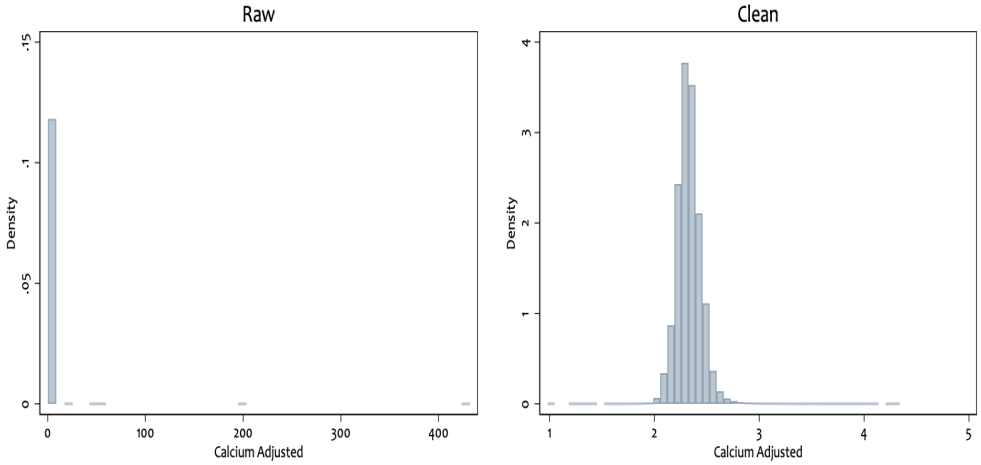
Calcium

After cleaning, 120244 (14.01%) of patients have a measurement of calcium in the 1 year prior to index



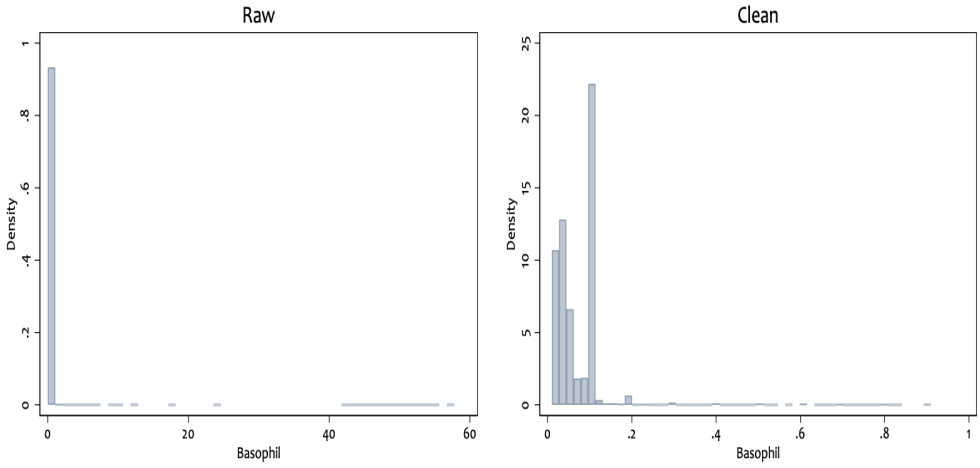
Calcium Adjusted

After cleaning, 107657 (13%) of patients have a measurement of calciumadjusted in the 1 year prior to index



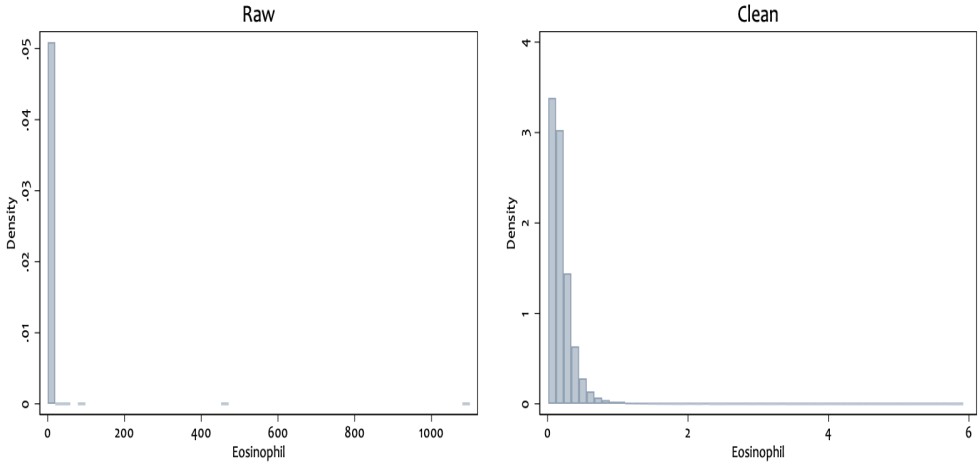
Basophil

After cleaning, 188444 (22%) of patients have a measurement of basophil in the 1 year prior to index



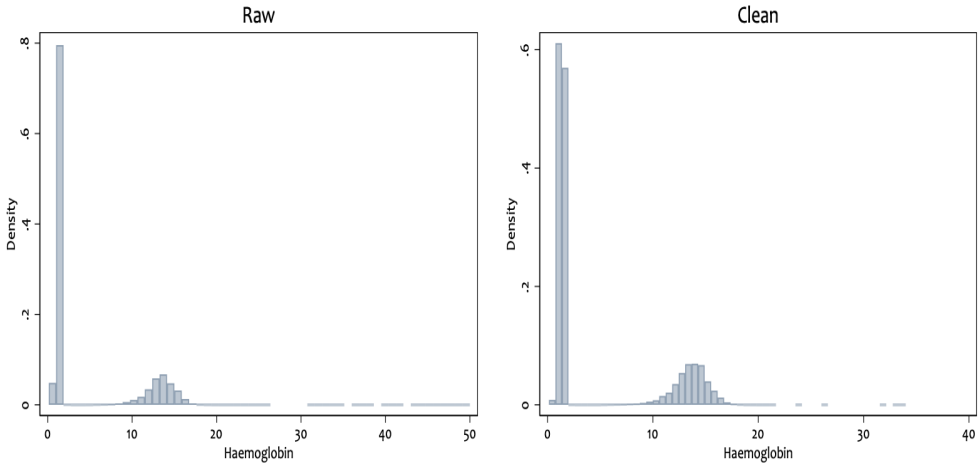
Eosinophil

After cleaning, 306517 (36%) of patients have a measurement of eosinophil in the 1 year prior to index



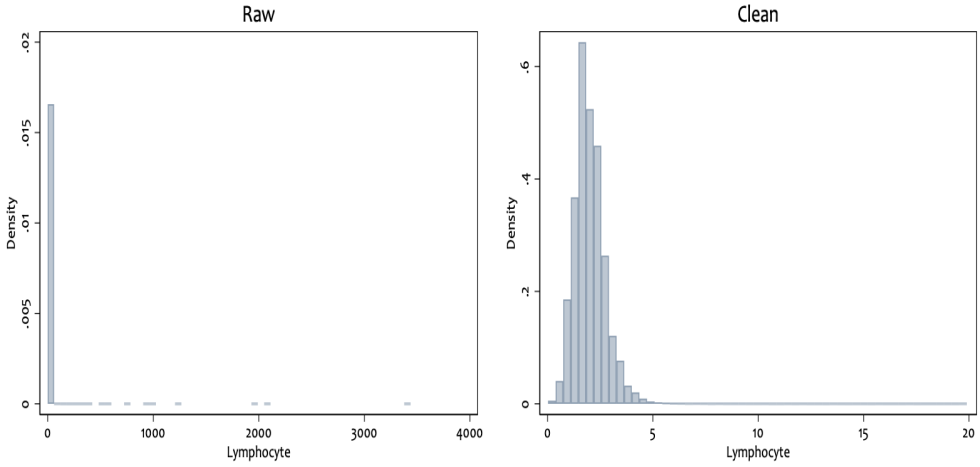
Haemoglobin

After cleaning, 343846 (40%) of patients have a measurement of haemoglobin in the 1 year prior to index



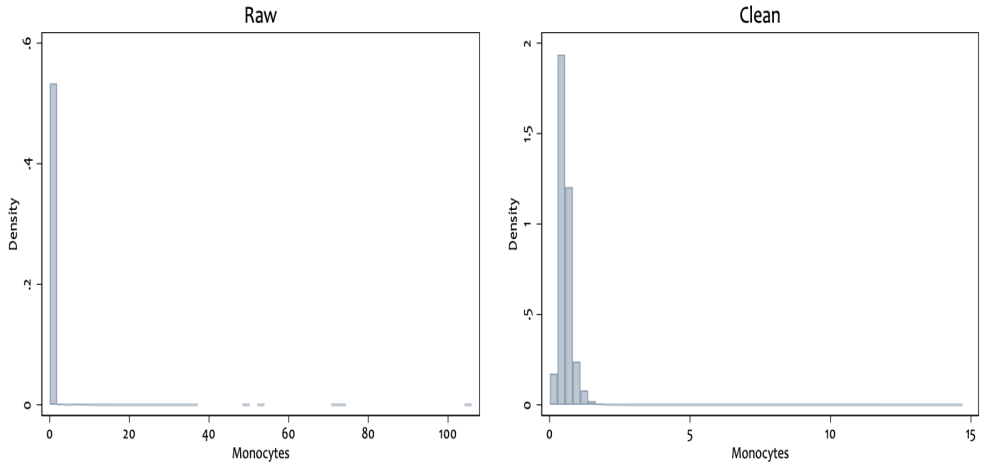
Lymphocyte

After cleaning, 318878 (37%) of patients have a measurement of lymphocyte in the 1 year prior to index



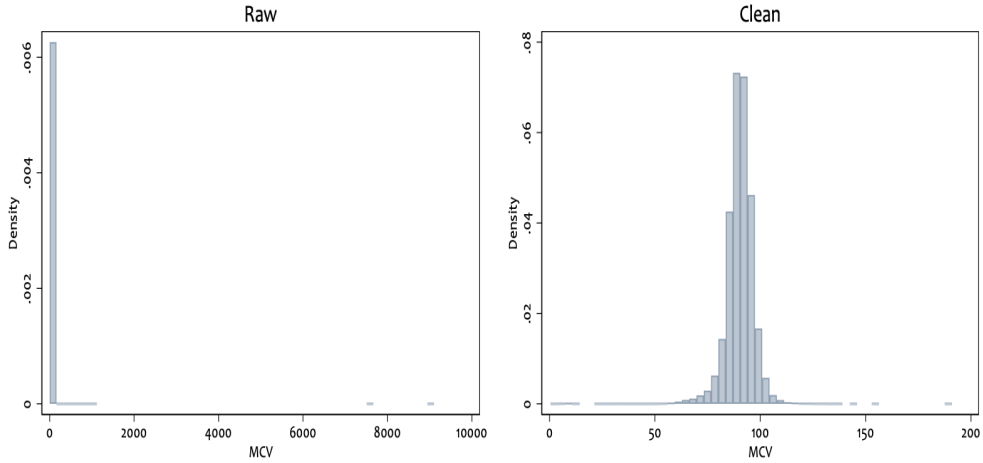
Monocytes

After cleaning, 315246 (37%) of patients have a measurement of monocytes in the 1 year prior to index



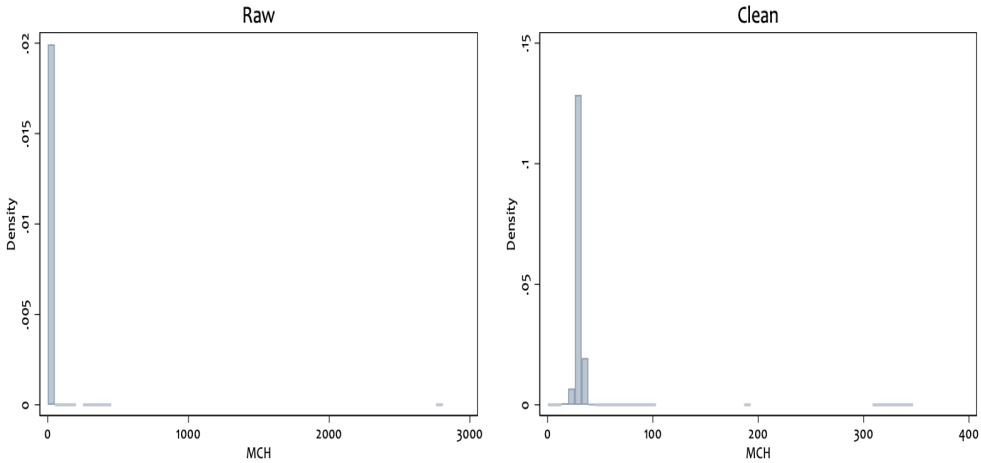
MCV

After cleaning, 334898 (39%) of patients have a measurement of mcv in the 1 year prior to index



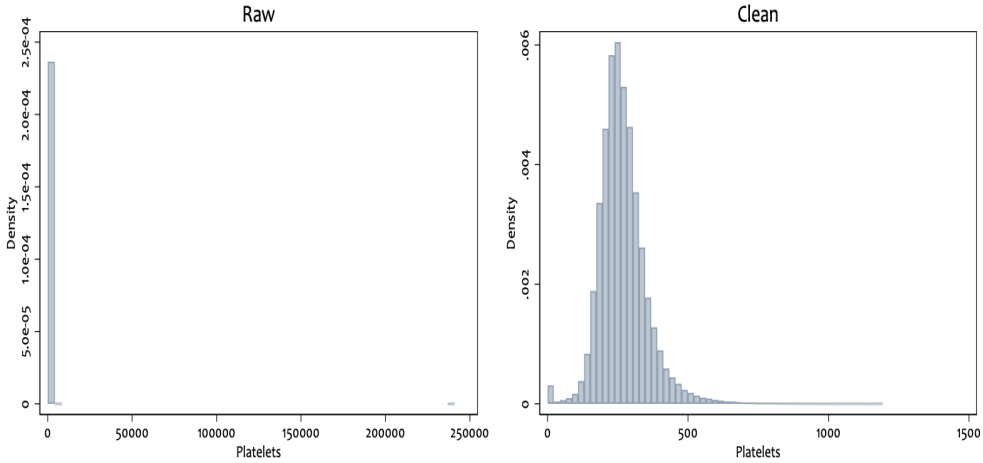
MCH

After cleaning, 311387 (36%) of patients have a measurement of mch in the 1 year prior to index



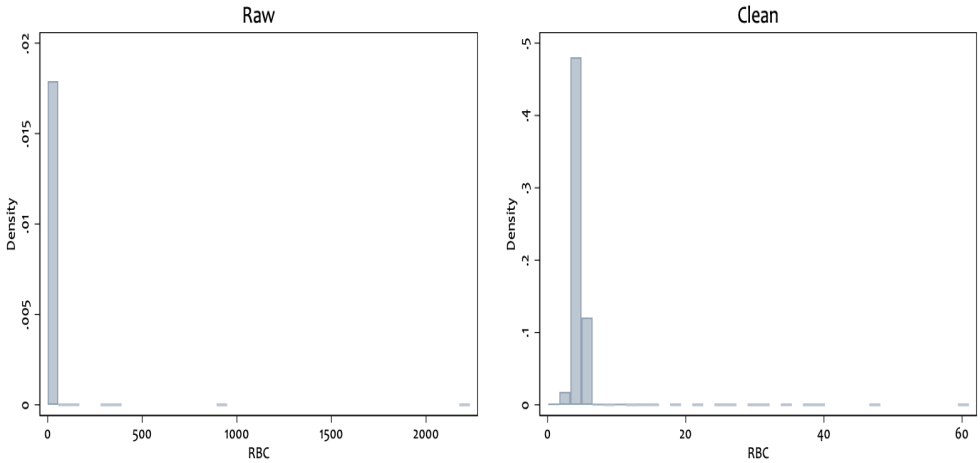
Platelets

After cleaning, 334552 (39%) of patients have a measurement of platelets in the 1 year prior to index



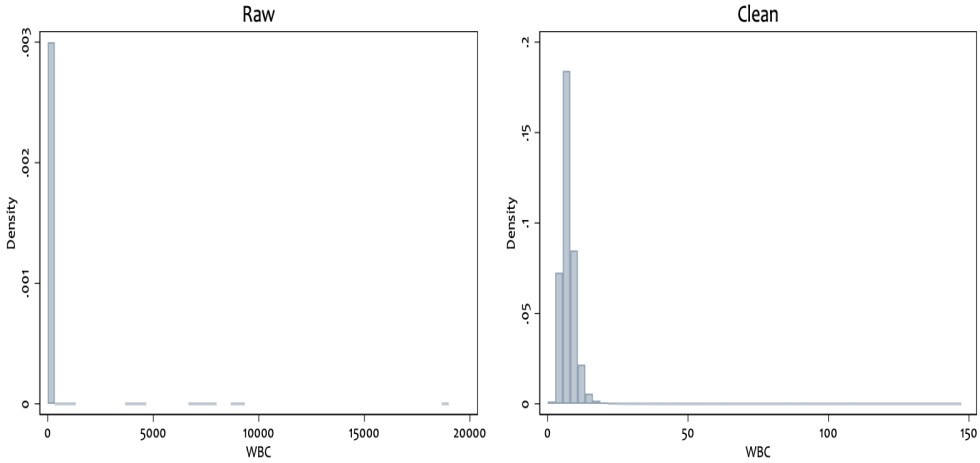
RBC

After cleaning, 7017 (1%) of patients have a measurement of rbc in the 1 year prior to index



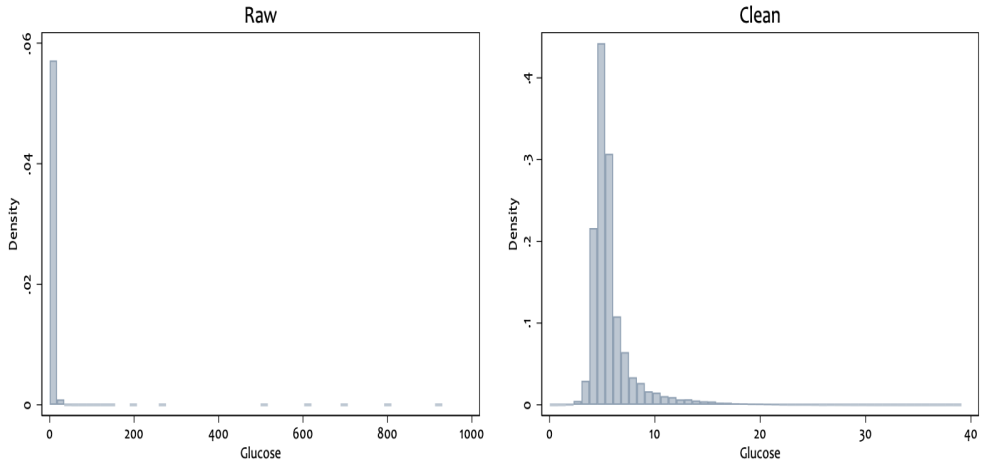
WBC

After cleaning, 330253 (38%) of patients have a measurement of wbc in the 1 year prior to index



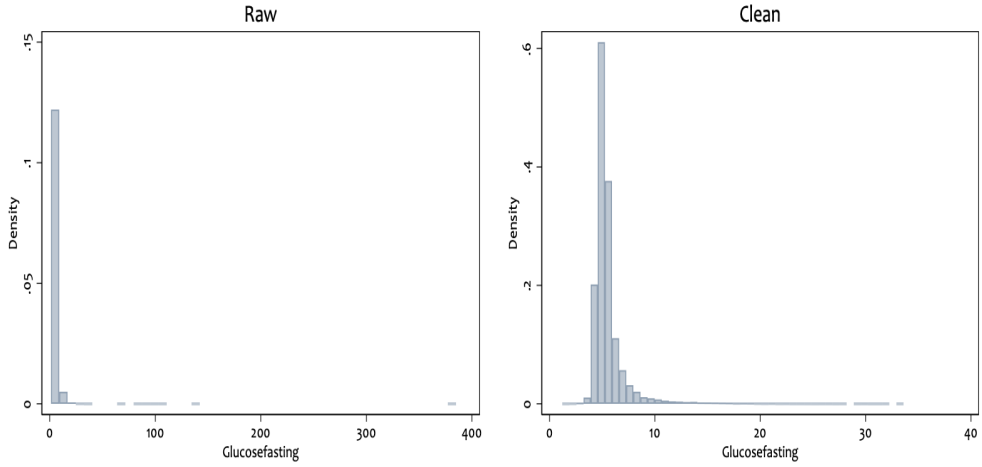
Glucose

After cleaning, 191916 (22%) of patients have a measurement of glucose in the 1 year prior to index



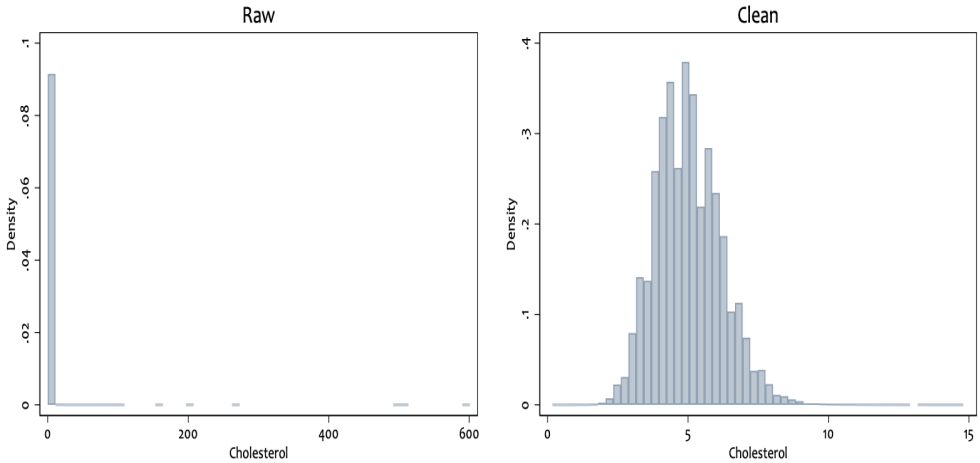
Glucosefasting

After cleaning, 71298 (8%) of patients have a measurement of glucosefasting in the 1 year prior to index



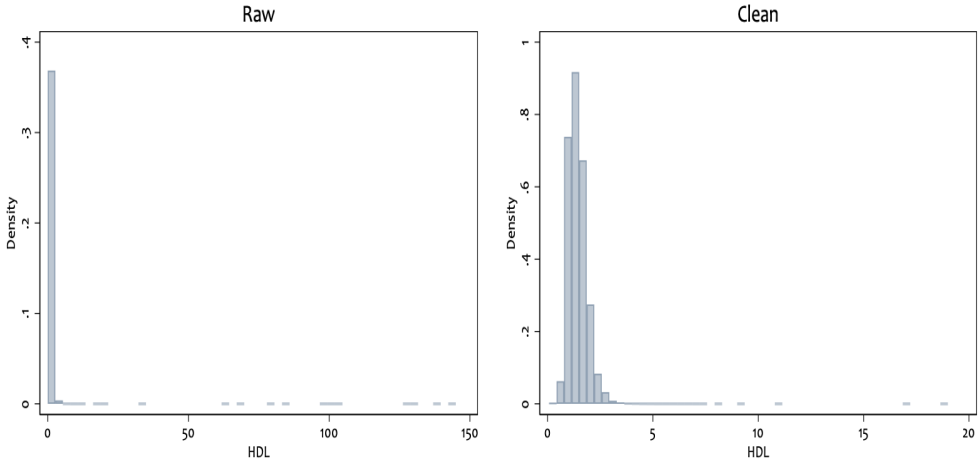
Cholesterol

After cleaning, 251515 (29%) of patients have a measurement of cholesterol in the 1 year prior to index



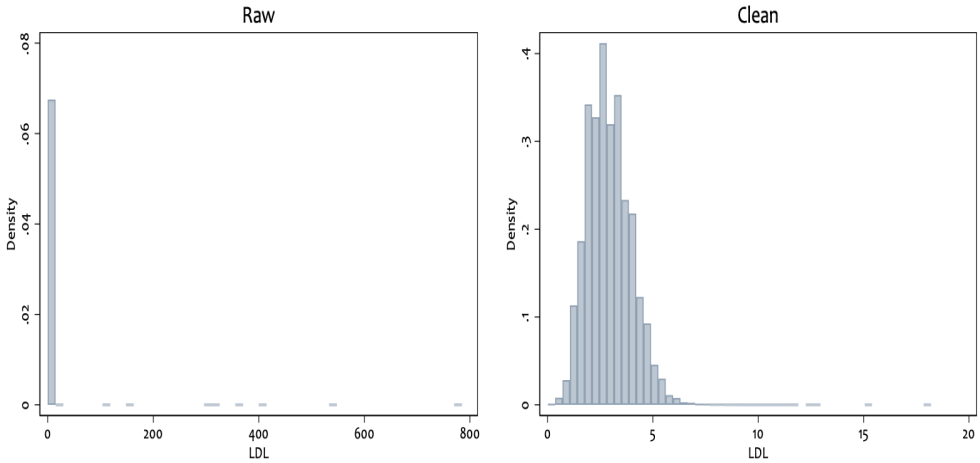
HDL

After cleaning, 201155 (23%) of patients have a measurement of hdl in the 1 year prior to index



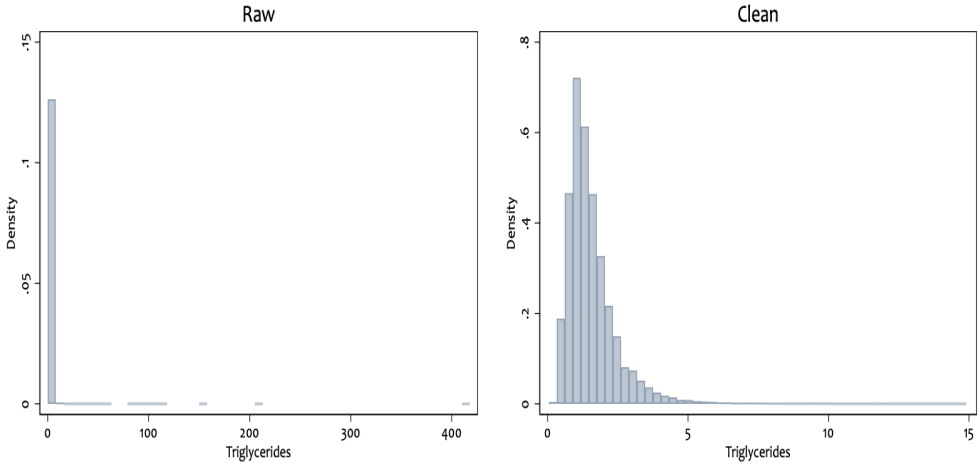
LDL

After cleaning, 161444 (19%) of patients have a measurement of ldl in the 1 year prior to index



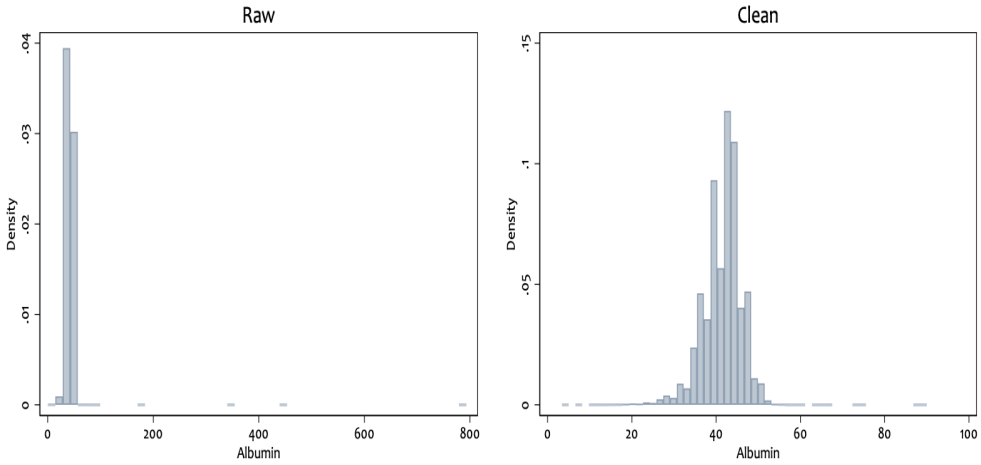
Triglycerides

After cleaning, 187138 (22%) of patients have a measurement of triglycerides in the 1 year prior to index



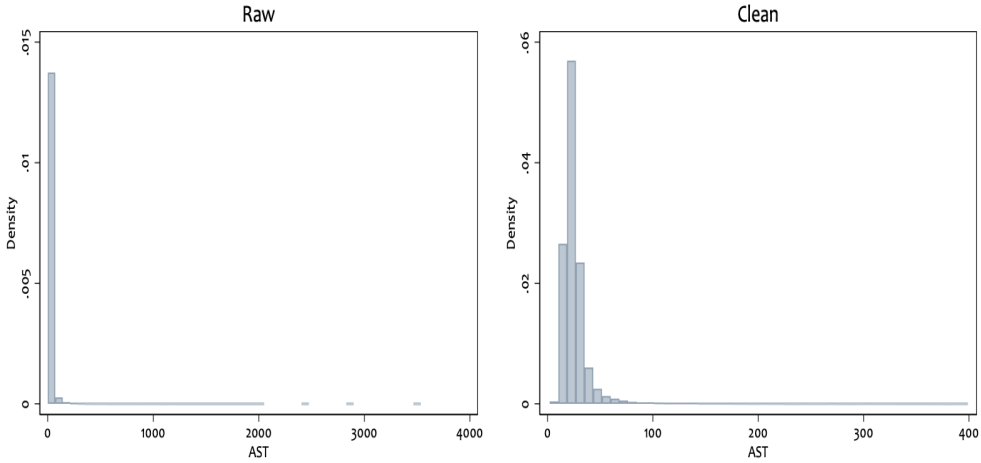
Albumin

After cleaning, 315127 (37%) of patients have a measurement of albumin in the 1 year prior to index



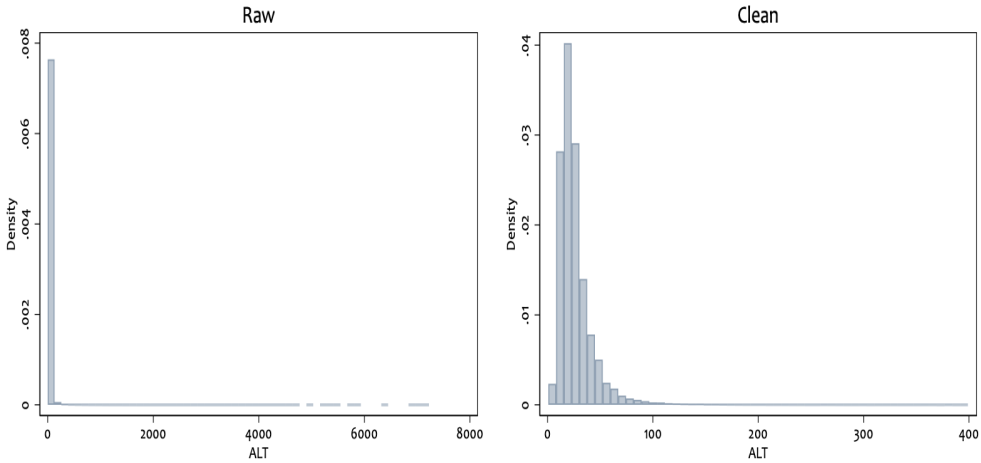
AST

After cleaning, 68153 (8%) of patients have a measurement of ast in the 1 year prior to index



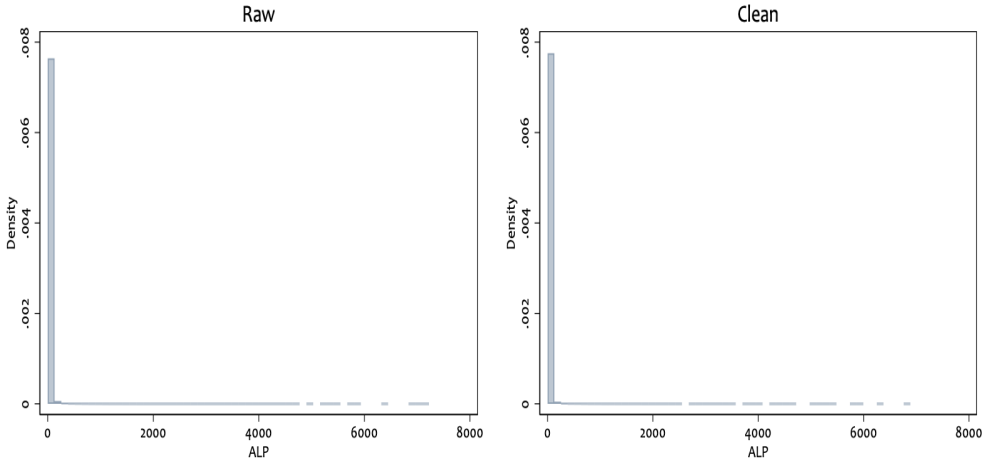
ALT

After cleaning, 262611 (31%) of patients have a measurement of alt in the 1 year prior to index



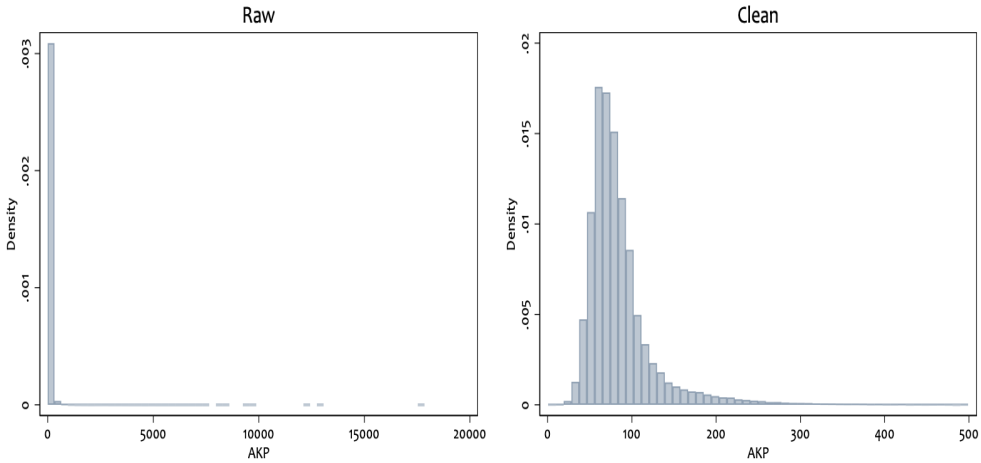
ALP

After cleaning, 265843 (31%) of patients have a measurement of alp in the 1 year prior to index



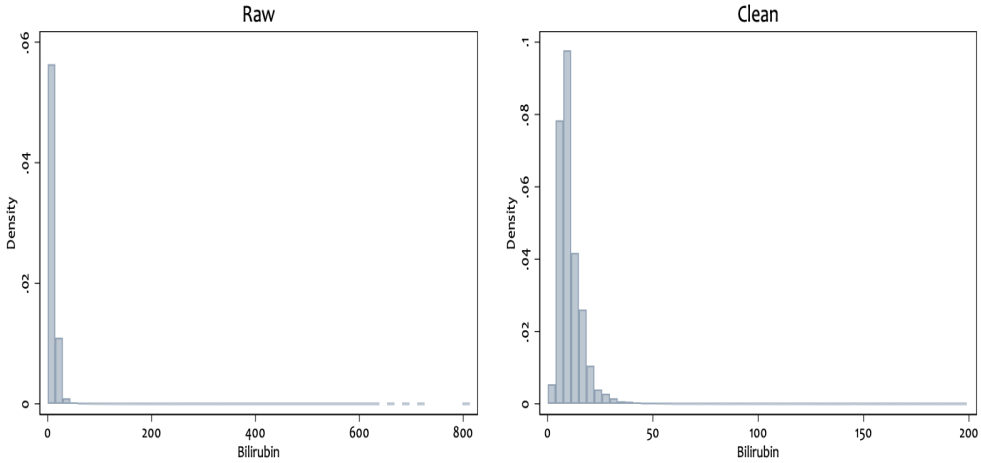
AKP

After cleaning, 313093 (36%) of patients have a measurement of akp in the 1 year prior to index



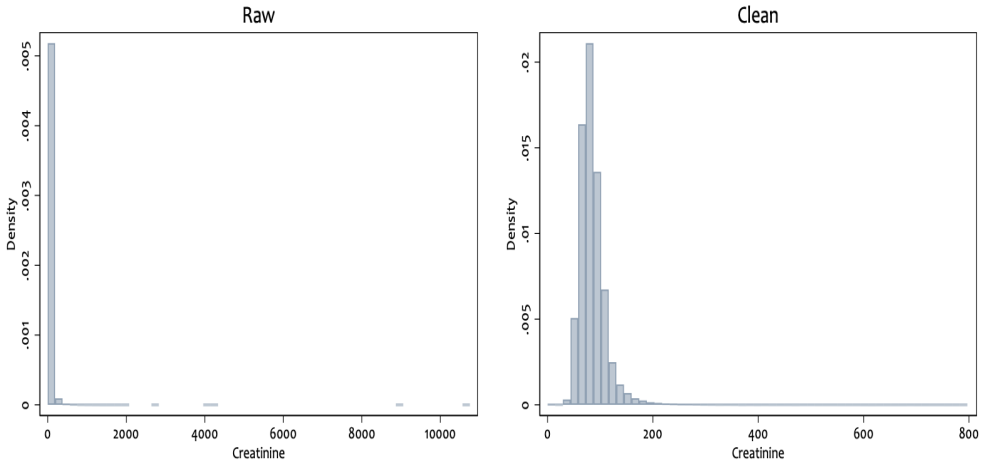
Bilirubin

After cleaning, 313489 (37%) of patients have a measurement of bilirubin in the 1 year prior to index



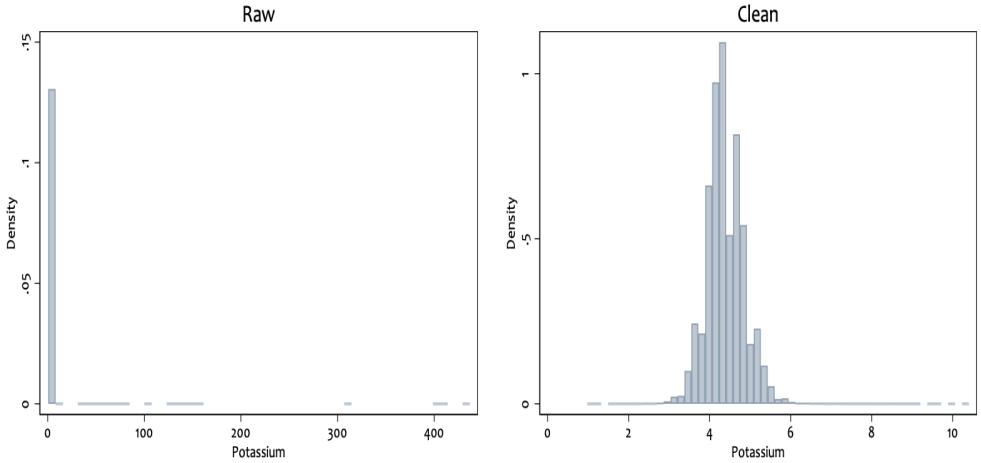
Creatinine

After cleaning, 365707 (43%) of patients have a measurement of creatinine in the 1 year prior to index



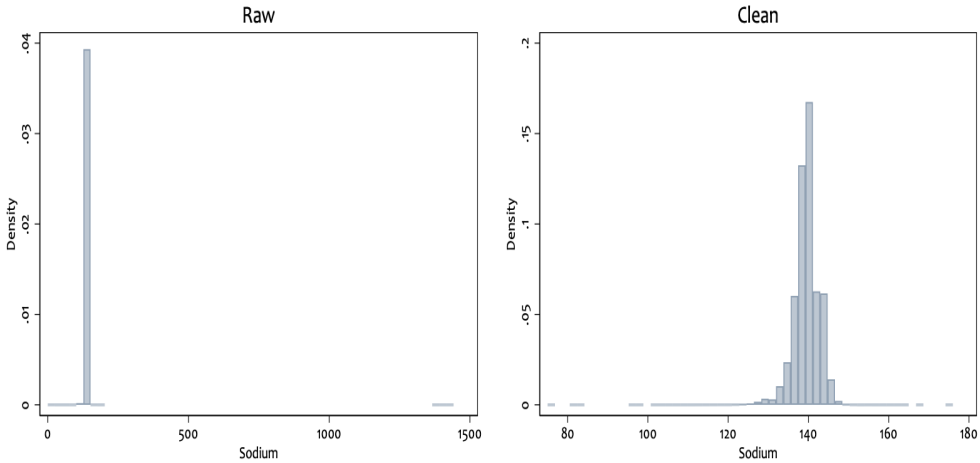
Potassium

After cleaning, 357462 (42%) of patients have a measurement of potassium in the 1 year prior to index



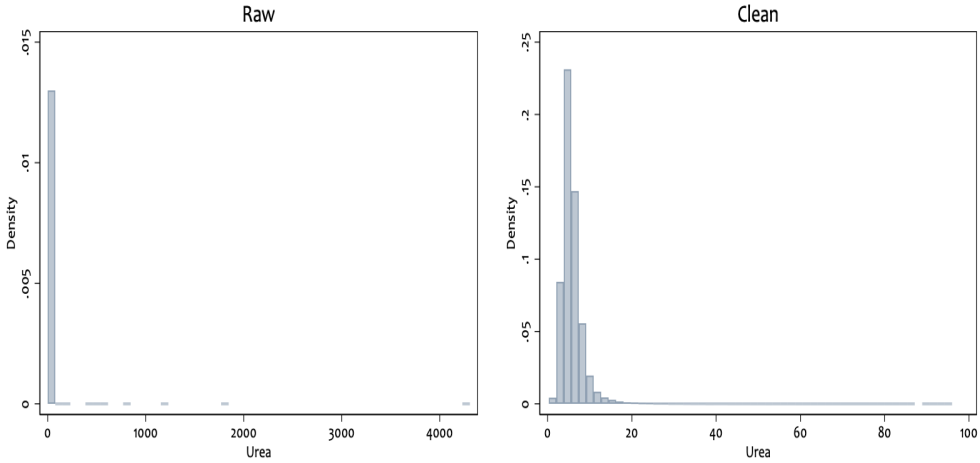
Sodium

After cleaning, 359888 (42%) of patients have a measurement of sodium in the 1 year prior to index



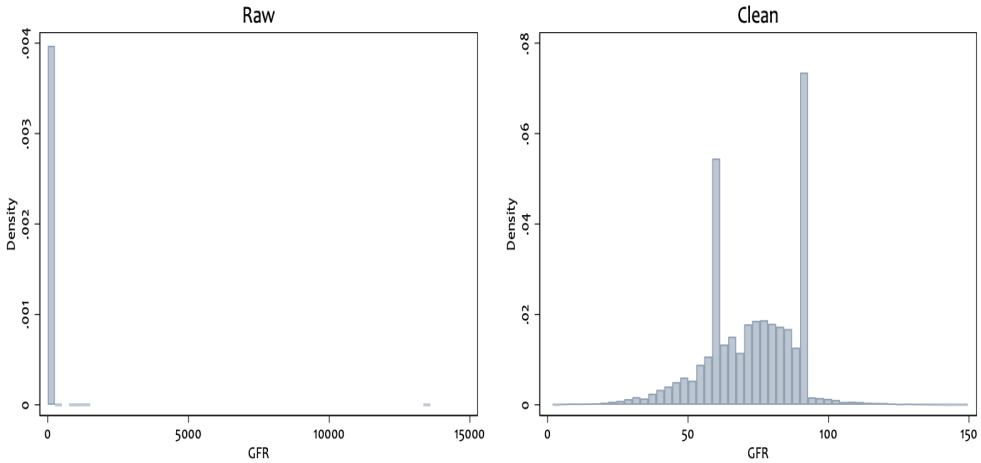
Urea

After cleaning, 289881 (34%) of patients have a measurement of urea in the 1 year prior to index



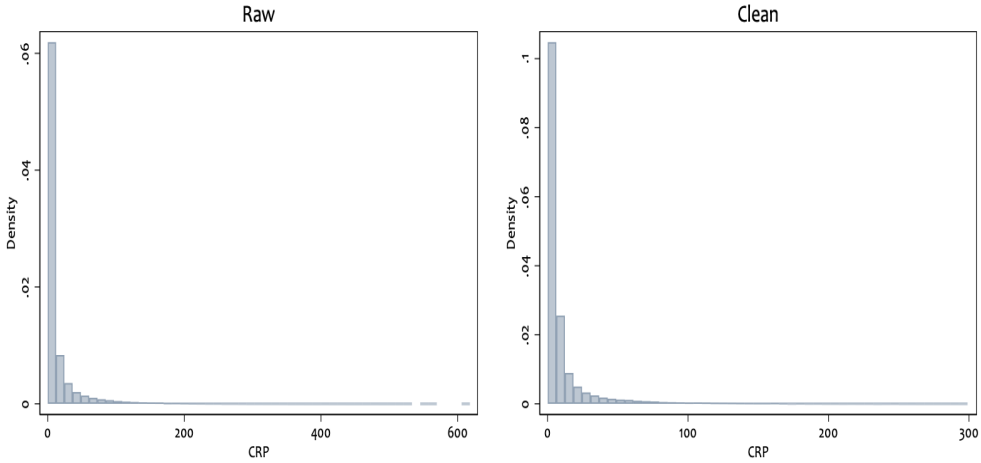
GFR

After cleaning, 159477 (19%) of patients have a measurement of gfr in the 1 year prior to index



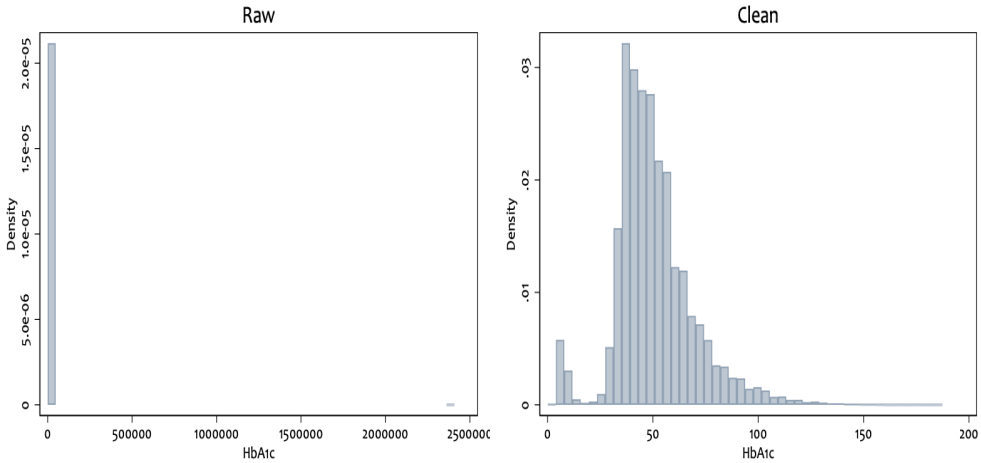
CRP

After cleaning, 94259 (11%) of patients have a measurement of crp in the 1 year prior to index



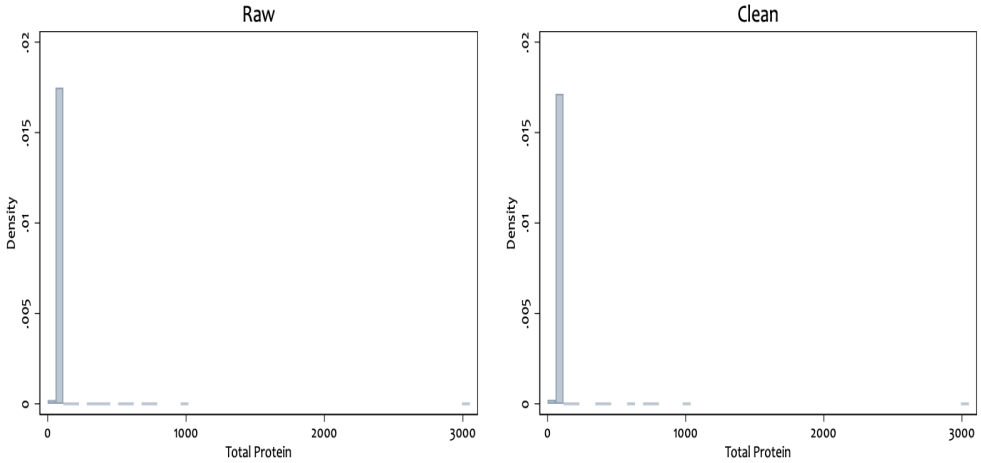
HbA1c

After cleaning, 72543 (8%) of patients have a measurement of hba1c in the 1 year prior to index



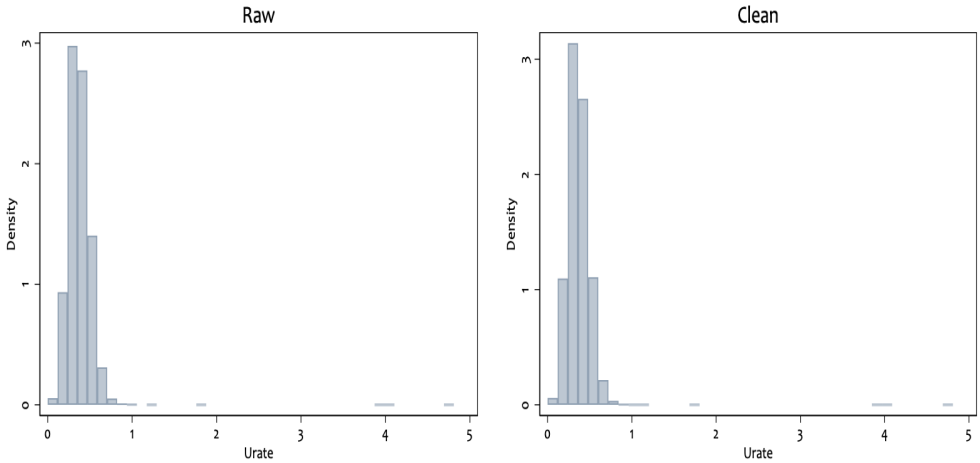
Total Protein

After cleaning, 204440 (24%) of patients have a measurement of totalprot in the 1 year prior to index



Urate

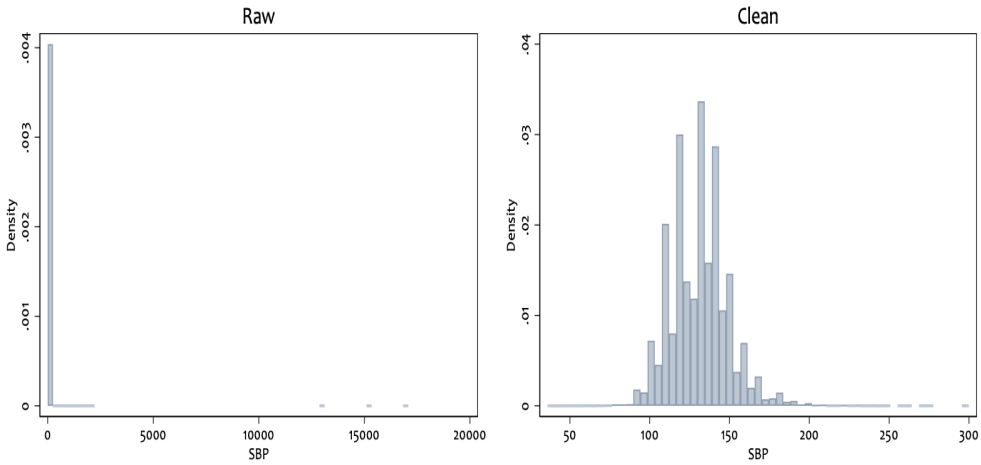
After cleaning, 11824 (1%) of patients have a measurement of urate in the 1 year prior to index


Raw
Clean
Density
Urate
0
1
2
3
4
5

Test results in the year prior to index

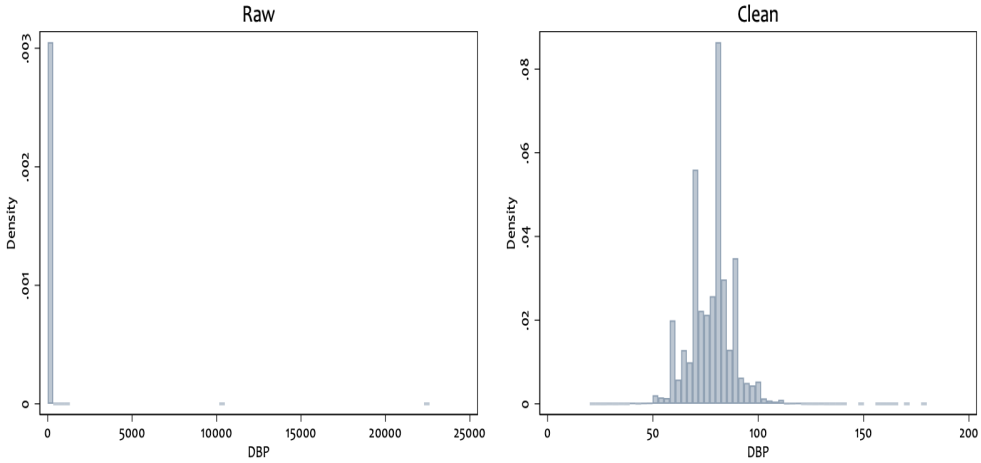
SBP

After cleaning, 816076 (95.08%) of patients have a measurement of sbp in the 1 year prior to index



DBP

After cleaning, 816076 (95.08%) of patients have a measurement of dbp in the 1 year prior to index



8.9.2 B: Continuous blood test results incorporated

The 19 continuous variables incorporated for the final analysis are as follows: Potassium, Sodium, Bilirubin, AKP, Creatinine, ALT, ALT, Lymphocyte, Urea, MCV, Platelets, Haemoglobin, WBC, Eosinophil, MCH, Cholesterol, Total protein, DBP, and Albumin

Chapter 9

Discussion

9.1 Overview

Summary

In the previous chapters, I have explored the use of high-dimensional propensity score (HDPS) approaches for confounder adjustment in UK electronic health records. In this chapter, I synthesis, review and discuss the key findings from the PhD, considering general strengths and limitations of this work. Finally, based on these findings I discuss the possible implications for current practice and outline directions for future work.

9.2 Summary of findings

The aim of this research was to investigate the use of high-dimensional propensity score (HDPS) approaches for data-driven confounder adjustment in UK electronic health records (EHRs). Central to this aim was understanding the potential benefit of HDPS approaches compared to more traditional investigator led approaches (based on a set of pre-defined covariates, chosen *a priori* based on clinical knowledge), whilst also providing practical guidance and improving accessibility of HDPS methods. In this section, I discuss the key findings of the thesis in context of the following objectives:

1. Describe UK EHRs and review relevant propensity score methodology (Chapter 2).
2. Propose modifications for implementing the underlying principles of the HDPS in UK EHRs (Chapter 3).
3. Apply the HDPS and proposed modifications in the context of UK EHRs (Chapters 3, 6, 7, 8).
4. Provide guidance surrounding diagnostic tools and reporting of HDPS analyses (Chapter 4).
5. Implement HDPS approaches in the Stata statistical software package (Chapter 5).
6. Investigate extensions to the HDPS framework that allow for the incorporation of laboratory test information (Chapter 8).

9.2.1 Objective 1: Describe UK EHRs and review relevant propensity score methodology

Large healthcare databases are increasingly used for non-interventional observational research studying the benefits and harms of medications (*Strom et al.*, 2013). In Chapter

2, I introduced the use of administrative claims databases and electronic health records (EHRs) in this context, describing the main similarities and differences, for example, surrounding the types of data typically available. Furthermore, I described the UK Clinical Practice Research Datalink (CPRD) (*Herrett et al.*, 2015), which is where the HDPS is applied throughout this thesis. One of the key features of the CPRD is the availability of data linkages and I described the relevant linkages used within the applied examples presented.

In Chapter 2, I highlighted that successful mitigation of confounding bias is a key challenge in UK EHRs (as with large healthcare databases more generally) (*Brookhart et al.*, 2010). In pharmacoepidemiological studies a key method for confounder adjustment is the propensity score (PS) and in this chapter I reviewed relevant methodology, in particular, highlighting the potential benefit of these approaches compared to traditional multivariable outcome regression (*Jackson et al.*, 2017; *Rosenbaum and Rubin*, 1983). Finally, I introduced the HDPS which is an extension to PS methodology tailored for use in large healthcare databases (*Schneeweiss et al.*, 2009). I described the various steps of the HDPS algorithm and provided a short critique outlining the key methodological and practical issues of this approach.

9.2.2 Objective 2: Propose modifications for implementing the underlying principles of the HDPS in UK EHRs

The second objective was to propose modifications for implementing HDPS principles in UK EHRs. The HDPS was developed in the context of administrative claims data (*Schneeweiss et al.*, 2009), however, there has been little work exploring how the HDPS should be applied outside this setting. Despite numerous examples applying the HDPS in UK EHRs, for example (*Schneeweiss*, 2018; *Suissa et al.*, 2017a,b; *Toh et al.*, 2011), none considered possible implications of directly applying the HDPS (the version widely applied in claims data) in this setting.

In Chapter 3, I elucidated the underlying principles of the HDPS algorithm. These principles allowed me to consider the steps of the HDPS algorithm in the context of

known differences between administrative claims data and UK EHRs (as described in Chapter 2) before proposing ways to better apply these underlying principles in UK EHR data. The modifications were applied to three case studies, presented in Chapters 3, 6 and 7. Based on the findings of this thesis, we recommend the following modifications.

Firstly, I mapped clinical and referral data from Read codes to the ICD-10 coding system (*Tazare et al.*, 2020). Since the Read coding system is not hierarchical the truncation of codes (for example, to the first 3-digits) does not capture distinct concepts. Manually grouping all Read codes was not a practical solution, I therefore mapped Read codes to the ICD-10 coding system (a hierarchical coding system) to allow for the grouping of medical concepts at a given granularity level. The key advantage to this approach is that it allows for proxy variables selected by the HDPS to be easily translated by researchers. This has important consequences since it can lead to increased understanding of epidemiological drivers of treatment initiation which might be omitted under a pre-specified model. Additionally, since this approach maps to SNOMED CT in an intermediary step, it allows for easy replication of the proposed approach in CPRD AURUM (where medical concepts are captured using this coding system) (*Wolf et al.*, 2019).

Since a large number of Read codes represent non-clinical and administrative information, these codes are dropped during the mapping procedure. However, the aim of this work was not to capture these concepts, but instead replicate the approach taken in claims data capturing homogeneous medically meaningful proxies. If administrative codes were deemed important, they could be included as a separate data dimension. Finally, the cross-map between Read and ICD-10 developed by NHS Digital is continually being updated and future studies applying the HDPS will benefit from the improved mapping between these coding systems.

Secondly, I proposed extending the lower frequency cut-off from 'Once' to 'Ever' (*Schneeweiss et al.*, 2009; *Tazare et al.*, 2020). The aim of this modification was to better capture recording practice in UK primary care where I hypothesised that the completeness of relevant information recorded at each consultation was likely to vary by data dimen-

sion. Therefore, the 'Ever' category captured whether a concept was recorded in a patient's entire medical history, rather than focusing on codes recorded within a covariate assessment window (for example, 1-year prior to cohort entry). This allowed for the incorporation of information on lifelong medical conditions (that may not be re-coded very frequently) that would otherwise be omitted under the traditional approach to assessing code recurrence. Furthermore, the resulting variables are defined comparably to how this information is typically defined in pre-specified models.

The findings from this thesis suggest that these proposed modifications can be usefully applied when implementing HDPS approaches in UK EHRs. They were not designed to outperform the traditional HDPS approach and, in the case studies presented, similar results have been obtained from the two approaches. Instead, these proposals aim to present a principled approach to applying the HDPS in UK EHRs which better captures characteristics of these data.

9.2.3 Objective 3: Apply the HDPS and proposed modifications in the context of UK EHRs

Throughout this thesis, I have focused on applying the proposed HDPS methodology to case studies where successful adjustment for confounding is likely to depend on the capture of subtle or hard to measure concepts relating to disease severity and frailty.

Chapter 3 presented a study investigating the interaction between clopidogrel and proton pump inhibitor (PPI) use on the risk of myocardial infarction. Whilst a previous cohort study had found that combined use was associated with an increased risk of myocardial infarction (*Douglas et al.*, 2012), results from randomised trials, genetic instrumental variable studies and a self-controlled case series indicated no evidence of an increased risk (*Bhatt et al.*, 2010; *Douglas et al.*, 2012; *Holmes et al.*, 2011). When applying the HDPS to the cohort study, both the standard and modified approaches obtained results closer to the expected null association (*Tazare et al.*, 2020). Furthermore, given the pattern of results and PS overlap plots, it was clear that the HDPS captured additional predictors of treatment initiation that were also causing confounding bias

(*Tazare et al.*, 2020).

Chapter 6 presented a study investigating the association between non-steroidal anti-inflammatory drugs (NSAIDs) and cyclo-oxygenase-2 (COX-2) inhibitor use on the risk of upper gastrointestinal bleeding (UGIB). This example was chosen given the availability of reliable trial evidence surrounding the protective effect of COX-2 inhibitor use (*Bombardier et al.*, 2000; *Silverstein et al.*, 2000) . Furthermore, it is a key case study in the HDPS literature and has been used to test the performance of the HDPS in a number of different databases used for pharmacoepidemiological research (*Garbe et al.*, 2013; *Hallas and Pottegard*, 2017; *Schneeweiss*, 2018; *Schneeweiss et al.*, 2009; *Toh et al.*, 2011) . In all examples, the HDPS appeared to improve confounder adjustment, in particular through the incorporation of subtle risk factors of UGIB (which is the hypothesised mechanism for residual confounding). I applied the modified HDPS and the pattern of results obtained was similar to those found by other studies and randomised controlled trials, although the 95% confidence intervals did not rule out an increased risk. Investigation of the HDPS covariates included also appeared to confirm the inclusion of risk factors for UGIB that had been omitted under the pre-specified model.

Chapter 7 presented a study investigating the association between PPIs and both all-cause and cause-specific mortality (*Brown et al.*, 2021). The standard HDPS was applied to the primary comparison between PPI users and H2-receptor antagonist users and, compared to a pre-specified PS model containing covariates selected *a priori*, the results obtained suggested better control for confounding based on the plausibility of associations and disease pathogeneis. However, we concluded that the HDPS was not able to fully eliminate confounding bias in this example. As highlighted by *Austin et al.* (2020), the HDPS is likely to balance factors which are similar to the proxies included in the HDPS model (for example, presence of a disease or receipt of a specific prescription). Therefore, in this instance, it is possible that residual confounding was related to differences between PPI and H2RA users that were not closely related to the proxies included (from clinical, referral and prescription dimensions).

PPIs are prescribed for a number of different indications and data suggest PPI users

often have worse underlying health status to comparator groups (*Brown et al.*, 2021). A key marker of underlying health status is laboratory test result information and in Chapter 8 I investigated the use of these data for further optimising confounder control in this study. I initially re-analysed the chronic pulmonary obstruction disorder specific mortality outcome (where we believe a causal association was biologically implausible) and obtained comparable results between the standard and modified-HDPS. Furthermore, incorporation of laboratory test information did not meaningfully alter the conclusions in this example. Despite this, almost half of the top 100 HDPS covariates (ranked by the Bross formula) were derived from test-based dimensions, indicating the potential importance of these data for confounder adjustment more generally.

The three case studies presented highlight the potential for HDPS approaches to improve confounder adjustment in UK EHRs, particularly in settings where residual confounding is hypothesised to be driven by proxies related to medical information available in the database under investigation. However, the PPI-Mortality study is a useful reminder that unmeasured confounding is a persistent issue in observational studies and no method is likely to be a silver bullet.

9.2.4 Objective 4: Provide guidance surrounding diagnostic tools and reporting of HDPS analyses

The fourth objective was to provide guidance surrounding diagnostic tools and reporting of HDPS analyses. Despite the popularity of HDPS methods (*Schneeweiss*, 2018), reporting of these approaches is often inadequate; especially surrounding implementation details and results of sensitivity analyses. This has led some to consider the HDPS a black-box approach.

In Chapter 4, I developed a set of diagnostic visualisations, sensitivity analyses and reporting suggestions to increase the transparency of HDPS analyses. Furthermore, these were illustrated in the context of the PPI and clopidogrel study presented in Chapter 3.

The reporting guidance consisted of 7 items surrounding the implementation of HDPS approaches, including the method for code prioritisation, number of variables selected and software used.

Graphical tools consisted of extensions of traditional diagnostic tools (for example, PS overlap plots) and those more tailored to HDPS analyses (*Granger et al.*, 2020). A key diagnostic surrounds checking the balance of covariates before and after inclusion of the HDPS covariates. This includes investigating whether the inclusion of several hundred HDPS covariates negatively impacts the balance in the pre-defined set of covariates. Furthermore, I highlighted the potential for the HDPS to identify and balance key confounders (as ranked by the Bross formula) that would be otherwise omitted and not balanced in an analysis including only the pre-defined covariates. Empirical and graphical tools for identifying instrumental-like variables were also proposed to help inform sensitivity analyses surrounding potentially influential variables (*Myers et al.*, 2011). Finally, I proposed several graphical tools for investigating the properties of selected HDPS covariates.

Suggested sensitivity analyses included varying the number of covariates selected and assessing the impact of instrumental-like covariates. The number of covariates selected was highlighted as an important decision especially since the optimal number to choose is usually unknown. In settings where results are particularly sensitive to this decision reporting a range of effect estimates may be more appropriate.

Finally, given the trend towards combining and comparing the performance of machine learning and HDPS methods (*Franklin et al.*, 2015; *Karim et al.*, 2018; *Tian et al.*, 2018; *Wyss et al.*, 2018b), the diagnostic tools presented will be important for highlighting differences between these various approaches.

9.2.5 Objective 5: Develop reusable software to implement HDPS approaches in the Stata statistical software package

The HDPS is typically applied using packages available in R and SAS (*Lendle*, 2017; *Rassen et al.*, 2020). These packages implement the generic steps of the HDPS procedure and allow investigators to specify tuning parameters surrounding the prevalence filter, method of prioritisation (exposure-based or Bross-based (*Rassen et al.*, 2011a; *Schneeweiss et al.*, 2009; *Wyss et al.*, 2018a)) and number of covariates selected.

Given modifications to the HDPS proposed in Chapter 3 surrounding the incorporation of 'Ever' information, I developed Stata do-files implementing both the standard and proposed modified-HDPS procedures (*Schneeweiss et al.*, 2009; *Tazare et al.*, 2020).

In Chapter 5, I used these do-files as the foundation for developing the `hdps` suite of commands in Stata. Stata is often used in the analysis of UK EHR databases and it was therefore important to provide a solution for researchers looking to apply these methods in this setting. The Stata package has similar functionality to the SAS and R packages, allowing users to specify the prevalence filter, method of prioritisation and number of covariates selected. However, it also has features currently unavailable in the SAS and R HDPS packages. The `hdps` Stata package allows users to apply the proposed modifications developed in Chapter 3 (*Tazare et al.*, 2020) and investigate the properties of covariates selected using the visualisations developed in Chapter 4.

The features available in the `hdps` suite are illustrated using simulated data which are freely available on GitHub, along with example code. This allows researchers to understand the required data structures and syntax for performing HDPS analyses in Stata.

9.2.6 Objective 6: Investigate extensions to the HDPS framework that allow for the incorporation of laboratory test information

The final objective of this thesis was to investigate extensions to the HDPS framework that allow for the incorporation of laboratory test information. In this work I focused on simple methods that were consistent with the philosophy of the HDPS approach. The two types of test data considered were tests requested and continuous test result values. Given the primary focus was on the proposed methods and to narrow the scope of this initial work, I focused only on blood tests when incorporating continuous values. This provided a set of tests that were likely to be prevalent in our study population and allowed me to draw on previous experience available in the research team (*Morales*, 2018b).

I initially incorporated these data by including an additional data dimension capturing whether certain tests were requested in a covariate assessment period prior to cohort entry. Tests requested are likely related to a number of factors but generally they are a marker of healthcare utilisation and underlying health status. This data dimension was incorporated within the HDPS procedure and the traditional cut-offs for assessing code recurrence were applied (as opposed to the use of 'Ever' information) since these data are likely to be complete.

I also investigated methods for incorporating continuous blood test result values within the HDPS procedure. These values are likely to be more reflective of an individual's underlying health status, however, given issues surrounding the data cleaning and missing data, the inclusion of these data is not straight forward. Furthermore, the HDPS algorithm does not readily support the inclusion of continuous values.

I started by using previously developed data cleaning rules to remove implausible blood test values and ensure the units of measurement were consistent (*Morales*, 2018b). These continuous values were then considered a separate data dimension in the HDPS procedure. The initial method for incorporating test result values was based on cut-

offs, similar in nature to the cut-offs implemented in the traditional HDPS procedure (*Schneeweiss et al.*, 2009). The key distinction is that, whilst traditionally cut-offs are based on code frequency, these were based on whether a continuous value was within a therapeutic range for a specific test result. In some instances, these therapeutic ranges were stratified by age or sex. These cut-offs were then included in the pool of binary covariates and prioritised and selected in the usual way. The key drawback of this approach is the loss of information and potential residual confounding arising from the categorisation of continuous values (*Groenwold et al.*, 2013; *Royston et al.*, 2006).

To address this limitation I also investigated how to incorporate the continuous values. For test cut-off covariates selected in the top 100 (as ranked by the Bross formula) I additionally included the continuous test value variable and a missing indicator in the PS model (*Carpenter and Kenward*, 2013). Given recent work by *Blake et al.* (2020), this approach is valid under a set of assumptions that are likely to hold in the context of UK EHRs. Furthermore, in comparison to multiple imputation (*Carpenter and Kenward*, 2013), this approach is less computationally intensive and more easily implemented within the existing HDPS framework.

Results incorporating these types of test data suggested that both can be important for mitigating confounding bias and are therefore important to consider within the HDPS procedure when available. Furthermore, in this particular example, a large proportion of the top 100 covariates selected (after prioritisation by the Bross formula) originated from either the tests requested or cut-offs dimensions.

Whilst developed for blood test results, an advantage of the proposed approach is the ability to easily incorporate additional types of test values in the future, for example, respiratory test results.

9.3 Strengths

9.3.1 Application of proposed approaches to applied studies

A key strength of this project is the application of the proposed HDPS approaches to relevant studies where residual confounding was a concern.

Statisticians are often able to test and develop new methodology under a set of ideal conditions, for example, through the use of simulation studies (*Morris et al.*, 2019). However, as highlighted in Chapter 2, large healthcare databases are inherently messy and complex which means that the performance of a method in a given setting is not guaranteed. Whilst frameworks for simulating healthcare databases exist, for example the plasmode simulation framework (*Franklin et al.*, 2014), the focus on resampling covariates rather than proposing data generating mechanisms can mean the generalisability of results is unclear.

Given these issues, empirical studies have historically guided methodological best practice in the field of pharmacoepidemiology, helping to inform settings and questions where particular methods might be expected to perform well. In this thesis, I have presented three diverse case studies in UK EHRs where the hypothesised mechanisms for residual confounding are different. The narrow focus on a particular data source has allowed for the investigation into the strengths and limitations of the HDPS in UK primary care data. Whilst this might limit the generalisability of some of the findings, it is possible the HDPS would perform comparably in data sources with similarly rich primary care data.

9.3.2 Accessibility of methods

Another strength of this work is the focus on accessibility of the methods and guidance developed. *Cadarette et al.* (2017) suggests recommendations for improving diffusion of methodological work in the field of pharmacoepidemiology. In the paragraphs below I discuss each of the five recommendations in the context of this thesis.

1. Clearly describe using foundational principles (simple language)

In Chapter 3, I elucidated the underlying principles of the HDPS approach, highlighting four principles for translating the HDPS to UK EHRs in a way that accurately characterises key features of these data.

2. Consider comparing results to established methods

In each of the applied examples presented (Chapters 3, 6, and 7) the HDPS has been compared to a traditional PS approach where a set of confounders has been identified based on clinical knowledge. The results obtained from these approaches have then been externally benchmarked against known biological mechanisms, randomised controlled trials, other pharmacoepidemiological studies and unconfounded genetic instrumental variable studies to evaluate the performance of the HDPS.

3. Provide sample data, code or calculation examples, and instructions

Throughout this thesis I have aimed to work in a transparent manner and make code and data available where possible. However, given licensing laws and data protection, it is often not possible to share analysis datasets from large healthcare databases. Whilst this is likely to remain the case, during the COVID-19 pandemic there has been increased focus on 'open science' and in this field the use of trusted research environments, such as OpenSAFELY, has facilitated increased sharing of analytical code between researchers improving the reproducibility and transparency of analyses (*Besançon et al.*, 2021; *Williamson et al.*, 2020). Below I describe the code and data sharing in this thesis.

Chapter 4 focused on the transparency of HDPS models and provided guidance surrounding diagnostic tools and reporting of these analyses. Whilst it has not been possible to share this data, code has been made available for all the diagnostic visualisations and is available at `https://github.com/johntaz/HDPS-Diagnostics`.

Chapter 5 introduced the `hdps` Stata package, implementing both the standard HDPS procedure and modifications introduced in Chapter 3. In this chapter, the data used to illustrate features of this package were entirely simulated. As well as detailed instructions highlighting various features of the commands in both the paper and Stata help files, the simulated data and example analysis scripts are available on GitHub at `https://github.com/johntaz/HDPS-Stata-Demo/`. Finally, the package itself is maintained on GitHub (`https://github.com/johntaz/hdps`) to allow for tracked version control and easy review of the underlying code.

4. Early communication, support and testing

Most of the work presented in this thesis has benefited from early communication at academic conferences. The work in Chapters 3 and 7 was presented at the *35th International Conference on Pharmacoepidemiology & Therapeutic Risk Management (ICPE) (2019)*. An initial version of the Stata package was presented at the *2019 UK Stata Conference*. Finally, work presented in Chapters 4 and 8 was presented at *ICPE All Access 2020*.

To support understanding surrounding the implementation of HDPS principles in UK EHRs, I delivered a lecture on this work at a two-day PS workshop held at Health Data Research UK, London.

Lastly, the `hdps` Stata package has benefited from review and testing from members of the EHR Research Team at LSHTM. This process helped to improve the usability of the commands presented.

5. Provide methodological and reporting guidance

Throughout this thesis case studies have been used to help support guidance relating to scenarios where HDPS analyses might outperform traditional investigator-led covariate adjustment methods. In Chapter 3, a proposed framework for applying the HDPS in UK EHRs was outlined and this was extended to provide guidance surrounding

incorporation of hospital admission data (Chapter 6) and laboratory test information (Chapter 8).

Chapter 4 focused on guidance surrounding diagnostic tools and reporting suggestions for HDPS analyses in general. This work provided guidance on visualisations and sensitivity analyses for investigating HDPS models and the robustness of results obtained. Furthermore, a seven item checklist outlined key information for the reporting of these analyses to aid transparency and reproducibility.

In Chapter 5, guidance was given surrounding the implementation of HDPS analyses in Stata. This work outlined ongoing methodological discussions (such as the use of a prevalence filter (*Schuster et al.*, 2015)) and how features of the commands presented could be used to implement variations of the HDPS procedure.

9.4 Limitations

9.4.1 Comparison with machine learning approaches

There is a growing number of studies comparing the HDPS with machine learning methods, both separately and through so-called hybrid approaches (*Schneeweiss*, 2018), for example, *Franklin et al.* (2015); *Karim et al.* (2018); *Schneeweiss et al.* (2017); *Tian et al.* (2018); *Wyss et al.* (2018b). There is understandable optimism around these methods since they can contribute to the HDPS procedure in a number of ways, including variable selection, model specification and regularization of coefficients in the PS model (*Schneeweiss*, 2018).

However, whilst machine learning methods have shown promise, they also add complexities relating to data management, computational burden and the need to specify additional tuning parameters (*Schneeweiss et al.*, 2017). Furthermore, despite having clear advantages in certain settings (for example, rare outcomes), evidence suggests they are not uniformly superior and often obtain similar results to the traditional HDPS (*Karim et al.*, 2018; *Schneeweiss*, 2018; *Schneeweiss et al.*, 2017).

In this thesis, I have focused on the principled application of HDPS methods in UK EHRs and have not attempted to make comparisons with alternative machine learning approaches. However, given the potential advantages of these approaches, including in UK EHR data (*Karim et al.*, 2018), further investigation of when machine learning approaches can be usefully applied is a natural next step to the work presented in this thesis.

9.4.2 Generalisability of results

Assessing the superiority of confounder adjustment methods in large health databases is a key challenge in pharmacoepidemiological research and, even with available simulation frameworks, the generalisability of results is unclear (*Franklin et al.*, 2014; *Tian et al.*, 2018). For example, in the plasmode simulation framework, the simulated data does not depend on the set of 'unmeasured' HDPS covariates, they are simply resampled to preserve the complexities present in the dataset used to create the plasmodes (*Franklin et al.*, 2014).

In this thesis, the performance of the HDPS compared to investigator led pre-specified PS models has been evaluated, in part, by which method has obtained results closest to an external gold-standard or are most plausible given known biological mechanisms. This type of evaluation has been an important part of the growing evidence base surrounding the HDPS and gives the following key insights (*Schneeweiss*, 2018). Firstly, it allows researchers to understand the scenarios where HDPS might be expected to perform well, both in terms of aetiological questions and confounding structures. Secondly, by reviewing the types of covariates selected by the HDPS, these analyses can add to our epidemiological understanding of the key drivers of treatment initiation in healthcare databases.

However, a key limitation is that these empirical studies do not guarantee the performance of the HDPS in a new setting, which is especially important when external gold standard data are unreliable or not available. Therefore, whilst the results from empirical studies can help researchers understand when these methods can perform well, they

should not lead to over-optimism surrounding the potential benefit of these methods in any given setting.

The HDPS aims to identify proxies of concepts that are unmeasured or hard to measure in the data source under investigation and likely to be important for mitigating confounding bias (*Schneeweiss et al.*, 2009). Therefore, as highlighted in Chapter 4, the ability of the HDPS (and importantly, any other PS model under consideration) to successfully balance the identified proxy covariates is an important metric. Reliability of the results obtained should then be framed based on whether the HDPS is likely to have captured the types of covariates hypothesised to be contributing to residual confounding. For example, when comparing the performance of the HDPS to gold-standard clinical data (where complete and detailed clinical information was available on key conditions and measures of health status), *Austin et al.* (2020) observed that the HDPS proxy covariates tended to be more correlated with previously unmeasured clinical and therapy concepts (in which good covariate balance was achieved) than continuous measures such as laboratory test results. Therefore, in this example, if the residual confounding was driven by these continuous measures, the HDPS would not necessarily lead to vastly improved confounder adjustment compared to a pre-specified PS model.

Finally, this highlights a further point that the data dimensions available are always going to be a limiting factor when judging the performance of the HDPS. In this thesis, I have shown that the incorporation of laboratory test information can lead to a high proportion of codes selected from these data dimensions. This highlights that in order to optimise the confounding control in a particular setting, investigators might benefit from incorporating a diverse range of data in the HDPS procedure. However, it is important to acknowledge that successful mitigation of confounding might depend on unmeasured factors not captured or balanced by the HDPS proxy covariates selected.

9.5 Future work

In this section, I discuss possible directions for future work based on the findings of this thesis.

9.5.1 Incorporating additional data available in UK EHRs

In this thesis, I have focused on data commonly included in applications of the HDPS more generally, including clinical, referral, therapy and hospitalisation data. However, as illustrated in pilot work presented in Chapter 8, the HDPS framework can be extended to incorporate other types of data that could be important for mitigating confounding bias in UK EHRs.

Chapter 8 proposed simple methods for including laboratory test information in the HDPS framework and future work could extend these ideas to better characterise the continuous data included. For example, splines and fractional polynomials could be investigated as methods for improving the modelling of these continuous variables (*Binder et al.*, 2013).

The incorporation of continuous laboratory test data also leads to a missing data problem. In Chapter 8, this was handled using the missing indicator approach, shown to be valid under a set of assumptions often likely to hold in UK EHRs (*Blake et al.*, 2020). However, the validity of these assumptions have not been fully investigated in the context of high-dimensional confounder adjustment.

Finally, free text information has been used to investigate whether the presence of specific text strings can be used within the HDPS framework to help mitigate confounding bias (*Rassen et al.*, 2013). Whilst this is theoretically possible in the CPRD, current data governance restrictions mean these data are not routinely available amid concerns surrounding the release of patient identifiable information (*Price et al.*, 2016; *Shah et al.*, 2019). However, were these data to become available in the future, this is a possible avenue for future work.

9.5.2 HDPS R package

In Chapter 6, I presented a study conducted in collaboration with GlaxoSmithKline. Given the availability of software on computer systems at GlaxoSmithKline, I developed R code implementing the modified-HDPS presented in Chapter 3. Whilst *Lendle* (2017) has developed an R package for running the standard HDPS, this can not apply the modifications to the HDPS presented in this thesis.

Therefore, whilst a Stata package exists (presented in Chapter 5), future work could develop and release a similar package in R; which has the advantage of being a free and open source software (*R Core Team*, 2020).

9.5.3 CPRD Aurum

Throughout this thesis the case studies presented have used the CPRD GOLD database (*Herrett et al.*, 2015). However, CPRD Aurum is a relatively new UK primary care EHR database that is rapidly being used in conjunction with CPRD GOLD in a wide range of research studies (*Wolf et al.*, 2019).

One key distinction is that CPRD Aurum records clinical information in SNOMED CT rather than the Read coding system (*Wolf et al.*, 2019). In Chapter 3 when identifying homogeneous clinical concepts, I mapped Read codes to ICD-10 codes via SNOMED CT. Therefore, the work developed in CPRD GOLD is also transferable to CPRD Aurum. Future work could investigate further practical and operational consequences of applying the HDPS in CPRD Aurum.

9.5.4 Empirical studies

This thesis presents three case studies illustrating proposed modifications and extensions for applying the HDPS in UK EHRs. Whilst the collective results indicate that HDPS approaches can help reduce residual confounding in these studies, more empirical

studies are needed to gain a better understanding of the scenarios and questions where the HDPS might perform well. In particular, studies might expand on work investigating the HDPS in settings where different strengths of association are expected and where the treatment and outcome are rare (*Patorno et al.*, 2014; *Rassen et al.*, 2011a; *Schneeweiss et al.*, 2017).

There are several questions that might be explored in these future empirical studies. Firstly, the HDPS could be compared to machine learning approaches to better understand when these might usefully augment or outperform the HDPS procedure in UK EHRs (*Franklin et al.*, 2015; *Karim et al.*, 2018; *Schneeweiss et al.*, 2017; *Tian et al.*, 2018; *Wyss et al.*, 2018b). Secondly, studies could investigate settings where the HDPS might be able to fully automate confounding control in UK EHRs (*Rassen and Schneeweiss*, 2012; *Schneeweiss*, 2018). For example, after defining treatment, outcome and demographic variables (e.g. age and sex), could the HDPS fully mitigate confounding bias (in the absence of other investigator defined covariates).

9.5.5 Prediction modelling

Finally, one area of future work relates to the wider use of the covariate generation steps of the HDPS procedure (*Schneeweiss et al.*, 2009), for example, to aid clinical prediction modelling (*Steryerberg*, 2009).

The HDPS algorithm highlights that when analysing large healthcare databases the covariates selected for inclusion in a model are just as important as the method used to estimate a quantity of interest (*Austin et al.*, 2020). Furthermore, the HDPS often identifies important covariates otherwise omitted by an investigator (*Schneeweiss*, 2018).

Despite designed as an extension to PS methodology, the initial steps (Chapter 2, Steps 1-3) of the HDPS are essentially a set of data management tools for deriving data-driven covariates in large healthcare databases (*Schneeweiss et al.*, 2009). However, these types of covariates are rarely derived more widely in the analysis of these databases.

There is increasing overlap and interchangeability of methods between the fields of causal inference and prediction (*Blakely et al.*, 2020). In the context of UK EHRs, there is interest in the use of methods that exploit the volume of the data available, however, covariates typically only measure the presence of a code rather than, for example, assessing the frequency of codes (as in the HDPS) (*Cowling et al.*, 2021).

Whilst there is potential for using the HDPS data management steps in the context of prediction modelling, this has not been fully investigated. One example exists predicting long-term adherence in US claims data (*Franklin et al.*, 2016), however, the performance of such an approach might differ in UK EHR data given differences in the richness of data available in these data sources (*Schneeweiss and Avorn*, 2005). Future work could therefore investigate the ability of HDPS-derived covariates to improve prediction modelling in UK EHRs.

9.6 Concluding remarks

In this thesis, I have investigated the use of the HDPS for data-driven confounder adjustment in UK EHRs. I have proposed modifications to the existing HDPS procedure and applied this modified-HDPS to a number of case studies. I have also extended this framework to incorporate hospitalisation and laboratory test data.

Additionally, this thesis contributes to the growing literature surrounding the HDPS more generally. Firstly, I presented diagnostic visualisations and reporting suggestions for increasing the transparency and interpretability of HDPS analyses. Secondly, I have developed and released a Stata package implementing both the standard and modified HDPS procedures to help improve the accessibility of these methods.

Whilst the HDPS has shown promise in scenarios where confounding is driven by hard to measure concepts, it is unlikely to outperform investigator led approaches when the confounding structure is relatively simple, the key drivers of confounding bias are well understood, or if key confounders are not captured by the proxy covariates. Furthermore, it is important to recognise that residual confounding may still remain after

adjustment for HDPS-derived covariates.

The collective findings of this thesis demonstrate the potential for HDPS approaches to overcome intractable confounding in UK EHRs and highlight its versatility as a data-driven method for confounder identification and selection.

Appendix A

LSHTM Ethical approval for PPI-Clopidogrel study

London School of Hygiene & Tropical Medicine
Keppel Street, London WC1E 7HT
United Kingdom
Switchboard: +44 (0)20 7636 8636

www.lshtm.ac.uk

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

Observational / Interventions Research Ethics Committee

Mr John Tazare
LSHTM

6 August 2019

Dear John

Study Title: Clopidogrel and Proton Pump Inhibitors: a propensity score adjusted cohort study to investigate issues with residual confounding

LSHTM Ethics Ref: 16780

Thank you for your application for the above research project which has now been considered by the Observational Committee via Chair's Action.

Confirmation of ethical opinion

On behalf of the Committee, I am pleased to confirm a favourable ethical opinion for the above research on the basis described in the application form, protocol and supporting documentation, subject to the conditions specified below.

Conditions of the favourable opinion

Approval is dependent on local ethical approval having been received, where relevant.

Approved documents

The final list of documents reviewed and approved is as follows:

Document Type	File Name	Date	Version
Investigator CV	ID CV Apr18 LEO	01/04/2019	1
Investigator CV	cv_elizabeth_williamson_22_02_19	01/04/2019	1
Investigator CV	jt_cv_jan19	01/04/2019	1
Consent form	17_194R_ISAC feedback	01/04/2019	1
Local Approval	Appendix_1_original_douglas_protocol	01/04/2019	1
Local Approval	Appendix_2_original_LSHTM_ethics_application	01/04/2019	1
Local Approval	Appendix_3_original_LSHTM_ethics_approval	01/04/2019	1
Protocol / Proposal	jt_protocol	23/07/2019	v1.0

After ethical review

The Chief Investigator (CI) or delegate is responsible for informing the ethics committee of any subsequent changes to the application. These must be submitted to the committee for review using an Amendment form. Amendments must not be initiated before receipt of written favourable opinion from the committee.

The CI or delegate is also required to notify the ethics committee of any protocol violations and/or Suspected Unexpected Serious Adverse Reactions (SUSARs) which occur during the project by submitting a Serious Adverse Event form.

An annual report should be submitted to the committee using an Annual Report form on the anniversary of the approval of the study during the lifetime of the study.

At the end of the study, the CI or delegate must notify the committee using the End of Study form.

All aforementioned forms are available on the ethics online applications website and can only be submitted to the committee via the website at: http://leo.lshtm.ac.uk.

Further information is available at: www.lshtm.ac.uk/ethics.

Yours sincerely,

[redacted]

ethics@lshtm.ac.uk

Appendix B

ISAC application & approval for PPI-Clopidogrel study

ISAC EVALUATION OF PROTOCOLS FOR RESEARCH INVOLVING CPRD DATA

FEEDBACK TO APPLICANTS

CONFIDENTIAL			***by e-mail***
PROTOCOL NO:	17_194R		
PROTOCOL TITLE:	Handling missing covariate data and changes in exposure status		
APPLICANT:	Ian Douglas, Associate Professor, London School of Hygiene and Tropical Medicine, ian.douglas@lshtm.ac.uk		
APPROVED ☒	**APPROVED WITH COMMENTS (resubmission not required)** ☐	**REVISION/ RESUBMISSION REQUESTED** ☐	**REJECTED** ☐
INSTRUCTIONS: ***Protocols with an outcome of 'Approved' or 'Approved with comments' do not require resubmission to the ISAC.*** **APPLICANT FEEDBACK:** **The Protocol is approved.**			
DATE OF ISAC FEEDBACK:	06/09/17		
DATE OF APPLICANT FEEDBACK:			

For protocols approved from 01 April 2014 onwards, applicants are required to include the ISAC protocol in their journal submission with a statement in the manuscript indicating that it had been approved by the ISAC (with the reference number) and made available to the journal reviewers. If the protocol was subject to any amendments, the last amended version should be the one submitted.

***** Please refer to the ISAC advice about protocol amendments provided below*****

Amendments to protocols approved by ISAC Version June 2015

During the course of some studies, it may become necessary to deviate from a protocol which has been approved by ISAC. Any deviation to an ISAC approved protocol should be clearly documented by the applicant but not all such amendments need be submitted for ISAC review and approval. The general principles to be applied in regard to the need for submission are as follows:

- Major amendments should be submitted
- Minor amendments need not be submitted (but must still be documented by the applicant and should normally be mentioned at the publication stage)

In cases of uncertainty, the applicant should contact the ISAC secretariat for advice quoting the original reference number and providing a brief explanation of the nature of the amendment(s) and underlying reason(s).

Major Amendments

We consider an amendment as major if it substantially changes the study design or analysis plan of the proposed research. An amendment should be considered major if it involves the following (although this is not necessarily an exhaustive list):

- A change to the primary hypothesis being tested in the research
- A change to the design of the study
- Additional outcomes or exposures unrelated to the main focus of the approved study*
- Non-trivial changes to the analysis strategy
- Not performing a primary outcome analysis
- Omissions from the analysis plan which may impact on important validity issues such as confounding
- Change of Chief Investigator
- Use of additional linkages to other databases
- Any new proposal involving contact with health professionals or patient or change in regard to such matters

* N.B. extensive changes in this respect will require a new protocol rather than an amendment - if in doubt please consult the Secretariat

Minor Amendments

Examples of amendments which can generally be considered minor include the following:

- Change of personnel other than the Chief Investigator (these should be notified to the Secretariat)
- A change to the definition of the study population, providing the change is mentioned and justified in the paper/output [NB previously major]
- Extension of the time period in relation to defining the study population
- Changes to the definitions of outcomes or exposures of interest, providing the change is mentioned and justified in the paper/output [NB previously major]
- Not using linked data which are part of the approved protocol, unless the linked data are considered critical in defining exposures or outcomes (in which case this would be a major amendment)

- Limited additional analysis suggested by unexpected findings, provided these are clearly presented as post-hoc
- Additional methods to further control for confounding or sensitivity analysis provided these are to be reported as secondary to the main findings
- Validation and data quality work provided additional information from GPs is not required

To submit an amendment of protocol to the ISAC, please submit the following documents to the ISAC mailbox (isac@cprd.com)
1. A covering letter providing justification for the request
2. A completed and, if necessary, updated application form with all changes highlighted; if new linkages are required the current version of the ISAC application form must be completed. Otherwise, the original form may be amended as necessary
3. **The updated protocol document containing the heading 'Amendment' at the end of it.** Please include all amendments to the protocol under this heading. No other changes should be made to the already approved document.

ISAC APPLICATION FORM

PROTOCOLS FOR RESEARCH USING THE CLINICAL PRACTICE RESEARCH DATALINK (CPRD)

For ISAC use only		
Protocol No. Submission date (DD/MM/YYYY)	....	***IMPORTANT*** *Please refer to the **guidance** for '**Completing the ISAC application form'** found on the CPRD website (www.cprd.com/isac). If you have any queries, please contact the ISAC Secretariat at isac@cprd.com.*

SECTION A: GENERAL INFORMATION ABOUT THE PROPOSED RESEARCH STUDY

1. Study Title[§] (*Please state the study title below)*
Handling missing covariate data and changes in exposure status:
[§]*Please note: This information will be published on the CPRD's website as part of its transparency policy.*

2. Has any part of this research proposal or a related proposal been previously submitted to ISAC?
Yes[*] ☒ No ☐

**If yes, please provide the previous protocol number/s below. Please also state in your current submission how this/these are related or relevant to this study.*

09_042R

3. Has this protocol been peer reviewed by another Committee? (e.g. grant award or ethics committee)
Yes[*] ☐ No ☒

**If Yes, please state the name of the reviewing Committee(s) below and provide an outline of the review process and outcome as an Appendix to this protocol :*

4. Type of Study (please tick all the relevant boxes which apply)

Adverse Drug Reaction/Drug Safety	☐	Drug Effectiveness	☐
Drug Utilisation	☐	Pharmacoeconomics	☐
Disease Epidemiology	☐	Post-authorisation Safety	☐
Health care resource utilisation	☐	Methodological Research	☒
Health/Public Health Services Research	☐	Other[*]	☐

[*]*If Other, please specify the type of study here and in the lay summary below:*

5. Health Outcomes to be Measured[§]
[§]*Please note: This information will be published on CPRD's website as part of its transparency policy.*

Please summarise below the primary/secondary health outcomes to be measured in this research protocol:

- Myocardial Infarction
- All-cause mortality
- Myocardial Infarction or all-cause mortality
-
-
-
-
-

[Please add more bullet points as necessary]

6. Publication: This study is intended for (please tick all the relevant boxes which apply)**:**

Publication in peer-reviewed journals	☒	Presentation at scientific conference	☒
Presentation at company/institutional meetings	☒	Regulatory purposes	☐
Other*	☐		

**If Other, please provide further information:*

SECTION B: INFORMATION ON INVESTIGATORS AND COLLABORATORS

7. Chief Investigator[§]

Please state the full name, job title, organisation name & e-mail address for correspondence - see guidance notes for eligibility. Please note that there can only be one Chief Investigator per protocol.

Ian Douglas, Associate Professor, London School of Hygiene and Tropical Medicine, ian.douglas@lshtm.ac.uk

[§]*Please note: The name and organisation of the Chief Investigator and will be published on CPRD's website as part of its transparency policy*

CV has been previously submitted to ISAC	☒	**CV number:**	157_15CESL
A new CV is being submitted with this protocol	☐		
An updated CV is being submitted with this protocol	☐		

8. Affiliation of Chief Investigator (full address)

London School of Hygiene and Tropical Medicine, Keppel St, London, WC1E 7HT.

9. Corresponding Applicant[§]

Please state the full name, affiliation(s) and e-mail address below:

John Tazare, London School of Hygiene and Tropical Medicine, john.tazare1@lshtm.ac.uk

[§]*Please note: The name and organisation of the corresponding applicant and their organisation name will be published on CPRD's website as part of its transparency policy*

Same as chief investigator	☐	
CV has been previously submitted to ISAC	☐	**CV number:**
A new CV is being submitted with this protocol	☒	
An updated CV is being submitted with this protocol	☐	

10. List of all investigators/collaborators[§]

Please list the full name, affiliation(s) and e-mail address* of all collaborators, other than the Chief Investigator below:

[§]*Please note: The name of all investigators and their organisations/institutions will be published on CPRD's website as part of its transparency policy*

Other investigator: John Tazare, LSHTM, john.tazare1@lshtm.ac.uk

CV has been previously submitted to ISAC	☐	**CV number:**	
A new CV is being submitted with this protocol	☒		
An updated CV is being submitted with this protocol	☐		

Other investigator: Elizabeth Williamson, LSHTM, Elizabeth.williamson@lshtm.ac.uk

CV has been previously submitted to ISAC	☒	**CV number:**	354_16S
A new CV is being submitted with this protocol	☐		
An updated CV is being submitted with this protocol	☐		

Other investigator: Liam Smeeth, LSHTM, Liam.smeeth@lshtm.ac.uk

CV has been previously submitted to ISAC	☒	**CV number:**	045_15CEPSL
A new CV is being submitted with this protocol	☐		
An updated CV is being submitted with this protocol	☐		

Other investigator:

[Please add more investigators as necessary]

Please note that your ISAC application form and protocol **must be copied to all e-mail addresses listed above at the time of submission of your application to the ISAC mailbox. Failure to do so will result in delays in the processing of your application.*

11. Conflict of interest statement*

Please provide a draft of the conflict (or competing) of interest (COI) statement that you intend to include in any publication which might result from this work

- Dr Douglas is funded by an unrestricted grant from, has consulted for and holds stock in GlaxoSmithKline.
- Professor Smeeth reports grants from Wellcome, MRC, NIHR, BHF, Diabetes UK, ESRC and the EU; grants and personal fees for advisory work from GSK, and personal fees for advisory work from AstraZeneca. He is a Trustee of the British Heart Foundation.

There are no conflicts of interest to declare

**Please refer to the International Committee of Medical Journal Editors (ICMJE) for guidance on what constitutes a COI.*

12. Experience/expertise available

Please complete the following questions to indicate the experience/ expertise available within the team of investigators/collaborators actively involved in the proposed research, including the analysis of data and interpretation of results.

	Previous GPRD/CPRD Studies	Publications using GPRD/CPRD data
None	☐	☐
1-3	☐	☐
> 3	☒	☒

Experience/Expertise available	Yes	No
Is statistical expertise available within the research team? *If yes, please indicate the name(s) of the relevant investigator(s)* Elizabeth Williamson, Ian Douglas, John Tazare	☒	☐
Is experience of handling large data sets (>1 million records) available within the research team? *If yes, please indicate the name(s) of the relevant investigator(s)* Elizabeth Williamson, Ian Douglas, Liam Smeeth	☒	☐
Is experience of practising in UK primary care available to or within the research team? *If yes, please indicate the name(s) of the relevant investigator(s)* Liam Smeeth	☒	☐

13. References relating to your study

Please list up to 3 references (most relevant) relating to your proposed study:

1. Douglas IJ, et al. Clopidogrel and interaction with proton pump inhibitors BMJ.2012;345:e4388.
2. Carpenter J, Kenward M. Multiple Imputation and its Application: Wiley; 2013.
3. Fewell, Z., M. A. Hernan, et al. Controlling for time-dependent confounding using marginal structural models. Stata Journal. 2004;4(4): 402-20.

14. Financial Sponsor of study[§]

[§]*Please note: The name of the source of funding will be published on CPRD's website as part of its transparency policy*

Pharmaceutical Industry	☐	*Please specify name and country:*
Academia	☒	*Please specify name and country:* LSHTM, UK.
Government / NHS	☐	*Please specify name and country:*
Charity	☐	*Please specify name and country:*
Other	☐	*Please specify name and country:*
None	☐	

15. Type of Institution conducting the research

Pharmaceutical Industry	☐	*Please specify name and country:*
Academia	☒	*Please specify name and country:* LSHTM
Government Department	☐	*Please specify name and country:*
Research Service Provider	☐	*Please specify name and country:*
NHS	☐	*Please specify name and country:*
Other	☐	*Please specify name and country:*

16. Data access arrangements

The financial sponsor/ collaborator* has a licence for CPRD GOLD and will extract the data ☐
The institution carrying out the analysis has a licence for CPRD GOLD and will extract the data** ☐
A data set will be provided by the CPRD[¥€] ☐
CPRD has been commissioned to extract the data and perform the analyses[€] ☐
Other: ☒
*If Other, please specify:*This study will use processed data already obtained for ISAC approved study 09_042R. Since the proposed study seeks to address methodological problems identified in the original study, it is important the same dataset is used.

**Collaborators supplying data for this study must be named on the protocol as co-applicants.*
***If data sources other than CPRD GOLD are required, these will be supplied by CPRD*
[¥]*Please note that datasets provided by CPRD are limited in size; applicants should contact CPRD (enquiries@cprd.com) if a dataset of >300,000 patients is required.*
[€]*Investigators must discuss their request with a member of the CPRD Research team before submitting an ISAC application. Please contact the CPRD Research Team on +44 (20) 3080 6383 or email (enquiries@cprd.com) to discuss your requirements. Please also state the name of CPRD Research team with whom you have discussed this request (provide the date of discussion and any relevant reference information):*

Name of CPRD Researcher Reference number (where available) Date of contact

17. Primary care data

Please specify which primary care data set(s) are required)
Vision only (Default for CPRD studies ☒ Both Vision and EMIS®* ☐
EMIS® only* ☐

Note: Vision and EMIS are different practice management systems. CPRD has traditionally collected data from Vision practice. Data collected from EMIS is currently under evaluation prior to wider release.
Investigators requiring the use of EMIS data **must discuss the study with a member of the CPRD Research team before submitting an ISAC application*

Please state the name of the CPRD Researcher with whom you have discussed your request for EMIS data:
Name of CPRD Researcher Reference number (where available) Date of contact

SECTION D: INFORMATION ON DATA LINKAGES

18. Does this protocol seek access to linked data

Yes* ☒ No☐ If No, please move to section E.

Please note that we are not seeking any new linkages, see the answer given to Question 16.

25 April 2017 Version 1.1

Research groups which have not previously accessed CPRD linked data resources **must discuss access to these resources with a member of the CPRD Research team, before submitting an ISAC application. Investigators requiring access to HES Accident and Emergency data, HES Diagnostic Imaging Dataset PROMS data and the Pregnancy Register **must** also discuss this with a member of the CPRD Research team before submitting an ISAC application. Please contact the CPRD Research Team on +44 (20) 3080 6383 or email enquiries@cprd.com to discuss your requirements **before** submitting your application.*

Please state the name of the CPRD Researcher with whom you have discussed your linkage request.

Name of CPRD Researcher Reference number (where available) Date of contact

Please note that as part of the ISAC review of linkages, your protocol may be shared - in confidence - with a representative of the requested linked data set(s) and summary details may be shared - in confidence - with the Confidentiality Advisory Group of the Health Research Authority.

19. Please select the source(s) of linked data being requested[§]

[§]Please note: This information will be published on the CPRD's website as part of its transparency policy.

- ☐ ONS Death Registration Data
- ☐ HES Admitted Patient Care
- ☐ HES Outpatient
- ☐ HES Accident and Emergency
- ☐ HES Diagnostic Imaging Dataset
- ☒ MINAP (Myocardial Ischaemia National Audit Project)
- ☐ Cancer Registration Data*
- ☐ PROMS (Patient Reported Outcomes Measure)**
- ☐ CPRD Mother Baby Link
- ☐ Pregnancy Register

- ☐ Practice Level Index of Multiple Deprivation (Standard)
- ☐ Practice Level Index of Multiple Deprivation (Bespoke)
- ☐ Patient Level Index of Multiple Deprivation***
- ☐ Patient Level Townsend Score ***
- ☐ Other**** *Please specify:*

**Applicants seeking access to cancer registration data must complete a Cancer Dataset Agreement form (available from CPRD). This should be submitted to the ISAC as an appendix to your protocol. Please also note that applicants seeking access to cancer registry data must provide consent for publication of their study title and study institution on the UK Cancer Registry website.*
***Assessment of the quality of care delivered to NHS patients in England undergoing four procedures: hip replacement, knee replacement, groin hernia and varicose veins. Please note that patient level PROMS data are only accessible by academics*
**** 'Patient level IMD and Townsend scores will not be supplied for the same study*
*****If "Other" is specified, please provide the name of the individual in the CPRD Research team with whom this linkage has been discussed.*

Name of CPRD Researcher Reference number (where available) Date of contact

20. Total number of linked datasets requested <u>including</u> CPRD GOLD

1

Number of linked datasets requested *(practice/ 'patient' level Index of Multiple Deprivation, Townsend Score, the CPRD Mother Baby Link and the Pregnancy Register should **not** be included in this count)*

Please note: Where ≥5 linked datasets are requested, approval may be required from the Confidentiality Advisory Group (CAG) to access these data

21. Is linkage to a <u>local[¥]</u> dataset with <1 million patients being requested?

Yes* ☐ No ☒

**If yes, please provide further details:*
[¥] Data from defined geographical areas i.e. non-national datasets.

22. If you have requested one or more linked data sets, please indicate whether the Chief Investigator or any of the collaborators listed in question 5 above, have access to these data in a patient identifiable form (e.g. full date of birth, NHS number, patient post code), or associated with an identifiable patient index.

Yes* ☐ No ☒

25 April 2017 Version 1.1

** If yes, please provide further details:*

23. Does this study involve linking to patient *identifiable* data (e.g. hold date of birth, NHS number, patient post code) from other sources?

Yes ☐ No ☒

SECTION E: VALIDATION/VERIFICATION

24. Does this protocol describe a purely observational study using CPRD data?

Yes* ☒ No** ☐

** Yes: If you will be using data obtained from the CPRD Group, this study does not require separate ethics approval from an NHS Research Ethics Committee.*
*** No: You may need to seek separate ethics approval from an NHS Research Ethics Committee for this study. The ISAC will provide advice on whether this may be needed.*

25. Does this protocol involve requesting any additional information from GPs?

Yes* ☐ No ☒

** If yes, please indicate what will be required:*

Completion of questionnaires by the GP[ψ] Yes ☐ No ☐
Is the questionnaire a validated instrument? Yes ☐ No ☐
If yes, has permission been obtained to use the instrument? Yes ☐ No ☐
Please provide further information:

Other (please describe)

[ψ] Any questionnaire for completion by GPs or other health care professional must be approved by ISAC before circulation for completion.

26. Does this study require contact with patients in order for them to complete a questionnaire?

Yes* ☐ No ☒

**Please note that any questionnaire for completion by patients must be approved by ISAC before circulation for completion.*

27. Does this study require contact with patients in order to collect a sample?

Yes* ☐ No ☒

** Please state what will be collected:*

SECTION F: DECLARATION

28. Signature from the Chief Investigator

- I have read the guidance on '***Completion of the ISAC application form***' and '***Contents of CPRD ISAC Research Protocols***' and have understood these;
- I have read the submitted version of this research protocol, including all supporting documents, and confirm that these are accurate.
- I am suitably qualified and experienced to perform and/or supervise the research study proposed.
- I agree to conduct or supervise the study described in accordance with the relevant, current protocol
- I agree to abide by all ethical, legal and scientific guidelines that relate to access and use of CPRD data for research
- I understand that the details provided in sections marked with ([§]) in the application form and protocol will be published on the CPRD website in line with CPRD's transparency policy.
- I agree to inform the CPRD of the final outcome of the research study: publication, prolonged delay, completion or termination of the study.

25 April 2017 Version 1.1

Name: Ian Douglas	Date: 27th July 2017	e-Signature (type name): Ian Douglas

PROTOCOL INFORMATION REQUIRED

The following sections below **must** be included in the CPRD ISAC research protocol. Please refer to the guidance on '***Contents of CPRD ISAC Research Protocols***' (www.cprd.com/isac) for more information on how to complete the sections below. Pages should be numbered. All abbreviations must be defined on first use.

Applicants must complete all sections listed below
Sections which do not apply should be completed as '*Not Applicable*'

A. Study Title[§]
[§]*Please note: This information will be published on CPRD's website as part of its transparency policy*

Handling missing covariate data and changes in exposure status in the analysis of Electronic Health Records

B. Lay Summary (Max. 200 words)[§]
[§]*Please note: This information will be published on CPRD's website as part of its transparency policy*

In recent years, data collected in routine practice by General Practitioners (GP) have become more widely used to investigate the safety and effectiveness of drugs. However, in some cases, the results are in conflict with results from more traditional studies such as randomized trials. This highlights the need to explore potential issues and biases in the analysis of routinely collected health data.

Two key issues that may be responsible for these inconsistent results are missing data and treatment switching. GPs record only health information relevant to the care of the patient, thus information required to fully address the research question may not always be available, resulting in missing data. Treatment switching refers to a patient swapping prescriptions from one treatment to another, which complicates the comparison between patients on the different treatments.

Using a recent example where the results were inconsistent with randomized trial results, we aim to apply novel statistical methods to handle the aforementioned problems to better understand when it is relevant to take account of such characteristics of the data. We will then provide guidance regarding the relative benefits of different methods of analysis for future studies.

C. Technical Summary (Max. 200 words)[§]
[§]*Please note: This information will be published on CPRD's website as part of its transparency policy*

Results from a recent study suggested between person confounding remained a problem when investigating a potential interaction between PPIs and clopidogrel; , as biologically implausible harmful associations were observed. Results from a self-controlled case series (SCCS) showing no increased risk of MI with PPI exposure were thought to be more reliable as this is not affected by between person confounding. SCCS limitations mean a more general solution is needed.

We identified treatment (PPI) switching and exclusion of potentially important confounders due to missing data as key issues.

To investigate treatment switching, we would perform "intent-to-treat", "per-protocol" and "as-treated" analyses using Cox models, incorporating probability weights accounting for participant differences between those who did and didn't change exposure status during follow-up. We will extend this idea by using more complex approaches such as marginal structural models, splitting data into, for example, 3-month intervals.

For missing data, we would initially incorporate information from confounders, previously omitted using missing categories approaches and then proceed to multiple imputation based analyses in the different analysis settings outlined above.

Finally, we would investigate methods to incorporate both issues.

11 April 2017 Version 1.1

Applicants must complete all sections listed below
Sections which do not apply should be completed as '*Not Applicable*'

Through this work we seek only to improve methodological approaches in future studies, not answer additional clinical questions.

D. Objectives, Specific Aims and Rationale

(i) Aim: To understand when different statistical methods to deal with missing data and treatment switching are relevant in the analysis of EHRs.

(ii) Objectives

- Investigate the sensitivity of results to the method used to handle treatment switching.
- Establish the robustness of results to different missingness mechanisms.
- Provide general guidance on the relative benefits of different missing data and treatment switching analyses.

(iii) Rationale

The use of EHRs has expanded considerably in recent years. Increased legislation (including by the European Union) has made it mandatory for pharmaceuticals, as part of drug licensing conditions in some circumstances, to conduct safety and effectiveness studies investigating the long-term and rare effects of medications in routine care settings. One of the main issues in observational research is adequate adjustment for confounding, and whilst there is a vast literature in more conventional settings guidance is scarce in the EHR setting where the validity of standard methods remains largely untested.

Using these data as an example we would investigate which aspects highlighted seem most vital to control for and provide guidance for other researchers based on our findings.

E. Study Background

Work from a previously approved CPRD study (protocol number: 09_042R), published in a peer reviewed journal by Douglas et[1] found a clinically important statistically significant increased risk of myocardial infarction associated with use of proton pump inhibitor among patients prescribed clopidogrel and aspirin. The pattern of associations found strongly suggested residual confounding between patients may have explained the results as they were not specific to MI and were found for both strong and weak inhibitors of cytochrome P450 3A4 (the mechanism proposed for the drug interaction). Furthermore, a self-controlled case series conducted on the same data found no evidence of increased risk. The authors concluded that the results from the cohort study reflect bias in the cohort estimate, and since the study was conducted, a meta-analysis of randomised trials has confirmed a lack of clinical effect of PPIs on MI risk, when used in combination with clopidogrel.[2]

The authors' hypothesised that inadequate adjustment for confounding is a big problem with the cohort study, which invites further methodological research. Inadequate adjustment is an unavoidable concern in observational research, however it is unclear how best to deal with it in Electronic Health Records. We plan to investigate methods that allow the inclusion of confounders previously omitted due to missing data and methods that allow for treatment switching in both regression adjustment and propensity score based approaches.[3]

Critical to missing data analyses is the careful consideration of the reasons for missingness and an investigation of the patterns of missing data. We will firstly consider simple approaches to handle missing data, such as restriction to complete record analyses and the addition of missing category indicators.[4] One approach proposed in the

Applicants must complete all sections listed below
Sections which do not apply should be completed as '*Not Applicable*'

setting of longitudinal EHRs is the two-fold conditional specification multiple imputation algorithm and we will investigate the use of this method in our data.[5] Within the propensity score framework we will look to draw from recent work by Leyrat et al[6] on the inclusion of partially observed covariates using fully conditional specification multiple imputation.

Treatment switching, or the ability to change exposure status during follow-up is a distinct characteristic of EHRs and the reason for treatment switching is often linked to underlying health. There is a vast literature of methods for dealing with time-dependent confounding and time-varying exposures in discrete longitudinal settings[7], however the application of these methods to EHRs (which are less structured) remains largely unexplored. Fewell et al[8] have looked at inverse probability weighted estimated of marginal structural models using pooled logistic regression and this would be the starting point for conducting more complicated methods where we split the data into increasingly small time intervals.

F. Study Type

Methodological study.

G. Study Design

Comparison of statistical methods to handle missing data and treatment switching in the setting of a cohort study.

H. Feasibility counts

24471 patients receiving clopidogrel and aspirin were included in the cohort used by Douglas et al[1]. Of these, 9111 (37%) also received a proton pump inhibitor from the date of first clopidogrel prescription. In total, 12439 (50%) patients received a proton pump inhibitor at some point during the study period.

I. Sample size considerations

In the cohort study published by Douglas et al, the fully adjusted hazard ratio for association between proton pump inhibitor use and incident myocardial infarction was 1.30 (95% CI: 1.12 to 1.50).[1]

We plan to use the same dataset from which this estimate was obtained, and will consider the methodological enhancements to be successful if we obtain a null result with comparable precision.

J. Data Linkage Required (if applicable):[§]

[§]*Please note that the data linkage/s requested in research protocols will be published by the CPRD as part of its transparency policy*

This study uses CPRD data linked with MINAP, approval has been granted for this linkage and confirmation is attached in Appendix 2.

K. Study population

Important Note for Sections K, L and M: We do not propose to extract data on a new study population for this study, but to re-use data already extracted for the original approved and now published study (09_042R, Douglas et al 2012). This is to ensure direct comparability between the original results, and results from the methodological development work we will conduct. Sections K, L and M purely describe what was done for the original study. All patients registered in the GPRD from 1 January 2003 receiving clopidogrel in combination with aspirin and with at least 12 months UTS observation before the first prescription for clopidogrel were eligible for inclusion. Last data collection was on the 31st July 2009. This resulted in 24471 patients receiving clopidogrel and aspirin being included in the final cohort.[1]

L. Selection of comparison group(s) or controls

Applicants must complete all sections listed below
Sections which do not apply should be completed as '*Not Applicable*'

Patients receiving clopidogrel in combination with aspirin and with at least 12 months UTS observation before the first prescription for clopidogrel were eligible for inclusion. They were classified as unexposed if they didn't receive a proton pump inhibitor in conjunction with the clopidogrel prescription. 15360 (63%) patients didn't receive a proton pump inhibitor with their first clopidogrel prescription. In total, 16900 (69%) patients had at least some follow-up time with no exposure to a proton pump inhibitor.[1]

M. Exposures, Health Outcomes[§] and Covariates

[§]*Please note: Summary information on health outcomes (as included on the ISAC application form above) will be published on CPRD's website as part of its transparency policy*

Primary Exposure: Any PPI in combination with aspirin and clopidogrel. Prescriptions for PPI's, aspirin and clopidogrel were identified by Douglas et al[1] using the code lists were outlined in the original ISAC protocol (protocol number: 09_042R, see Appendix 1).

Primary Outcome: Incident myocardial infarction (MI) was determined by Douglas et al using Myocardial Ischaemia National Audit Project (MINAP) records.[1]

Covariates: Douglas et al examined the confounding effects of the following covariates: age, sex, smoking status, alcohol status, body mass index (BMI) categorised as <20, 20-25, or >25, diabetes, coronary heart disease, peripheral vascular disease, ischaemic stroke, and cancer. Patients status for each covariate was updated as relevant at any change in exposure to a proton pump inhibitor. Other covariates considered but ultimately omitted due to missing data were blood pressure, pulse rate, lipids, HbA1c, cholesterol and NSAIDS.

N. Data/ Statistical Analysis

Originally Douglas et al used Cox regression adjusted for the covariates outlined in Section M, updating covariates as relevant at any change in exposure to proton pump inhibitor, to compare the hazard of MI amongst PPI users and non-users.[1] For our further work we will use multivariable adjusted and propensity score based Cox models, considering the covariates outline in Section M for inclusion. Where the propensity score is used, it will be constructed using the principle that predictors of the exposure and outcome, or outcome only should be included. As previously stated we will consider the methodological enhancements to be successful if we obtain a null result with comparable precision to the original cohort study result (1.30, 95% CI: 1.12 to 1.50).

To investigate treatment switching, we will perform "intent-to-treat", "per-protocol" and "as-treated" analyses, incorporating probability weights to account for participant differences between those who did and did not change exposure status during follow-up. These weights will be obtained using logistic regression and marginal structural models will be estimated as described by Fewell et al[8], fitting a pooled logistic regression model and using robust variance estimators to calculate 95% confidence intervals. We will explore splitting the follow up of patients into different intervals based on the average follow-up and see how sensitive are results are to these changes. As previously mentioned, these methods do enforce an artificial structure to the data. In EHRs information on covariates are not collected at planned intervals (of say, 3months) so methods where time is treated continuously will also be explored and comparisons made with the methods assuming discrete time points.

In the original final analysis, the following potentially important confounders were omitted due to missing data: blood pressure, pulse rate, lipids, HbA1c, cholesterol and NSAIDS. We intend to explore methods to adequately incorporate these variables and will start by applying missing data techniques within the "intent-to-treat" setting, using only the baseline covariate values. We will investigate the patterns of missing data and examine the percentage missing for each variable. Using missing category and missing indicator techniques we will look at simple ways of including the missing data and compare these results to a complete records analysis. Following this, we intend to use fully conditional specification multiple imputation and in particular the two-fold fully conditional specification algorithm implemented by Welch et al[5]. Finally, we will compare results with the use of fully conditional specification multiple imputation with the propensity score[6]. From this basis, we will be able to apply missing data methods in the more complicated but realistic "as-treated" setting. The validity of these missing data methods is related to often untestable assumptions about the nature of missingness. We will carefully consider whether these data are likely to be missing not at random (MNAR) and establish sensitivity analyses for this case since the guidance in this setting is currently not established.

Applicants must complete all sections listed below
Sections which do not apply should be completed as '*Not Applicable*'

We will finally look to combine methods of time-varying confounding and missing data into an optimal analysis, making adjustment for both. We believe this is an area which has yet to be explored in the EHR context and will help us gain an understanding of the estimands being estimated in different sensitivity analyses. This will help us provide both methods for combined adjustment as well as simpler sensitivity analyses.

O. Plan for addressing confounding

We intend to use multivariable adjustment and propensity score techniques as described in Section N. Additional potential confounders with large amounts of missing data previously excluded from the final analysis will also be included as described in Sections N and P. Furthermore, we will adjust for time-dependent confounding using methods described in Section N.

P. Plans for addressing missing data

We will investigate the patterns of missing data and explore reasons for missingness. We will conduct complete records, missing categories and multiple imputation based analyses, as described in more detail in Section N. Furthermore, we will conduct sensitivity analyses under differing assumptions.

Q. Patient or user group involvement (if applicable)

Due to the purely methodological nature of this project, and the lack of novel clinical questions to be answered, we have not sought patient involvement.

R. Plans for disseminating and communicating study results, including the presence or absence of any restrictions on the extent and timing of publication

We intend to publish full results of the study in a peer reviewed epidemiological journal.

S. Limitations of the study design, data sources, and analytic methods

There is the possibility of residual confounding, but by accounting for more potential confounders than the original final analysis we hope to minimise this.

As in all observational studies of drug use, fundamental differences between those exposed and not exposed to a drug can make comparisons difficult or invalid. We will use propensity scores to determine how similar patients prescribed PPIs are to those not prescribed PPIs amongst the cohort of aspirin and clopidogrel users. It is possible that valid comparisons may be restricted to a small subset of these patients, raising issues of generalisability. These issues will be explored and acknowledged.

Assumptions of missing data techniques are often untestable but we plan to perform sensitivity analyses to see if our inferences are valid to extreme deviations from assumptions. Complicated and numerous missingness patterns can also make the handling of such missing data more complex and this will need to be further explored.

The ability for the treatment switching methods to correctly account for confounding will rely on the amount of treatment switching present in the data. Lack of switching could lead to these methods having little impact. The use of marginal structural models also somewhat artificially imposes a structure on the data, is it yet unclear whether these methods can be applied in the EHR context.

T. References

1. Douglas IJ, et al. Clopidogrel and interaction with proton pump inhibitors BMJ.2012;345:e4388.
2. Melloni C, et al. Conflicting results between randomized trials and observation studies on the impact of proton pump inhibitors on cardiovascular events when coadministered with dual antiplatelet therapy. Circ. Cardiovasc. Qual. Outcomes. 2015;8:47-55.

Applicants must complete all sections listed below Sections which do not apply should be completed as '*Not Applicable*'
3. Austin PC. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. Multivariate Behav Res. 2011 May; 46(3): 399–424. 4. Carpenter J, Kenward M. Multiple Imputation and its Application: Wiley; 2013. 5. Welch C, et al. Application of multiple imputation using the two-fold fully conditional specification algorithm in longitudinal clinical data. Stata J. 2014;14(2):418-31. 6. Leyrat C, et al. Propensity score analysis with partially observed covariates: How should multiple imputation be used? Stat Methods Med Res. 2017: 1:962280217713032. 7. Daniel RM, et al. Methods for dealing with time-dependent confounding. Statist. Med. 2013;32:1584-1618. 8. Fewell, Z., Hernan MA, et al. Controlling for time-dependent confounding using marginal structural models. Stata J. 2004;4(4): 402-20.
List of Appendices *(Submit all appendices as separate documents to this application)* 1. Original Douglas et al protocol. 2. Original LSHTM ethics application 3. Original LSHTM ethics approval 4. Linkage approval. 5. Codelists.

Appendix C

LSHTM Ethical approval for NSAID-COX2i study

London School of Hygiene & Tropical Medicine
Keppel Street, London WC1E 7HT
United Kingdom
Switchboard: +44 (0)20 7636 8636

www.lshtm.ac.uk

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

Observational / Interventions Research Ethics Committee

Mr John Tazare
LSHTM

28 April 2020

Dear John,

Study Title: Comparison of the prevalent new user and active comparator new user designs for assessing real world safety and effectiveness of medications

LSHTM Ethics Ref: 21992

Thank you for your application for the above research project which has now been considered by the Observational Committee via Chair's Action.

Confirmation of ethical opinion

On behalf of the Committee, I am pleased to confirm a favourable ethical opinion for the above research on the basis described in the application form, protocol and supporting documentation, subject to the conditions specified below.

Conditions of the favourable opinion

Approval is dependent on local ethical approval having been received, where relevant.

Approved documents

The final list of documents reviewed and approved is as follows:

Document Type	File Name	Date	Version
Protocol / Proposal	LSHTM Protocol V1.0	10/04/2020	v1.0
Investigator CV	JT_CV_2020	10/04/2020	v1.0
Investigator CV	cv_elizabeth_williamson	10/04/2020	v1.0
Investigator CV	ID CV LEO	10/04/2020	v1.0
Local Approval	19_273_ISAC feedback	10/04/2020	v1.0
Covering Letter	Cover Letter	27/04/2020	1.0

After ethical review

The Chief Investigator (CI) or delegate is responsible for informing the ethics committee of any subsequent changes to the application. These must be submitted to the committee for review using an Amendment form. Amendments must not be initiated before receipt of written favourable opinion from the committee.

The CI or delegate is also required to notify the ethics committee of any protocol violations and/or Suspected Unexpected Serious Adverse Reactions (SUSARs) which occur during the project by submitting a Serious Adverse Event form.

An annual report should be submitted to the committee using an Annual Report form on the anniversary of the approval of the study during the lifetime of the study.

At the end of the study, the CI or delegate must notify the committee using the End of Study form.

All aforementioned forms are available on the ethics online applications website and can only be submitted to the committee via the website at: http://leo.lshtm.ac.uk.

Further information is available at: www.lshtm.ac.uk/ethics.

Yours sincerely,

[redacted]

ethics@lshtm.ac.uk
http://www.lshtm.ac.uk/ethics/

Appendix D

ISAC application & approval for NSAID-COX2i study

Medicines & Healthcare products Regulatory Agency

CPRD

INDEPENDENT SCIENTIFIC ADVISORY COMMITTEE (ISAC) PROTOCOL APPLICATION FORM

PART 1: APPLICATION FORM

IMPORTANT

Both parts of this application must be completed in accordance with the guidance note 'Completion of the ISAC Protocol Application Form', which can be found on the CPRD website (https://cprd.com/research-applications).

FOR ISAC USE ONLY	
Protocol No. -	**Submission date -**

GENERAL INFORMATION ABOUT THE PROPOSED RESEARCH STUDY

1. Study Title (Max. 255 characters including spaces)

Comparison of the prevalent new user and active comparator new user designs for assessing the real-world safety and effectiveness of medications

2. Research Area (place '**X**' in all boxes that apply)

Drug Safety	X	Economics	
Drug Utilisation		Pharmacoeconomics	
Drug Effectiveness	X	Pharmacoepidemiology	X
Disease Epidemiology		Methodological	X
Health Services Delivery			

3. Chief Investigator

Title:	Dr
Full name:	Daniel C Gibbons
Job title:	Manager, VEO Data, Methods & Analytics
Affiliation/organisation:	GlaxoSmithKline
Email address:	Daniel.c.gibbons@gsk.com
CV Number (if applicable):	292_19
Will this person be analysing the data? (Y/N)	Y

4. Corresponding Applicant

Title:	Mr
Full name:	John Tazare
Job title:	PhD Candidate (London School of Hygiene & Tropical Medicine, LSHTM)

Medicines & Healthcare products Regulatory Agency

CPRD

	UK University Worker (GlaxoSmithKline)
Affiliation/organisation:	LSHTM / GSK
Email address:	John.tazare1@lshtm.ac.uk
CV Number (if applicable):	448_17
Will this person be analysing the data? (Y/N)	Y

Medicines & Healthcare products Regulatory Agency

CPRD

5. List of all investigators/collaborators

Title:	Dr
Full name:	John Logie
Job title:	Director, VEO Data, Methods & Analytics
Affiliation/organisation:	GlaxoSmithKline
Email address:	John.w.logie@gsk.com
CV Number (if applicable):	049_16CESL
Will this person be analysing the data? (Y/N)	N

Title:	Dr
Full name:	Elizabeth Williamson
Job title:	Associate Professor of Medical Statistics
Affiliation/organisation:	LSHTM
Email address:	Elizabeth.williamson@lshtm.ac.uk
CV Number (if applicable):	354_16S
Will this person be analysing the data? (Y/N)	N

Title:	Dr
Full name:	M Sanni Ali
Job title:	Assistant Professor of Epidemiology
Affiliation/organisation:	LSHTM
Email address:	Sanni.ali@lshtm.ac.uk
CV Number (if applicable):	070_15CS
Will this person be analysing the data? (Y/N)	N

Title:	Professor
Full name:	Ian Douglas
Job title:	Professor of Pharmacoepidemiology
Affiliation/organisation:	LSHTM
Email address:	Ian.douglas@lshtm.ac.uk
CV Number (if applicable):	157_15CESL
Will this person be analysing the data? (Y/N)	N

Title:	Professor
Full name:	Liam Smeeth
Job title:	Professor of Clinical Epidemiology
Affiliation/organisation:	LSHTM
Email address:	Liam.smeeth@lshtm.ac.uk
CV Number (if applicable):	045_15CEPSL
Will this person be analysing the data? (Y/N)	N

Medicines & Healthcare products Regulatory Agency

CPRD

6. Experience/expertise available

List below the member(s) of the research team who have experience with CPRD data.

Name(s):
Daniel Gibbons
John Tazare
John Logie

List below the member(s) of the research team who have statistical expertise.

Name(s):
John Tazare
M Sanni Ali
Elizabeth Williamson

List below the member(s) of the research team who have experience of handling large datasets (greater than 1 million records).

Name(s):
Daniel Gibbons
Ian Douglas
John Tazare

List below the member(s) of the research team, or supporting the research team, who have experience of practicing in UK primary care.

Name(s):
Liam Smeeth

ACCESS TO THE DATA

7. Sponsor of the study

Institution/Organisation:	London School of Hygiene & Tropical Medicine
Address:	Keppel Street, London, WC1E 7HT

8. Funding source for the study

Same as Sponsor?	Yes	X	No	
Institution/Organisation:	MRC National Productivity Investment Fund via LSHTM			
Address:	Keppel Street, London, WC1E 7HT, UK			

Medicines & Healthcare products Regulatory Agency

CPRD

9. Institution conducting the research

Same as Sponsor?	Yes		No	**X**
Institution/Organisation:	GlaxoSmithKline R&D			
Address:	Stockley Park West, 1-3 Ironbridge Road, Uxbridge, UB11 1BT, UK			

10. Data Access Arrangements

Indicate with an '**X**' the method that will be used to access the data for this study:

Study-specific Dataset Agreement	

Institutional Multi-study Licence	X
Institution Name	GlaxoSmithKline R&D
Institution Address	Stockley Park West, 1-3 Ironbridge Road, Uxbridge, UB11 1BT, UK

Will the dataset be extracted by CPRD?

Yes		No	X

If yes, provide the reference number:

11. Data Processor(s):

Processing	X
Accessing	X
Storing	X

Processing area (UK/EEA/Worldwide)	Worldwide
Organisation name	GlaxoSmithKline R&D
Organisation address	Stockley Park West, 1-3 Ironbridge Road, Uxbridge, UB11 1BT, UK

INFORMATION ON DATA

12. Primary care data (place '**X**' in all boxes that apply)

CPRD GOLD	X	CPRD Aurum	

Reference number (if applicable):

13. Please select any linked data or data products being requested

Patient Level Data (place '**X**' in all boxes that apply)

ONS Death Registration Data	X

Medicines & Healthcare products Regulatory Agency

CPRD

HES Admitted Patient Care	X		
HES Outpatient			
HES Accident and Emergency		NCRAS Cancer Registration Data	
HES Diagnostic Imaging Dataset		NCRAS Cancer Patient Experience Survey (CPES) data	
HES PROMS (Patient Reported Outcomes Measure)		NCRAS Systemic Anti-Cancer Treatment (SACT) data	
CPRD Mother Baby Link		NCRAS National Radiotherapy Dataset (RTDS) data	
Pregnancy Register		NCRAS Quality of Life Cancer Survivors Pilot (QOLP)	
Mental Health Data Set (MHDS)		NCRAS Quality of Life Colorectal Cancer Survivors (QOLC)	

Area Level Data (place '**X**' in one Practice / Patient level box that may apply)

Practice level (UK)		**Patient level (England only)**	
Practice Level Index of Multiple Deprivation		Patient Level Index of Multiple Deprivation	X
Practice Level Index of Multiple Deprivation (index other than the most recent)		Patient Level Index of Multiple Deprivation Domains	
Practice Level Index of Multiple Deprivation Domains		Patient Level Carstairs Index for 2011 Census	
Practice Level Carstairs Index for 2011 Census (Excluding Northern Ireland)		Patient Level Townsend Score	
2011 Rural-Urban Classification at LSOA level		2011 Rural-Urban Classification at LSOA level	

Reference / Protocol number (where applicable):

14. Are you requesting linkage to a dataset not listed above?

Yes	**X**	No	

If yes, provide the Non-Standard Linkage reference number:

2004 Rural-Urban Classification at LSOA level

15. Does any person named in this application already have access to any of these data in a patient identifiable form, or associated with an identifiable patient index?

Yes		No	**X**

Medicines & Healthcare products Regulatory Agency

CPRD

If yes, provide further details:

VALIDATION/VERIFICATION

16. Does this protocol describe an observational study using purely CPRD data?

Yes	**X**	No	

17. Does this protocol involve requesting any additional information from GPs, or contact with patients?

Yes		No	**X**

If yes, provide the reference number:

Medicines & Healthcare products Regulatory Agency

CPRD

PART 2: PROTOCOL INFORMATION

Applicants must complete all sections listed below Applications with sections marked 'Not applicable' without justification will be returned as invalid
A. Study Title (Max. 255 characters, including spaces) Comparison of the prevalent new user and active comparator new user designs for assessing the real-world safety and effectiveness of medications
B. Lay Summary (Max. 250 words) Investigating the real-world safety and effectiveness of medications is an important, post-licensing, stage of drug development. These studies give us more in-depth knowledge, especially surrounding the long-term and rare effects of medications, and help us to characterise patient experiences outside the controlled setting of a clinical trial. Frequently, we will want to compare users of a newly licensed drug to users of an older drug which is prescribed for similar reasons. Traditionally, studies of this type are designed as follows. We identify a population of patients who have no prior usage of the two drugs and compare initiators of the newer drug to initiators of the older drug. Patients are then followed up and the risk of a particular outcome compared between the two groups. However, this approach excludes a large number of initiators of the newer drug that previously received the older drug. This can lead to concerns surrounding the representativeness of the study population compared to the usage of the drugs in practice. A recently proposed study design aims to address this limitation and explicitly accounts for patients who switch from the older drug to the newer drug. In this study, we will provide an assessment of the results obtained from the traditional design against variations of the newly developed study design. This will help to inform us of the potential for this design to be used in future studies.

Medicines & Healthcare products Regulatory Agency

CPRD

C. Technical Summary (Max. 300 words)

The active comparator new user (ACNU) study design has become the gold-standard for conducting cohort studies to assess the real-world safety and effectiveness of medications [1]. The attractiveness of the ACNU approach is largely due to the baseline washout period (mimicking a clinical trial) and using an active comparator to reduce confounding by indication [1].

However, one issue with the ACNU is the selection of a suitable comparator drug. Often the comparator is an older drug that has been on the market for a long time. In the patient population, this means that many new users of the study drug are not, in fact, treatment naïve but have instead switched to the newer drug from the old comparator. The ACNU would typically exclude those who switched from the comparator drug to the study drug and this can result in investigators mischaracterising the real world patient population [2]. The newly proposed Prevalent New User (PNU) design aims to address this limitation by incorporating patients who have switched from the older drug to the newer drug[2].

In this study, we aim to provide an assessment of the PNU design and proposed variations of defining exposure sets compared to the existing ACNU design[2,3]. Furthermore, we will investigate the use of high-dimensional propensity scores (hd-PS) for confounder adjustment in the context of PNU designs and provide initial guidance to investigators planning to combine these approaches[4].

D. Outcomes to be Measured

This is a methodological study. We assess several study designs in the context of a cohort study, with the following outcome:

Upper gastrointestinal bleeding leading to hospitalisation or death. .

Medicines & Healthcare products Regulatory Agency

CPRD

E. Objectives, Specific Aims and Rationale

Rationale:
The overall aim of this study is to provide guidance for the application of PNU design variations.

Aim 1: To provide an assessment of the PNU design and proposed variations compared to the ACNU design.

Objectives:

- We will compare results obtained from prescription-based and time-based exposure set approaches to implementing the PNU design to those obtained by the ACNU design.
    - Question of interest: To examine the risk of upper gastrointestinal (GI) bleed leading to hospitalization or death in COX-2 inhibitor (newer/study drug) versus NSAID (older/comparator drug) users.
    - ACNU:
        - New users of COX-2 inhibitors versus new users of NSAIDs, with no prior treatment of either drugs in the 12 months before initiation.
    - PNU:
        - Incident new users of COX-2 inhibitors (i.e. without prior use of NSAIDs) and prevalent new users of COX-2 inhibitors (i.e. COX-2 inhibitor users with prior use of NSAIDs) versus new users of NSAIDs.

- We will investigate the utility of a recent development to the PNU design, so-called hybrid approaches, which propose extending the existing methods by simultaneously considering both duration of prior treatment and cumulative prior dose.
- Describe the populations identified by different study design approaches

Aim 2: To empirically investigate different methods for confounder adjustment in the context of PNU design.

Objectives:

- We will compare the results obtained from the use of high-dimensional propensity score (hd-PS) approaches for confounder adjustment to that of a propensity score model including only investigator-led covariates.

Medicines & Healthcare products Regulatory Agency

CPRD

F. Study Background

The active comparator new user (ACNU) study design is frequently used to examine the real-world safety and effectiveness of medications[1]. ACNU compares initiators of two therapies, that are both indicated and prescribed for the same indication, with no prior use of the drugs of interest.

However, ACNU has been criticised for excluding a large number of initiators of the newer study drug that were previously on the older comparator treatment. This can lead to concerns surrounding the representativeness of the study population compared to the real-world use of the drugs in practice.

The prevalent new user (PNU) design has been proposed to address this limitation by also including initiators of the new drug who were previously on the older treatment thereby aiming to provide a more comprehensive assessment of relative drug effects (and implemented in CPRD)[2]. Whilst the PNU design has the potential to answer a wide range of questions, given its infancy, there are currently only a few examples of this study design being implemented[2,5,6]. There is also a lack of clear guidance on applying this study design to new settings. Furthermore, given the added computational and resource cost of implementing the PNU design compared to the ACNU design, it is important to explore whether this approach gives different conclusions that may affect interpretation of evidence, especially given the increasing number of PNU design variations being proposed[5,7,8].

A separate issue with PNU design is how to adequately account for confounding. The PNU uses propensity scores that incorporate time-varying patient information measured at carefully defined points in time (so-called time-conditional propensity scores) to account for confounding. However, this approach relies on the correct identification and specification of confounders by investigators. The clinical decision when switching a patient between treatments is undoubtedly complex and capturing the reasons behind switching is likely to be a challenge, especially if hard to measure concepts (e.g. frailty) are deemed important. We will investigate the use of high-dimensional propensity score (hd-PS)[4] approaches for improving confounder adjustment in PNU designs.

This study will provide guidance for the application of PNU design variations.

The association between non-steroidal anti-inflammatory drug (NSAIDs) and cyclo-oxygenase-2 (COX-2) inhibitor use on the risk of upper GI bleeding leading to hospitalisation (simplified to GI bleeding throughout the protocol) will be used as an illustrative example throughout this study. Evidence accrued through randomised trial and observational data strongly suggest a decreased risk of GI bleeding associated with COX-2 inhibitor use compared to NSAIDs[9–12]. The established association between these drugs and GI bleeding will serve as a useful benchmark for the results obtained to be meaningfully compared to.

G. Study Type

This is a methodological study focusing on two aspects of applying PNU designs:

1. Assessing the proposed variations of the PNU design to the standard ACNU study design.
2. Comparing methods for confounder adjustment in this setting: traditional investigator-led time-conditional propensity scores versus time-conditional high-dimensional propensity scores.

Medicines & Healthcare products Regulatory Agency

CPRD

H. Study Design

This work is focussed on the comparison of the ACNU and PNU variations of cohort study design.

- The ACNU design measures the effect of initiating COX-2 inhibitors at time zero versus initiating NSAIDs at time zero.
- Conditional on prior usage of NSAIDs, the PNU study design measure the effect of initiating COX-2 inhibitors versus continuing NSAIDs.

I. Feasibility counts

We conducted an internal feasibility count using CPRD data. Over the study period, 2000-2004, there is a sharp increase in COX-2 inhibitor prescriptions following their introduction in 1999. In 2000, there were similar numbers of COX-2 inhibitor and NSAID prescriptions (~12,500) per quarter. By the end of 2004, there were approximately 4 times as many COX-2 inhibitor prescriptions (~100,000) compared to NSAID prescriptions (~25,000) per quarter.

J. Sample size considerations

In terms of the methodological work outlined, the numbers obtained from our feasibility count more than adequately allow this work to be successfully carried out. These numbers reflect conservative estimates of our realised sample size since we will draw information from a 4-year period. Furthermore, by design, the PNU incorporates patients who switch from NSAIDs to COX-2 inhibitors as opposed to solely initiators of the two medications, which increases the sample size.

Medicines & Healthcare products Regulatory Agency

CPRD

K. Planned use of linked data (if applicable):

We require linkage to ONS Death Registration Data, HES Admitted Patient Care, HES Accident and Emergency, Patient Level Index of Multiple Deprivation and 2004 Rural-Urban Classification at LSOA level.

ONS Death Registration Data:
- Ascertaining deaths due to upper GI bleeding

HES Admitted Patient Care:
- Ascertaining outcome
- Defining propensity score covariates
- High-dimensional propensity score data dimension

HES Accident and Emergency:
- Ascertaining outcome
- Defining propensity score covariates
- High-dimensional propensity score data dimension

Patient Level Index of Multiple Deprivation:
- Defining propensity score covariates

2004 Rural-Urban Classification at LSOA level:
- Defining propensity score covariates

Medicines & Healthcare products Regulatory Agency

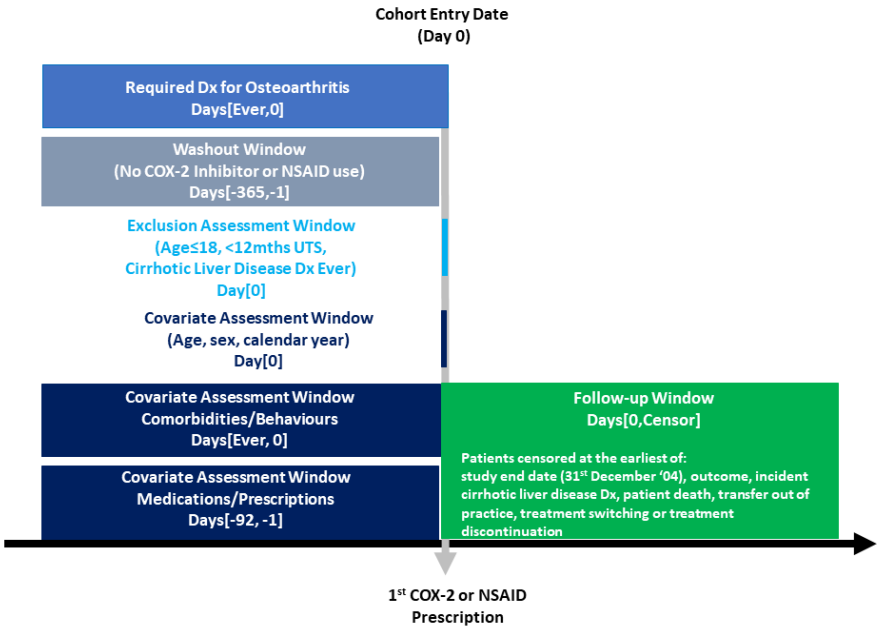
CPRD

L. Definition of the Study population

The study population will be osteoarthritis patients (see Appendix C for codelist) aged 18 years or older, who initiate NSAIDs or COX-2 inhibitors between 1st January 2000 and 31st December 2004. Patients will need to have at least 12 months of up-to-standard data available prior to cohort entry. This will allow us to assess baseline confounder information and distinguish new users of medications from prevalent users or switchers.

ACNU Analysis:
Patients enter the study as incident new users of either COX-2 inhibitors or NSAIDs. A washout window is defined prior to index date and patients are excluded if they have had a prescription for NSAIDs or COX-2 inhibitors in the previous 365 days. Patients are also excluded if they have had a diagnosis for cirrhotic liver disease (defined by the presence of code in code list, see Appendix D) ever in their medical history prior to cohort entry. Finally, patients are excluded if they are aged ≤18 or have < 12 months up-to-standard follow up available prior to index. Patients are then followed and censored at the earliest of outcome, study end date, death, incident cirrhotic liver disease, transfer out of practice, treatment switching or treatment discontinuation. Treatment discontinuation is defined as absence of a refill prescription 30 days after the end of the previous prescription.
Baseline conditions will be established as outlined in Section N. This approach is summarised in Figure L-1.

Figure L-1: Schematic showing active comparator study design for a study of NSAID and COX-2 inhibitor use on upper GI bleeding risk



PNU Analysis:

Medicines & Healthcare products Regulatory Agency

CPRD

Once the exposure sets are formed, patients are matched based on the time-conditional propensity score. This results in each study drug user having a matched comparator.

Cohort entry is the date of the 1st prescription for the study drug and the corresponding prescription date for the matched comparator drug user [2]. Patients are then followed-up and censored at the earliest of outcome, study end date, death, incident cirrhotic liver disease, transfer out of practice, treatment switching (i.e. including if a study drug user switches back to the comparator drug) or treatment discontinuation (defined as described in ACNU analysis).

Applying exclusion criteria is more challenging in PNU designs and the potential for selection bias is high [2]. To avoid this, matched comparators are identified in chronological order. The first (in calendar time) new study drug user is matched first and at this time they are assessed for whether they have any of the exclusion criteria. If they do, they are excluded from any further selection into the cohort analysis [2]. If they do not, they are matched to the comparator with the closest propensity score. If this comparator has a history of the exclusion event, they are excluded from all future exposure sets and the next closest match (without a history of the exclusion event) is chosen [2].

ACNU versus PNU Analysis:
The PNU aims to recover data otherwise lost to censoring and baseline exclusions (Figure L-2).

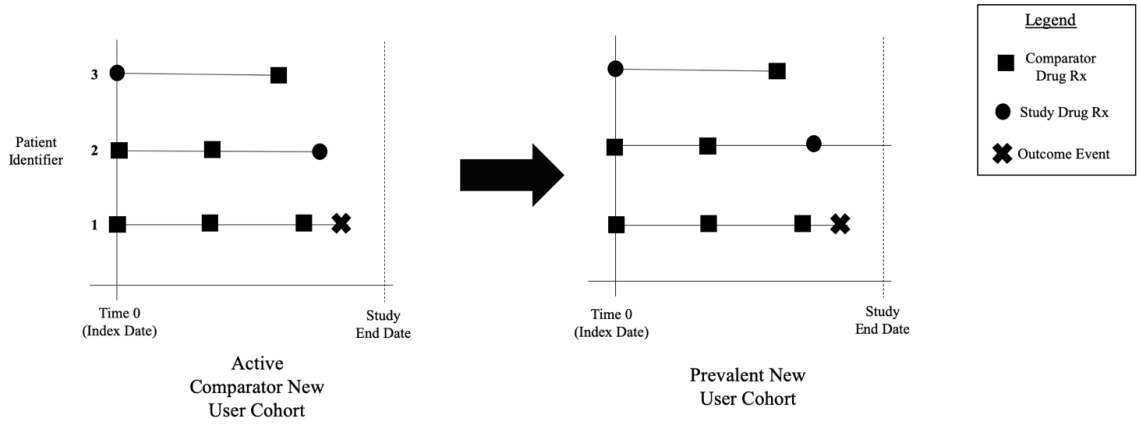
We will now briefly describe how the PNU design changes the information recorded for the 3 hypothetical patients in Figure L-2.

- Patient 1: Continues to receive comparator (older drug) prescriptions throughout the study period before having an event after their third prescription. This patient would be treated the same in ACNU and PNU designs.
- Patient 2: Would be censored in an ACNU design when switching from the comparator drug to the study drug (newer drug). However, in a PNU study this patient would be a 'Switcher' and we would incorporate follow-up time after this switch until the earliest of study end date, outcome or a censoring event.
- Patient 3: Switches from the study drug to the comparator drug. In both ACNU and PNU designs this patient is censored at this point.

Medicines & Healthcare products Regulatory Agency

CPRD

Figure L-2: Illustrative example of how the PNU design can make use of more of the available data compared to ACNU.



M. Selection of comparison group(s) or controls

Throughout this study, the comparison cohort will be patients prescribed an NSAID (see Appendix A2 for a codelist). In ACNU analyses, these patients will be incident users only, with no recorded NSAID or COX-2 inhibitor use prior to the recruitment period. In PNU analyses these patients will be a mix of incident and prevalent users.

ACNU Analysis:

Each COX-2 inhibitor user will be matched to an NSAID user using propensity score matching via a nearest-neighbour matching algorithm.

PNU Analysis:
Once the exposure sets are formed, patients are matched 1:1 based on the time-conditional propensity score (explained in Section O). This results in each COX-2 inhibitor user being matched to an NSAID user.

Comparison of confounder adjustment method:

The same procedures described above will apply but COX-2 and NSAID users will be matched on high-dimensional propensity score (further details of how these are applied can be found in Section O).

Medicines & Healthcare products Regulatory Agency

CPRD

N. Exposures, Outcomes and Covariates

Exposure definition:

All patients receiving first prescription for NSAIDs or COX-2 inhibitors (Defined using Appendices A1 & A2 for codelist) between 1st January 2000 and 31st December 2004.

We identified product codes referring to use of either NSAIDs or COX-2 inhibitors (See Appendices A1 & A2 for codelist).

Continuous exposure will be defined as following:
1st Prescription Date + Duration of Prescription + Duration of any successive overlapping prescriptions (of same drug) + 30 days

Outcome definition:

The study outcome will be first occurrence of an upper GI bleed (Defined using codelist in Appendix B) leading to hospitalisation or death. This will be defined as a binary variable which will be analysed in a time-to-event framework.

We will use linked HES and ONS mortality data to define this outcome using a mixture ICD-10 codes (see Appendix B).

Covariates

We will consider including the following covariates in our propensity score model (see Section O):

- **Demographics:** Age, sex, index of multiple deprivation score rank decile, body mass index, smoking status, alcohol consumption
- **Comorbidities/ behaviours (any recording in patient history on or prior to cohort entry):** Hypertension, chronic renal failure, inflammatory bowel disease, gastrointestinal tract tumours, coagulopathies, gastro-oesophageal reflux disease, diabetes, heart failure, previous upper GI bleed (defined in Read and ICD-10), number of admissions to A&E in previous 6 months.
- **Medications/therapies (any recording in the 3 months prior to cohort entry):** anticoagulants, systemic corticosteroids, proton pump inhibitors, H2 antagonists, coronary angioplasty, selective serotonin reuptake inhibitors, statins and clopidogrel
- **Other:** Calendar year

Medicines & Healthcare products Regulatory Agency

CPRD

O. Data/ Statistical Analysis

1. Comparison of study design approaches:
   - Main Analysis
   a) ACNU

We will analyse the HR for the association between COX-2 inhibitor and NSAID use on upper GI bleeding risk using Cox models, adjusting for confounders using propensity scores [13]. The propensity score will be estimated using multivariable logistic regression to model the relationship between treatment and potential confounders. A propensity score-matched sample will be created by matching each COX-2 inhibitor user to an NSAID user using a nearest-neighbour matching algorithm.

   b) PNU

For each of the exposure set definitions under investigation (time-based, prescription-based, hybrid)[2,3], we will analyse the HR for the association between COX-2 inhibitor and NSAID use on upper GI bleeding risk using Cox models. Furthermore, we will apply robust standard errors to account or the fact that patients may be used as both comparators and study drug users.

- Sensitivity Analysis

We will conduct a sensitivity analysis excluding patients at high risk of an upper GI bleed. In addition to existing criteria the following exclusions applied:
- History of a coagulopathy
- Currently prescribed anticoagulants

As a sensitivity analysis, we will propose trimming with cut points at the 1st and 99th percentiles of the PS distribution in the treated and untreated patients, respectively[14].

In both the ACNU and PNU approaches patients will be censored at the occurrence of cirrhotic liver disease, treatment switching (only COX-2 inhibitor to NSAID switching in PNU designs) and discontinuation. These censoring events are unlikely to be random and this violates the non-informative censoring assumption of the Cox model. In practice, it is unclear how violations of this assumption are likely to affect these designs. To investigate the possible consequences, we will conduct a sensitivity analysis incorporating inverse probability of censoring weights[15].

2. Comparison of approaches to confounder adjustment:
   - Main Analysis

The ACNU and PNU approaches mainly focus on handling issues arising in the design stage of a study. Whilst the use of active comparators mitigates some confounding bias, residual confounding may remain even after adjustment for the set of investigator-chosen factors.

The high-dimensional propensity score (hd-PS) is a semi-automated approach to confounder selection in large healthcare databases [4] . The hd-PS supplements investigator chosen covariates by including a number (usually several hundred) of proxy variables which aim to capture underlying constructs that are important for confounder adjustment, e.g. frailty. The hd-PS has become widely used in ACNU designs in

Medicines & Healthcare products Regulatory Agency

CPRD

a number of settings [16] and singularly implemented in the context PNU designs [6]. However, how best to apply this approach in the PNU setting is unclear. Furthermore, recent developments have indicated the potential benefit of hd-PS approaches for improving confounder adjustment in UK EHRs [17].

There are a number of investigator decisions to be made when using this approach and we provide a summary of how we plan to implement hd-PS alongside the steps of the algorithm below:

Step 1: Identify p dimensions which reflect aspects of care.
Previous work by Tazare et al identified three dimensions in Clinical Practice Research Datalink (CPRD) separating sign, symptoms and diagnoses, referrals and prescribing patterns [17]. In order to obtain clinically meaningful proxies we mapped information recorded in the Read coding system to ICD-10 (i.e. the Clinical and Referral dimensions) using crossmaps developed by NHS Digital [18]. In this study we will also have HES data available. We will incorporate this extra information by forming dimensions separating discharge and A&E information. Dimensions and their coding systems are outlined below:

- CPRD Dimensions:
  Clinical (Read transcoded to ICD-10): Signs, symptoms and diagnoses
  Referral (Read transcoded to ICD-10 codes): Indicate a possible escalation in care
  Prescriptions (BNF codes): Patterns of drug usage
- HES Dimensions:
  Discharge (ICD-10): Diagnoses/disposition information recorded on discharge
  A&E (ICD-10): A&E diagnoses

Step 2: Sort codes by prevalence within each dimension.
- The top 200 most prevalent codes are selected from each dimension

Step 3: Assess recurrence of codes
- The recurrence of codes is assessed using 3 indicators of frequency denoting whether a patient has a code measured once, sporadically or frequently [4].
- We extend the definition of the once category to better characterise the recording of UK EHR data [17]. This means capturing if a code is recorded Ever in a patient's entire medical history for the lowest category. All other categories are assessed within the time-window defined (typically 12 months).

Step 4: Prioritise the resulting hd-PS covariates
- Covariates are prioritised using the Bross Formula, which prioritises covariates with most potential to bias the treatment-outcome relationship of interest [19,20]

Step 5: Select the top k covariates
- The top 500 covariates are included in the propensity score alongside the investigator-chosen covariates. Sensitivity analyses will assess the robustness of the results to the number of covariates chosen.

Step 6: Standard propensity score analysis conducted
- Estimated propensity score are incorporated to estimate the desired treatment effect

We propose to further adjust our primary analyses for 500 empirically chosen covariates using the hd-PS

Medicines & Healthcare products Regulatory Agency

CPRD

procedure outlined above. Results of this secondary analysis will inform us of the possible utility of hd-PS approaches in the context of PNU designs.

- Sensitivity Analysis

We will vary the time-window (12 months is the default) used to identify covariates to 6 and 24 months. We will also investigate the robustness of results to the number of covariates chosen (250, 500 and 750).

3. Descriptive Statistics:

We will use simple counts to describe how many COX-2 inhibitor new users, previously NSAID users (i.e. switchers) are excluded under the traditional ACNU design.

It will be important to investigate the comparability between patients in the COX-2 inhibitor and NSAID treatment groups both between and within the ACNU and PNU approaches.

For the ACNU design we will compare the baseline characteristics between incident new COX-2 inhibitor users and incident new NSAID users before and after propensity score matching. Whenever we compare characteristics before and after propensity score matching, absolute standardised differences (ASD) will be used to assess the balance of characteristics achieved. ASDs less than 0.1 levels typically indicate good balance [21].

For the PNU design we will compare the baseline characteristics between each of the methods for deriving exposure sets (i.e. Time-based, prescription-based and hybrid), before and after propensity matching. We will randomly sample one comparator (NSAID) prescription, representing an NSAID user, from each exposure set before propensity score matching to generate the unmatched NSAID user group formed from all exposure sets [5,6].

It is also of interest to compare the matched group of patients generated from each of the methods for deriving exposure sets. After propensity score matching, we will compare the baseline characteristics between COX-2 inhibitor and NSAID stratifying by incident or prevalent new user status.

Medicines & Healthcare products Regulatory Agency

CPRD

P. Plan for addressing confounding

Propensity scores (time-conditional when referring to PNU designs) will be used to adjust for confounding[2,13].

PNU studies generate exposure sets for each of the switchers in the study. Each exposure set comprises of the study drug user (switcher) and a set of comparator drug users identified via one of the three exposure set approaches (time-based, prescription-based or hybrid) [2,5].

In PNU study designs, patient confounder information is identified within each of the exposure sets before a single conditional logistic regression model is fitted over all the exposure sets to estimate the propensity score. Since individuals will often appear in several exposure sets, the phrase "time-conditional" is used to acknowledge the use of time-varying confounder information measured within different exposure sets [2]. As outlined by Suissa et al [2], given the size and number of exposure sets there can be a computational challenge fitting this model. Sampling can be used to overcome this issue. We will select random samples of 100 comparators to estimate the time-conditional propensity scores. The coefficients from this model will then be applied to all patients in the exposure sets, not just those who are sampled [2].

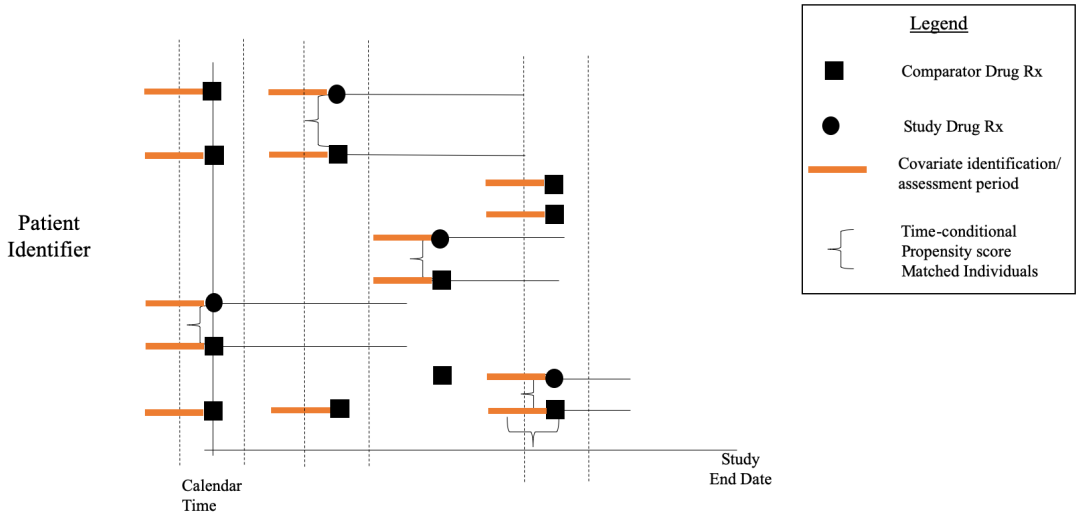
The estimated time-conditional propensity score for new users is the probability of a patient initiating COX-2 inhibitors. Whereas for switchers, the time-conditional propensity score refers to the probability of a patient switching from NSAIDs to COX-2 inhibitors.

These time-conditional propensity scores are used, for each exposure set, to identify and match the individual with the closet value of the propensity score to the switcher. This process is performed chronologically and once a patient has been matched (and selected into the comparator group) they are not considered for further inclusion [2]. This results in a study cohort containing both incident and prevalent new users matched a comparator drug user (see Figure P-1).

Figure P-1: Illustrative example of covariate identification and time-conditional propensity score matching in PNU designs.

Medicines & Healthcare products Regulatory Agency

CPRD



Prevalent New User Cohort

The ACNU and PNU approaches mainly focus on handling issues arising in the design stage of a study. Whilst the use of active comparators mitigates some confounding bias, residual confounding may remain even after adjustment for the set of investigator-chosen factors. This motivates our secondary analysis, outlined above, which will investigate the use of hd-PS approaches in the context of PNU designs to further account for residual confounding.

Q. Plans for addressing missing data

Given our previous experience using CPRD and HES data, we do not anticipate missing data on the outcome of interest.

However, we do expect a small amount of missing data on body mass index, smoking status and alcohol consumption. We will handle missing data using a missing indicator approach (recent developments indicate that this is often a sensible approach in studies involving electronic health record)[22] and run a sensitivity analysis restricting to complete-case only.

R. Patient or user group involvement

This work is focussed on outstanding methodological questions relating to the PNU design. Given this focus, we have chosen to investigate these questions in the context of a known association supported by a vast existing literature. For these reasons, patient/user groups will not be engaged. This is not to say that future studies utilising PNU designs will not engage with patient/user groups.

Medicines & Healthcare products Regulatory Agency

CPRD

S. Plans for disseminating and communicating study results, including the presence or absence of any restrictions on the extent and timing of publication

We plan to publish at least one article in a peer-reviewed scientific journal. Results will also be presented at conferences and institutional meetings.

Conflict of interest statement:
IJD has research grants from and holds shares in GlaxoSmithKline
DCG and JWL are employees of and hold shares in GlaxoSmithKline

T. Limitations of the study design, data sources, and analytic methods

Whilst we will adjust for confounders via propensity scores and supplement these using hd-PS identified covariates, the observational nature of this study means we cannot rule out residual confounding, however these concerns are not unique to this study. Given the methodological focus of our work we have chosen a drug-outcome relationship with a vast literature to benchmark our results against. This will help us understand if results from this exploratory work on PNU designs are in keeping with existing literature.

Our outcome requires hospitalisation and as such, information on treatment received during hospital stay will not be available in primary care records. Our outcome requires hospitalisation and as such, information on treatment received during hospital stay will not be available in primary care records. This means we will likely miss non-severe upper GI bleeds. Another consequence of using HES data is that this data source does not contain any information on prescriptions issued in hospital. This will affect ascertainment of prescribed treatment received and we will not be able to incorporate medications exclusively prescribed by hospital specialists.

As NSAIDs are an over the counter (OTC) medication we cannot exclude the possibility that patients may have been chronically self-medicating prior to study entry. It is unknown whether such patients would have a different likelihood of receiving one treatment over another but we have no evidence to support a differential lead time bias between groups of patients as defined by initiation of COX-2 and NSAIDs.

Furthermore, we consider it likely that individuals who have a need for chronic NSAID use would be likely to engage with primary care and receive prescriptions for treatment, rather than obtaining treatment via OTC routes. Whilst these issues would be of legitimate concern for an observational study that seeks to explore differentials in risk between two compounds, such limitations are likely to have applied to a number of extant studies in the considerable body of work describing NSAIDs and COX-2 inhibitors.

As with any study design involving matching there are likely to be prescient issues when dealing with rare exposures, outcomes and/or small sample sizes. However, it is likely that this study will have the luxury of a large sample size, increasing the likelihood of finding suitable matches.

Our study period spans 2000-2004 which covers the introduction of the Quality of Outcomes Framework (QOF) to UK primary care in 2004. This may introduce a bias insofar as the recording of certain incentivised comorbidities or other directly or indirectly QOF-induced changes in provider behaviour during the latter part of our observation period. Furthermore, several COX-2s were withdrawn from the market in 2003. Again, whilst this may introduce bias, it exhibits a realistic scenario in which PNU study designs are likely to be applied.

Medicines & Healthcare products Regulatory Agency

CPRD

U. References

1. Lund JL, Richardson DB, Sturmer T. The active comparator, new user study design in pharmacoepidemiology: historical foundations and contemporary application. *Curr Epidemiol Rep*. 2015;2(4):221-228. doi:10.1007/s40471-015-0053-5
2. Suissa S, Moodie EEM, Dell'Aniello S. Prevalent new-user cohort designs for comparative drug effect studies by time-conditional propensity scores. *Pharmacoepidemiol Drug Saf*. 2017;26(4):459-468. doi:10.1002/pds.4107
3. Lin H-MD, Lai CL, Dong Y-H, Tu Y-K, Chan KA, Suissa S. Re - evaluating Safety and Effectiveness of Dabigatran Versus Warfarin in a Nationwide Data Environment : A Prevalent New - User Design Study. *Drugs - Real World Outcomes*. 2019;6(3):93-104. doi:10.1007/s40801-019-0156-2
4. Schneeweiss S, Rassen JA, Glynn RJ, Avorn J, Mogun H, Brookhart MA. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology*. 2009;20(4):512-522. doi:10.1097/EDE.0b013e3181a663cc
5. Min H, Lin D, Lai CL, et al. Re - evaluating Safety and Effectiveness of Dabigatran Versus Warfarin in a Nationwide Data Environment : A Prevalent New - User Design Study. *Drugs - Real World Outcomes*. 2019;6(3):93-104. doi:10.1007/s40801-019-0156-2
6. Douros A, Dell'Aniello S, Hoi O, et al. Sulfonylureas as second line drugs in type 2 diabetes and the risk of cardiovascular and hypoglycaemic events : population based cohort study. *BMJ*. 2018;362:k2693. doi:10.1136/bmj.k2693
7. Webster-Clark M, Sturmer T, Edwards JK, Simpson R, Poole C, Lund JL. Classes and Prevalence of Prevalent New Users: An Example in Medicare. In: *International Conference of Pharmacoepidemiology & Therapeutic Risk Management*. ; 2019.
8. Yang C, Kuo S, Yang C-T, Lai EC-C, Ou H-T. Development and Application of a Hybrid Matching Algorithm to Refine the Prevalent New-User Cohort Design for Comparative Drug Effect Studies. In: *International Conference of Pharmacoepidemiology & Therapeutic Risk Management*. ; 2019.
9. Silverstein FE, Faich G, Goldstein JL, et al. Gastrointestinal Toxicity With Celecoxib vs Nonsteroidal Anti-inflammatory Drugs for Osteoarthritis and Rheumatoid ArthritisThe CLASS Study: A Randomized Controlled Trial. *JAMA*. 2000;284(10):1247-1255. doi:10.1001/jama.284.10.1247
10. Eisen GM, Goldstein JL, Hanna DB, Rublee DA. Meta-analysis: upper gastrointestinal tolerability of valdecoxib, a cyclooxygenase-2-specific inhibitor, compared with nonspecific nonsteroidal anti-inflammatory drugs among patients with osteoarthritis and rheumatoid arthritis. *Aliment Pharmacol Ther*. 2005;21(5):591-598. doi:doi:10.1111/j.1365-2036.2005.02383.x
11. Schneeweiss S, Solomon DH, Wang PS, Rassen J, Brookhart MA. Simultaneous assessment of short-term gastrointestinal benefits and cardiovascular risks of selective cyclooxygenase 2 inhibitors and nonselective nonsteroidal antiinflammatory drugs: An instrumental variable analysis. *Arthritis Rheum*. 2006;54(11):3390-3398. doi:doi:10.1002/art.22219
12. Chen Y-F, Jobanputra P, Barton P, et al. Cyclooxygenase-2 selective non-steroidal anti-inflammatory drugs (etodolac, meloxicam, celecoxib, rofecoxib, etoricoxib, valdecoxib and lumiracoxib) for osteoarthritis and rheumatoid arthritis: a systematic review and economic evaluation. *Health Technol Assess*. 2008;12(11):1-278, iii.
13. Williamson E, Morley R, Lucas A, Carpenter J. Propensity scores: from naive enthusiasm to intuitive understanding. *Stat Methods Med Res*. 2012;21(3):273-293. doi:10.1177/0962280210394483
14. Brookhart MA, Wyss R, Layton JB, Stürmer T. Propensity score methods for confounding control

Medicines & Healthcare products Regulatory Agency

CPRD

in nonexperimental research. *Circ Cardiovasc Qual Outcomes*. 2013;6(5):604-611. doi:10.1161/CIRCOUTCOMES.113.000359

15. Penning de Vries BBL, Groenwold RHH. Cautionary note: propensity score matching does not account for bias due to censoring. *Nephrol Dial Transplant*. 2017;33(6):914-916. doi:10.1093/ndt/gfx198
16. Schneeweiss S. Automated data-adaptive analytics for electronic healthcare data to study causal treatment effects. *J Clin Epidemiol*. 2018;10:771-788.
17. Tazare J, Smeeth L, Evans SJW, Williamson E, Douglas IJ. Implementing high-dimensional Propensity Score (hd-PS) Principles to UK Electronic Health Records. *[Submitted]*.
18. NHS Digital. Coding Cross Maps. https://isd.digital.nhs.uk/.
19. Bross IDJ. Spurious effects from an extraneous variable. *J Chronic Dis*. 1966;19:637-647.
20. Wyss R, Fireman B, Rassen JA, Schneeweiss S. Erratum: High-dimensional Propensity Score Adjustment in Studies of Treatment Effects Using Health Care Claims Data. *Epidemiology*. 2018;29(6). https://journals.lww.com/epidem/Fulltext/2018/11000/Erratum___High_dimensional_Propensity_Score.34.aspx.
21. Austin PC. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivar Behav Res*. 2011;46(3):399-424. doi:10.1080/00273171.2011.568786
22. Blake H, Leyrat C, Mansfield K, et al. Propensity scores using missingness pattern information: a practical guide. *arXiv Prepr arXiv190103981*. January 2019.

List of Appendices

Appendix A1: Code list for COX-2 inhibitors
Appendix A2: Code list for NSAIDs
Appendix B: Code list for upper GI bleed
Appendix C: Code list for osteoarthritis
Appendix D: Code list for cirrhotic liver disease

Medicines & Healthcare products Regulatory Agency

CPRD

Amendment – 6th February 2020

We have deselected HES Accident and Emergency Linkage (in Part 1) since this linkage is unavailable for the period of our study. This has the following impact on Part 2 of our ISAC application:

1. Section K: Remove HES Accident and Emergency as well as the reasons for requesting this linkage.

"We require linkage to ONS Death Registration Data, HES Admitted Patient Care, Patient Level Index of Multiple Deprivation and 2004 Rural-Urban Classification at LSOA level.

ONS Death Registration Data:
- Ascertaining deaths due to upper GI bleeding

HES Admitted Patient Care:
- Ascertaining outcome
- Defining propensity score covariates
- High-dimensional propensity score data dimension

,
Patient Level Index of Multiple Deprivation:
- Defining propensity score covariates

2004 Rural-Urban Classification at LSOA level:
- Defining propensity score covariates

2. Section N: 'Number of admissions to A&E in previous 6 months' will no longer be considered as a potential confounder in our propensity score model.

- **Comorbidities/ behaviours (any recording in patient history on or prior to cohort entry):** Hypertension, chronic renal failure, inflammatory bowel disease, gastrointestinal tract tumours, coagulopathies, gastro-oesophageal reflux disease, diabetes, heart failure, previous upper GI bleed (defined in Read and ICD-10).

3. Section O: Remove A&E as a HES Dimension in the high-dimensional propensity scores

We will incorporate this extra information by forming a dimension of discharge information. Dimensions and their coding systems are outlined below:

- <u>CPRD Dimensions:</u>
  Clinical (Read transcoded to ICD-10): Signs, symptoms and diagnoses
  Referral (Read transcoded to ICD-10 codes): Indicate a possible escalation in care
  Prescriptions (BNF codes): Patterns of drug usage
- <u>HES Dimensions:</u>
  Discharge (ICD-10): Diagnoses/disposition information recorded on discharge

.

Medicines & Healthcare products
Regulatory Agency

CPRD

ISAC EVALUATION OF PROTOCOLS FOR RESEARCH INVOLVING CPRD DATA

FEEDBACK TO APPLICANTS

CONFIDENTIAL			***by e-mail***
PROTOCOL NO:	19_273		
PROTOCOL TITLE:	Comparison of the prevalent new user and active comparator new user designs for assessing the real-world safety and effectiveness of medications		
APPLICANT:	Dr Daniel C Gibbons GlaxoSmithKline Daniel.c.gibbons@gsk.com		
APPROVED ☐	**APPROVED WITH COMMENTS (resubmission not required)** ☒	**REVISION/ RESUBMISSION REQUESTED** ☐	**REJECTED** ☐

INSTRUCTIONS:

Protocols with an outcome of 'Approved' or 'Approved with comments' do not require resubmission to the ISAC.

REVIEWER COMMENTS:

Reviewer 1
DISCRETIONARY COMMENTS TO APPLICANTS
Technical Summary
Please do not use referencing in your technical summary

Reviewer 2
DISCRETIONARY COMMENTS TO APPLICANTS
Feasibility Counts
While they have indicated the volume of prescribing, no estimates have been provided with regards individual patient numbers on
- Expected number of osteoarthritis patients (case definition in section L)
- Incident new users of either COX-2 inhibitors or NSAIDs for the (smaller) ACNU analysis cohort
- Number of expected outcomes during the study period.

Planned use of linked data (if applicable)
Please note that according to documentation produced by CPRD that "*The collection of HES A&E was first started in April 2007 on an experimental basis*". Thus, I cannot see any role for the A&E data in this study.

Technical Summary
Should this section have references i.e. how are they viewable when published on website

Sample size considerations
More detail in the previous section I (feasibility counts) would make their first sentence more convincing.

Selection of comparison group(s) or controls
Is practice accounted for in the matching process, or is not feasible? They are including an Urban-Rural indicator. Would a regional indicator also be worth considering if practice is not feasible?

Data/Statistical Analysis

A&E data is not primarily coded via ICD-10 but this will be a moot point (see section K comment).

COMMENTS
Generally, very clear. Two main points

- A&E data is requested but I'm not sure that this is feasible as my understanding is that the collection period for this dataset (2007) is after the end of your study (2004)
- The sections on feasibility counts and sample size could be improved by providing some patients estimates for theproposed cohorts and number of expected outcomes.

Please consider the role of practice in the proposed matching strategies.

General comment:

It is essential that consideration is given to preserving confidentiality at the reporting stage. The possibility of unintentional (deductive) disclosure arises when cells with small numbers of patients are quoted. Please note that, when reporting the data, CPRD policy is that no cell should contain <5 events and where necessary 'protect' these counts with secondary suppression. Please contact CPRD for further information if you encounter this issue during publication.

APPLICANT FEEDBACK:

DATE OF ISAC FEEDBACK:	08/01/20
DATE OF APPLICANT FEEDBACK:	

For protocols approved from 01 April 2014 onwards, applicants are required to include the ISAC protocol in their journal submission with a statement in the manuscript indicating that it had been approved by the ISAC (with the reference number) and made available to the journal reviewers. If the protocol was subject to any amendments, the last amended version should be the one submitted.

Guidance on resubmitting applications, or making amendments to approved protocols, can be found on the CPRD website at https://cprd.com/research-applications.

Appendix E

LSHTM Ethical approval for PPI-Mortality study

London School of Hygiene & Tropical Medicine
Keppel Street, London WC1E 7HT
United Kingdom
Switchboard: +44 (0)20 7636 8636

www.lshtm.ac.uk

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

Observational / Interventions Research Ethics Committee

Dr Ian Douglas
Associate Professor
Department of Non-communicable Disease Epidemiology (NCDE)
LSHTM

19 April 2018

Dear Dr Douglas

Study Title: Proton pump inhibitors and mortality: a cohort study

LSHTM Ethics Ref: 15655

Thank you for your application for the above research project which has now been considered by the Observational Committee via Chair's Action.

Confirmation of ethical opinion

On behalf of the Committee, I am pleased to confirm a favourable ethical opinion for the above research on the basis described in the application form, protocol and supporting documentation, subject to the conditions specified below.

Conditions of the favourable opinion

Approval is dependent on local ethical approval having been received, where relevant.

Approved documents

The final list of documents reviewed and approved is as follows:

Document Type	File Name	Date	Version
Protocol / Proposal	ISAC_Protocol_V1_25.10.2017	25/10/2017	1
Investigator CV	ID CV Apr18 LEO	12/04/2018	2

After ethical review

The Chief Investigator (CI) or delegate is responsible for informing the ethics committee of any subsequent changes to the application. These must be submitted to the committee for review using an Amendment form. Amendments must not be initiated before receipt of written favourable opinion from the committee.

The CI or delegate is also required to notify the ethics committee of any protocol violations and/or Suspected Unexpected Serious Adverse Reactions (SUSARs) which occur during the project by submitting a Serious Adverse Event form.

An annual report should be submitted to the committee using an Annual Report form on the anniversary of the approval of the study during the lifetime of the study.

At the end of the study, the CI or delegate must notify the committee using the End of Study form.

All aforementioned forms are available on the ethics online applications website and can only be submitted to the committee via the website at: http://leo.lshtm.ac.uk.

Further information is available at: www.lshtm.ac.uk/ethics.

Yours sincerely,

[redacted]

ethics@lshtm.ac.uk
http://www.lshtm.ac.uk/ethics/

Appendix F

ISAC application & approval for PPI-Mortality study

ISAC APPLICATION FORM

PROTOCOLS FOR RESEARCH USING THE CLINICAL PRACTICE RESEARCH DATALINK (CPRD)

For ISAC use only		
Protocol No. Submission date (DD/MM/YYYY)	....	***IMPORTANT*** *Please refer to the **guidance** for '**Completing the ISAC application form'** found on the CPRD website (www.cprd.com/isac). If you have any queries, please contact the ISAC Secretariat at isac@cprd.com.*

SECTION A: GENERAL INFORMATION ABOUT THE PROPOSED RESEARCH STUDY

1. Study Title[§] (*Please state the study title below)*

Proton pump inhibitors and mortality: a cohort study

[§]*Please note: This information will be published on the CPRD's website as part of its transparency policy.*

2. Has any part of this research proposal or a related proposal been previously submitted to ISAC?

Yes[*] ☐ No ☒

**If yes, please provide the previous protocol number/s below. Please also state in your current submission how this/these are related or relevant to this study.*

3. Has this protocol been peer reviewed by another Committee? (e.g. grant award or ethics committee)

Yes[*] ☐ No ☒

**If Yes, please state the name of the reviewing Committee(s) below and provide an outline of the review process and outcome as an Appendix to this protocol :*

4. Type of Study (please tick all the relevant boxes which apply)

Adverse Drug Reaction/Drug Safety	☒	Drug Effectiveness	☐
Drug Utilisation	☐	Pharmacoeconomics	☐
Disease Epidemiology	☐	Post-authorisation Safety	☐
Health care resource utilisation	☐	Methodological Research	☐
Health/Public Health Services Research	☐	Other[*]	☐

[*]*If Other, please specify the type of study here and in the lay summary below:*

5. Health Outcomes to be Measured[§]

[§]*Please note: This information will be published on CPRD's website as part of its transparency policy.*

Please summarise below the primary/secondary health outcomes to be measured in this research protocol:

- All-cause mortality
- Cause-specific mortality
-
-
-
-
-
-

[Please add more bullet points as necessary]

6. Publication: This study is intended for (please tick all the relevant boxes which apply)**:**

Publication in peer-reviewed journals	☒	Presentation at scientific conference	☒
Presentation at company/institutional meetings	☒	Regulatory purposes	☐
Other*	☐		

**If Other, please provide further information:*

SECTION B: INFORMATION ON INVESTIGATORS AND COLLABORATORS

7. Chief Investigator[§]

Please state the full name, job title, organisation name & e-mail address for correspondence - see guidance notes for eligibility. Please note that there can only be one Chief Investigator per protocol.

Dr Ian Douglas, Associate Professor of Pharmacoepidemiology, LSHTM
Ian.douglas@lshtm.ac.uk

[§]*Please note: The name and organisation of the Chief Investigator and will be published on CPRD's website as part of its transparency policy*

CV has been previously submitted to ISAC	☒	**CV number:** 157_15CESL
A new CV is being submitted with this protocol	☐	
An updated CV is being submitted with this protocol	☐	

8. Affiliation of Chief Investigator (full address)

London School of Hygiene and Tropical Medicine,
Keppel Street
London
WC1E 7HT
United Kingdom

9. Corresponding Applicant[§]

Please state the full name, affiliation(s) and e-mail address below:

Jeremy Brown, Research Fellow in Pharmacoepidemiology, LSHTM
Jeremy.brown@lshtm.ac.uk

[§]*Please note: The name and organisation of the corresponding applicant and their organisation name will be published on CPRD's website as part of its transparency policy*

Same as chief investigator	☐	
CV has been previously submitted to ISAC	☐	**CV number:**
A new CV is being submitted with this protocol	☒	
An updated CV is being submitted with this protocol	☐	

10. List of all investigators/collaborators[§]

Please list the full name, affiliation(s) and e-mail address* of all collaborators, other than the Chief Investigator below:

[§]*Please note: The name of all investigators and their organisations/institutions will be published on CPRD's website as part of its transparency policy*

Other investigator: Dr Krishnan Bhaskaran, Associate Professor in Statistical Epidemiology, LSHTM
Krishnan.bhaskaran@lshtm.ac.uk

CV has been previously submitted to ISAC	☒	**CV number:** 156_15CESL
A new CV is being submitted with this protocol	☐	
An updated CV is being submitted with this protocol	☐	

Other investigator: Jeremy Brown, Research Fellow in Pharmacoepidemiology, LSHTM
Jeremy.brown@lshtm.ac.uk

CV has been previously submitted to ISAC ☐ **CV number:**
A new CV is being submitted with this protocol ☒
An updated CV is being submitted with this protocol ☐

Other investigator: Dr Kathryn Mansfield, Research Fellow, LSHTM
Kathryn.Mansfield@lshtm.ac.uk

CV has been previously submitted to ISAC ☒ **CV number:** 319_15S
A new CV is being submitted with this protocol ☐
An updated CV is being submitted with this protocol ☐

Other investigator: Dr Adrian Root, Research Fellow in Epidemiology, LSHTM
Adrian.root@lshtm.ac.uk

CV has been previously submitted to ISAC ☒ **CV number:** 357_16P
A new CV is being submitted with this protocol ☐
An updated CV is being submitted with this protocol ☐

Other investigator: Professor Liam Smeeth, Professor of Clinical Epidemiology, LSHTM
Liam.smeeth@lshtm.ac.uk

CV has been previously submitted to ISAC ☒ **CV number:** 045_15CEPSL
A new CV is being submitted with this protocol ☐
An updated CV is being submitted with this protocol ☐

Other investigator: Dr Laurie Tomlinson, Associate Professor, LSHTM
Laurie.Tomlinson@lshtm.ac.uk

CV has been previously submitted to ISAC ☒ **CV number:** 271_15CESL
A new CV is being submitted with this protocol ☐
An updated CV is being submitted with this protocol ☐

Other investigator: Dr Elizabeth Williamson, Associate Professor of Medical Statistics, LSHTM
Elizabeth.williamson@lshtm.ac.uk

CV has been previously submitted to ISAC ☒ **CV number:** 354_16S
A new CV is being submitted with this protocol ☐
An updated CV is being submitted with this protocol ☐

Other investigator: Dr Kevin Wing, Assistant Professor, LSHTM
Kevin.wing@lshtm.ac.uk

CV has been previously submitted to ISAC ☒ **CV number:** 497_16ES
A new CV is being submitted with this protocol ☐
An updated CV is being submitted with this protocol ☐

Other investigator: Professor Stephens Evans, Professor of Pharmacoepidemiology, LSHTM
Stephen.Evans@lshtm.ac.uk

CV has been previously submitted to ISAC ☒ **CV number:** 158_15CESL
A new CV is being submitted with this protocol ☐
An updated CV is being submitted with this protocol ☐

Other investigator: John Tazare, PhD student, LSHTM
john.tazare1@lshtm.ac.uk

CV has been previously submitted to ISAC ☒ **CV number:** 448_17
A new CV is being submitted with this protocol ☐
An updated CV is being submitted with this protocol ☐

Other investigator: Dr Corentin Segalas, Research Fellow in Statistical Methodology, LSHTM
Corentin.Segalas@lshtm.ac.uk

CV has been previously submitted to ISAC ☐ **CV number:**
A new CV is being submitted with this protocol ☒
An updated CV is being submitted with this protocol ☐

Other investigator: Christopher Lee, MSc Student, LSHTM
Christopher.Lee1@student.lshtm.ac.uk

CV has been previously submitted to ISAC ☐ **CV number:**
A new CV is being submitted with this protocol ☒
An updated CV is being submitted with this protocol ☐

Other investigator: Ikpemesi Olubor, MSc Student, LSHTM
Ikpemesi.Olubor1@student.lshtm.ac.uk

CV has been previously submitted to ISAC ☐ **CV number:**
A new CV is being submitted with this protocol ☒
An updated CV is being submitted with this protocol ☐

Other investigator: Jack Collis, MSc Student, LSHTM
Jack.Collis1@student.lshtm.ac.uk

CV has been previously submitted to ISAC ☐ **CV number:**
A new CV is being submitted with this protocol ☒
An updated CV is being submitted with this protocol ☐

Other investigator: Professor Chris Frost, Professor of Medical Statistics, LSHTM
chris.frost@lshtm.ac.uk

CV has been previously submitted to ISAC ☐ **CV number:**
A new CV is being submitted with this protocol ☒
An updated CV is being submitted with this protocol ☐

Other investigator: Professor Isabel Dos Santos Silva, Professor of Epidemiology, LSHTM
Isabel.Silva@lshtm.ac.uk

CV has been previously submitted to ISAC ☒ **CV number: 615_16S**
A new CV is being submitted with this protocol ☐
An updated CV is being submitted with this protocol ☐

Please note that your ISAC application form and protocol **must be copied to all e-mail addresses listed above at the time of submission of your application to the ISAC mailbox. Failure to do so will result in delays in the processing of your application.*

11. Conflict of interest statement*
Please provide a draft of the conflict (or competing) of interest (COI) statement that you intend to include in any publication which might result from this work

IJD has consulted for and holds stock in GlaxoSmithKline and is funded by an unrestricted grant from GlaxoSmithKline. JB is funded by the Association of the British Pharmaceutical Industry for an unrelated project. All other authors declare no conflict of interest.

**Please refer to the International Committee of Medical Journal Editors (ICMJE) for guidance on what constitutes a COI.*

12. Experience/expertise available

Please complete the following questions to indicate the experience/ expertise available within the team of investigators/collaborators actively involved in the proposed research, including the analysis of data and interpretation of results.

	Previous GPRD/CPRD Studies	Publications using GPRD/CPRD data
None	☐	☐
1-3	☐	☐
> 3	☒	☒

Experience/Expertise available	Yes	No
Is statistical expertise available within the research team? *If yes, please indicate the name(s) of the relevant investigator(s)* Krishnan Bhaskaran, Elizabeth Williamson	☒	☐
Is experience of handling large data sets (>1 million records) available within the research team? *If yes, please indicate the name(s) of the relevant investigator(s)* Ian Douglas, Krishnan Bhaskaran	☒	☐
Is experience of practising in UK primary care available to or within the research team? *If yes, please indicate the name(s) of the relevant investigator(s)* Adrian Root, Liam Smeeth	☒	☐

13. References relating to your study

Please list up to 3 references (most relevant) relating to your proposed study:

1. Xie Y, Bowe B, Li T, et al. Risk of death among users of proton pump inhibitors: a longitudinal observational cohort study of United States veterans. BMJ open 2017; 7(6):e015735.
2. Othman F, Card TR, Crooks CJ. Proton pump inhibitor prescribing patterns in the UK: a primary care database study. Pharmacoepidemiology and drug safety 2016; 25(9):1079-87.
3. Schoenfeld AJ, Grady D. Adverse effects associated with proton pump inhibitors. JAMA internal medicine. 2016 Feb 1; 176(2):172-4.

SECTION C: ACCESS TO THE DATA

14. Financial Sponsor of study[§]

[§]*Please note: The name of the source of funding will be published on CPRD's website as part of its transparency policy*

Pharmaceutical Industry	☐	*Please specify name and country:*
Academia	☒	*Please specify name and country:* LSHTM, UK
Government / NHS	☐	*Please specify name and country:*
Charity	☐	*Please specify name and country:*
Other	☐	*Please specify name and country:*
None	☐	

15. Type of Institution conducting the research

Pharmaceutical Industry	☐	*Please specify name and country:*
Academia	☒	*Please specify name and country:* LSHTM, UK
Government Department	☐	*Please specify name and country:*
Research Service Provider	☐	*Please specify name and country:*

NHS ☐ *Please specify name and country:*
Other ☐ *Please specify name and country:*

16. Data access arrangements

The financial sponsor/ collaborator* has a licence for CPRD GOLD and will extract the data ☒
The institution carrying out the analysis has a licence for CPRD GOLD and will extract the data** ☐
A data set will be provided by the CPRD[¥€] ☐
CPRD has been commissioned to extract the data and perform the analyses[€] ☐
Other: ☐
If Other, please specify:

**Collaborators supplying data for this study must be named on the protocol as co-applicants.*
***If data sources other than CPRD GOLD are required, these will be supplied by CPRD*
[¥]Please note that datasets provided by CPRD are limited in size; applicants should contact CPRD (enquiries@cprd.com) if a dataset of >300,000 patients is required.
[€]Investigators must discuss their request with a member of the CPRD Research team before submitting an ISAC application. Please contact the CPRD Research Team on +44 (20) 3080 6383 or email (enquiries@cprd.com) to discuss your requirements. Please also state the name of CPRD Research team with whom you have discussed this request (provide the date of discussion and any relevant reference information):

Name of CPRD Researcher Reference number (where available) Date of contact

17. Primary care data
Please specify which primary care data set(s) are required)
Vision only (Default for CPRD studies ☒ Both Vision and EMIS®* ☐
EMIS® only* ☐

Note: Vision and EMIS are different practice management systems. CPRD has traditionally collected data from Vision practice. Data collected from EMIS is currently under evaluation prior to wider release.
Investigators requiring the use of EMIS data **must discuss the study with a member of the CPRD Research team before submitting an ISAC application*

Please state the name of the CPRD Researcher with whom you have discussed your request for EMIS data:
Name of CPRD Researcher Reference number (where available) Date of contact

SECTION D: INFORMATION ON DATA LINKAGES

18. Does this protocol seek access to linked data

Yes* ☒ No☐ If No, please move to section E.

Research groups which have not previously accessed CPRD linked data resources **must discuss access to these resources with a member of the CPRD Research team, before submitting an ISAC application. Investigators requiring access to HES Accident and Emergency data, HES Diagnostic Imaging Dataset PROMS data and the Pregnancy Register **must** also discuss this with a member of the CPRD Research team before submitting an ISAC application. Please contact the CPRD Research Team on +44 (20) 3080 6383 or email enquiries@cprd.comto discuss your requirements **before** submitting your application.*

Please state the name of the CPRD Researcher with whom you have discussed your linkage request.

Name of CPRD Researcher Reference number (where available) Date of contact

Please note that as part of the ISAC review of linkages, your protocol may be shared - in confidence - with a representative of the requested linked data set(s) and summary details may be shared - in confidence - with the Confidentiality Advisory Group of the Health Research Authority.

19. Please select the source(s) of linked data being requested[§]

[§]Please note: This information will be published on the CPRD's website as part of its transparency policy.

- ☒ ONS Death Registration Data
- ☒ HES Admitted Patient Care
- ☐ HES Outpatient
- ☐ HES Accident and Emergency
- ☐ HES Diagnostic Imaging Dataset
- ☐ MINAP (Myocardial Ischaemia National Audit Project)
- ☐ Cancer Registration Data*
- ☐ PROMS (Patient Reported Outcomes Measure)**
- ☐ CPRD Mother Baby Link
- ☐ Pregnancy Register

- ☐ Practice Level Index of Multiple Deprivation (Standard)
- ☐ Practice Level Index of Multiple Deprivation (Bespoke)
- ☒ Patient Level Index of Multiple Deprivation***
- ☐ Patient Level Townsend Score ***
- ☐ Other**** *Please specify:*

**Applicants seeking access to cancer registration data must complete a Cancer Dataset Agreement form (available from CPRD). This should be submitted to the ISAC as an appendix to your protocol. Please also note that applicants seeking access to cancer registry data must provide consent for publication of their study title and study institution on the UK Cancer Registry website.*
***Assessment of the quality of care delivered to NHS patients in England undergoing four procedures: hip replacement, knee replacement, groin hernia and varicose veins. Please note that patient level PROMS data are only accessible by academics*
**** 'Patient level IMD and Townsend scores will not be supplied for the same study*
*****If "Other" is specified, please provide the name of the individual in the CPRD Research team with whom this linkage has been discussed.*

Name of CPRD Researcher | Reference number (where available) | Date of contact

20. Total number of linked datasets requested <u>including</u> CPRD GOLD

Number of linked datasets requested (*practice/ 'patient' level Index of Multiple Deprivation, Townsend Score, the CPRD Mother Baby Link and the Pregnancy Register should **<u>not</u>** be included in this count)*

3

Please note: Where ≥5 linked datasets are requested, approval may be required from the Confidentiality Advisory Group (CAG) to access these data

21. Is linkage to a <u>local[¥]</u> dataset with <1 million patients being requested?

Yes* ☐ No ☒

**If yes, please provide further details:*
[¥] Data from defined geographical areas i.e. non-national datasets.

22. If you have requested one or more linked data sets, please indicate whether the Chief Investigator or any of the collaborators listed in question 5 above, have access to these data in a patient identifiable form (e.g. full date of birth, NHS number, patient post code), or associated with an identifiable patient index.

Yes* ☐ No ☒

** If yes, please provide further details:*

23. Does this study involve linking to patient *identifiable* data (e.g. date of birth, NHS number, patient post code) from other sources?

Yes ☐ No ☒

SECTION E: VALIDATION/VERIFICATION

24. Does this protocol describe a purely observational study using CPRD data?

Yes* ☒ No** ☐

** Yes: If you will be using data obtained from the CPRD Group, this study does not require separate ethics approval from an NHS Research Ethics Committee.*
*** No: You may need to seek separate ethics approval from an NHS Research Ethics Committee for this study. The ISAC will provide advice on whether this may be needed.*

25. Does this protocol involve requesting any additional information from GPs?

Yes* ☐ No ☒

** If yes, please indicate what will be required:*

Completion of questionnaires by the GP[ψ] Yes ☐ No ☐
Is the questionnaire a validated instrument? Yes ☐ No ☐
If yes, has permission been obtained to use the instrument? Yes ☐ No ☐
Please provide further information:

Other (please describe)

[ψ] Any questionnaire for completion by GPs or other health care professional must be approved by ISAC before circulation for completion.

26. Does this study require contact with patients in order for them to complete a questionnaire?

Yes* ☐ No ☒

**Please note that any questionnaire for completion by patients must be approved by ISAC before circulation for completion.*

27. Does this study require contact with patients in order to collect a sample?

Yes* ☐ No ☒

** Please state what will be collected:*

SECTION F: DECLARATION

28. Signature from the Chief Investigator

- I have read the guidance on '***Completion of the ISAC application form***' and '***Contents of CPRD ISAC Research Protocols***' and have understood these;
- I have read the submitted version of this research protocol, including all supporting documents, and confirm that these are accurate.
- I am suitably qualified and experienced to perform and/or supervise the research study proposed.
- I agree to conduct or supervise the study described in accordance with the relevant, current protocol
- I agree to abide by all ethical, legal and scientific guidelines that relate to access and use of CPRD data for research
- I understand that the details provided in sections marked with (§) in the application form and protocol will be published on the CPRD website in line with CPRD's transparency policy.
- I agree to inform the CPRD of the final outcome of the research study: publication, prolonged delay, completion or termination of the study.

Name: Ian Douglas Date: 20/10/17 e-Signature (type name): Ian Douglas

PROTOCOL INFORMATION REQUIRED

The following sections below **must** be included in the CPRD ISAC research protocol. Please refer to the guidance on '***Contents of CPRD ISAC Research Protocols***' (www.cprd.com/isac) for more information on how to complete the sections below. Pages should be numbered. All abbreviations must be defined on first use.

Applicants must complete all sections listed below
Sections which do not apply should be completed as '*Not Applicable*'

A. Study Title[§]
[§]*Please note: This information will be published on CPRD's website as part of its transparency policy*

Proton pump inhibitors and mortality: a cohort study

B. Lay Summary (Max. 200 words)[§]
[§]*Please note: This information will be published on CPRD's website as part of its transparency policy*

Proton pump inhibitors are a group of drugs that reduce the amount of acid produced by your stomach. They are used to treat a number of conditions including indigestion and heartburn.

They are effective drugs and, as a result, are prescribed frequently by doctors. However, there are concerns that they are associated with serious negative health outcomes. Research studies have indicated that usage of proton pump inhibitors might increase the risk of fractures, dementia and other negative outcomes including death.

It is not entirely clear, however, whether proton pump inhibitors cause these negative health outcomes. It may be that these associations occur because people who are prescribed proton pump inhibitors have on average poorer health when they start treatment.

It is important that we investigate the effects of proton pump inhibitors in depth given that they are frequently prescribed and given that the negative health outcomes suggested to be associated with usage are serious.

In our study, we will investigate the relationship between proton pump inhibitors and risk of death. We will estimate the risk of death associated with usage of proton pump inhibitors overall and by cause of death.

C. Technical Summary (Max. 200 words)[§]
[§]*Please note: This information will be published on CPRD's website as part of its transparency policy*

Proton pump inhibitors (PPIs) have been associated with a range of adverse outcomes including death. It is likely that many of these associations are not causal but due to confounding. Concern about adverse effects may reduce prescribing of this important class of medications. Therefore, we will examine cause-specific mortality including outcomes unlikely to be causally associated with PPIs.

We will use a cohort study design to estimate the risk of death in new users of PPIs compared to new users of H2 receptor antagonists (H2RAs) and, in a secondary analysis, to non-users of acid suppression therapy.

The outcomes of interest are all-cause and cause-specific mortality. For cause-specific mortality we will include both broad categories of cause of death (i.e. neoplasms) and selected specific causes. We will include specific causes related to adverse events previously found to be associated with PPI usage (e.g. pneumonia) and "control" outcomes (e.g. liver cancer mortality), which may be associated with frailty, but which have not been previously linked to PPIs.

Inverse probability of treatment weighting by propensity score, calculated using logistic regression, will be used to control for confounding. Cox regression will be used to estimate the effect of PPI prescription on risk of death.

Applicants must complete all sections listed below
Sections which do not apply should be completed as '*Not Applicable*'

D. Objectives, Specific Aims and Rationale

The general objective of this study is to examine the association between usage of proton pump inhibitors (PPIs) and mortality. More specifically, we aim to estimate the association between PPI usage and:

- all-cause mortality (primary outcome)
- cause-specific mortality with cause of death grouped into broad categories (secondary outcome)
- cause-specific mortality by individual cause of death for selected causes (secondary outcome)

There is some evidence to suggest that PPI use is associated with increased all-cause mortality.[1] This finding has not yet been validated and it is not known which specific causes of death are leading to the overall association. A true causal association between PPI usage and mortality could have major implications for prescribing practice and public health in the UK and internationally.

E. Study Background

Proton pump inhibitors (PPIs) are a group of drugs used to supress gastric acid production. They are prescribed for a variety of indications including the treatment of dyspepsia, peptic ulcers, and gastro-oesophageal reflux disease (GORD), the eradication of *H. pylori*, and prophylaxis to prevent drug-induced (e.g. non-steroidal anti-inflammatory drugs and corticosteroids) gastrointestinal damage. PPIs are some of the most commonly prescribed drugs in the UK. It was estimated that 15% of adults in the Clinical Practice Research Datalink (CPRD) were prescribed a PPI in 2014.[2]

An older alternative group of acid suppressing drugs, H2 receptor antagonists (H2RAs), is also available. The indications for the two groups of drugs are very similar.[3] However, H2RAs, unlike PPIs, are not typically used in *H. pylori* eradication. PPIs are more effective acid suppressors, with a longer duration of effect, and are now prescribed more frequently than H2RAs.[4]

Though PPIs are widely seen as safe and effective medicines, there are increasing concerns over adverse effects.[5] Observational studies have indicated associations between PPI usage and a number of adverse health outcomes. These outcomes include fractures, *Clostridium difficile* infection, community-acquired pneumonia, vitamin B12 deficiency, hypomagnesemia, cardiovascular events, dementia and chronic kidney disease.[6-12] Strength of evidence varies by outcome, and is in some cases inconsistent. For instance, observational studies indicated an association between PPI usage and cardiovascular events due to a suggested drug interaction between PPIs and clopidogrel, but randomised controlled trials found no such effect.[13]

Based on the prescribing experience of clinicians in the study team, we believe that PPIs are widely prescribed to people with a broad range of underlying illnesses. They may also be given during the early stages of serious diseases that have not been recognised, but which have led to gastrointestinal symptoms. For these reasons, we suspect that observational studies may readily detect non-causal associations between PPIs and adverse outcomes.

A recent study (Xie et al. 2017) identified an increased risk of all-cause mortality in users of proton pump inhibitors, compared with people taking H2RAs, and similarly when compared with people not taking acid suppression therapy.[1] We plan to use CPRD data to independently examine the risk of mortality in PPI users. CPRD data has been used to estimate the association between PPI usage and a number of negative health outcomes (ISAC protocols - 16_165R2, 16_149, 16_123RA, and 15_210).[11 14-23] Arana and colleagues investigated the risk of sudden cardiac death in users of PPIs versus domperidone users.[14] We previously examined the risk of death or myocardial infarction with PPI usage in patients receiving clopidogrel and aspirin.[16] There has not however yet been a study in the CPRD examining in depth the risk of death associated with general PPI usage.

Furthermore, while Xie's study had many strengths, there were important limitations. Their study included mostly older white males, which limited generalisability. They did not have information on important confounders, including BMI, smoking status, alcohol consumption, and the concomitant use of medications such as non-steroidal anti-inflammatory drugs and anticoagulants. Furthermore, they did not have information on cause of death. Using CPRD

Applicants must complete all sections listed below **Sections which do not apply should be completed as '*Not Applicable*'**
data we will be able to overcome many of the limitations of Xie's study and will be able to estimate the risk of cause-specific mortality with PPI prescription. Attempts to account for confounding in previous studies may not be sufficient and to guard against incorrect conclusions we propose a study that incorporates a range of "control" outcomes to aid interpretation.
F. Study Type The study will primarily be hypothesis testing, comparing the risk of all-cause mortality in PPI users to H2 receptor antagonist (H2RA) users. For the secondary outcomes of cause-specific mortality, the study will be exploratory/hypothesis generating.
G. Study Design This will be a cohort study comparing new users of PPIs to: 1) new users of an alternative acid-suppression therapy - H2RAs (main analysis) 2) non-users of acid suppression therapy (secondary analysis)
H. Feasibility counts We calculated feasibility counts using CPRD data (July 2017 data release). We counted the number of eligible new-users of PPIs and H2RAs who started treatment between 02/01/1998 and 31/12/2015. For a description of eligibility criteria, please see the Sections K (Study population) and L (Selection of comparison group(s) or controls). There were approximately 735,000 eligible new-users of PPIs and 125,000 eligible new-users of H2RAs.

I. Sample size considerations

Sample size calculations were performed based on the log-rank test for a difference in all-cause mortality between cohorts. We assumed a mean of 3 years follow-up from treatment initiation and estimated the probability of death by three years (using 2006-2008 Office of National Statistics life tables) weighted by age of PPI users in the CPRD.[2]

Table 1: Mortality estimates by age group for sample size calculation

Age group	Proportion of PPI users in CPRD in age group	Three-year mortality at mid-point of age group*
18-30	0.084	0.001
31-40	0.141	0.002
41-50	0.191	0.005
51-60	0.203	0.013
61-70	0.194	0.032
71-80	0.136	0.087
80†	0.047	0.263

*Assuming a 55:45 female to male ratio as found in PPI users by Othman et al.[2]
† Three-year mortality at age 85 was calculated to produce a conservative estimate of mortality in the 80+ age group

Overall estimated probability of death over three years was 0.034. This is likely to be a conservative estimate given that individuals who are prescribed acid suppression therapy are likely to have poorer health than the general population.[1]

Based on this assumed probability of death during follow-up, the estimated minimum sample size per group to detect a hazard ratio of 1.25 with 90% power at $\alpha=0.05$ is 11,153. We chose a hazard ratio of 1.25 as this was the primary estimate produced by Xie et al. Given that the feasibility count identified 735,000 PPI new-users and 125,000 H2RA new-users we should have adequate power for the primary outcome in our study.

For cause-specific mortality, as we are hypothesis generating rather than hypothesis testing, a formal power calculation is not applicable. Precision of estimates will vary by cause, but given the large number of PPI users and H2RA users we have identified, we should be able to generate informative estimates.

J. Data Linkage Required (if applicable):[§]

[§]*Please note that the data linkage/s requested in research protocols will be published by the CPRD as part of its transparency policy*

We require linkage to Office of National Statistics (ONS) mortality data, Hospital Episode Statistics (HES) admitted patient care data, and patient level Index of Multiple Deprivation (IMD) data. ONS mortality data is required in order to accurately ascertain date and cause of death. We will use HES admitted patient care data to estimate the number of hospital admissions in the 6 months prior to baseline, which we will adjust for in our propensity score model. Similarly, we will use IMD data in order to adjust for socioeconomic status.

K. Study population

The source population will be individuals aged 18 years and over who have been flagged as acceptable by the CPRD and are eligible for linkage. The study will run from 2nd January 1998 to 17th April 2017, and individuals will be recruited for inclusion in the study up to 31st December 2015 (allowing for the possibility of up to two years follow up for those recruited at the end of the recruitment period).

The recruitment period will start at the latest of 02/01/1998 (start of ONS mortality coverage), patient's 18th birthday, one year after up-to-standard date, or one year after current registration date. The end of the recruitment period will be the earliest of transfer out date, CPRD death date, ONS death date, practice last collection date, or 31/12/2015 (the recruitment period is within the HES admitted patient care data coverage period - 01/04/1997 to 31/03/2016). Adding one year to the current registration date reduces bias due to retrospective recording by GPs following registration.[24] The addition of one year to the up-to-standard date ensures that higher quality data is available, for defining covariates, for at least a year pre-baseline for all patients.

From the source population, patients with a new prescription for a PPI during the recruitment period will be selected (see Appendix A for a code list of PPIs). Only patients with no recorded acid suppression therapy (PPI or H2RA) prior to the recruitment period, and with no H2RA prescription prior to first PPI initiation during the recruitment period, will be eligible for inclusion.

The start of follow-up for eligible PPI users will be the date of first PPI prescription and the end of follow-up will be the earliest of transfer out date (except when transfer out reason is death), ONS death date, practice last collection date, or end of ONS coverage date (17/04/2017).

L. Selection of comparison group(s) or controls

The primary comparison cohort will be patients prescribed a H2RA as a new-user during the recruitment period (see Appendix A for a code list of H2RAs). Only patients with no recorded acid suppression therapy (PPI or H2RA) prior to the recruitment period, and with no PPI prescription prior to first H2RA initiation during the recruitment period, will be eligible for inclusion.

A second comparison cohort will be comprised of non-users of acid suppression therapy. For each PPI user, we will match up to five non-users of acid suppression therapy, matching on age (within 2 years), sex, practice and being registered with an up-to-standard practice in the CPRD on the date the PPI user was first prescribed a PPI (the index date). Non-users of acid suppression therapy must not have received a PPI or H2RA before the index date, but could receive one later. Furthermore, to be eligible for matching, non-users must have had at least one GP appointment prior to the index date since current registration date to ensure engagement with medical services.

The start of follow-up for eligible H2RA users will be the date of first H2RA prescription. For non-users of acid suppression therapy the start of follow-up will be the index date on which they were matched with the PPI user. For H2RA users and acid suppression therapy non-users the end of follow-up will be the earliest of PPI prescription date, H2RA prescription date (acid suppression therapy non-users only), transfer out date (except when transfer out reason is death), ONS death date, practice last collection date, or end of ONS coverage date (17/04/2017).

M. Exposures, Health Outcomes[§] and Covariates

[§]*Please note: Summary information on health outcomes (as included on the ISAC application form above)will be published on CPRD's website as part of its transparency policy*

The exposure of interest is prescription of a PPI: omeprazole, lansoprazole, pantoprazole, esomeprazole or rabeprazole.

The outcomes of interest are all-cause mortality and cause-specific mortality ascertained using ONS death registration data. Cause of death will be defined using ICD9 (International Classification of Disease 9) and ICD10.

We will categorise underlying cause of death into broad categories (Table 2) using the Global Burden of Disease categorisation system, as we have done in a separate study in which we are investigating the association between BMI and risk of death in the CPRD (ISAC – Protocol 16_174).[25]

Table 2: Broad categories of cause of death and corresponding ICD codes

Cause of death outcomes	Relevant ICD-10 chapters/codes	Relevant ICD-9 codes
Top-level outcomes		
Communicable diseases	A, B, J00-22	1-139, 460-469, 480-488
Non-communicable diseases	C through R	140-459, 470-479, 490-799
Injuries/external	S through Y	800-999, E001-E999
Second-level outcomes		
Neoplasms	C	140-239
Cardiovascular/circulatory	I	390-459
Chronic respiratory diseases	J23-99	470-478, 490-519
Liver cirrhosis	K70.3, K71.7, K74.3-6	571.2, 571.5, 571.6
Digestive other than cirrhosis	K except codes above	520-579 (except 571.2, 571.5 & 571.6)
Neurological	G	320-359, 290
Mental and behavioural	F	291-319
Diabetes, urogenital, blood and endocrine	D50-89, E, N	240-289, 580-629
Musculoskeletal	M	710-739

We will also include selected individual underlying causes of death as outcomes, as specified in Table 3. These include causes that are related to adverse events previously associated with PPI usage. In order to investigate potential residual confounding by frailty, we will also include causes that are associated with frailty, but have not been found to be associated with PPI usage.

Table 3: Specific causes of death and corresponding ICD-10 codes

Cause of death outcomes	Relevant ICD-10 chapters/codes	Relevant ICD-9 codes	Plausible development timeframe
Events that have been linked to PPI usage			
Events that could plausibly be affected causally by PPI usage in short term after treatment initiation			
Pneumonia	J10.0, J11.0, J12-J18	480-486, 487.0, 514	Short or long term
Acute kidney injury	N17	584	Short or long term
Enterocolitis due to *Clostridium difficile*	A04.7	008.45	Short or long term
Atrial fibrillation/flutter	I48	427.3	Short or long term
Heart failure	I50	428	Short or long term
Aortic Aneurysm	I71	441	Short or long term
Events that if causally associated would not likely be affected in the short term after treatment initiation			
Dementia and Alzheimer's disease	F00, F01, F03, G30	290, 294.2, 331	Long term
Chronic kidney disease	N18	585	Long term
Hypertensive heart disease	I11	402	Long term
Ischaemic heart disease	I20-I25	410-414	Long term
Events that have not been linked to PPI usage i.e. "control" outcomes			
Accidental trauma (excluding falls)	V01-X59 (excluding W00-W19), Y86, Y86	E800-E928 (excluding E870-E888)	Short or long term
Pulmonary embolism	I26	415.1-415.19	Short or long term
Lung cancer	C34	162 (except 162.0 & 162.2)	Long term
Mesothelioma	C45	163	Long term
Breast cancer	C50	174	Long term
Liver cancer	C22	155	Long term
Prostate cancer	C61	185	Long term
COPD	J40-J44	490-492, 496	Long term
Alcoholic liver disease	K70	571.0-571.3	Long term

In terms of covariates we will include in our propensity score model (see Section N), as recommended by simulation studies, variables that are likely to be associated with the outcome. In particular, we have selected covariates that are likely to be associated with both the outcome, all-cause mortality, and treatment assignment.[26 27]

Specifically, we will include:

- **Demographic and lifestyle variables at start of follow-up:** Age, sex, index of multiple deprivation score (IMD), body mass index (BMI), smoking status, alcohol consumption
- **Potential indications for PPI treatment in 6 months prior to start of follow-up**: Prescription for NSAID, aspirin, clopidogrel, oral anticoagulant or corticosteroid, upper gastrointestinal endoscopy, gastric cancer, GORD, peptic ulcers, upper GI bleeding, Zollinger-Ellison syndrome, pancreatitis, cirrhosis, oesophagitis, Barrett's oesophagus, *H. pylori* infection, and *H. pylori* treatment
- **Indicators of frailty in 6 months prior to start of follow-up:** number of hospital admissions in prior six months (based on HES admitted patient care), number of GP appointments in prior 6 months, number of different drugs types prescribed in prior six months (based on distinct BNF chapters)
- **Comorbidities**: hypertension, cardiovascular disease, peripheral artery disease, cerebrovascular disease, chronic lung disease, cancer, chronic liver disease, HIV, chronic kidney disease (CKD) and CKD stage, dementia, and diabetes mellitus.
- **Other:** calendar year

N. Data/ Statistical Analysis

Propensity scores will be used to adjust for differences in baseline covariates. Propensity scores will be calculated using a multivariable logistic regression model for PPI prescription versus H2RA prescription (primary comparison) and multivariable conditional logistic regression for PPI prescription versus no acid suppression therapy (secondary comparison). Each patient will be weighted by the inverse probability of the received treatment. After propensity score weighting we will assess balance using standardised differences and by graphing cumulative distribution functions and boxplots.[28]

Hazard ratios will be estimated using inverse probability weighted univariable Cox regression models based on time from first prescription (PPI users/H2RA users) or index date (non-users of acid suppression therapy) to event (all-cause and cause-specific mortality).

Kaplan-Meier curves for all-cause mortality, and cumulative incidence functions for cause-specific mortality, will be used to estimate the absolute difference in risk of death associated with PPI prescription.[29] We will assess proportionality using graphical approaches, tests based on Schoenfeld residuals, and by adding an interaction between time and PPI prescription. If there is evidence of non-proportionality, we will report hazard ratios at a number of different time points (e.g. 3 months, 6 months, 1 years, 3 years and 5 years post-baseline).[30]

For all outcomes we will examine the hazard ratio for the first year of treatment, and the period specific hazard ratio from one year after treatment initiation onwards. For outcomes that we do not expect to be affected in the short term after treatment initiation if there were a true causal association (see Table 3), an early increase in outcome-specific mortality may indicate residual confounding.

We will also investigate potential effect modification by age and sex.

Secondary/sensitivity analyses

1. We will investigate the effect of censoring follow-up in survival analysis at first treatment break and in PPI users on prescription of a H2RA.
2. We will examine the effect of including all recorded causes of death, rather than just underlying cause, on the hazard ratios for cause-specific mortality. We will use a binary outcome for each participant indicating whether that cause of death was recorded, or not recorded, as an underlying or secondary cause.
3. We will investigate whether there is any difference in effect on mortality between the two most commonly prescribed proton pump inhibitors: omeprazole and lansoprazole. If the class as a whole appears to be associated with specific causes of death, it would be important to know if this is truly a class effect, or whether it is associated with a specific PPI. From previous work we anticipate lansoprazole and omeprazole will account for the vast majority of PPI usage in the UK and so will assess them separately.
4. We will investigate the effect of duration of usage by using a time-updated duration of usage variable.
5. We will examine our findings using direct adjustment for covariates instead of inverse probability of treatment weighting using propensity scores. We do not expect there to be a significant difference in results using the two different methods.

O. Plan for addressing confounding

Propensity scores with inverse probability of treatment weighting will be used to adjust for measured confounding covariates. We believe residual confounding may remain a problem in this study as the subtle reasons why an individual is prescribed a PPI, whilst a seemingly exchangeable person is not, may not be captured in the data. For this reason, we have included a range of “control” outcomes and “control” time windows within which we will measure any association with PPIs. Harmful associations detected in these analyses will point towards residual confounding as a possible explanation.

P. Plans for addressing missing data

Given our experience of using CPRD and ONS mortality data, we do not expect there to be missing data on underlying cause of death, but we do expect there to be missing data on body mass index, smoking status and alcohol consumption.

We plan to conduct a complete case analysis, which relies on the assumption that the probability of these data being missing is independent of mortality risk, conditional on covariates; given the small amount of anticipated missing data, any violation of the assumption is unlikely to importantly affect the results.[31]

Q. Patient or user group involvement (if applicable)

None

R. Plans for disseminating and communicating study results, including the presence or absence of any restrictions on the extent and timing of publication

From this study, we plan to publish at least one peer-reviewed article in a scientific journal. We may also present at conferences and institutional meetings.

S. Limitations of the study design, data sources, and analytic methods

Although we will adjust for confounders using propensity scores, given the observational nature of the study design it is not possible to prove causality. Furthermore, limitations with available reference groups mean that there is likely to be residual confounding by indication. However, this study will still provide useful evidence on whether, as Xie and colleagues suggested, proton pump inhibitors increase risk of death. Moreover, our secondary and sensitivity analyses deliberately attempt to investigate whether/how unadjusted confounding may be a driver for associations with PPIs.

There are limitations to death certificate data. Cause of death is not always known for certain and, furthermore, assignation of a single underlying cause is often an oversimplification.[32] For this reason we examine multiple causes of death, rather than just the recorded underlying cause of death, in sensitivity analyses.

T. References

1. Xie Y, Bowe B, Li T, et al. Risk of death among users of Proton Pump Inhibitors: a longitudinal observational cohort study of United States veterans. *BMJ open* 2017;7(6):e015735.
2. Othman F, Card TR, Crooks CJ. Proton pump inhibitor prescribing patterns in the UK: a primary care database study. *Pharmacoepidemiology and drug safety* 2016;25(9):1079-87.
3. Excellence; NIfHaC. A to Z of Drugs | BNF Provided by NICE [Available from: https://bnf.nice.org.uk/drug/ accessed 04/10/2017.
4. Savarino V, Di Mario F, Scarpignato C. Proton pump inhibitors in GORD: an overview of their pharmacology, efficacy and safety. *Pharmacological research* 2009;59(3):135-53.
5. Schoenfeld AJ, Grady D. Adverse effects associated with proton pump inhibitors. *JAMA internal medicine* 2016;176(2):172-74.
6. Maes ML, Fixen DR, Linnebur SA. Adverse effects of proton-pump inhibitor use in older adults: a review of the evidence. *Therapeutic Advances in Drug Safety* 2017;8(9):273-97.
7. Elaine WY, Bauer SR, Bain PA, et al. Proton pump inhibitors and risk of fractures: a meta-analysis of 11 international studies. *The American journal of medicine* 2011;124(6):519-26.
8. Janarthanan S, Ditah I, Adler DG, et al. Clostridium difficile-associated diarrhea and proton pump inhibitor therapy: a meta-analysis. *The American journal of gastroenterology* 2012;107(7):1001.
9. Lambert AA, Lam JO, Paik JJ, et al. Risk of community-acquired pneumonia with outpatient proton-pump inhibitor therapy: a systematic review and meta-analysis. *PloS one* 2015;10(6):e0128004.
10. Cheungpasitporn W, Thongprayoon C, Kittanamongkolchai W, et al. Proton pump inhibitors linked to hypomagnesemia: a systematic review and meta-analysis of observational studies. *Renal failure* 2015;37(7):1237-41.
11. Gomm W, von Holt K, Thomé F, et al. Association of proton pump inhibitors with risk of dementia: a pharmacoepidemiological claims data analysis. *JAMA neurology* 2016;73(4):410-16.
12. Lazarus B, Chen Y, Wilson FP, et al. Proton pump inhibitor use and the risk of chronic kidney disease. *JAMA internal medicine* 2016;176(2):238-46.
13. Melloni C, Washam JB, Jones WS, et al. Conflicting results between randomized trials and observational studies on the impact of proton pump inhibitors on cardiovascular events when coadministered with dual antiplatelet therapy. *Circulation: Cardiovascular Quality and Outcomes* 2015:CIRCOUTCOMES. 114.001177.
14. Arana A, Johannes CB, McQuay LJ, et al. Risk of out-of-hospital sudden cardiac death in users of domperidone, proton pump inhibitors, or metoclopramide: a population-based nested case-control study. *Drug safety* 2015;38(12):1187-99.
15. Bradley M, Murray L, Cantwell M, et al. Proton pump inhibitors and histamine-2-receptor antagonists and pancreatic cancer risk: a nested case–control study. *British journal of cancer* 2012;106(1):233.
16. Douglas IJ, Evans SJ, Hingorani AD, et al. Clopidogrel and interaction with proton pump inhibitors: comparison between cohort and within person study designs. *BMJ* 2012;345:e4388.
17. Filion KB, Chateau D, Targownik LE, et al. Proton pump inhibitors and the risk of hospitalisation for community-acquired pneumonia: replicated cohort studies with meta-analysis. *Gut* 2013:gutjnl-2013-304738.
18. Kaye JA, Jick H. Proton pump inhibitor use and risk of hip fractures in patients without major risk factors. *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy* 2008;28(8):951-59.
19. Leonard CE, Freeman CP, Newcomb CW, et al. Proton pump inhibitors and traditional nonsteroidal anti-inflammatory drugs and the risk of acute interstitial nephritis and acute kidney injury. *Pharmacoepidemiology and drug safety* 2012;21(11):1155-72.
20. Rodríguez LG, Johansson S, Soriano LC. Use of clopidogrel and proton pump inhibitors after a serious acute coronary event: risk of coronary events and peptic ulcer bleeding. *Thromb Haemost* 2013;110(5):1014-24.
21. Tran-Duy A, Vanmolkot F, Souverein P, et al. Co-Administration of Proton Pump Inhibitors in Chronic Aspirin Users and the Risk of Adverse Cardiovascular Events: a Population-Based Cohort Study. *Value in health: the journal of the International Society for Pharmacoeconomics and Outcomes Research* 2015;18(7):A378.
22. Verhaegh B, Vries F, Masclee A, et al. High risk of drug-induced microscopic colitis with concomitant use of NSAIDs and proton pump inhibitors. *Alimentary pharmacology & therapeutics* 2016;43(9):1004-13.
23. Yang YX, Hennessy S, Propert K, et al. Chronic proton pump inhibitor therapy and the risk of colorectal cancer. *Gastroenterology* 2007;133(3):748-54.
24. Lewis JD, Bilker WB, Weinstein RB, et al. The relationship between time since registration and measured incidence rates in the General Practice Research Database. *Pharmacoepidemiology and drug safety* 2005;14(7):443-51.
25. Abubakar I, Tillmann T, Banerjee A. Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death, 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013. *Lancet* 2015;385(9963):117-71.
26. Brookhart MA, Schneeweiss S, Rothman KJ, et al. Variable selection for propensity score models. *American journal of epidemiology* 2006;163(12):1149-56.

27. Austin PC, Grootendorst P, Anderson GM. A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. *Statistics in medicine* 2007;26(4):734-53.
28. Austin PC, Stuart EA. Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in medicine* 2015;34(28):3661-79.
29. Austin PC, Lee DS, Fine JP. Introduction to the analysis of survival data in the presence of competing risks. *Circulation* 2016;133(6):601-09.
30. Hernan MA. The hazards of hazard ratios. *Epidemiology* 2010;21(1):13-5. doi: 10.1097/EDE.0b013e3181c1ea43
31. White IR, Carlin JB. Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Statistics in medicine* 2010;29(28):2920-31.
32. Fedeli U, Zoppini G, Goldoni CA, et al. Multiple causes of death analysis of chronic diseases: the example of diabetes. *Population health metrics* 2015;13(1):21.

List of Appendices *(Submit all appendices as separate documents to this application)*

Appendix A: Code list for PPIs and H2RAs

Accepted amendment - 01/07/2019

We would like to request a minor amendment to study 17_252, "Proton pump inhibitors and mortality: a cohort study". The minor amendments that we would like to make are:

1. To exclude users of proton pump inhibitors (PPIs) and H2 receptor antagonists (H2RAs) who do not have a GP appointment prior to treatment initiation since current registration date.

2. To use a missingness category approach, rather than complete case approach to handle missing data on alcohol consumption, smoking and BMI.

3. To add gastric cancer mortality as an outcome.

4. To use average effect of treatment in the treated weights (ATT) rather than inverse probability weights (IPW).

5. To add, as an additional analysis to further investigate potential confounding, weighting based on high-dimensional, rather than conventional, propensity scores.

6. To add quantitative bias analysis for unmeasured confounding.

7. To add Kevin Wing, Stephen Evans and John Tazare as co-investigators.

In the original protocol, we wrote that we would exclude non-users who did not have a GP appointment. We planned this so that we would include only non-users who engaged with their general practice, who would thereby be more similar to PPI/H2RA users who we presumed would present to a GP to receive a PPI/H2RA prescription. There were, however, a very small number of PPI/H2RA users without a GP appointment on or before treatment initiation. We would, therefore, like to exclude these patients, to increase comparability with the non-user cohort, and thereby reduce confounding.

We planned originally to handle missing data using a complete case approach. However, we came to realise during data processing that a considerable fraction of the study population (17% of total study population) were missing information on one of alcohol consumption, smoking or BMI. We expect that the data are not missing completely at random, and therefore that a complete case approach could lead to bias. Statisticians in the group recommended an alternative approach, a missingness category approach, which they have found to be unbiased for propensity score analysis under less strict assumptions.

In our analysis we planned to include cause-specific mortality from outcomes that have previously been associated with PPI usage (e.g. pneumonia). Since we wrote the protocol, a study has been published identifying an association between proton pump inhibitors and gastric cancer (Cheung et al Gut. 2018). Given that this association has been identified, we would like to include gastric cancer mortality as one of the 20 cause-specific mortality outcomes that we will investigate.

We initially planned to adjust for propensity score using inverse probability weighting. Inverse probability weighting leads to an estimate of the effect in the overall study population. However, in order to investigate possible confounding by indication, we have two separate populations (PPI versus H2RA users, and PPI users versus non-users). If we used an IPW approach it would be unclear whether differences in our effect estimates were due to differences in the study population, rather than differences in confounding by indication. For this reason, we would like to adjust for propensity score using the related, but slightly different, average effect of treatment in the treated (ATT) weights. These weights lead to an estimate of the effect in the treated population (in our study PPI users), which as the PPI users are shared by both comparisons (PPI vs H2RA, and PPI vs non-user), should make the two analyses more comparable.

Our preliminary results strongly indicate residual confounding by indication. We would like to further investigate this confounding through the use of a secondary analysis using high dimensional propensity scores (HDPS). A reduction in the effect estimates, towards the null, in an analysis using HDPS would be a further indicator that confounding is driving observed associations.

Quantitative bias analysis is a sensitivity analysis technique that aims to quantify the extent to which an unmeasured confounder could be contributing to observed results (Ding and VanderWeele Epidemiology. 2016). We would like to use quantitative bias analysis in order to provide an estimate of how strong an unmeasured confounder/s would have to be to explain our observed results.

In order to conduct the HDPS additional analysis, we would like to include a statistician with expertise in this area, John Tazare (CV number: 448_17) as a co-investigator. To bring additional expertise in pharmacoepidemiology for the interpretation of study results, we would like to add the pharmacoepidemiologists Dr Kevin Wing and Dr Stephen Evans (CV numbers: 497_16ES and 158_15CESL) as co-investigators.

Amendment - 07/04/2020

Preliminary findings indicate an association between proton pump inhibitors (PPI) and both all-cause and cause-specific mortality. However, these results also indicate residual confounding is likely. Based on these preliminary results we plan to conduct a number of secondary analyses:

1. To estimate the association between PPI prescription and incidence of disease outcomes.
2. To compare associations estimated using a cohort study to those estimated using a study design less prone to confounding, the self-controlled case series.

Study Design [additional text]

As a secondary analysis we will compare associations with PPI prescription between two study designs: cohort study and self-controlled case series. This will provide a further indication of the presence of residual confounding, which we expect may be present in the cohort study. Self-controlled study designs, such as the self-controlled case series, control for between-person confounding by design (Petersen, Douglas and Whitaker 2016). Presence of an association in the cohort and absence in the self-controlled case series will provide an indication of confounding

Self-controlled case series are not suitable for outcomes that terminate follow-up such as mortality. As such we will compare (between the cohort study and self-controlled case series) incidence, rather than mortality, of two of the causes of mortality studied: ischaemic heart disease and alcoholic liver disease.

Exposures, Health Outcomes[§] and Covariates [additional text]

In order to better understand the association between PPI prescription and cause-specific mortality in a secondary analysis we will estimate the association between PPI prescription and incidence of selected causes: cancer overall and by sub-type, pneumonia, chronic obstructive pulmonary disease, ischaemic heart disease and alcoholic liver disease.

It may be that increased cause-specific mortality among PPI users is due to increased incidence, or increased mortality among those with disease, or a combination of both factors. This additional secondary analysis will allow us to investigate whether increased incidence is contributing to increased cause-specific mortality.

Data/ Statistical Analysis [additional text]

Secondary/sensitivity analyses [continued]

6. We will estimate the association between PPI prescription and incidence of causes of mortality studied. This will enable us to investigate if increased cause-specific mortality is related to increased disease incidence.
7. We will use a self-controlled case series to investigate for potential residual confounding. Self-controlled study designs control for between-person confounding by design. A discrepancy between the findings of the cohort study and self-controlled case series may indicate residual confounding in the cohort study.

References [additional reference]

Petersen I, Douglas I, Whitaker H. Self-controlled case series methods: an alternative to standard epidemiological study designs. BMJ. 2016 Sep 12;354:i4515.

ISAC EVALUATION OF PROTOCOLS FOR RESEARCH INVOLVING CPRD DATA

FEEDBACK TO APPLICANTS

CONFIDENTIAL			***by e-mail***
PROTOCOL NO:	17_252		
PROTOCOL TITLE:	Proton pump inhibitors and mortality: a cohort study		
APPLICANT:	Dr Ian Douglas, Associate Professor of Pharmacoepidemiology, LSHTM Ian.douglas@lshtm.ac.uk		
APPROVED ☒	**APPROVED WITH COMMENTS (resubmission not required)** ☐	**REVISION/ RESUBMISSION REQUESTED** ☐	**REJECTED** ☐
INSTRUCTIONS: ***Protocols with an outcome of 'Approved' or 'Approved with comments' do not require resubmission to the ISAC.*** **REVIEWER COMMENTS:** This was a pleasure to review, a very well thought out proposal with a clear public health benefit.			
DATE OF ISAC FEEDBACK:	15/11/2017		
DATE OF APPLICANT FEEDBACK:			

For protocols approved from 01 April 2014 onwards, applicants are required to include the ISAC protocol in their journal submission with a statement in the manuscript indicating that it had been approved by the ISAC (with the reference number) and made available to the journal reviewers. If the protocol was subject to any amendments, the last amended version should be the one submitted.

***** Please refer to the ISAC advice about protocol amendments provided below*****

Amendments to protocols approved by ISAC Version June 2015

During the course of some studies, it may become necessary to deviate from a protocol which has been approved by ISAC. Any deviation to an ISAC approved protocol should be clearly documented by the applicant but not all such amendments need be submitted for ISAC review and approval. The general principles to be applied in regard to the need for submission are as follows:

- Major amendments should be submitted
- Minor amendments need not be submitted (but must still be documented by the applicant and should normally be mentioned at the publication stage)

In cases of uncertainty, the applicant should contact the ISAC secretariat for advice quoting the original reference number and providing a brief explanation of the nature of the amendment(s) and underlying reason(s).

Major Amendments

We consider an amendment as major if it substantially changes the study design or analysis plan of the proposed research. An amendment should be considered major if it involves the following (although this is not necessarily an exhaustive list):

- A change to the primary hypothesis being tested in the research
- A change to the design of the study
- Additional outcomes or exposures unrelated to the main focus of the approved study*
- Non-trivial changes to the analysis strategy
- Not performing a primary outcome analysis
- Omissions from the analysis plan which may impact on important validity issues such as confounding
- Change of Chief Investigator
- Use of additional linkages to other databases
- Any new proposal involving contact with health professionals or patient or change in regard to such matters

* N.B. extensive changes in this respect will require a new protocol rather than an amendment - if in doubt please consult the Secretariat

Minor Amendments

Examples of amendments which can generally be considered minor include the following:

- Change of personnel other than the Chief Investigator (these should be notified to the Secretariat)
- A change to the definition of the study population, providing the change is mentioned and justified in the paper/output [NB previously major]
- Extension of the time period in relation to defining the study population
- Changes to the definitions of outcomes or exposures of interest, providing the change is mentioned and justified in the paper/output [NB previously major]
- Not using linked data which are part of the approved protocol, unless the linked data are considered critical in defining exposures or outcomes (in which case this would be a major amendment)

- Limited additional analysis suggested by unexpected findings, provided these are clearly presented as post-hoc
- Additional methods to further control for confounding or sensitivity analysis provided these are to be reported as secondary to the main findings
- Validation and data quality work provided additional information from GPs is not required

To submit an amendment of protocol to the ISAC, please submit the following documents to the ISAC mailbox (isac@cprd.com)
1. A covering letter providing justification for the request
2. A completed and, if necessary, updated application form with all changes highlighted; if new linkages are required the current version of the ISAC application form must be completed. Otherwise, the original form may be amended as necessary
3. **The updated protocol document containing the heading 'Amendment' at the end of it.** Please include all amendments to the protocol under this heading. No other changes should be made to the already approved document.

Appendix G

License Agreement for Papers A & D

Paper A

Paper A (https://doi.org/10.1002/pds.5121) is an open access article published under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

Paper D

Paper D (https://doi.org/10.1111/bcp.14728) is an an open access article published under the terms of the Creative Commons Attribution-NonCommercial License (https://creativecommons.org/licenses/by-nc/4.0/), which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

Bibliography

Abubakar, I., Tillmann, T. and Banerjee, A. Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death, 1990-2013: a systematic analysis for the global burden of disease study 2013. 385(9963):117–171, 2015.

Alexander, S.P. et al. The concise guide to pharmacology 2019/20: G protein-coupled receptors. 176: S21–S141, 2019a.

Alexander, S.P. et al. The concise guide to pharmacology 2019/20: Transporters. 176:S397–S493, 2019b.

Alhazzani, W. et al. Efficacy and safety of stress ulcer prophylaxis in critically ill patients: a network meta-analysis of randomized trials. 44(1):1–11, 2018.

Ali, M.S. et al. Reporting of covariate selection and balance assessment in propensity score analysis is suboptimal: a systematic review. *Journal of clinical epidemiology*, 68(2):112–121, 2015. ISSN 1878-5921 (Electronic). doi: 10.1016/j.jclinepi.2014.08.011.

Alshamsi, F. et al. Efficacy and safety of proton pump inhibitors for stress ulcer prophylaxis in critically ill patients: a systematic review and meta-analysis of randomized trials. 20(1):120, 2016.

Austin, P.C. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in medicine*, 28(25):3083–3107, 2009a. ISSN 1097-0258 (Electronic). doi: 10.1002/sim.3697.

Austin, P.C. The relative ability of different propensity score methods to balance measured covariates between treated and untreated subjects in observational studies. *Medical decision making : an international journal of the Society for Medical Decision Making*, 29(6):661–677, 2009b. ISSN 1552-681X (Electronic). doi: 10.1177/0272989X09341755.

Austin, P.C. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate Behav Res*, 46(3):399–424, 2011. ISSN 1532-7906 (Electronic) 0027-3171 (Linking). doi: 10.1080/00273171.2011.568786.

Austin, P.C. and Stuart, E.A. Moving towards best practice when using inverse probability of treatment weighting (iptw) using the propensity score to estimate causal treatment effects in observational studies. 34(28):3661–3679, 2015.

Austin, P.C., Wu, C.F., Lee, D.S. and Tu, J.V. Comparing the high-dimensional propensity score for use with administrative data with propensity scores derived from high-quality clinical data. *Statistical Methods in Medical Research*, 29(2):568–588, 2020. doi: 10.1177/0962280219842362. PMID: 30975044.

Avorn, J. In Defense of Pharmacoepidemiology — Embracing the Yin and Yang of Drug Research. *New England Journal of Medicine*, 357(22):2219–2221, 2007. ISSN 0028-4793. doi: 10.1056/NEJMp0706892.

Azoulay, L., Eberg, M., Benayoun, S. and Pollak, M. 5alpha-Reductase Inhibitors and the Risk of Cancer-Related Mortality in Men With Prostate Cancer. *JAMA Oncol*, 1(3):314–320, 2015. ISSN

2374-2445 (Electronic) 2374-2437 (Linking). doi: 10.1001/jamaoncol.2015.0387.

Bartlett, J.W., Harel, O. and Carpenter, J.R. Asymptotically Unbiased Estimation of Exposure Odds Ratios in Complete Records Logistic Regression. *American journal of epidemiology*, 182(8):730–736, 2015. ISSN 1476-6256 (Electronic). doi: 10.1093/aje/kwv114.

Batchelor, R. et al. Systematic review with meta-analysis: risk of adverse cardiovascular events with proton pump inhibitors independent of clopidogrel. 48(8):780–796, 2018.

Benson, V.S. et al. Associations between gastro-oesophageal reflux, its management and exacerbations of chronic obstructive pulmonary disease. 109(9):1147–54, 2015.

Besançon, L. et al. Open science saves lives: lessons from the COVID-19 pandemic. *BMC Medical Research Methodology*, 21(1):117, 2021. ISSN 1471-2288. doi: 10.1186/s12874-021-01304-y.

Bhaskaran, K. and Smeeth, L. What is the difference between missing completely at random and missing at random? *International journal of epidemiology*, 43(4):1336–1339, 2014. ISSN 1464-3685 (Electronic). doi: 10.1093/ije/dyu080.

Bhaskaran, K., dos Santos-Silva, I., Leon, D.A., Douglas, I.J. and Smeeth, L. Association of bmi with overall and cause-specific mortality: a population-based cohort study of 3· 6 million adults in the uk. 6(12):944–953, 2018.

Bhatt, D.L. et al. Clopidogrel with or without Omeprazole in Coronary Artery Disease. *New England Journal of Medicine*, 363(20):1909–1917, 2010. doi: 10.1056/NEJMoa1007964.

Binder, H., Sauerbrei, W. and Royston, P. Comparison between splines and fractional polynomials for multivariable model building with continuous covariates: a simulation study with continuous response. *Statistics in medicine*, 32(13):2262–2277, 2013. ISSN 1097-0258 (Electronic). doi: 10.1002/sim.5639.

Blake, H.A. et al. Propensity scores using missingness pattern information: a practical guide. *Statistics in Medicine*, 39(11):1641–1657, 2020. ISSN 0277-6715. doi: https://doi.org/10.1002/sim.8503.

Blakely, T., Lynch, J., Simons, K., Bentley, R. and Rose, S. Reflection on modern methods: when worlds collide—prediction, machine learning and causal inference. *International Journal of Epidemiology*, 49(6):2058–2064, 2020. ISSN 0300-5771. doi: 10.1093/ije/dyz132.

Bombardier, C. et al. Comparison of upper gastrointestinal toxicity of rofecoxib and naproxen in patients with rheumatoid arthritis. VIGOR Study Group. *The New England journal of medicine*, 343(21):1520–8, 2 p following 1528, 2000. ISSN 0028-4793 (Print). doi: 10.1056/NEJM200011233432103.

Brenner, H. and Blettner, M. Controlling for continuous confounders in epidemiologic research. *Epidemiology (Cambridge, Mass.)*, 8(4):429–434, 1997. ISSN 1044-3983 (Print).

Brookhart, M.A. et al. Variable selection for propensity score models. *Am J Epidemiol*, 163(12):1149–1156, 2006.

Brookhart, M.A., Sturmer, T., Glynn, R.J., Rassen, J. and Schneeweiss, S. Confounding control in healthcare database research: challenges and potential approaches. *Med Care*, 48(6 Suppl):S114–20, 2010. ISSN 1537-1948 (Electronic) 0025-7079 (Linking). doi: 10.1097/MLR.0b013e3181dbebe3.

Bross, I. Spurious effects from an extraneous variable. *J Chronic Dis*, 19:637–647, 1966.

Brown, H.K. et al. Association Between Serotonergic Antidepressant Use During Pregnancy and Autism Spectrum Disorder in Children. *JAMA*, 317(15):1544–1552, 2017. ISSN 0098-7484. doi: 10.1001/jama.2017.3415.

Brown, J.P. et al. Proton pump inhibitors and risk of all-cause and cause-specific mortality: A cohort study. *British Journal of Clinical Pharmacology*, n/a(n/a), 2021. doi: https://doi.org/10.1111/bcp.14728.

Cadarette, S.M. et al. Diffusion of Innovations model helps interpret the comparative uptake of two methodological innovations: co-authorship network analysis and recommendations for the integration of novel methods in practice. *Journal of clinical epidemiology*, 84:150–160, 2017. ISSN 1878-5921 (Electronic). doi: 10.1016/j.jclinepi.2016.12.006.

Carpenter, J.R. and Kenward, M.G. *Multiple Imputation and its Application*. Wiley, 2013.

Charlot, M. et al. Proton-pump inhibitors are associated with increased cardiovascular risk independent of clopidogrel use a nationwide cohort study. 153(6):378–386, 2010.

Cheema, E. Investigating the association of proton pump inhibitors with chronic kidney disease and its impact on clinical practice and future research: a review. 12(1):6, 2019.

Cheung, K.S. et al. Long-term proton pump inhibitors and risk of gastric cancer development after treatment for helicobacter pylori: a population-based study. 67(1):28–35, 2018.

Collins, R., Bowman, L., Landray, M. and Peto, R. The Magic of Randomization versus the Myth of Real-World Evidence. *New England Journal of Medicine*, 382(7):674–678, 2020. ISSN 0028-4793. doi: 10.1056/NEJMsb1901642.

Council of European Union. Council regulation (EU) no 1235/2010. Technical report, 2010.

Cowling, T.E., Cromwell, D.A., Bellot, A., Sharples, L.D. and van der Meulen, J. Logistic regression and machine learning predicted patient mortality from large sets of diagnosis codes comparably. *Journal of clinical epidemiology*, 133:43–52, 2021. ISSN 1878-5921 (Electronic). doi: 10.1016/j.jclinepi.2020.12.018.

CPRD. Clinical practice ressearch datalink linked data. Technical report.

CPRD. CPRD Online Bibliography. Technical report, 2021.

Crump, R.K., Hotz, V.J., Imbens, G.W. and Mitnik, O.A. Dealing with limited overlap in estimation of average treatment effects. 96(1):187–199, 2009.

de Vries, F., de Vries, C., Cooper, C., Leufkens, B. and van Staa, T.P. Reanalysis of two studies with contrasting results on the association between statin use and fracture risk: the General Practice Research Database. *International journal of epidemiology*, 35(5):1301–1308, 2006. ISSN 0300-5771 (Print). doi: 10.1093/ije/dyl147.

Demcsák, A. et al. Ppis are not responsible for elevating cardiovascular risk in patients on clopidogrel—a systematic review and meta-analysis. 9:1550, 2018.

Desai, R.J. and Franklin, J.M. Alternative approaches for confounding adjustment in observational studies using weighting based on the propensity score: a primer for practitioners. *BMJ*, 367, 2019.

ISSN 0959-8138. doi: 10.1136/bmj.l5657.

Dial, S., Delaney, J.A.C., Barkun, A.N. and Suissa, S. Use of gastric acid-suppressive agents and the risk of community-acquired clostridium difficile-associated disease. 294(23):2989–2995, 2005.

Ding, P. and VanderWeele, T.J. Sensitivity analysis without assumptions. 27(3):368–77, 2016.

Douglas, I.J. et al. Clopidogrel and interaction with proton pump inhibitors: comparison between cohort and within person study designs. *BMJ*, 345:e4388, 2012. ISSN 1756-1833 (Electronic) 0959-535X (Linking). doi: 10.1136/bmj.e4388.

Dultz, G., Piiper, A., Zeuzem, S., Kronenberger, B. and Waidmann, O. Proton pump inhibitor treatment is associated with the severity of liver disease and increased mortality in patients with cirrhosis. 41(5):459–466, 2015.

Eberg, M., Platt, R.W., Reynier, P. and Filion, K.B. Estimation of high-dimensional propensity scores with multiple exposure levels. *Pharmacoepidemiology and Drug Safety*, 29(S1):53–60, 2020. ISSN 1053-8569. doi: https://doi.org/10.1002/pds.4890.

Enders, D., Ohlmeier, C. and Garbe, E. The Potential of High-Dimensional Propensity Scores in Health Services Research: An Exemplary Study on the Quality of Care for Elective Percutaneous Coronary Interventions. *Health services research*, 53(1):197–213, 2018. ISSN 1475-6773 (Electronic). doi: 10.1111/1475-6773.12653.

Farmer, R. et al. Promises and pitfalls of electronic health record analysis. *Diabetologia*, 61(6):1241–1248, 2018. ISSN 1432-0428. doi: 10.1007/s00125-017-4518-6.

Fedeli, U. et al. Multiple causes of death analysis of chronic diseases: the example of diabetes. 13(1): 21, 2015.

Franklin, J.M., Schneeweiss, S., Polinski, J.M. and Rassen, J.A. Plasmode simulation for the evaluation of pharmacoepidemiologic methods in complex healthcare databases. *Comput Stat Data Anal*, 72: 219–226, 2014. doi: 10.1016/j.csda.2013.10.018.

Franklin, J.M., Eddings, W., Glynn, R.J. and Schneeweiss, S. Regularized regression versus the high-dimensional propensity score for confounding adjustment in secondary database analyses. *Am J Epidemiol*, 28:237–248, 2015.

Franklin, J.M. et al. Observing versus Predicting: Initial Patterns of Filling Predict Long-Term Adherence More Accurately Than High-Dimensional Modeling Techniques. *Health services research*, 51(1):220–239, 2016. ISSN 1475-6773 (Electronic). doi: 10.1111/1475-6773.12310.

Franklin, J.M., Glynn, R.J., Martin, D. and Schneeweiss, S. Evaluating the Use of Nonrandomized Real-World Data Analyses for Regulatory Decision Making. *Clinical pharmacology and therapeutics*, 105(4):867–877, 2019. ISSN 1532-6535 (Electronic). doi: 10.1002/cpt.1351.

Freemantle, N. et al. Making inferences on treatment effects from real world data: propensity scores, confounding by indication, and other perils for the unwary in observational research. *BMJ*, 347: f6409, 2013. ISSN 1756-1833 (Electronic) 0959-535X (Linking). doi: 10.1136/bmj.f6409.

Funk, M.J. et al. Doubly Robust Estimation of Causal Effects. *American Journal of Epidemiology*, 173(7):761–767, 2011. ISSN 0002-9262. doi: 10.1093/aje/kwq439.

Gangji, A.S., Cukierman, T., Gerstein, H.C., Goldsmith, C.H. and Clase, C.M. A systematic review and meta-analysis of hypoglycemia and cardiovascular events: a comparison of glyburide with other secretagogues and with insulin. *Diabetes care*, 30(2):389–394, 2007. ISSN 0149-5992 (Print). doi: 10.2337/dc06-1789.

Garbe, E., Kloss, S., Suling, M., Pigeot, I. and Schneeweiss, S. High-dimensional versus conventional propensity scores in a comparative effectiveness study of coxibs and reduced upper gastrointestinal complications. *European journal of clinical pharmacology*, 69(3):549–557, 2013. ISSN 1432-1041 (Electronic). doi: 10.1007/s00228-012-1334-2.

Gisbert, J.P. et al. Meta-analysis: proton pump inhibitors vs. H2-receptor antagonists–their efficacy with antibiotics in Helicobacter pylori eradication. *Alimentary pharmacology & therapeutics*, 18(8): 757–766, 2003. ISSN 0269-2813 (Print). doi: 10.1046/j.1365-2036.2003.01766.x.

Glynn, R.J., Schneeweiss, S. and Sturmer, T. Indications for Propensity Scores and Review of their Use in Pharmacoepidemiology. *Basic Clin Pharmacol Toxicol*, 98(3):253–259, 2006.

Granger, E., Sergeant, J.C. and Lunt, M. Avoiding pitfalls when combining multiple imputation and propensity scores. *Statistics in Medicine*, 38(26):5120–5132, 2019a. ISSN 0277-6715. doi: https://doi.org/10.1002/sim.8355.

Granger, E., Sergeant, J.C. and Lunt, M. Avoiding pitfalls when combining multiple imputation and propensity scores. *Statistics in Medicine*, 38(26):5120–5132, 2019b. doi: https://doi.org/10.1002/sim.8355.

Granger, E., Watkins, T., Sergeant, J.C. and Lunt, M. A review of the use of propensity score diagnostics in papers published in high-ranking medical journals. *BMC medical research methodology*, 20(1):132, 2020. ISSN 1471-2288. doi: 10.1186/s12874-020-00994-0.

Greenland, S. The effect of misclassification in the presence of covariates. *American journal of epidemiology*, 112(4):564–569, 1980. ISSN 0002-9262 (Print). doi: 10.1093/oxfordjournals.aje.a113025.

Greenland, S. Invited commentary: Variable selection versus shrinkage in the control of multiple confounders. *American Journal of Epidemiology*, 167(5):523–529, 2008. ISSN 00029262. doi: 10.1093/aje/kwm355.

Greenland, S. and Finkle, W.D. A critical look at methods for handling missing covariates in epidemiologic regression analyses. *American journal of epidemiology*, 142(12):1255–1264, 1995. ISSN 0002-9262 (Print). doi: 10.1093/oxfordjournals.aje.a117592.

Greenland, S. and Robins, J.M. Identifiability, exchangeability, and epidemiological confounding. *International journal of epidemiology*, 15(3):413–419, 1986. ISSN 0300-5771 (Print). doi: 10.1093/ije/15.3.413.

Greenland, S., Pearl, J. and Robins, J.M. Causal Diagrams for Epidemiologic Research. *Epidemiology*, 10(1), 1999. ISSN 1044-3983.

Groenwold, R.H. et al. Missing covariate data in clinical research: when and when not to use the missing-indicator method for analysis. *CMAJ*, 184(11):1265–1269, 2012. ISSN 0820-3946. doi: 10.1503/cmaj.110977.

Groenwold, R.H.H. et al. Adjustment for continuous confounders: an example of how to prevent residual confounding. *CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne*, 185(5):401–406, 2013. ISSN 1488-2329. doi: 10.1503/cmaj.120592.

Guo, S. and Fraser, M.W. *Propensity score analysis: statistcal methods applications*. Sage, 2010.

Haghish, E.F. Developing, maintaining, and hosting stata statistical software on github. *The Stata Journal*, 20(4):931–951, 2020. doi: 10.1177/1536867X20976323.

Hallas, J. and Pottegard, A. Performance of the High-dimensional Propensity Score in a Nordic Healthcare Model. *Basic Clin Pharmacol Toxicol*, 120:312–317, 2017.

Hennessy, S. Use of health care databases in pharmacoepidemiology. *Basic & Clinical Pharmacology & Toxicology*, 98(3):311–313, 2006. doi: https://doi.org/10.1111/j.1742-7843.2006.pto_368.x.

Herbert, A., Wijlaars, L., Zylbersztejn, A., Cromwell, D. and Hardelid, P. Data resource profile: Hospital episode statistics admitted patient care (hes apc). 46(4):1093–1093i, 2017.

Hernán, M. and Robins, J. *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC., 2020.

Hernan, M.A. The hazards of hazard ratios. *Epidemiology*, 21(1):13–15, 2010. ISSN 1531-5487 (Electronic) 1044-3983 (Linking). doi: 10.1097/EDE.0b013e3181c1ea43.

Hernan, M.A. and Robins, J.M. Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. *American journal of epidemiology*, 183(8):758–764, 2016. ISSN 1476-6256 (Electronic). doi: 10.1093/aje/kwv254.

Hernán, M.A., Brumback, B. and Robins, J.M. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology (Cambridge, Mass.)*, 11(5): 561–570, 2000. ISSN 1044-3983 (Print). doi: 10.1097/00001648-200009000-00012.

Herrett, E., Smeeth, L., Walker, L., Weston, C. and Group, M.A. The Myocardial Ischaemia National Audit Project (MINAP). *Heart*, 96(16):1264–1267, 2010a. ISSN 1468-201X (Electronic) 1355-6037 (Linking). doi: 10.1136/hrt.2009.192328.

Herrett, E., Thomas, S.L., Schoonen, W.M., Smeeth, L. and Hall, A.J. Validation and validity of diagnoses in the general practice research database: a systematic review. 69(1):4–14, 2010b.

Herrett, E. et al. Data Resource Profile: Clinical Practice Research Datalink (CPRD). *Int J Epidemiol*, 44(3):827–836, 2015. ISSN 1464-3685 (Electronic) 0300-5771 (Linking). doi: 10.1093/ije/dyv098.

Ho, D.E., Imai, K., King, G. and Stuart, E.A. Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference. *Political Analysis*, 15(3):199–236, 2007. ISSN 1047-1987. doi: DOI:10.1093/pan/mpl013.

Holland, P.W. Statistics and Causal Inference. *J Am Stat Ass.*, 81(396):945–960, 1986.

Holmes, M.V., Perel, P., Shah, T., Hingorani, A.D. and Casas, J.P. CYP2C19 genotype, clopidogrel metabolism, platelet function, and cardiovascular events: a systematic review and meta-analysis. *JAMA*, 306(24):2704–2714, 2011. ISSN 1538-3598 (Electronic). doi: 10.1001/jama.2011.1880.

Iwagami, M. et al. Validity of estimated prevalence of decreased kidney function and renal replacement therapy from primary care electronic health records compared with national survey and registry data in the United Kingdom. *Nephrology, dialysis, transplantation : official publication of the European*

Dialysis and Transplant Association - European Renal Association, 32(suppl_2):ii142–ii150, 2017. ISSN 1460-2385. doi: 10.1093/ndt/gfw318.

Jackson, J.W., Schmid, I. and Stuart, E.A. Propensity Scores in Pharmacoepidemiology: Beyond the Horizon. *Current Epidemiology Reports*, 4(4):271–280, 2017. ISSN 2196-2995. doi: 10.1007/s40471-017-0131-y.

Ju, C. et al. Scalable collaborative targeted learning for high-dimensional data. *Statistical methods in medical research*, 28(2):532–554, 2019. ISSN 1477-0334 (Electronic). doi: 10.1177/0962280217729845.

Karim, M.E., Pang, M. and Platt, R.W. Can we train machine learning methods to outperform the high-dimensional propensity score algortihm. *Epidemiology*, page Epub ahead of print, 2017. doi: 10.1097/EDE.0000000000000787.

Karim, M.E., Pang, M. and Platt, R.W. Can We Train Machine Learning Methods to Outperform the High-dimensional Propensity Score Algorithm? *Epidemiology*, 29(2):191–198, 2018. ISSN 15315487. doi: 10.1097/EDE.0000000000000787.

Krieger, N. and Davey Smith, G. The tale wagged by the DAG: broadening the scope of causal inference and explanation for epidemiology. *International Journal of Epidemiology*, 45(6):1787–1808, 2016. ISSN 0300-5771. doi: 10.1093/ije/dyw114.

Laheij, R.J.F. et al. Risk of community-acquired pneumonia and use of gastric acid-suppressive drugs. 292(16):1955–1960, 2004.

Langan, S.M. et al. The reporting of studies conducted using observational routinely collected health data statement for pharmacoepidemiology (RECORD-PE). *BMJ*, 363:k3532, 2018. ISSN 17561833. doi: 10.1136/bmj.k3532.

Lee, S.W. et al. Proton pump inhibitors did not increase risk of pneumonia in patients with chronic obstructive pulmonary disease. 7(11):880, 2015.

Lendle, S. R Code for high-dimensional propensity score. 2017.

Lester, H. The uk quality and outcomes framework. *BMJ*, 337, 2008. ISSN 0959-8138. doi: 10.1136/bmj.a2095.

Leyrat, C. et al. Propensity score analysis with partially observed covariates: How should multiple imputation be used? *Stat Methods Med Res*, 28(1):3–19, 2019. ISSN 1477-0334 (Electronic) 0962-2802 (Linking). doi: 10.1177/0962280217713032.

Li, M., Luo, Z., Yu, S. and Tang, Z. Proton pump inhibitor use and risk of dementia: Systematic review and meta-analysis. 98(7):e14422, 2019.

Liu, W., Brookhart, M.A., Schneeweiss, S., Mi, X. and Setoguchi, S. Implications of M Bias in Epidemiologic Studies: A Simulation Study. *American Journal of Epidemiology*, 176(10):938–948, 2012. ISSN 0002-9262. doi: 10.1093/aje/kws165.

Llorente, C. et al. Gastric acid suppression promotes alcoholic liver disease by inducing overgrowth of intestinal enterococcus. 8(1):837, 2017.

Lunt, M. Selecting an appropriate caliper can be essential for achieving good balance with propen-

sity score matching. *American journal of epidemiology*, 179(2):226–235, 2014. ISSN 1476-6256 (Electronic). doi: 10.1093/aje/kwt212.

MacDonald, T.M., Morant, S.V., Goldstein, J.L., Burke, T.A. and Pettitt, D. Channelling bias and the incidence of gastrointestinal haemorrhage in users of meloxicam, coxibs, and older, non-specific non-steroidal anti-inflammatory drugs. *Gut*, 52(9):1265–1270, 2003. ISSN 0017-5749. doi: 10.1136/gut.52.9.1265.

Martin, R.M., Biswas, P. and Mann, R.D. The incidence of adverse events and risk factors for upper gastrointestinal disorders associated with meloxicam use amongst 19,087 patients in general practice in England: cohort study. *British journal of clinical pharmacology*, 50(1):35–42, 2000. ISSN 0306-5251 (Print). doi: 10.1046/j.1365-2125.2000.00229.x.

Mathur, R. et al. Ethnic disparities in initiation and intensification of diabetes treatment in adults with type 2 diabetes in the UK, 1990–2017: A cohort study. *PLOS Medicine*, 17(5):e1003106, 2020.

Moayyedi, P. et al. Safety of proton pump inhibitors based on a large, multi-year, randomized trial of patients receiving rivaroxaban or aspirin. 157(3):682–691. e2, 2019.

Morales, D.R. Rapid response to: Temporal trends in use of tests in UK primary care, 2000-15: retrospective analysis of 250 million tests. *BMJ*, 363:k4666, 2018a. doi: 10.1136/bmj.k4666.

Morales, D.R. Challenges in interpreting trends in testing for α_1-antitrypsin deficiency in COPD patients from UK primary care. *European Respiratory Journal*, 52(6):1801986, 2018b. doi: 10.1183/13993003.01986-2018.

Morris, T.P., White, I.R. and Crowther, M.J. Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, 38(11):2074–2102, 2019. ISSN 0277-6715. doi: https://doi.org/10.1002/sim.8086.

Myers, J.A. et al. Effects of adjusting for instrumental variables on bias and precision of effect estimates. *American journal of epidemiology*, 174(11):1213–1222, 2011. ISSN 1476-6256. doi: 10.1093/aje/kwr364.

Neugebauer, R. et al. High-dimensional propensity score algorithm in comparative effectiveness research with time-varying interventions. *Statistics in medicine*, 34(5):753–781, 2015. ISSN 1097-0258 (Electronic). doi: 10.1002/sim.6377.

NHS. Blood test reference ranges. Technical report, NHS Scotland, 2017.

NHS. Lower layer super output area. Technical report, 2020.

NHS Digital. Coding Cross Maps. Technical report, 2019a.

NHS Digital. BNF Classification. Technical report, 2019b.

NHS Digital. Hospital Episode Statistics (HES). Technical report, 2020.

O'Sullivan, J.W. et al. Temporal trends in use of tests in UK primary care, 2000-15: retrospective analysis of 250 million tests. *BMJ*, 363:k4666, 2018. doi: 10.1136/bmj.k4666.

Padmanabhan, S. et al. Approach to record linkage of primary care data from Clinical Practice Research Datalink to other health-related patient data: overview and implications. *European journal of epidemiology*, 34(1):91–99, 2019. ISSN 1573-7284 (Electronic). doi: 10.1007/s10654-018-0442-4.

Patorno, E., Glynn, R.J., Hernández-díaz, S., Liu, J. and Schneeweiss, S. Studies with Many Covariates and Few Outcomes. *Epidemiology*, 25(2):268–278, 2014. doi: 10.1097/EDE.0000000000000069.

Patrick, A.R. et al. The implications of propensity score variable selection strategies in pharmacoepidemiology: an empirical illustration. *Pharmacoepidemiology and drug safety*, 20(6):551–559, 2011. ISSN 1099-1557 (Electronic). doi: 10.1002/pds.2098.

Pearl, J. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995. ISSN 0006-3444. doi: 10.1093/biomet/82.4.669.

Pearl, J., Glymour, M. and Jewell, N. *Causal Inference in Statistics: A Primer*. Chichester, UK:Wiley, 2016.

Petersen, I. et al. Health indicator recording in UK primary care electronic health records: key implications for handling missing data. *Clinical epidemiology*, 11:157–167, 2019. ISSN 1179-1349 (Print). doi: 10.2147/CLEP.S191437.

Pham, K. and Hirschberg, R. Global safety of coxibs and NSAIDs. *Current topics in medicinal chemistry*, 5(5):465–473, 2005. ISSN 1568-0266 (Print). doi: 10.2174/1568026054201640.

Pitt, B. et al. The effect of spironolactone on morbidity and mortality in patients with severe heart failure. Randomized Aldactone Evaluation Study Investigators. *The New England journal of medicine*, 341(10):709–717, 1999. ISSN 0028-4793 (Print). doi: 10.1056/NEJM199909023411001.

Pitt, B. et al. Eplerenone, a selective aldosterone blocker, in patients with left ventricular dysfunction after myocardial infarction. *The New England journal of medicine*, 348(14):1309–1321, 2003. ISSN 1533-4406 (Electronic). doi: 10.1056/NEJMoa030207.

Platt, K.D., Saini, S.D. and Kurlander, J.E. Selecting the appropriate patients for proton pump inhibitor discontinuation: A teachable moment. 179(9):1276–1277, 2019.

Platt, R. et al. The U.S. Food and Drug Administration's Mini-Sentinel program: status and direction. *Pharmacoepidemiology and drug safety*, 21 Suppl 1:1–8, 2012. ISSN 1099-1557 (Electronic). doi: 10.1002/pds.2343.

Polinski, J.M., Schneeweiss, S., Glynn, R.J., Lii, J. and Rassen, J.A. Confronting "confounding by health system use" in Medicare Part D: comparative effectiveness of propensity score approaches to confounding adjustment. *Pharmacoepidemiology and drug safety*, 21 Suppl 2(Suppl 2):90–98, 2012. ISSN 1099-1557 (Electronic). doi: 10.1002/pds.3250.

Price, S.J., Stapley, S.A., Shephard, E., Barraclough, K. and Hamilton, W.T. Is omission of free text records a possible source of data loss and bias in Clinical Practice Research Datalink studies? A case–control study. *BMJ Open*, 6(5):e011664, 2016. doi: 10.1136/bmjopen-2016-011664.

Puts, M.T., Lips, P. and Deeg, D.J. Sex differences in the risk of frailty for mortality independent of disability and chronic diseases. 53(1):40–47, 2005.

R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing., 2020.

Rafi, Z. and Greenland, S. Semantic and cognitive tools to aid statistical science: replace confidence and significance by compatibility and surprise. *BMC Medical Research Methodology*, 20(1):244,

2020. ISSN 1471-2288. doi: 10.1186/s12874-020-01105-9.

Rassen, J. et al. Automated use of electronic health record text data to improve validity in pharmacoepidemiology studies. *Pharmacoepidemiol Drug Saf*, 22(S1):376, 2013.

Rassen, J., Doherty, M., Huang, W. and Schneeweiss, S. Pharmacoepidemiology Toolbox, 2020.

Rassen, J.A. and Schneeweiss, S. Using high-dimensional propensity scores to automate confounding control in a distributed medical product safety surveillance system. *Pharmacoepidemiology and drug safety*, 21 Suppl 1:41–49, 2012. ISSN 1099-1557 (Electronic). doi: 10.1002/pds.2328.

Rassen, J.A., Glynn, R.J., Brookhart, M.A. and Schneeweiss, S. Covariate Selection in High-Dimensional Propensity Score Analyses of Treatment Effects in Small Samples. 173(12):1404–1413, 2011a. doi: 10.1093/aje/kwr001.

Rassen, J.A., Glynn, R.J., Brookhart, M.A. and Schneeweiss, S. Covariate selection in high-dimensional propensity score analyses of treatment effects in small samples. *American Journal of Epidemiology*, 173(12):1404–1413, 2011b. ISSN 00029262. doi: 10.1093/aje/kwr001.

Rassen, J.A. et al. One-to-many propensity score matching in cohort studies. *Pharmacoepidemiology and drug safety*, 21 Suppl 2:69–80, 2012. ISSN 1099-1557 (Electronic). doi: 10.1002/pds.3263.

Ray, W.A. Evaluating Medication Effects Outside of Clinical Trials: New-User Designs. *American Journal of Epidemiology*, 158(9):915–920, 2003. ISSN 0002-9262. doi: 10.1093/aje/kwg231.

Robins, J.M. Data, design, and background knowledge in etiologic inference. *Epidemiology (Cambridge, Mass.)*, 12(3):313–320, 2001. ISSN 1044-3983 (Print). doi: 10.1097/00001648-200105000-00011.

Roland, M. and Guthrie, B. Quality and outcomes framework: what have we learnt? *BMJ*, 354, 2016. doi: 10.1136/bmj.i4060.

Rosenbaum, P.R. and Rubin, D.B. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Rothnie, K.J., Chandan, J.S., Goss, H.G., Müllerová, H. and Quint, J.K. Validity and interpretation of spirometric recordings to diagnose COPD in UK primary care. *International journal of chronic obstructive pulmonary disease*, 12:1663–1668, 2017. ISSN 1178-2005. doi: 10.2147/COPD.S133891.

Royston, P., Altman, D.G. and Sauerbrei, W. Dichotomizing continuous predictors in multiple regression: a bad idea. *Statistics in medicine*, 25(1):127–141, 2006. ISSN 0277-6715 (Print). doi: 10.1002/sim.2331.

Rubin, D.B. Inference and Missing Data. *Biometrika*, 63(3):581–592, 1976. ISSN 00063444. doi: 10.2307/2335739.

Rubin, D.B. On principles for modeling propensity scores in medical research., 2004. ISSN 1053-8569 (Print).

Schneeweiss, S. Automated data-adaptive analytics for electronic healthcare data to study causal treatment effects. *J Clin Epidemiol*, 10:771–788, 2018.

Schneeweiss, S. Theory meets practice: a commentary on VanderWeele's 'principles of confounder selection'. *European Journal of Epidemiology*, 34(3):221–222, 2019. ISSN 1573-7284. doi: 10.1007/s10654-019-00495-5.

Schneeweiss, S. and Avorn, J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. *J Clin Epidemiol*, 58(4):323–337, 2005. ISSN 0895-4356 (Print) 0895-4356 (Linking). doi: 10.1016/j.jclinepi.2004.10.012.

Schneeweiss, S. and Glynn, R.J. Real-World Data Analytics Fit for Regulatory Decision-Making. *American journal of law & medicine*, 44(2-3):197–217, 2018. ISSN 0098-8588 (Print). doi: 10.1177/0098858818789429.

Schneeweiss, S. et al. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology*, 20(4):512–522, 2009. ISSN 1531-5487 (Electronic) 1044-3983 (Linking). doi: 10.1097/EDE.0b013e3181a663cc.

Schneeweiss, S. et al. Supplementing claims data with outpatient laboratory test results to improve confounding adjustment in effectiveness studies of lipid-lowering treatments. *BMC Medical Research Methodology*, 12(1):180, 2012. ISSN 1471-2288. doi: 10.1186/1471-2288-12-180.

Schneeweiss, S. et al. Variable Selection for Confounding Adjustment in High-dimensional Covariate Spaces When Analyzing Healthcare Databases. *Epidemiology*, 28(2):237–248, 2017. ISSN 15315487. doi: 10.1097/EDE.0000000000000581.

Schuster, T., Pang, M. and Platt, R.W. On the role of marginal confounder prevalence - implications for the high-dimensional propensity score algorithm. *Pharmacoepidemiology and drug safety*, 24(9): 1004–1007, 2015. ISSN 1099-1557 (Electronic). doi: 10.1002/pds.3773.

Shah, A.D. et al. Natural language processing for disease phenotyping in UK primary care records for research: a pilot study in myocardial infarction and death. *Journal of Biomedical Semantics*, 10(1): 20, 2019. ISSN 2041-1480. doi: 10.1186/s13326-019-0214-4.

Shah, B.R., Laupacis, A., Hux, J.E. and Austin, P.C. Propensity score methods gave similar results to traditional regression modeling in observational studies: a systematic review. *Journal of Clinical Epidemiology*, 58(6):550–559, 2005. ISSN 0895-4356. doi: https://doi.org/10.1016/j.jclinepi.2004.10.016.

Silverstein, F.E. et al. Gastrointestinal Toxicity With Celecoxib vs Nonsteroidal Anti-inflammatory Drugs for Osteoarthritis and Rheumatoid ArthritisThe CLASS Study: A Randomized Controlled Trial. *JAMA*, 284(10):1247–1255, 2000. ISSN 0098-7484. doi: 10.1001/jama.284.10.1247.

Smeeth, L., Douglas, I., Hall, A.J., Hubbard, R. and Evans, S. Effect of statins on a wide range of health outcomes: a cohort study validated by comparison with randomized trials. *Br J Clin Pharmacol*, 67(1):99–109, 2009. ISSN 1365-2125 (Electronic) 0306-5251 (Linking). doi: 10.1111/j.1365-2125.2008.03308.x.

Sooriakumaran, P. et al. Mortality in men with advanced prostate cancer appears to be reduced with radical treatment compared to androgen deprivation alone. *Eur Urol Suppl*, 1(13):e974–eb, 2014.

StataCorp. Stata Statistical Software: Release 14. College Station, TX:StataCorp LP. 2015.

StataCorp. Stata Statistical Software: Release 15. College Station, TX: StataCorp LLC. 2017.

Sterne, J.A. et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ*, 338:b2393, 2009a. ISSN 1756-1833 (Electronic) 0959-535X (Linking).

doi: 10.1136/bmj.b2393.

Sterne, J.A.C. et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *British Medical Journal*, 339, 2009b. ISSN 0959-8146. doi: ARTNb239310. 1136/bmj.b2393.

Steryerberg, E. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. Springer-Verlag New York, 2009.

Strom, B.L., Kimmel, S.E. and Hennessy, S. *Textbook of Pharmacoepidemiology*. Wiley, 2013. ISBN 9781118344842.

Stuart, E.A. Matching methods for causal inference: A review and a look forward. *Statistical science : a review journal of the Institute of Mathematical Statistics*, 25(1):1–21, 2010. ISSN 0883-4237 (Print). doi: 10.1214/09-STS313.

Sturmer, T. et al. A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. *J Clin Epidemiol*, 59(5):437–447, 2006.

Suissa, S. Statistical methods in pharmacoepidemiology: advances and challenges. *Statistical methods in medical research*, 18(1):3–6, 2009. ISSN 0962-2802 (Print). doi: 10.1177/0962280208099879.

Suissa, S., Dell'Aniello, S. and Ernst, P. Long-Acting Bronchodilator Initiation in COPD and the Risk of Adverse Cardiopulmonary Events: A Population-Based Comparative Safety Study. *Chest*, 151(1): 60–67, 2017a. ISSN 1931-3543 (Electronic) 0012-3692 (Linking). doi: 10.1016/j.chest.2016.08.001.

Suissa, S., Dell'Aniello, S. and Ernst, P. Concurrent use of long-acting bronchodilators in COPD and the risk of adverse cardiovascular events. *The European respiratory journal*, 49(5), 2017b. ISSN 1399-3003 (Electronic). doi: 10.1183/13993003.02245-2016.

Targownik, L.E. et al. Use of proton pump inhibitors and risk of osteoporosis-related fractures. 179 (4):319–326, 2008.

Tazare, J., Smeeth, L., Evans, S.J.W., Williamson, E. and Douglas, I.J. Implementing high-dimensional propensity score principles to improve confounder adjustment in UK electronic health records. *Pharmacoepidemiology and Drug Safety*, 29(11):1373– 1381, 2020. ISSN 1053-8569. doi: 10.1002/pds.5121.

Tian, Y., Schuemie, M.J. and Suchard, M.A. Evaluating large-scale propensity score performance through real-world and synthetic data experiments. *International Journal of Epidemiology*, (June): 2005–2014, 2018. ISSN 0300-5771. doi: 10.1093/ije/dyy120.

Toh, S. Pharmacoepidemiology in the era of real-world evidence. *Current epidemiology reports*, 4(4): 262–265, 2017. ISSN 2196-2995. doi: 10.1007/s40471-017-0123-y.

Toh, S., Garcia Rodriguez, L.A. and Hernan, M.A. Confounding adjustment via a semi-automated high-dimensional propensity score algorithm: an application to electronic medical records. *Pharmacoepidemiol Drug Saf*, 20(8):849–857, 2011. ISSN 1099-1557 (Electronic) 1053-8569 (Linking). doi: 10.1002/pds.2152.

Tvingsholm, S.A., Dehlendorff, C., Osterlind, K., Friis, S. and Jaattela, M. Proton pump inhibitor

use and cancer mortality. 143(6):1315–1326, 2018.

US FDA. Guidance for industry postmarketing studies and clinical trials implementation of section 505(o)(3) of the federal food, drug, and cosmetic act. 2011.

Van Pinxteren, B. et al. Short-term treatment of gastroesophageal reflux disease: A systematic review and meta-analysis of the effect of acid-suppressant drugs in empirical treatment and in endoscopy-negative patients. 18(9):755–763, 2003.

VanderWeele, T.J. Principles of confounder selection. *European Journal of Epidemiology*, 34(3):211–219, 2019. ISSN 1573-7284. doi: 10.1007/s10654-019-00494-6.

VanderWeele, T.J. and Shpitser, I. On the definition of a confounder. *Annals of statistics*, 41(1): 196–220, 2013. ISSN 0090-5364 (Print). doi: 10.1214/12-aos1058.

Virdee, P.S., Fuller, A., Jacobs, M., Holt, T. and Birks, J. Assessing data quality from the Clinical Practice Research Datalink: a methodological approach applied to the full blood count blood test. *Journal of Big Data*, 7(1):96, 2020. ISSN 2196-1115. doi: 10.1186/s40537-020-00375-w.

Webster-Clark, M. et al. Using propensity scores to estimate effects of treatment initiation decisions: State of the science. *Statistics in Medicine*, n/a(n/a), 2020. ISSN 0277-6715. doi: https://doi.org/10.1002/sim.8866.

Wettermark, B. The intriguing future of pharmacoepidemiology. *European journal of clinical pharmacology*, 69 Suppl 1:43–51, 2013. ISSN 1432-1041 (Electronic). doi: 10.1007/s00228-013-1496-6.

Whitaker, H.J., Farrington, C.P., Spiessens, B. and Musonda, P. Tutorial in biostatistics: the self-controlled case series method. *Stat Med*, 0:1–31, 2005.

Williamson, E., Morley, R., Lucas, A. and Carpenter, J. Propensity scores: from naive enthusiasm to intuitive understanding. *Stat Methods Med Res*, 21(3):273–293, 2012. ISSN 1477-0334 (Electronic) 0962-2802 (Linking). doi: 10.1177/0962280210394483.

Williamson, E.J. and Forbes, A. Introduction to propensity scores. *Respirology*, 19(5):625–635, 2014. ISSN 1440-1843 (Electronic) 1323-7799 (Linking). doi: 10.1111/resp.12312.

Williamson, E.J., Forbes, A. and White, I.R. Variance reduction in randomised trials by inverse probability weighting using the propensity score. *Statistics in Medicine*, 33(5):721–737, 2014. ISSN 02776715. doi: 10.1002/sim.5991.

Williamson, E.J. et al. Factors associated with COVID-19-related death using OpenSAFELY. *Nature*, 584(7821):430–436, 2020. ISSN 1476-4687. doi: 10.1038/s41586-020-2521-4.

Wolf, A. et al. Data resource profile: Clinical Practice Research Datalink (CPRD) Aurum. *International journal of epidemiology*, 48(6):1740–1740g, 2019. ISSN 1464-3685 (Electronic). doi: 10.1093/ije/dyz034.

World Health Organisation. International classification of diseases. Technical report, 2019.

Wyss, R., Fireman, B., Rassen, J.A. and Schneeweiss, S. Erratum: High-dimensional Propensity Score Adjustment in Studies of Treatment Effects Using Health Care Claims Data. *Epidemiology*, 29(6): e63–e64, 2018a. ISSN 1044-3983.

Wyss, R. et al. Using Super Learner Prediction Modeling to Improve High-dimensional Propensity

Score Estimation. *Epidemiology*, 29(1):96–106, 2018b.

Xie, Y. et al. Risk of death among users of proton pump inhibitors: a longitudinal observational cohort study of united states veterans. 7(6):e015735, 2017.

Xie, Y. et al. Estimates of all cause mortality and cause specific mortality associated with proton pump inhibitors among us veterans: cohort study. 365:l1580, 2019.

Yang, Y. et al. Proton-pump inhibitors use, and risk of acute kidney injury: a meta-analysis of observational studies. 11:1291–1299, 2017.

Yuan, J. et al. Regular use of proton pump inhibitor and risk of rheumatoid arthritis in women: a prospective cohort study. 52(3):449–458, 2020.

Zannad, F. et al. Eplerenone in patients with systolic heart failure and mild symptoms. *The New England journal of medicine*, 364(1):11–21, 2011. ISSN 1533-4406 (Electronic). doi: 10.1056/NEJMoa1009492.

Zhou, M. et al. Sentinel modular program for propensity score-matched cohort analyses: Application to Glyburide, Glipizide, and Serious Hypoglycemia. *Epidemiology*, 28(6):838–846, 2017. ISSN 15315487. doi: 10.1097/EDE.0000000000000709.