













RESEARCH ARTICLE

OpenSAFELY: Representativeness of electronic health record platform OpenSAFELY-TPP data compared to the population of England [version 1; peer review: 2 approved]

Colm Andrews ¹, Anna Schultze ², Helen Curtis ¹, William Hulme ¹, John Tazare ², Stephen Evans ², Amir Mehrkar¹, Sebastian Bacon¹, George Hickman ¹, Christopher Bates³, John Parry ³, Frank Hester³, Sam Harper³, Jonathan Cockburn³, David Evans ¹, Tom Ward¹, Simon Davy¹, Peter Inglesby¹, Ben Goldacre¹, Brian MacKenna¹, Laurie Tomlinson ², Alex Walker¹

¹Bennett Institute of Applied Data Science, Nuffield Department of Primary Care Health Sciences, University of Oxford, Oxford, Oxon, OX26GG,, UK

²London School of Hygiene & Tropical Medicine, London, WC1E 7HT, UK

³TPP, TPP House, Leeds, Yorkshire, LS18 5PX, UK

V1 First published: 18 Jul 2022, 7:191
<https://doi.org/10.12688/wellcomeopenres.18010.1>
 Latest published: 18 Jul 2022, 7:191
<https://doi.org/10.12688/wellcomeopenres.18010.1>

Abstract

Background: Since its inception in March 2020, data from the OpenSAFELY-TPP electronic health record platform has been used for more than 20 studies relating to the global COVID-19 emergency. OpenSAFELY-TPP data is derived from practices in England using SystemOne software, and has been used for the majority of these studies. We set out to investigate the representativeness of OpenSAFELY-TPP data by comparing it to national population estimates.



Methods: With the approval of NHS England, we describe the age, sex, Index of Multiple Deprivation and ethnicity of the OpenSAFELY-TPP population compared to national estimates from the Office for National Statistics. The five leading causes of death occurring between the 1st January 2020 and the 31st December 2020 were also compared to deaths registered in England during the same period.


Results: Despite regional variations, TPP is largely representative of the general population of England in terms of IMD (all within 1.1 percentage points), age, sex (within 0.1 percentage points), ethnicity and causes of death. The proportion of the five leading causes of death is broadly similar to those reported by ONS (all within 1 percentage point).


Conclusions: Data made available via OpenSAFELY-TPP is broadly representative of the English population. Users of OpenSAFELY must

Open Peer Review

Approval Status  

	1	2
version 1 18 Jul 2022	 view	 view

1. **Tom Fahey** , Royal College of Surgeons in Ireland, Dublin, Ireland

2. **Evangelos Kontopantelis** , University of Manchester, Manchester, UK

Any reports and responses or comments on the article can be found at the end of the article.

consider the issues of representativeness, generalisability and external validity associated with using TPP data for health research. Although the coverage of TPP practices varies regionally across England, TPP registered patients are generally representative of the English population as a whole in terms of key demographic characteristics.

Keywords

Representativeness, Covid research, OpenSafely

Corresponding author: Colm Andrews (colm.andrews@phc.ox.ac.uk)

Author roles: **Andrews C:** Conceptualization, Data Curation, Formal Analysis, Methodology, Project Administration, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Schultze A:** Methodology, Writing – Review & Editing; **Curtis H:** Methodology, Writing – Review & Editing; **Hulme W:** Methodology, Writing – Review & Editing; **Tazare J:** Methodology, Writing – Review & Editing; **Evans S:** Methodology, Writing – Review & Editing; **Mehrkar A:** Project Administration, Resources, Writing – Review & Editing; **Bacon S:** Data Curation, Resources, Software; **Hickman G:** Data Curation, Resources, Software; **Bates C:** Data Curation, Resources, Software; **Parry J:** Data Curation, Resources, Software; **Hester F:** Data Curation, Resources, Software; **Harper S:** Data Curation, Resources, Software; **Cockburn J:** Data Curation, Resources, Software; **Evans D:** Data Curation, Resources, Software; **Ward T:** Data Curation, Resources, Software; **Davy S:** Data Curation, Resources, Software; **Inglesby P:** Data Curation, Resources, Software; **Goldacre B:** Conceptualization, Funding Acquisition, Methodology, Supervision, Writing – Review & Editing; **MacKenna B:** Conceptualization, Supervision, Writing – Review & Editing; **Tomlinson L:** Conceptualization, Methodology, Supervision, Writing – Review & Editing; **Walker A:** Conceptualization, Methodology, Supervision, Writing – Review & Editing

Competing interests: BG has received research funding from the Laura and John Arnold Foundation, the NHS National Institute for Health Research (NIHR), the NIHR School of Primary Care Research, the NIHR Oxford Biomedical Research Centre, the Mohn-Westlake Foundation, NIHR Applied Research Collaboration Oxford and Thames Valley, the Wellcome Trust, the Good Thinking Foundation, Health Data Research UK, the Health Foundation, the World Health Organisation, UKRI, Asthma UK, the British Lung Foundation, and the Longitudinal Health and Wellbeing strand of the National Core Studies programme; he also receives personal income from speaking and writing for lay audiences on the misuse of science.

Grant information: This work was jointly funded by UKRI [COV0076;MR/V015737/1] NIHR and Asthma UK-BLF and the Longitudinal Health and Wellbeing strand of the National Core Studies programme. The OpenSAFELY data science platform is funded by the Wellcome Trust. BG's work on better use of data in healthcare more broadly is currently funded in part by: the Wellcome Trust, NIHR Oxford Biomedical Research Centre, NIHR Applied Research Collaboration Oxford and Thames Valley, the Mohn-Westlake Foundation; all DataLab staff are supported by BG's grants on this work. LS reports grants from Wellcome, MRC, NIHR, UKRI, British Council, GSK, British Heart Foundation, and Diabetes UK outside this work. AS is employed by LSHTM on a fellowship sponsored by GSK. KB holds a Wellcome Senior Research Fellowship (220283/Z/20/Z). BMK is also employed by NHS England working on medicines policy and clinical lead for primary care medicines data. The views expressed are those of the authors and not necessarily those of the NIHR, NHS England, Public Health England or the Department of Health and Social Care. Funders had no role in the study design, collection, analysis, and interpretation of data; in the writing of the report; and in the decision to submit the article for publication.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2022 Andrews C *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Andrews C, Schultze A, Curtis H *et al.* **OpenSAFELY: Representativeness of electronic health record platform OpenSAFELY-TPP data compared to the population of England [version 1; peer review: 2 approved]** Wellcome Open Research 2022, 7:191 <https://doi.org/10.12688/wellcomeopenres.18010.1>

First published: 18 Jul 2022, 7:191 <https://doi.org/10.12688/wellcomeopenres.18010.1>

Abbreviations

EHR - Electronic Health Record

ICD - International Classification of Diseases

IMD - Index of Multiple Deprivation

NUTS - Nomenclature of Territorial Units for Statistics

ONS - Office for National Statistics

SUS - Secondary Uses Service

Background

OpenSAFELY is a secure health analytics platform created by our team on behalf of NHS England. OpenSAFELY provides a secure software interface allowing analysis of pseudonymised primary care patient records from England in near real-time within highly secure data environments. OpenSAFELY software is currently deployed within the secure data centres of the two largest electronic health record providers in the NHS: EMIS and TPP, and is delivering federated analytics where the same data curation and analysis code executes in each environment. To date more than 20 publications have used the OpenSAFELY platform, focused on delivering vital and urgent results related to the global COVID-19 emergency. Some of these papers have used the federated analytics functionality more latterly available in OpenSAFELY to deliver combined analyses across 58 million patients' data in both OpenSAFELY-EMIS and OpenSAFELY-TPP¹⁻⁵; however the majority of analyses published during the pandemic specifically used OpenSAFELY-TPP which covers 40% of general practices in England, those using SystemOne Electronic Health Record (EHR) software produced by TPP.

The use of data from EHR providers is an invaluable tool for health research, however data is primarily collected for clinical use and not specifically with research in mind. As these datasets are not a random sample of the population of interest, it is important to understand the representativeness of the data. The deployment of TPP SystemOne software is known to be geographically clustered⁶, and factors such as sex, age, ethnicity and levels of deprivation, which are important clinical risk factors for death from COVID-19⁷, show regional variability across England⁸. Key outcomes such as causes of death also vary by region⁹. However, little is currently known about how the characteristics of patients in TPP practices compare to the population at large.

In order to aid the interpretation of ongoing COVID-19 research projects in OpenSAFELY-TPP we therefore set out to compare key demographic characteristics of patients registered with TPP practices to national estimates from the Office for National Statistics (ONS). We also compared the distribution of the five leading causes of death registered in ONS between 1st January 2020 and 31st December 2020 to deaths registered in TPP during the same period.

Methods

Study design

Data source. Primary care records managed by the GP software provider TPP were linked to ONS death data through [OpenSAFELY-TPP](#), a data analytics platform created by our team on behalf of NHS England to address urgent COVID-19 research questions. Similarly pseudonymized datasets from other data providers are securely provided to the EHR vendor and linked to the primary care data. The dataset analysed within OpenSAFELY-TPP is based on 24 million people currently registered with GP surgeries using TPP SystemOne software. **It includes pseudonymized data such as coded diagnoses, medications and physiological parameters. No free text data are included.** Further details on our information governance can be found in the [information governance and ethics section](#).

The UK Census collects individual and household-level demographic data every 10 years for the whole UK population. Data on ethnicity were obtained from the 2011 UK Census for England. In addition to census data, ONS release annual mid-year estimates of the resident population of England produced using a cohort component method¹⁰. Data on IMD, Age and sex were obtained from the 2020-mid year estimates and estimates of the 5 most common causes of death in 2020 were obtained from ONS mortality statistics published via NOMIS¹¹.

Study population

For demography and coverage analyses, patients were included in the study if they were registered at an English general practice using a TPP SystemOne clinical information system on 30th June 2020. For analysis of causes of death, patients were included if they were registered with an English general practice using a TPP SystemOne clinical information system on the day of a death registered on ONS between 1st January 2020 and 31st December 2020.

Demographic data

Ethnicity: The primary care recorded ethnicity, supplemented where missing with ethnicity data from the Secondary Uses Service (SUS), was collapsed into the five high-level and 16 detailed census categories of White (White British, White Irish, other White), South Asian (Indian, Pakistani, Bangladeshi, other South Asian), Black (African, Caribbean, other Black), other (Chinese, all others), and mixed (White and Asian, White and African, White and Caribbean, other mixed) with an additional unknown ethnicity category included.

Age: Patients' age was calculated as of 30th June 2020 and grouped into 5 year bands.

Sex: We used categories "male" and "female", matching the ONS recorded categories; patients with any other/unknown sex were included as "unknown".

Deprivation: Deprivation was measured by the Index of Multiple Deprivation (IMD) derived from the patient's postcode at

lower super output area level. IMD was divided into quintiles, with higher values indicating greater deprivation.

Causes of death

Patients were flagged if they had any death certified and registered in England or Wales between 1st January 2020 and 31st December 2020 and where applicable grouped into the 5 most common underlying causes of death (Table 1).

Statistical methods

We investigated the representativeness of TPP data by comparing OpenSAFELY-TPP-derived figures for 2020 with the following: (a) ONS IMD for all of England, (b) ONS age, sex (2020 estimates) and ethnicity (2011 census) across NHS England operating regions, and (c) causes of death (Malignant neoplasm of trachea, bronchus and lung, Ischaemic heart diseases, Dementia and Alzheimer disease, COVID-19 and Cerebrovascular diseases) in 2020 across NHS England operating regions. Proportions of each age group, sex, IMD band, ethnicity and cause of death were calculated and compared to the corresponding ONS data. For mortality analysis the denominator was the total number of deaths in 2020 and the numerator was the number of patients with the relevant ICD10 code (Table 1) as the underlying cause. TPP coverage was calculated as the proportion of TPP registered patients compared to ONS estimated populations within each Nomenclature of Territorial Units for Statistics (NUTS 1) region.

Software and reproducibility

Data management was performed using Python 3.8, with analysis carried out using R. Code for data management and analysis as well as codelists are openly available [online](#) at for inspection and re-use by anyone.

Patient and public involvement

We have developed a publicly available [website](#) through which we invite any patient or member of the public to contact us regarding this study or the broader OpenSAFELY project.

Table 1. Most common underlying causes of death occurring in England 2020¹².

Cause of Death	ICD 10 codes	N (% of all deaths)
COVID-19	U07	73680 (13%)
Dementia and Alzheimer disease	F01-F03, G30	70035 (12%)
Ischaemic heart diseases	I20-I25	55690 (10%)
Cerebrovascular diseases	I60-I69	29680 (5%)
Malignant neoplasm of trachea, bronchus and lung	C33-C34	28720 (5%)

Results

TPP coverage

The population of active TPP patients (alive and registered on 30th June 2020) was 24 million representing 42.6% of the total UK population (based on the UK 2020 mid-year population estimate of 56 million). TPP coverage as a proportion of the ONS population was highest in the East of England (90.5%) and East Midlands (86.1%) and lowest in the West Midlands (16.8%), South East England (17.6%) and London (18.7%) (Figure 1).

IMD

Overall the proportion of IMD groups was similar, with only small differences between the TPP and ONS populations: In those with a recorded IMD there was a slightly higher proportion of TPP patients in the most deprived IMD group 1 (20.5%) and IMD group 3 (21.1) compared to national ONS estimates (20.0 and 20.3 respectively). TPP practices underrepresented patients in the least deprived IMD group 5 (18.3%) compared to ONS (19.4%) (Figure 2). IMD was missing for 2.3% of the TPP records.

Sex

For those with sex recorded as either male or female there was a similar proportion of women in the English population (50%) compared to ONS (50.1%) (Figure 3). The South West of England had the highest proportion of Females in TPP (50.8%) with London having the lowest proportion (48.8%). The difference in proportion of women between TPP and ONS estimates was within 0.1 percentage points for all regions (Figure 4).

Age

There was a higher proportion of TPP patients in the age range 25–59 compared to ONS nationally, with a lower proportion of those under 25 years old (Figure 5). The difference in age distribution between TPP and ONS estimates was highest in London and the age distribution of the South West most closely resembled the ONS estimates (Figure 6).

Age and sex

Across England as a whole the higher proportion of TPP patients in the 35–59 age range compared to ONS estimates was largely due to a higher proportion of men in this age group in TPP. There was a higher proportion of women aged 20–29 in TPP and a lower proportion of men aged 20–29 compared to ONS estimates (Figure 7, Figure 8).

Causes of death

Across England there was a lower proportion of all five of the leading causes of deaths in TPP compared with ONS data (Figure 9). The biggest difference was in COVID-19 (12.2% in TPP, 12.9% in ONS) and the smallest difference was in Malignant neoplasm of trachea, bronchus and lung (4.9% in TPP, 5.0% in ONS). The difference in proportions of all 5 leading causes of death compared to ONS varied by region (Figure 10). COVID was overrepresented in TPP in all regions other than the North West (14.9% in TPP, 14.9% in ONS) and South East (7.5%, 10.0%).

TPP population coverage per NUTS 1 Region

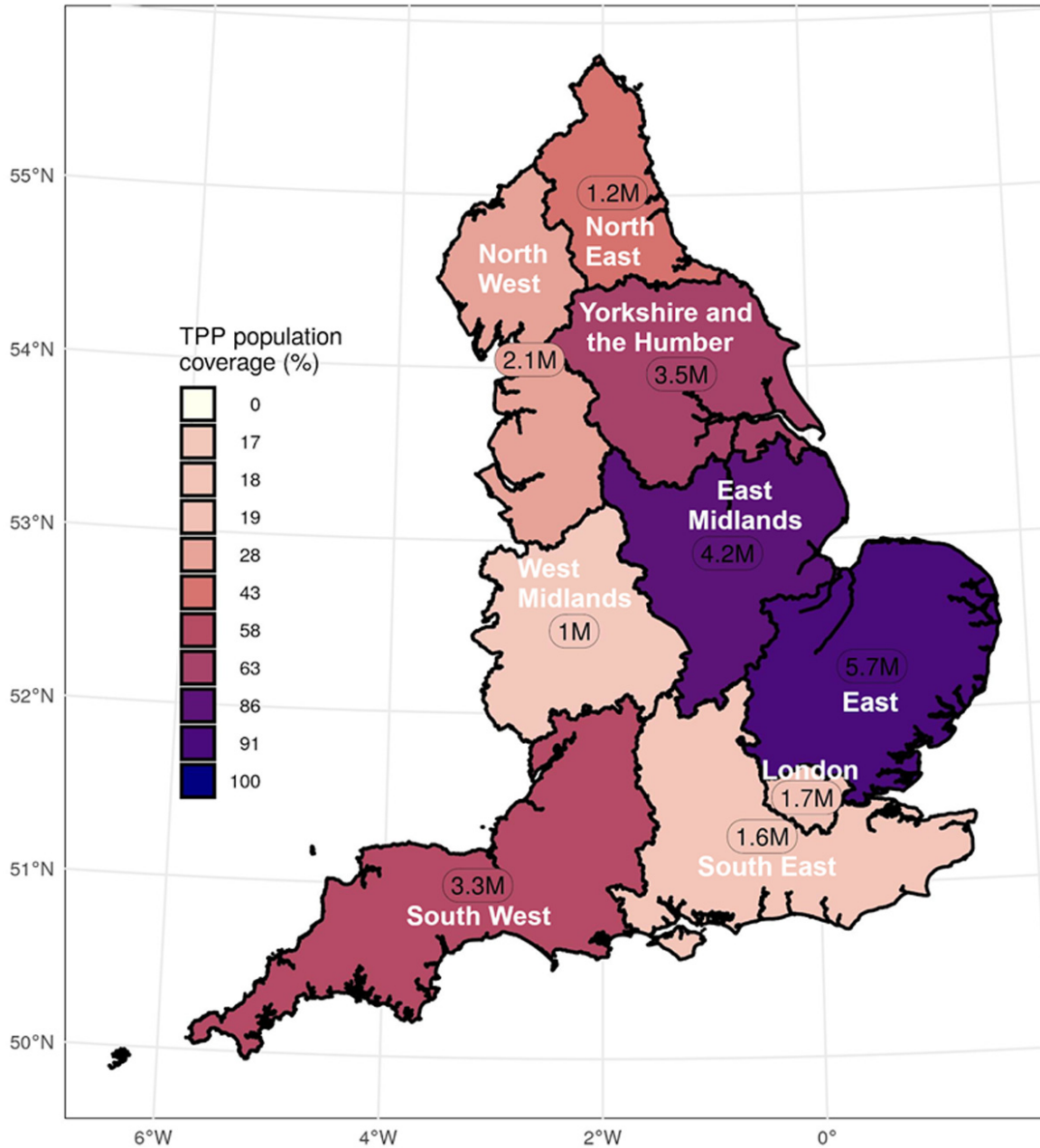


Figure 1. Population coverage map showing coverage of each Nomenclature of Territorial Units for Statistics (NUTS-1) region. Population coverage based on ONS estimates covered by TPP with the number of patients in TPP per region.

Ethnicity (5 groups)

Of those with a recorded ethnicity the proportion of each ethnic group was within 1 percentage point of the ONS estimate across England as a whole for the 5 group ethnicity (Figure 11A, Figure 12). The White population was underrepresented in all regions other than the North West (93.3%, 90.2%) (Figure 13). The Asian population was overrepresented in all regions other than the North West (3.5%, 5.5%) and South East (3.9%, 4.6%)

(Figure 14). Ethnicity was not recorded for 9.4% of the TPP population.

Ethnicity (16 groups)

Of those with a recorded ethnicity there was a lower proportion of White British people in TPP (74.8%) compared to ONS (79.8%) and higher proportion of Other White patients (9.6% TPP, 4.7% ONS). There was a lower proportion of both

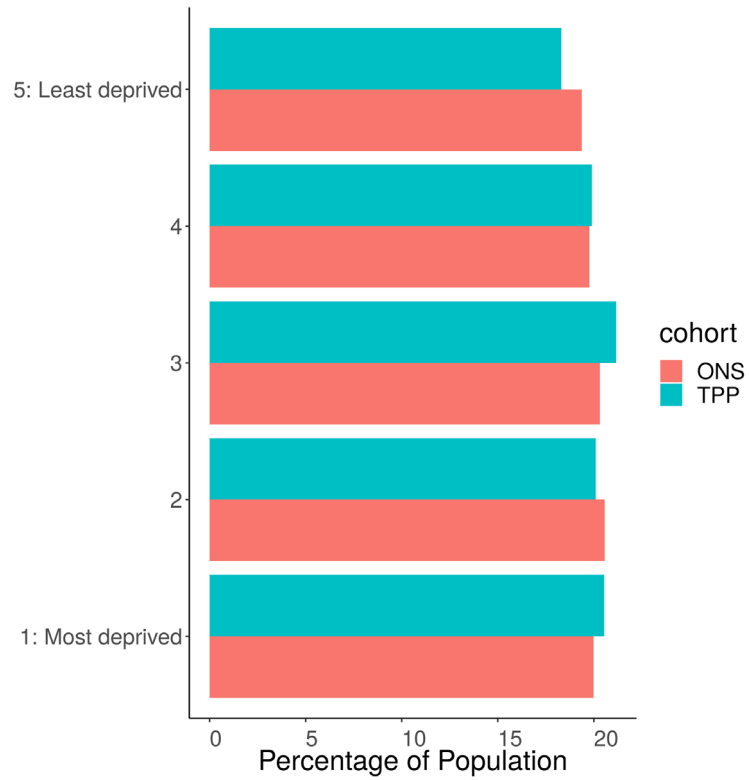


Figure 2. Barplot showing the proportion of ONS and TPP populations per IMD Quintile. The TPP population excludes 2.3% of patients without a recorded IMD.

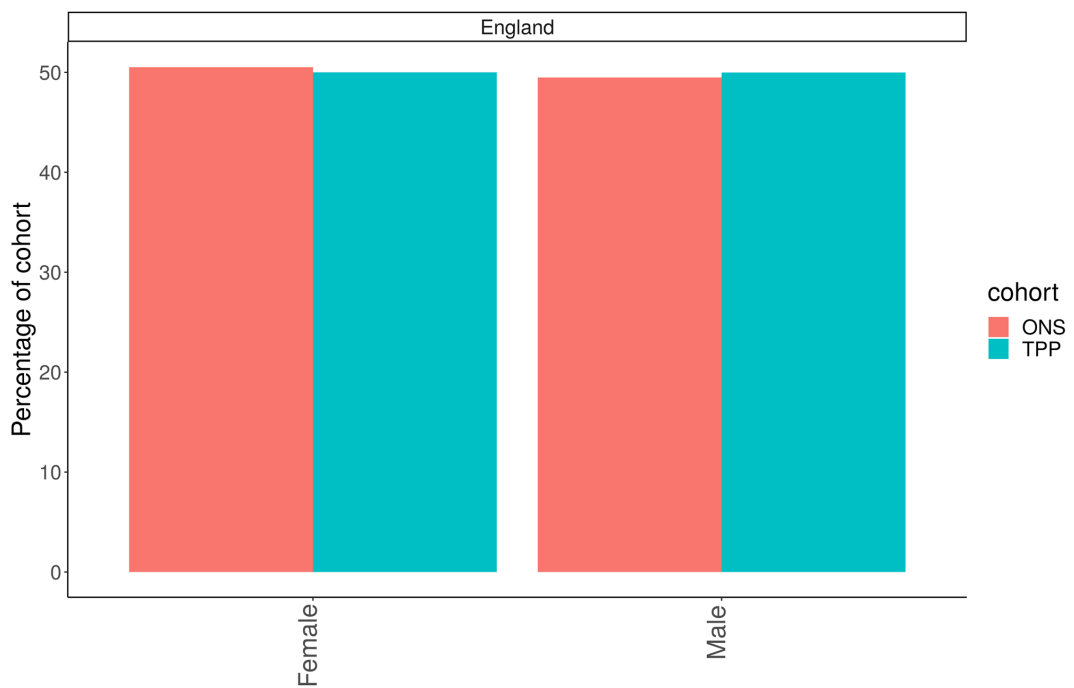


Figure 3. Barplot showing the proportion of ONS and TPP populations by Sex.



Figure 4. Barplot showing the proportion of ONS and TPP populations by Sex per NUTS-1 region.

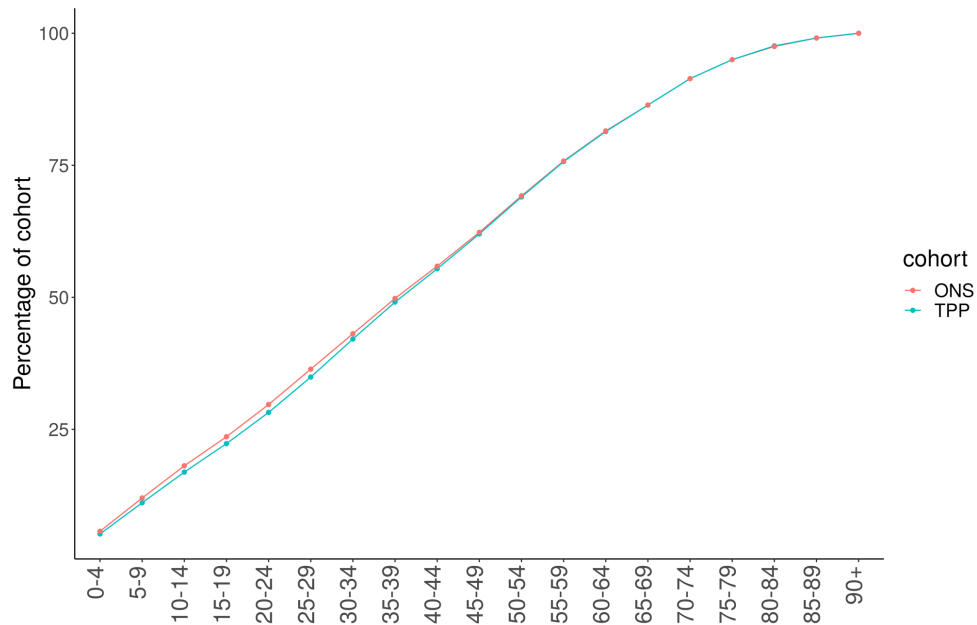


Figure 5. Cumulative frequency graph of ONS and TPP populations by age band.

African (1.5%, 1.8%) and Caribbean (0.6%, 1.1%) patients and a higher proportion of Other black patients (0.6%, 0.5%) (Figure 11B). There was clear regional variation in both the ethnic makeup of populations and the representativeness of ethnicity in TPP (Figure 15, Figure 16).

Discussion

Summary

This study has shown that TPP data made available via OpenSAFELY-TPP is broadly representative of the English population. Though there is high regional variability in the

coverage of the OpenSAFELY-TPP data amongst English general practices, we nonetheless found broad similarity within regions, with only occasional discrepancies which should be considered when designing studies and interpreting outcomes from OpenSAFELY-TPP. Particularly notable was the over-representation of 25–50 year olds in London for both males and females. This may have contributed to a slight overall under-representation of under 25 year olds and over-representation of 25 –59 year olds nationally. We await the 2021 Census results as the assumption that ONS mid-year estimates nearly 10 years after the Census are more accurate than the TPP data may not be true.

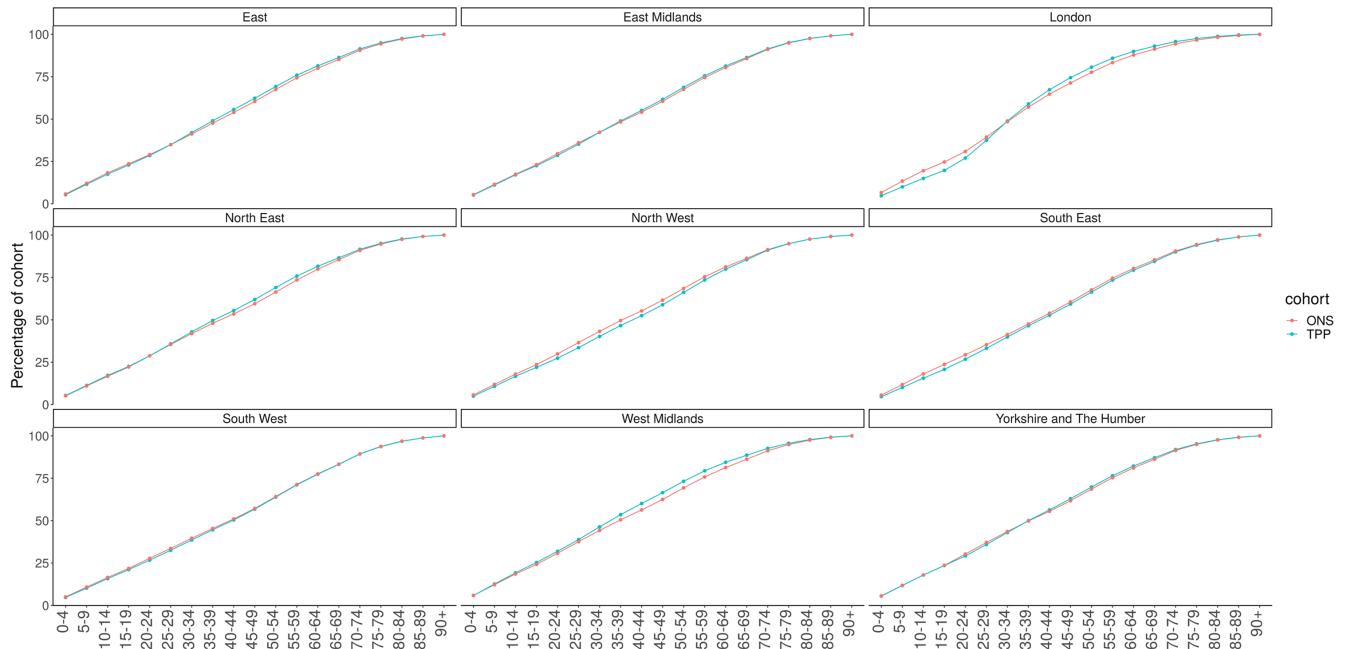


Figure 6. Cumulative frequency graph of ONS and TPP populations by age band per NUTS-1 region.

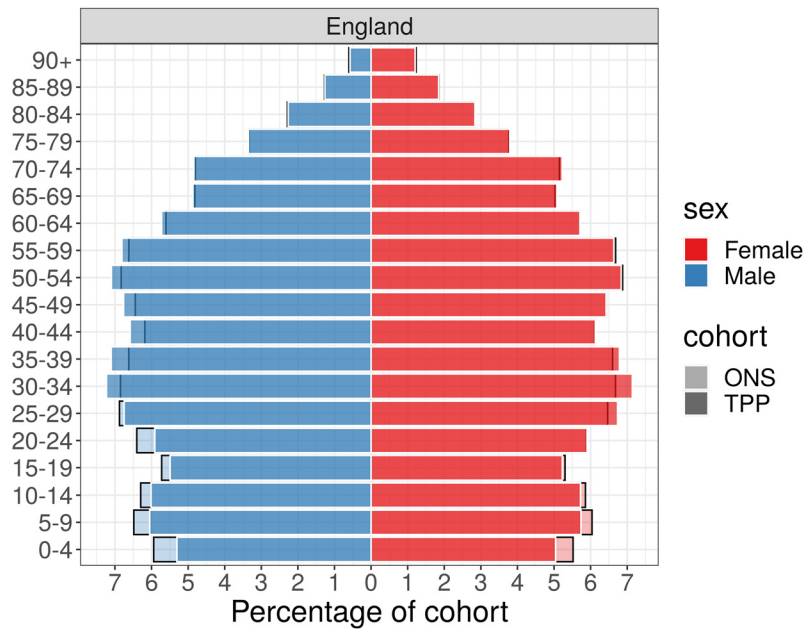


Figure 7. Barplot showing the proportion of ONS and TPP populations by sex and age band.

Strengths and weaknesses

This study provides an overview of the representativeness of the OpenSAFELY-TPP cohort with regard to a variety of key characteristics in comparison with the general UK. The key strengths of this study are the use of high quality data from the

EHR of all patients registered with a TPP practice which enabled us to compare participation rates for key sociodemographic characteristics (age, sex, IMD, and geographic location) and the comparison of ONS data held in OpenSAFELY-TPP to the matching national ONS data.



Figure 8. Barplot showing the proportion of ONS and TPP populations by sex and age band per NUTS-1 region.

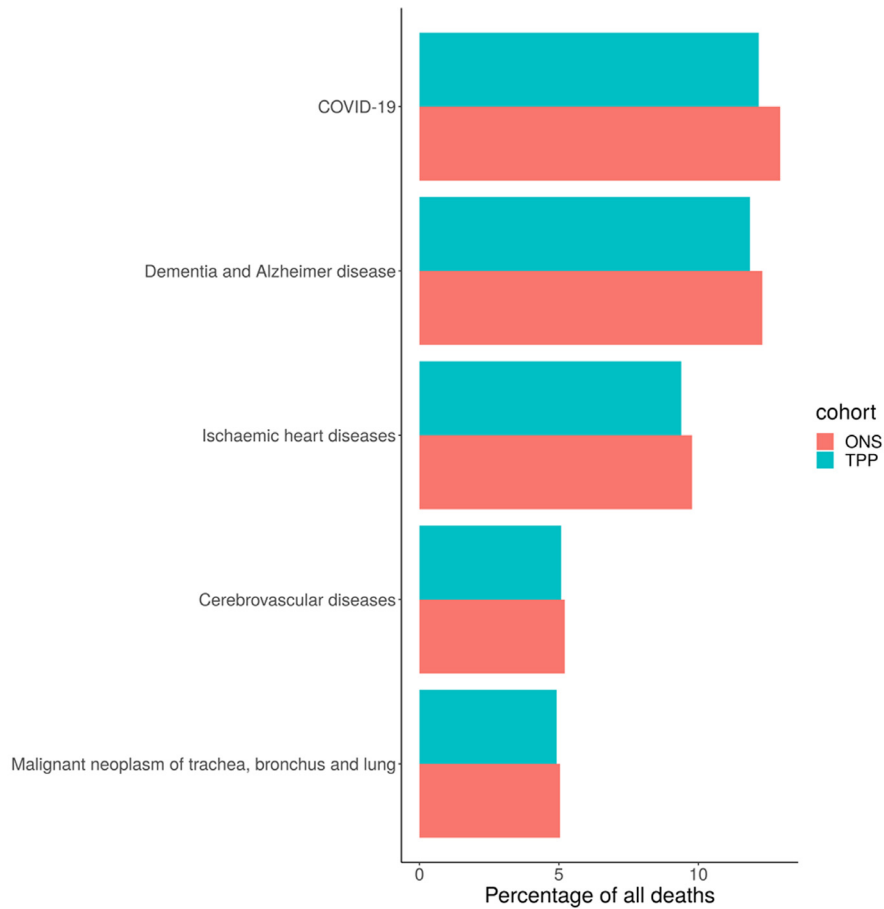


Figure 9. Barplot showing the proportion of the 5 most common causes of deaths occurring in ONS and TPP in 2020.

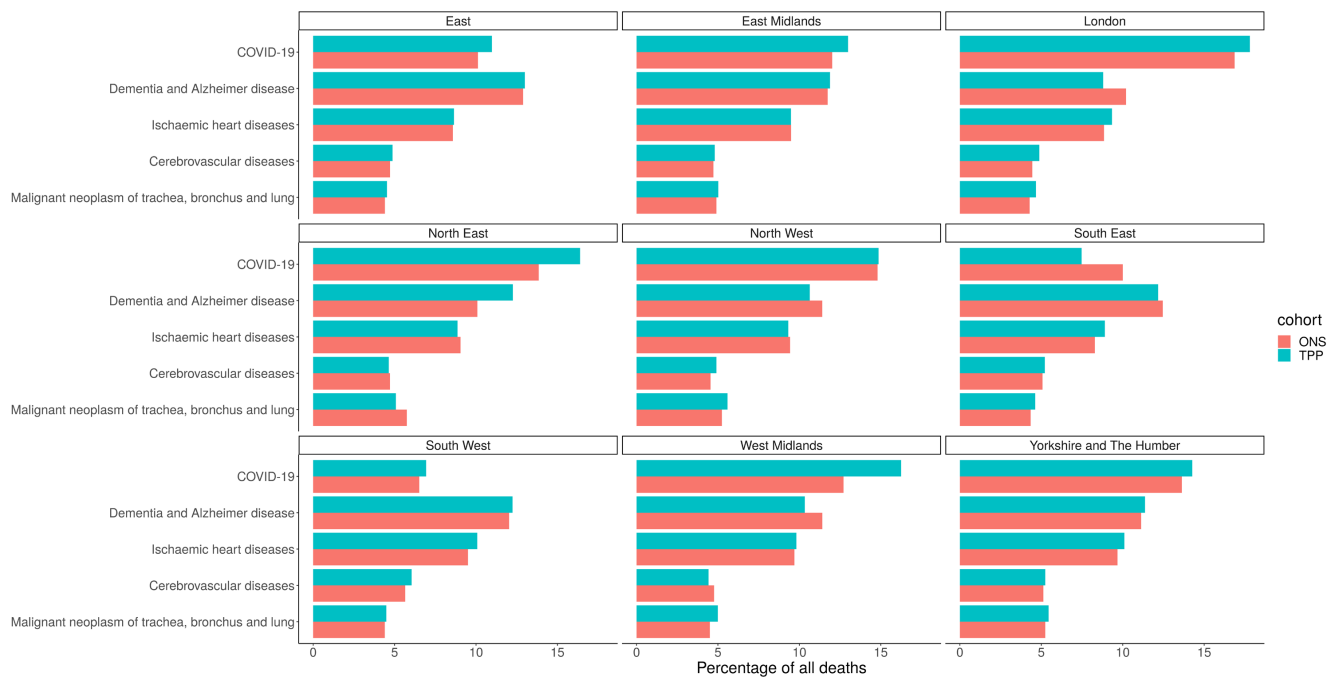


Figure 10. Barplot showing the proportion of the 5 most common causes of deaths occurring in ONS and TPP in 2020 per NUTS-1 region.

The true population of England is not known¹³, so we compared the OpenSAFELY-TPP data to ONS estimates as it forms the official population estimates of the UK¹⁴. However, ONS estimates may over- or under-estimate population sizes based on regional differences. The mid-year population estimates also have limitations; they heavily rely on the use of health service data to estimate internal migration, in a process that may not be detecting all aspects of population change^{13,14}. The total GP registered population is always higher than the ONS population nationally, possibly due to over-counting in GP practice registers; under-counting in population estimates and different definitions of who counts as ‘resident’ in the country, but can be lower in certain regions¹³. We have, therefore, probably overestimated the overall percentage of the total UK population covered by TPP (42.6%). Compared to the total GP registered population (60 million)¹⁵, TPP covers 39.9% of the UK population. If we considered smaller geographic regions it would be possible to get areas with over 100% coverage.

OpenSAFELY-TPP may have included deaths that were registered in Wales for patients registered in an English practice using TPP, whereas the ONS data was restricted to deaths registered in England.

The most up-to-date formal estimates of the population by ethnic group currently available are from the 2011 Census. The ethnic makeup may have changed substantially from this point. OpenSAFELY-TPP was missing ethnicity for 10% of patients, and the missingness of ethnicity data in EHRs may not be

random⁸. The 2011 census used multiple imputation to account for missing ethnicity¹⁶.

We investigated the top 5 causes of death (accounting for ~45% of all deaths) but did not look at others, or at health status more generally e.g. number of long term conditions. Regions are very large and more detailed regional analysis would be informative. We looked at one point in time and representativeness could change with time (e.g. TPP may take on or lose practices) or vary from year to year (e.g. the vaccine campaign may have prompted duplicated patients to be identified and deregistered).

Findings in context

Over 20 studies have been conducted using the OpenSAFELY framework. However, the sheer scale of data made available in OpenSAFELY-TPP alone does not guarantee that the findings are generalisable to the English population at large. While at least one study has shown the large degree of geographical variation in coverage between EHR software providers¹⁷, to our knowledge this is the first time that the representativeness of the population covered by the EHR software provider TPP has been systematically reported. We found that patients in TPP practices are broadly representative of England in terms of age, sex, IMD and ethnicity. The proportion of the five leading causes of death was broadly similar to those reported by ONS. The importance of a representative sample depends on the study question^{18,19}; careful consideration of this issue is warranted at the design, analysis and interpretation stage of every epidemiological

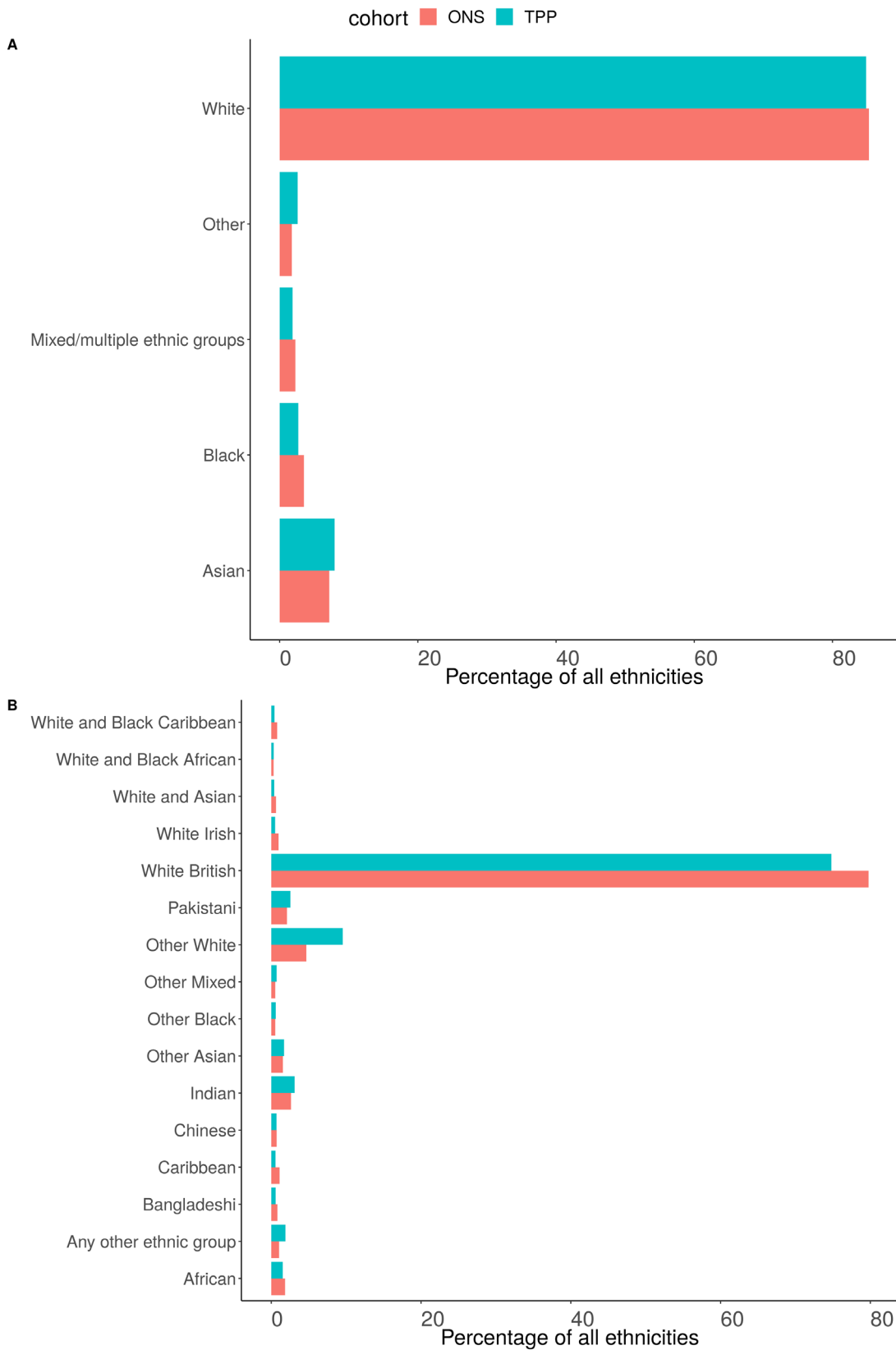


Figure 11. Barplot showing the proportion of ONS and TPP populations per ethnicity grouped into **A)** 5 and **B)** 16 groups. The TPP population excludes the 9.4% without a recorded ethnicity.

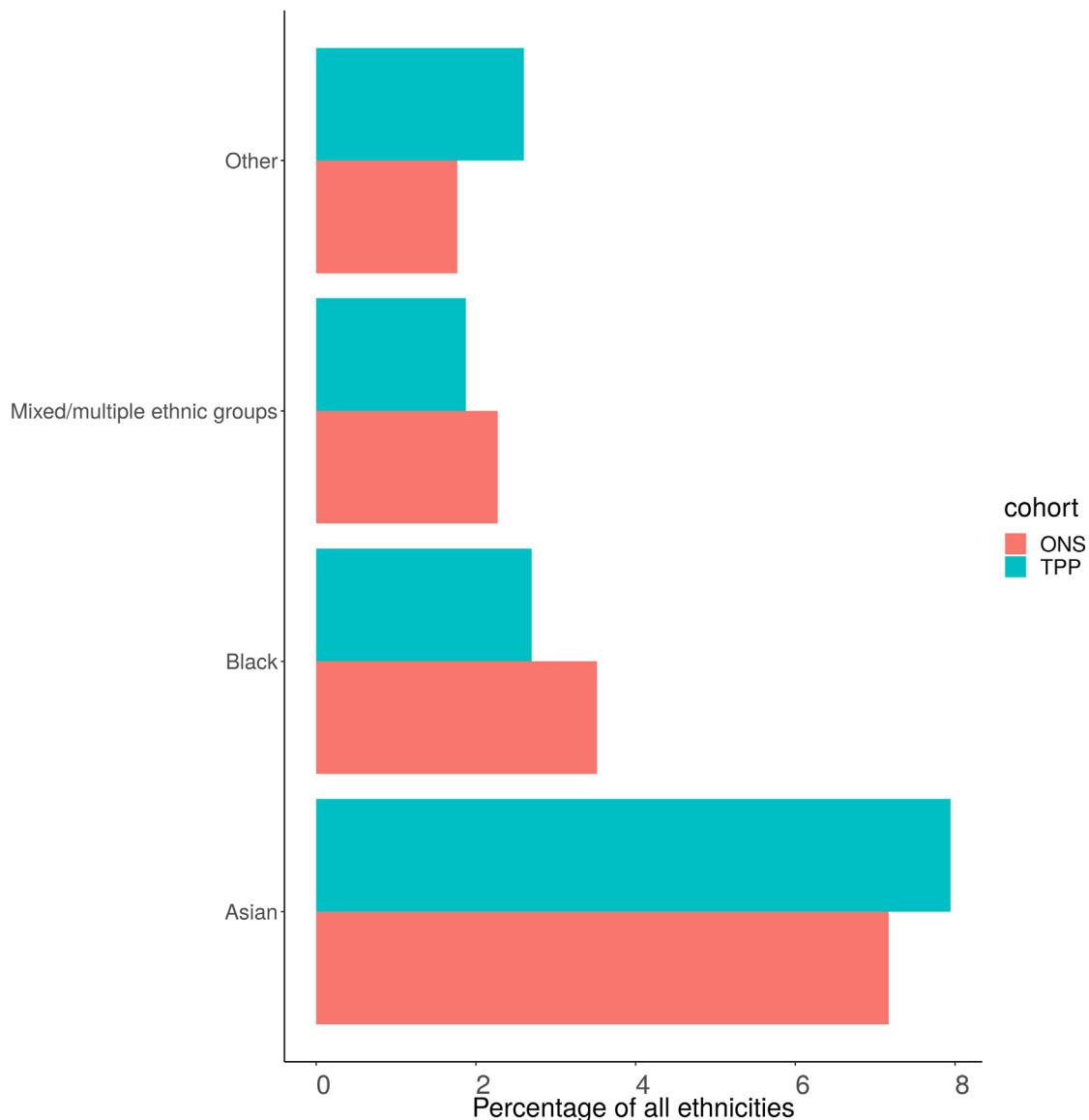


Figure 12. Barplot showing the proportion of ONS and TPP populations per ethnicity (excluding White group).

study²⁰. We have previously described how differences in the design of EHR user interfaces can affect clinical coding² and the prescribing of certain medications^{21,22}. Additionally, investigators may wish to consider representativeness of TPP when assessing variation in delivery of healthcare services due to variation in NHS service delivery and TPPs geographical coverage.

Policy implications and interpretation

The breadth of coverage provided by OpenSAFELY-TPP, 24 million patients across England representing 43% of the total English population (based on the ONS 2020 mid-year population estimate of 56 million), provides an unprecedented

opportunity to support urgent research into the COVID-19 emergency. Users of OpenSAFELY must consider the issues of representativeness, generalisability and external validity associated with using TPP data for health research: overall, as this analysis shows, TPP registered patients are a representative sample of the English population as a whole.

This paper is principally to inform interpretation of the numerous analyses completed and published using OpenSAFELY-TPP. However, OpenSAFELY is now also implemented in the data analysis environment of EMIS, a primary care electronic health record system supplier covering 55% of all practices in England. In addition, OpenSAFELY also supports federated

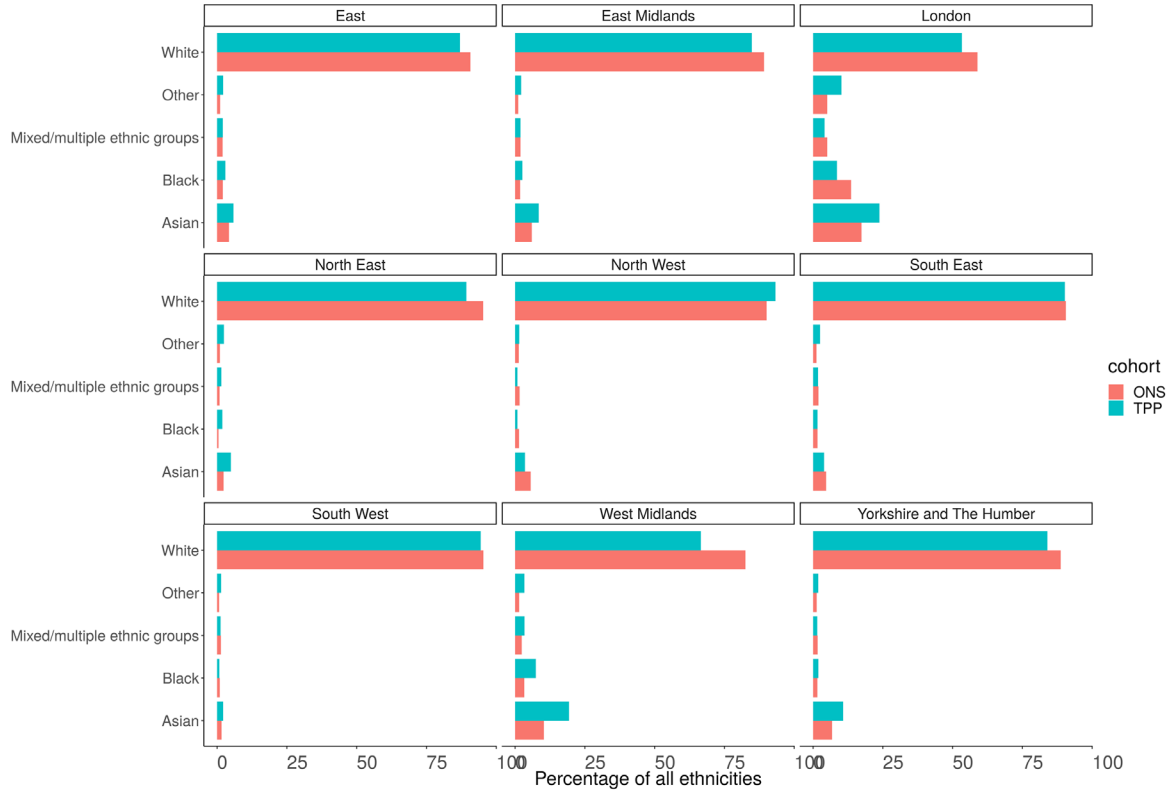


Figure 13. Barplot showing the proportion of ONS and TPP populations per ethnicity grouped into 5 groups per NUTS-1 region.

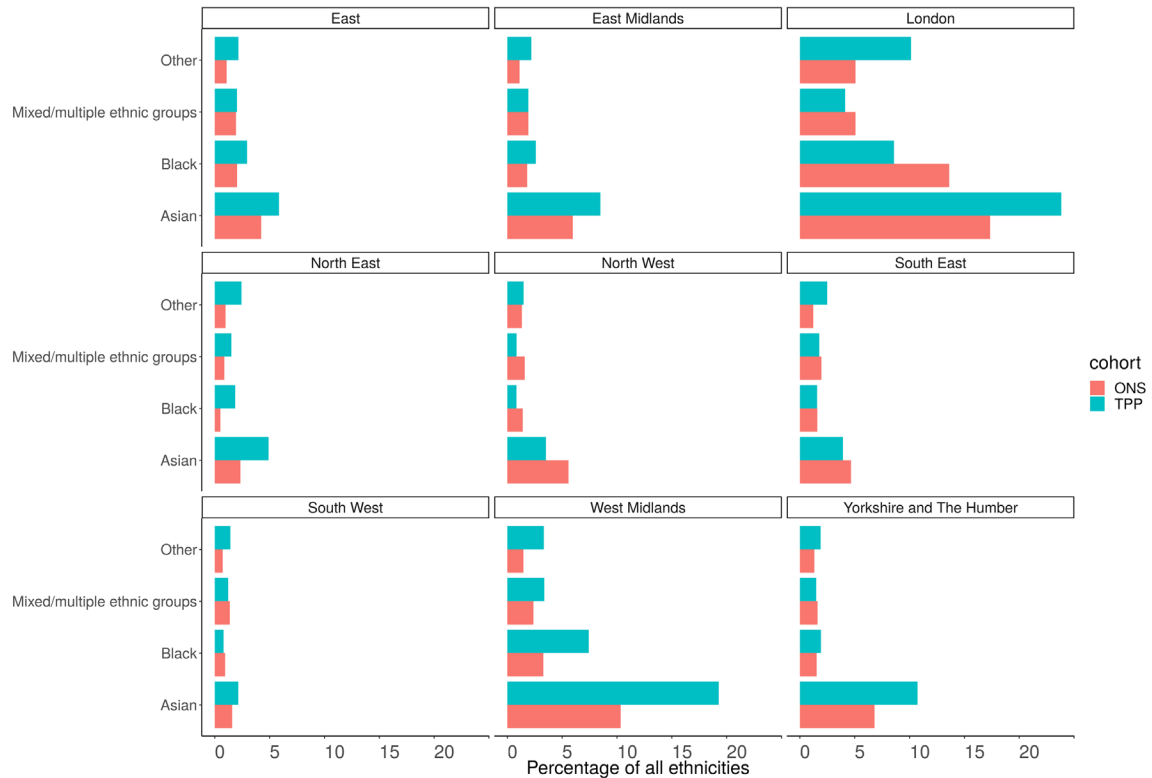


Figure 14. Barplot showing the proportion of ONS and TPP populations per ethnicity grouped into 5 groups per NUTS-1 region (excluding White group).

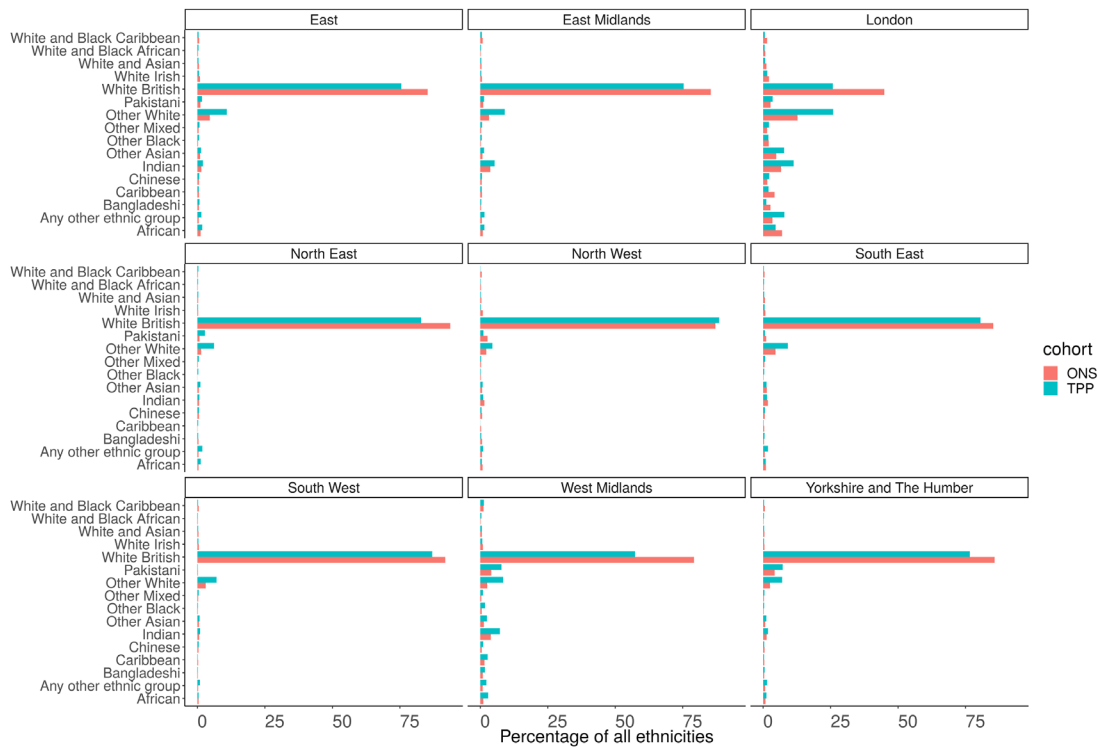


Figure 15. Barplot showing the proportion of ONS and TPP populations per ethnicity grouped into 16 groups per NUTS-1 region.

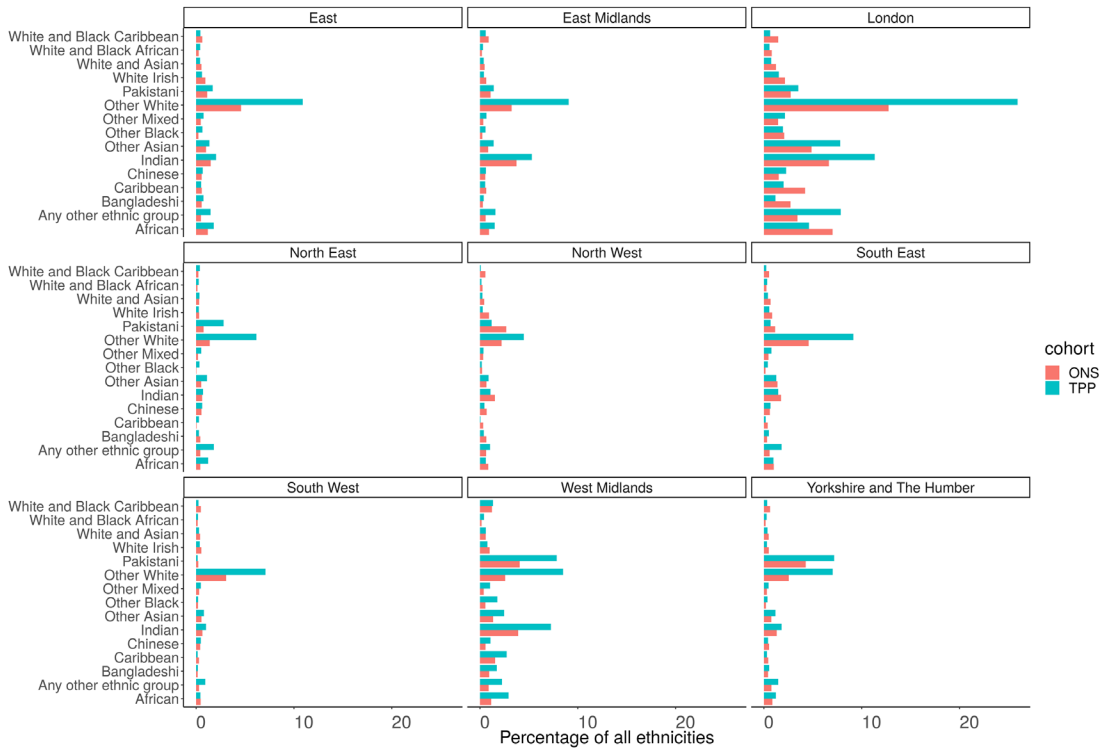


Figure 16. Barplot showing the proportion of ONS and TPP populations per ethnicity grouped into 16 groups per NUTS-1 region (Excluding White British Group).

analytics, where the same data preparation and analysis code is sent to OpenSAFELY-TPP and OpenSAFELY-EMIS to execute the same curation and analysis in each setting, with the results then combined into a single analysis, with a variety of papers already published using this approach¹⁻⁵. Nonetheless in the future it may still be more convenient or proportionate to execute analyses only in OpenSAFELY-TPP; therefore the high degree of representativeness demonstrated in this paper provides strong reassurance that such analyses present no interpretive or generalisability challenges.

Conclusions

Despite regional variations, data from OpenSAFELY-TPP is largely representative of the general population of England in terms of IMD, age, sex, ethnicity and causes of death. Following the use of OpenSAFELY-TPP data for a large number of COVID-19 studies since March 2020, it is reassuring to find that the overall representativeness of the population in the datasets is high.

Key messages

- There is regional variability across England in terms of key population characteristics
- Users of OpenSAFELY should carefully consider the issues of representativeness, generalisability and external validity associated with using TPP data for health research.
- TPP registered patients are a representative sub-sample of the English population as a whole in terms of age, sex, IMD and ethnicity.
- The proportions of the five leading causes of death in TPP in 2020 are broadly similar to those reported by ONS.

Data availability

Access to the underlying identifiable and potentially re-identifiable pseudonymised electronic health record data is tightly governed by various legislative and regulatory frameworks, and restricted by best practice. The data in OpenSAFELY is drawn from General Practice data across England where TPP is the Data Processor. TPP developers (CB, JC, JP, FH, and SH) initiate an automated process to create pseudonymised records in the core OpenSAFELY database, which are copies of key structured data tables in the identifiable records. These are linked onto key external data resources that have also been pseudonymised via SHA-512 one-way hashing of NHS numbers using a shared salt. DataLab developers and PIs (BG, LS, CEM, SB, AJW, KW, WJH, HJC, DE, PI, SD, GH, BBC, RMS, ID, KB, SE, EJW and CTR) holding contracts with NHS England have access to the OpenSAFELY pseudonymised data tables as needed to develop the OpenSAFELY tools. These tools in turn enable researchers with OpenSAFELY Data Access Agreements to write and execute code for data management and data analysis without direct access to the underlying raw

pseudonymised patient data, and to review the outputs of this code. All code for the full data management pipeline—from raw data to completed results for this analysis—and for the OpenSAFELY platform as a whole is available for review at github.com/OpenSAFELY.

The data management and analysis code for this paper was led by CDA and contributed to by WJH and AJW.

Information governance and ethical approval

NHS England is the data controller; TPP is the data processor; and the researchers on OpenSAFELY are acting with the approval of NHS England. This implementation of OpenSAFELY is hosted within the TPP environment which is accredited to the ISO 27001 information security standard and is NHS IG Toolkit compliant^{23,24}; patient data has been pseudonymised for analysis and linkage using industry standard cryptographic hashing techniques; all pseudonymised datasets transmitted for linkage onto OpenSAFELY are encrypted; access to the platform is via a virtual private network (VPN) connection, restricted to a small group of researchers; the researchers hold contracts with NHS England and only access the platform to initiate database queries and statistical models; all database activity is logged; only aggregate statistical outputs leave the platform environment following best practice for anonymisation of results such as statistical disclosure control for low cell counts²⁵. The OpenSAFELY research platform adheres to the obligations of the UK General Data Protection Regulation (GDPR) and the Data Protection Act 2018. In March 2020, the Secretary of State for Health and Social Care used powers under the UK Health Service (Control of Patient Information) Regulations 2002 (COPI) to require organisations to process confidential patient information for the purposes of protecting public health, providing healthcare services to the public and monitoring and managing the COVID-19 outbreak and incidents of exposure; this sets aside the requirement for patient consent²⁶. Taken together, these provide the legal bases to link patient datasets on the OpenSAFELY platform. GP practices, from which the primary care data are obtained, are required to share relevant health information to support the public health response to the pandemic, and have been informed of the OpenSAFELY analytics platform.

This study was approved by the Health Research Authority (REC reference 20/LO/0651) and by the LSHTM Ethics Board (reference 21863).

Acknowledgements

We are very grateful for all the support received from the TPP Technical Operations team throughout this work, and for generous assistance from the information governance and database teams at NHS England / NHSX.

An earlier version of this article can be found on medRxiv (<https://doi.org/10.1101/2022.06.23.22276802>)

References

1. The OpenSAFELY Collaborative, Curtis HJ, Inglesby P, *et al.*: **Trends and clinical characteristics of COVID-19 vaccine recipients: a federated analysis of 57.9 million patients' primary care records in situ using OpenSAFELY.** *medRxiv.* 2021.
[Publisher Full Text](#)
2. Walker AJ, MacKenna B, Inglesby P, *et al.*: **Clinical coding of long COVID in English primary care: a federated analysis of 58 million patient records in situ using OpenSAFELY.** *Br J Gen Pract.* 2021; **71**(712): e806–e814.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
3. OpenSAFELY Collaborative, Fisher L, Speed V, *et al.*: **Potentially inappropriate prescribing of DOACs to people with mechanical heart valves: A federated analysis of 57.9 million patients' primary care records in situ using OpenSAFELY.** *Thromb Res.* 2022; **211**: 150–153.
[PubMed Abstract](#) | [Publisher Full Text](#)
4. Fisher L, Hopcroft LEM, Rodgers S, *et al.*: **Changes in English medication safety indicators throughout the COVID-19 pandemic: a federated analysis of 57 million patients' primary care records in situ using OpenSAFELY.** *medRxiv.* 2022.
[Publisher Full Text](#)
5. Curtis HJ, Inglesby P, Morton CE, *et al.*: **Trends and clinical characteristics of COVID-19 vaccine recipients: a federated analysis of 57.9 million patients' primary care records in situ using OpenSAFELY.** *Br J Gen Pract.* 2022; **72**(714): e51–e62.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
6. Herrett E, Gallagher AM, Bhaskaran K, *et al.*: **Data Resource Profile: Clinical Practice Research Datalink (CPRD).** *Int J Epidemiol.* 2015; **44**(3): 827–36.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
7. Williamson EJ, Walker AJ, Bhaskaran K, *et al.*: **Factors associated with COVID-19-related death using OpenSAFELY.** *Nature.* 2020; **584**(7821): 430–436.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
8. Mathur R, Bhaskaran K, Chaturvedi N, *et al.*: **Completeness and usability of ethnicity data in UK-based primary care and hospital databases.** *J Public Health (Oxf).* 2014; **36**(4): 684–92.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
9. Steel N, Ford JA, Newton JN, *et al.*: **Changes in health in the countries of the UK and 150 English Local Authority areas 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016.** *Lancet.* 2018; **392**(10158): 1647–1661.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
10. **Mid-year population estimates QMI.** [cited 2022 Jan 14].
[Reference Source](#)
11. **Nomis - Nomis - Official Labour Market Statistics.** [cited 2021 Dec 13].
[Reference Source](#)
12. **Mortality statistics - underlying cause, sex and age - Nomis - Official Labour Market Statistics.** [cited 2022 Jan 28].
[Reference Source](#)
13. Baker C: **Population estimates & GP registers: why the difference?** 2016 [cited 2021 Dec 16].
[Reference Source](#)
14. Park N: **Population estimates for the UK England and Wales, Scotland and Northern Ireland - Office for National Statistics.** Office for National Statistics; 2021 [cited 2021 Dec 16].
[Reference Source](#)
15. **Patients Registered at a GP Practice January 2020; Special Topic.** NHS Digital. [cited 2022 Jan 28].
[Reference Source](#)
16. Wardman L, Aldrich S, Rogers S: **Census item edit and imputation process.** 2011.
[Reference Source](#)
17. Kontopantelis E, Stevens RJ, Helms PJ, *et al.*: **Spatial distribution of clinical computer systems in primary care in England in 2016 and implications for primary care electronic medical record databases: a cross-sectional population study.** *BMJ Open.* 2018; **8**(2): e020738.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
18. Rothman KJ, Gallacher JEJ, Hatch EE: **Why representativeness should be avoided.** *Int J Epidemiol.* 2013; **42**(4): 1012–4.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
19. Richiardi L, Pizzi C, Pearce N: **Commentary: Representativeness is usually not necessary and often should be avoided.** *Int J Epidemiol.* 2013; **42**(4): 1018–22.
[PubMed Abstract](#) | [Publisher Full Text](#)
20. Ebrahim S, Davey Smith G: **Commentary: Should we always deliberately be non-representative?** *Int J Epidemiol.* 2013; **42**(4): 1022–6.
[PubMed Abstract](#) | [Publisher Full Text](#)
21. MacKenna B, Bacon S, Walker AJ, *et al.*: **Impact of Electronic Health Record Interface Design on Unsafe Prescribing of Ciclosporin, Tacrolimus, and Diltiazem: Cohort Study in English National Health Service Primary Care.** *J Med Internet Res.* 2020; **22**(10): e17003.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
22. MacKenna B, Curtis HJ, Walker AJ, *et al.*: **Suboptimal prescribing behaviour associated with clinical software design features: a retrospective cohort study in English NHS primary care.** *Br J Gen Pract.* 2020; **70**(698): e636–e643.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
23. **BETA - Data Security Standards - NHS Digital.** NHS Digital. [cited 2020 Apr 30].
[Reference Source](#)
24. **Data Security and Protection Toolkit - NHS Digital.** NHS Digital. [cited 2020 Apr 30].
[Reference Source](#)
25. **ISB1523: Anonymisation Standard for Publishing Health and Social Care Data - NHS Digital.** NHS Digital. [cited 2020 Apr 30].
[Reference Source](#)
26. Secretary of State for Health and Social Care -UK Government: **Coronavirus (COVID-19): notification to organisations to share information.** 2020.
[Reference Source](#)

Open Peer Review

Current Peer Review Status:  

Version 1

Reviewer Report 02 August 2022

<https://doi.org/10.21956/wellcomeopenres.19969.r51579>

© 2022 Kontopantelis E. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

 **Evangelos Kontopantelis** 

Faculty of Biology, Medicine & Health, Division of Informatics, Imaging and Data Sciences, University of Manchester, Manchester, UK

This is a clearly written paper that described the characteristics of the TPP population, compared to national (England) estimates, mid-year 2020. I only have a few minor points to raise.

1. Explain that you created 5-year bands in the TPP patients, so they could be compared to ONS age groups.
2. Clarify IMD year when first described, and also explain for an international reader (e.g. what domains it includes, how many indicators e.g. see <https://jech.bmj.com/content/72/2/140> for some useful references.²
3. There is an issue with ghost patients, which hopefully is less of an issue following 2013 (there was an initiative to clear GP lists of ghost patients). Practices get paid on patient numbers so there is not much of an incentive to purge, which has led to some practices experiencing delays in death registrations. Although the authors do not seem to have many 120 year olds, it would be good to report if that has been checked. See more on the issue here: <https://jech.bmj.com/content/72/6/532>.² This also relates to the TPP coverage, which is almost certainly an overestimation.
4. The graphs look a bit basic but they convey the message.

References

1. Kontopantelis E, Mamas MA, van Marwijk H, Ryan AM, et al.: Geographical epidemiology of health and overall deprivation in England, its changes and persistence from 2004 to 2015: a longitudinal spatial population study. *J Epidemiol Community Health*. **72** (2): 140-147 [PubMed Abstract](#) | [Publisher Full Text](#)
2. Burch P, Doran T, Kontopantelis E: Regional variation and predictors of over-registration in English primary care in 2014: a spatial analysis. *J Epidemiol Community Health*. **72** (6): 532-538

[PubMed Abstract](#) | [Publisher Full Text](#)

Is the work clearly and accurately presented and does it cite the current literature?

Yes

Is the study design appropriate and is the work technically sound?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Yes

If applicable, is the statistical analysis and its interpretation appropriate?

Yes

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions drawn adequately supported by the results?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Professor in Data Science and Health Services Research

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 21 July 2022

<https://doi.org/10.21956/wellcomeopenres.19969.r51575>

© 2022 Fahey T. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



Tom Fahey 

HRB Centre for Primary Care Research, Department of General Practice, Royal College of Surgeons in Ireland, Dublin, Ireland

Thank you for asking me to review this paper. It examines the representativeness of the OpenSAFELY-TPP dataset, referenced to the population of England using ONS data.

The paper is clear and well written. The methods are appropriate and the conclusions are consistent with the data presented. OpenSAFELY-TPP is broadly representative in terms of deprivation (IMD), age, gender, ethnicity and cause of death. These are important findings in relation to generalisability/external validity of the data and any subsequent papers that will use

OpenSAFELY-TPP.

Is the work clearly and accurately presented and does it cite the current literature?

Yes

Is the study design appropriate and is the work technically sound?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Yes

If applicable, is the statistical analysis and its interpretation appropriate?

Yes

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions drawn adequately supported by the results?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Primary care epidemiology- observational, RCTs and systematic reviews

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.
