

Survival of patients with non-Hodgkin lymphoma in
England: investigating the socioeconomic
inequalities

LONDON
SCHOOL *of*
HYGIENE
& TROPICAL
MEDICINE



Matthew J. Smith

Thesis submitted in accordance with the requirements for the degree of Doctor of
Philosophy

2021

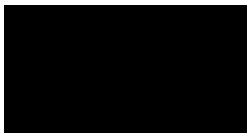
Funded by Cancer Research UK

Department of Non-Communicable Disease Epidemiology
Faculty of Epidemiology and Population Health
London School of Hygiene and Tropical Medicine
University of London

Declaration

I, Matthew J. Smith, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

This is a research paper style thesis. Three papers have been submitted and are under review. I declare that the work presented in this thesis is my own. I am lead author for all three papers, which are included in the Appendix. I am corresponding author for two of the three papers. As lead author, I carried out the literature review, application of the methods, writing of the code, data analysis, interpretation of the study findings, and preparation for all drafts of each paper included in this thesis. Co-authors advised the study design, provided feedback, gave suggestions on the application of the methods, critically reviewed the manuscripts, and approved the submitted versions.



Matthew J. Smith

Tuesday 16th November, 2021

Acknowledgements

It takes a community to raise a child, that child to write a thesis, and that thesis to help others. In this community I was raised by Edmund Njeru Njagi's calm and relaxed demeanor, Miguel Angel Luque Fernandez's endless thought-provoking conversations, Yuki Alencar's uncanny ability to see the bigger picture, Bernard Rachet's father-like atmosphere, Aurelien Belot's firm but fair encouragement, and Audrey Bonaventure's emotional intelligence and sensitive clinical guidance. Through them I found persistence to help cancer patients. I am grateful to Sara Benitez Majano, Matteo Quartagno and Susan Gachau for their invaluable insights from their own in-depth experiences. Finally, I am forever thankful to those who have supported me and provided refreshing and uplifting experiences outside of academia.

Funding

This work was funded by the Cancer Research UK Studentship in Cancer Survival to the Inequalities in Cancer Outcomes Network at the London School of Hygiene and Tropical Medicine.

Dedication

*“To those suffering, to those surviving, to those silencing... I hope this work
alleviates, acknowledges, and admires your experience with cancer... I will
continue my pursuit of diminishing inequalities so that your experiences will no
longer be lived again.”*

—MJS

Survival of patients with non-Hodgkin lymphoma in England: investigating the socioeconomic inequalities

Matthew J. Smith

Supervisors: Edmund Njeru Njagi, Bernard Rachet, Miguel Angel Luque Fernandez

Advisory panel: Audrey Bonaventure, Stijn Vansteelandt

Abstract

Non-Hodgkin lymphoma is a heterogeneous group of malignancies characterised by various behaviours and prognoses. Two of the most common subtypes are diffuse large B-cell lymphoma and follicular lymphoma, where patients with either can have markedly different health outcomes. Survival probability is commonly used to measure the performance of a healthcare system in managing cancer patient health outcomes in a population, such as England. In England, the National Health Service is responsible for the care and management of patients and their health outcomes, and is committed to providing equal access to healthcare regardless of the patient's underlying characteristics.

However, although cancer patients are now more likely to live to 5 years after diagnosis, there are vast inequalities in survival between patient characteristics. Socioeconomic inequalities in survival, for all cancers, have narrowed since the late 20th century but these socioeconomic-gaps in survival persist. Comorbidity, the presence of a chronic disease unrelated to the cancer, is more prevalent amongst individuals living in more deprived areas. These socioeconomic gaps in survival may be explained by the presence of comorbid conditions or by the interaction between patients and the healthcare system.

The aim of this PhD is to investigate the inequalities in survival of patients with non-Hodgkin lymphoma in England using population-based cancer registry data linked to other population-based health outcomes databases. This thesis includes one paper investigating the association between patient and healthcare pathway characteristics and long-term survival probabilities, another paper that focuses on inequalities in short-term survival probability, and a final paper on inequalities in diagnostic delay. An additional paper was written concurrently to this thesis that provides a tutorial on the methods, amongst others, that were used for the paper investigating short-term survival.

Contents

List of Abbreviations	9
List of Tables	10
List of Figures	11
Aims	12
1 Background	13
1.1 Cancer	13
1.2 Non-Hodgkin lymphoma	14
1.2.1 Tumour characteristics	15
1.2.2 Diagnosis	17
1.2.3 Incidence patterns	19
1.2.4 Risk factors	20
1.2.5 Treatment	24
1.2.6 Clinical management	25
1.2.7 Prognostic factors	26
1.3 Cancer survival	29
1.3.1 Measures of disease burden	29
1.3.2 Non-Hodgkin lymphoma survival	30
1.4 Health inequality	30
1.4.1 Public health policies	30
1.4.2 Classifying social groups	31
1.4.3 Classifying comorbidity status	33

1.5	Inequality and non-Hodgkin lymphoma	34
1.6	Study rationale	35
2	Material	55
2.1	Tumour data	55
2.1.1	Tumour data cleaning and manipulation	57
2.1.2	Tumour data description	57
2.2	Deprivation data	60
2.3	Comorbidity data	63
2.4	Population and mortality data	66
3	Methods	72
3.1	Survival analyses	72
3.1.1	Data settings	72
3.1.2	Net survival measure	73
3.1.3	Strategy of survival estimation	77
3.1.4	Multilevel excess hazard models	77
3.2	Causal inference	79
3.3	Dependent discrete data	83
3.4	Missing data analysis	85
3.4.1	Missing data mechanisms	86
3.4.2	Multiple imputation	89
3.4.3	Number of imputations	92
3.4.4	Hypothesis testing after multiple imputation	92
3.4.5	Pitfalls	93
3.4.6	Available software packages	94
3.5	Conclusion	94

4	Patient characteristics and survival	99
4.1	Differences in survival by patient characteristics	99
4.1.1	Introduction	99
4.1.2	Overview of paper	99
4.1.3	Conclusion	102
4.2	Association between patient characteristics and survival	103
4.2.1	Introduction	103
4.2.2	Overview of paper	103
4.2.3	Conclusion	106
5	The role of comorbidity on survival	110
5.1	The role of comorbidity in explaining cancer survival differences	110
5.1.1	Introduction	110
5.1.2	Overview of paper	110
5.1.3	Conclusion	112
5.2	Impact of comorbidity on patient's access to the healthcare system	113
5.2.1	Introduction	113
5.2.2	Overview of paper	113
5.2.3	Conclusion	116
6	Discussion	120
6.1	Summary	120
6.2	Interpretations	121
6.2.1	Socioeconomic status	121
6.2.2	Comorbidity status	122
6.2.3	Healthcare pathway	122
6.3	Implications	123

6.3.1	Health policies	123
6.4	Limitations	125
6.4.1	Data	125
6.4.2	Unmeasured variables	128
6.4.3	Missing data	129
6.4.4	Methods	130
6.5	Recommendations	137
6.5.1	Inequalities	137
6.5.2	Methods	138
6.5.3	Coronavirus	138
6.6	Conclusion	139
A	Appendix	150
A.1	Details of ethics approvals obtained	150
A.2	Details of copyright approvals	152
A.3	Comorbidity algorithm	153
A.4	Publications	155
A.5	Research papers	156
A.5.1	Descriptive survival of patients with non-Hodgkin lymphoma	156
A.5.2	Association between patient and healthcare pathway characteristics on survival of patients with non-Hodgkin lymphoma	187
A.5.3	Association between comorbidity and short-term mortality	232
A.5.4	Introduction to computational causal inference	270
A.5.5	Association between patient characteristics and delayed diagnosis of patients with non-Hodgkin lymphoma	303
A.6	R code for the Approximate F-test of Inference for Vector β	340

List of Abbreviations

CAR-T	Chimeric Antigen Receptor T-cell
CCG	Clinical Commissioning Groups
CMAR	Covariate-dependent missing at random
HDI	Human Development Index
HIV	Human Immunodeficiency Virus
HTLV-1	Human T-cell Leukemia/Lymphoma Virus type 1
ICD-O	International Classification of Diseases for Oncology
MAR	Missing at random
MCAR	Missing completely at random
MNAR	Missing not at random
NHL	Non-Hodgkin Lymphoma
NHS	National Health Service
NICE	National Institute for Health and Care Excellence
NS	Net Survival
OMAR	Outcome-dependent missing at random
PCT	Primary Care Trusts
PET/CT	Positron Emission Tomography - Computed Tomography
RS	Relative Survival
RTD	Route to Diagnosis
WHO	World Health Organisation

List of Tables

1	Non-Hodgkin lymphoma numbers and mean ages at diagnosis: 2005-2013	16
2	Proportion of not otherwise specified cases of non-Hodgkin lymphoma by cancer registry	17
3	Ann-Arbor classification of stage at diagnosis for non-Hodgkin lymphoma	19
4	The eight routes to diagnosis defined by Elliss-Brookes <i>et al.</i> (2012)	26
5	Patient data linkage between sources of data sets	55
6	Variables in each of the data sets	56
7	Distribution of non-Hodgkin lymphoma subtypes for patients in England diagnosed from 2005-2013, with respective morphology and topography ICD-O-3 codes	58
8	Weights of each domain that comprises the Index of Multiple Deprivation	61
9	Proportion of cancer patients diagnosed with NHL for each IMD update by deprivation quintile	62
10	Number of cancer patients having an agreeable deprivation quintile between the measures directly before or after their diagnosis. 1 is least deprived, 5 is most deprived. Numbers shaded in gray indicate agreement in deprivation level between different time periods of IMD measures.	63
11	Proportion of patients diagnosed within each cancer registry by deprivation quintile	64
12	Royal College of Surgeons Charlson Score indicating International Classification of Disease tenth revision code for 14 categories	65
13	Monotone missing data pattern	85
14	Non-monotone missing data pattern	86

List of Figures

1	Distribution of the lymphatic system and dissection of a cancerous lymph node	14
2	Flowchart for the development of B- and T-cells and the corresponding lymphomas that can possibly arise	15
3	Microscopic image of a lymph node biopsy showing abnormal, large cells spread diffusely (characteristic of DLBCL)	18
4	Age-standardised rate of NHL diagnosis (per 100,000 people) between 2005 and 2013 in England, bounded by Clinical Commissioning Groups.	21
5	Number of patients diagnosed for each category of NHL subtype in England, 2005-2013.	59
6	Box plot of age at diagnosis by gender (m: male, f: female) for each category of NHL subtype in England, 2005-2013.	60
7	Prevalence of comorbidity score amongst non-Hodgkin lymphoma patients in England diagnosed 2005-2013. Diabetes, hemi/paraplegia, renal disease and AIDS/HIV are automatically scored as 2 or more. COPD: Chronic obstructive pulmonary disease.	67
8	Probability of comorbidity score by deprivation level amongst non-Hodgkin lymphoma patients in England diagnosed 2005-2013.	67
9	Probability of comorbidity score amongst non-Hodgkin lymphoma patients in England diagnosed 2009-2013. Diabetes, hemi/paraplegia, renal disease and AIDS/HIV are automatically scored as 2 or more. COPD: Chronic obstructive pulmonary disease. Note: using a 6-year optimal time window for comorbidities to be recorded.	153
10	Probability of comorbidity score by deprivation level amongst non-Hodgkin lymphoma patients in England diagnosed 2005-2013. Note: using a 6-year optimal time window for comorbidities to be recorded.	154

Aims and objectives

1. Contribute to research in the description of survival of patients with non-Hodgkin lymphoma (NHL)
 - (a) Estimate the survival of patients with NHL by patient characteristics
 - (b) Compare 5-year survival estimates between patient characteristics: focusing on comorbidity status and deprivation level
2. Quantify the association between patient characteristics and survival of NHL
 - (a) Build an excess mortality hazard model adjusting for patient characteristics
 - (b) Incorporate parameters to estimate the non-linear and time-dependent effects of patient characteristics on the excess mortality hazard
 - (c) Expand the excess hazard model to incorporate correlation between NHL patients
3. Evaluate the comorbidity and socioeconomic inequalities in short-term mortality amongst patients with NHL
 - (a) Develop a model for the short-term mortality risk standardised to the distribution of patient characteristics
 - (b) Predict and compare the cumulative mortality hazard between comorbidity status and deprivation levels
4. Investigate the variation in access to the health care system amongst patients with NHL
 - (a) Assess the association between diagnostic delay and patient characteristics
 - (b) Describe patterns in diagnostic delay by population density

1 Background

“Cancer’s life is a recapitulation of the body’s life, its existence a pathological mirror of our own.”

—Siddhartha Mukherjee, *The Emperor of All Maladies*

1.1 Cancer

Cancer is an overarching term to describe a group of diseases that arise when a biological cell ceases to carry out its normal function of programmed cell death. The dysfunction of a cell leads to an uncontrolled proliferation of cells and results in the formation of a neoplasm. The dysfunction of a cell is caused by a mutation in the genetic structure of the cell. Tumour suppressor genes (genes that control the timing of cell division) and proto-oncogenes (genes that regulate the rate of cell division) are commonly found to be mutated in a number of tumours. The mutation of proto-oncogenes, and tumour suppressor genes, can be inherited but is most commonly due to external factors, called carcinogens.

A normal cell has growth factors and receptors that inform the cell to stop dividing upon contact of other cells, such as during hyperplasia. A cancerous cell secretes a large amount of growth factors, which means they are constantly active and dividing. Consequently, cancerous cells resist the inhibitory signals, which is one of the hallmarks of cancer: loss of contact inhibition. This loss of contact inhibition leads to the formation of a neoplasm.

Neoplasms are distinguishable by two different formations: benign and malignant. Benign neoplasms, also termed noncancerous, are a proliferation of cells that do not invade nearby tissues. The danger is that these tumours may grow too large for the space they occupy and compress on vital organs or transport channels. In some diagnoses, such as adenomas, benign tumours may transform into malignant tumours. Malignant tumours are cancerous cells capable of infiltrating the basement membrane and invading local tissues and organs. This progression of invasive cancerous cells via the vascular or lymphatic systems, that eventually settle in other organs, is termed a metastasis.

The diagnosis of cancer is based on the type of cancer cell, system, or organ in which it began. The primary cancer can be a: carcinoma (cells of the skin and the lining of organs), sarcoma (bone and soft tissues), myeloma (plasma cells in the immune system), leukemia (blood cells, originating in the bone marrow), or lymphoma (cells of the immune system). Upon pathological investigation of a metastasis in a distant organ, for example in the brain, it is possible to ascertain the location of the primary cancer due to the characteristics of the cancerous cell.

1.2 Non-Hodgkin lymphoma

The lymphatic system is the body's highway for transporting and removing cellular waste material. The system (figure 1) comprises of lymph vessels intermittently connected by lymph nodes: found in abundance at the surface-level in the neck, armpit, and groin; and at a deeper level saturating the lungs, heart, and spleen. Lymph, the fluid within the system, carries waste material to the lymph nodes for filtration. It is within these such lymph nodes that leukocytes (white blood cells) mature into B- or T-lymphocytes. The uncontrolled proliferation of a lymphocyte is termed a lymphoma.

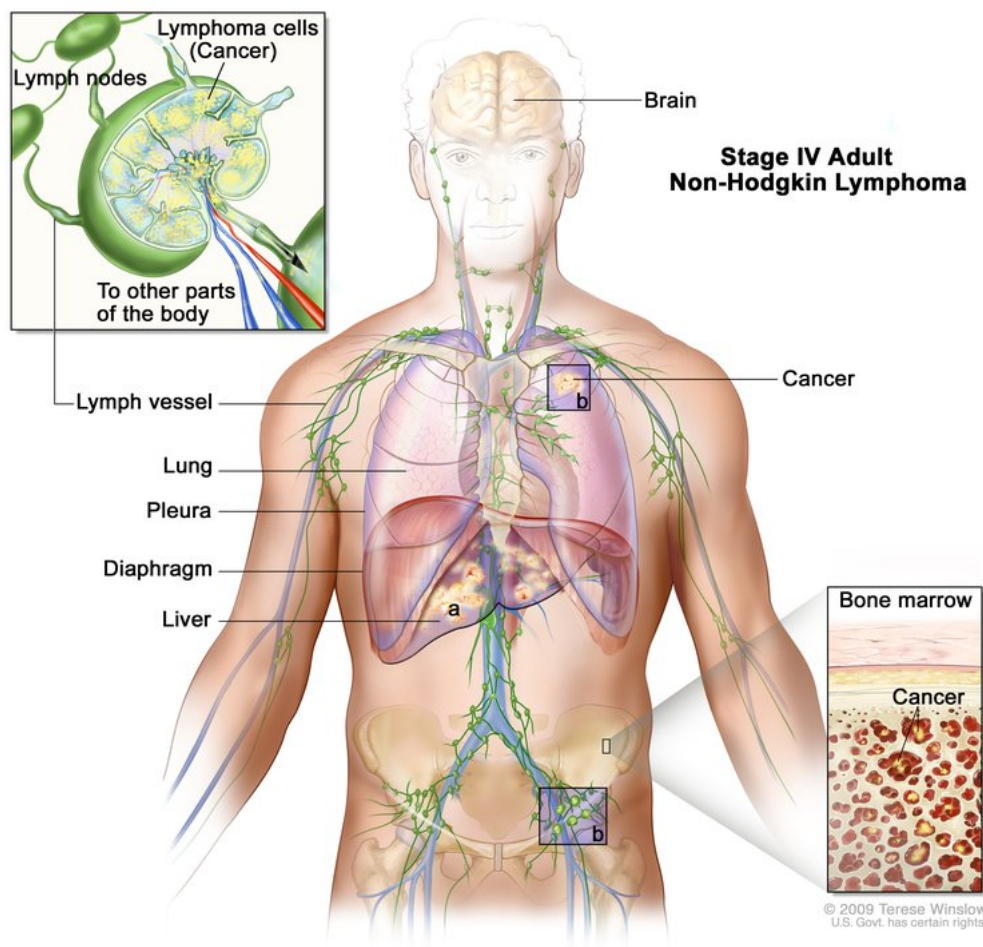


Figure 1: Distribution of the lymphatic system and dissection of a cancerous lymph node

For the National Cancer Institute (copyright)(2009) Terese Winslow LLC, U.S. Govt. has certain rights. See Appendix A.2.¹

Lymphoma can further be classified as Hodgkin lymphoma or non-Hodgkin lymphoma. Hodgkin lymphoma, which is less common, is diagnosed in the presence of Hodgkin's cells, such as multinucleated Reed-Sternberg cells. Non-Hodgkin lymphoma (NHL) is a heterogeneous group of malignancies categorised into over sixty subtypes,^{2,3} the majority of which are generated from B-lymphocytes.⁴ The main symptom of NHL is a painless, swollen

lymph node around the neck, armpit, or groin; other symptoms include: abdominal pain or swelling, fatigue, night sweats, unintentional weight loss, fever, feeling of breathlessness, and persistent itching of the skin.^{5,6} The classification of NHL depends on the B- or T-cell lineage, location of primary tumour, and degree of differentiation (indicative of the grade) determined by a pathological investigation.⁷

1.2.1 Tumour characteristics

Histology

Non-Hodgkin lymphoma is comprised of several histological subtypes characterised by cellular origin, morphology, immunophenotype, and cytogenetic and molecular abnormalities. Accounting for 95% of NHL diagnoses, B-cell lymphomas are far more common than the rarer T-cell lymphomas.⁸ The formation of a B-cell has several important stages (figure 2):⁹

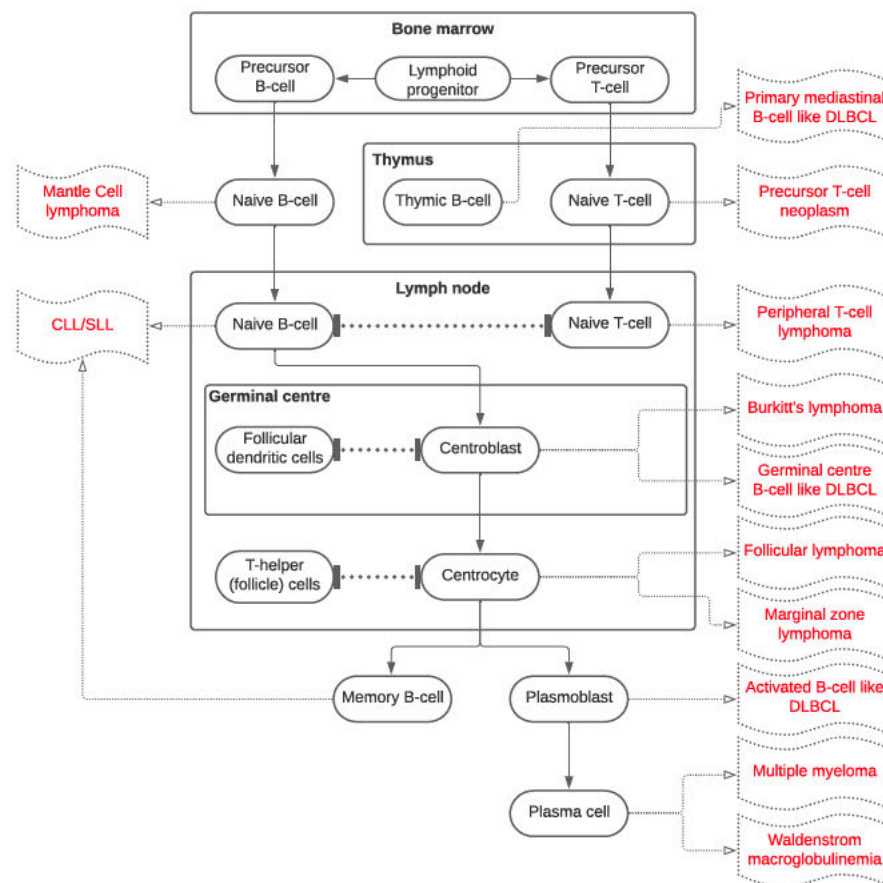


Figure 2: Flowchart for the development of B- and T-cells and the corresponding lymphomas that can possibly arise

1. B-cells begin their journey in the bone marrow as a precursor B-cell and are transported via the lymphatic system as naive B-cells;

2. They then undergo various cell surface transformations to become experienced B-cells, before proceeding into the lymph node as a B-lymphocyte;
3. Whilst in the lymph node, B-cells, activated by T-cells, proliferate into centroblasts;
4. Centroblasts move through the lymph node and bind with follicular dendritic cells to form centrocytes;
5. Finally, centrocytes, activated by T-helper cells, proliferate into memory B-cells or plasmoblasts: eventually maturing into plasma cells.

At every interaction along the multistage pathway, B-cells have the potential to proliferate into a cancerous cell, which is a reason why B-cell lymphomas are more common than T-cell lymphomas (table 1). In stage 1, naive B-cells can proliferate into mantle cell (MC) lymphoma.¹⁰ In stage 2, experienced B-cells can proliferate into small lymphocytic lymphoma (SLL), which is the same as chronic lymphocytic leukemia (CLL).¹¹ In stage 3, where somatic hypermutation occurs, centroblasts can proliferate into Burkitt lymphoma (a rare form),¹² or diffuse large B-cell lymphoma (DLBCL).¹³ In stage 4, centrocytes can proliferate into follicular lymphoma (the second most common subtype) or marginal zone lymphoma.^{14,15} In stage 5, memory B-cells can proliferate into CLL/SLL,¹¹ plasmoblasts into another form of DLBCL,¹³ and plasma cells into multiple myeloma or waldenstrom macroglobulinemia.^{16,17} There is a third form of DLBCL arising from thymic B-cells, located in the thymus gland; this form is called primary mediastinal B-cell like lymphoma.¹³

Table 1: Non-Hodgkin lymphoma numbers and mean ages at diagnosis: 2005-2013

Histological category	N.o. (%)	Age (SD)
All non-Hodgkin lymphoma	84,504 (100%)	67.0 (15.0)
Diffuse large B-cell lymphoma	30,750 (36.4%)	67.3 (15.3)
Follicular lymphoma	15,624 (18.5%)	63.9 (13.7)
Mature T-cell	6,066 (7.2%)	63.2 (16.7)
Marginal Zone lymphoma	4,615 (5.5%)	67.8 (14.1)
SLL/CLL ¹	4,043 (4.8%)	69.6 (12.6)
Mantle cell	3,549 (4.2%)	70.1 (11.5)
Waldenstrom macroglobulinemia	2,453 (2.9%)	71.5 (11.3)
Burkitt lymphoma	1,077 (1.3%)	53.6 (19.6)
Not otherwise specified ²	16,327 (19.3%)	69.4 (15.3)

¹ Small lymphocytic lymphoma / Chronic lymphocytic leukemia.

² No pathological information was available for these patients.

SD: standard deviation.

Data obtained from the National Cancer Registry and Analysis Service (Public Health England) on patients diagnosed and recorded within England cancer registries.

Not otherwise specified non-Hodgkin lymphoma

Not otherwise specified (NOS) is a subcategory in disease classification systems (e.g., International Classification of Diseases) that is used to indicate a disease where the symptoms

or haematopathological investigation were sufficient to make a general diagnosis of non-Hodgkin lymphoma, but a specific subtype could not be attributed. For example, DLBCL was considered a single subtype of NHL but research has shown that DLBCL is clinically heterogeneous and, through gene expression profiling and other diagnostic tools, can be further defined based on whether they formed from the lymph node’s germinal center (i.e., germinal center B-cell [GCB] DLBCL) or as activated B-cell (ACB) DLBCL. GCB and ABC represent roughly 50% and 35% of not otherwise specified DLBCL cases, where the remaining 15% remains unclassifiable.¹⁸

The distribution of ‘not otherwise specified’ cases of non-Hodgkin lymphoma differs by geographical location, in this case represented by cancer registries (table 2). Of those patients with NHL not otherwise specified, the highest proportion occurred within the Thames, followed by the South and West, cancer registries. The lowest proportion of these cases were in the Oxford, followed by the Trent, cancer registries.

Table 2: Proportion of not otherwise specified cases of non-Hodgkin lymphoma by cancer registry

Cancer Registry	<i>N</i>	(%)
Northern and Yorkshire	1,104	10.7
North Western	1,630	15.8
Trent	530	5.1
West Midlands	1,135	11.0
Eastern	885	8.6
Oxford	407	4.0
South and West	1,729	16.8
Thames	2,888	28.0
Total	10,308	100.0

1.2.2 Diagnosis

For the majority of non-Hodgkin lymphoma cases, diagnosis is carried out by a trained specialist in a haematopathology laboratory with expertise in morphological interpretation.¹⁹ The specialist studies a biopsy (a tissue sample from the affected lymph node), which was taken from the patient by a trained surgeon, and is assessed for the type of lymphoma and how fast the lymphoma is growing. Figure 3 shows a biopsy of diffuse large B-cell lymphoma characterised by abnormal, large cells that are spread diffusely. If lymphoma cells are present, the subtype, and therefore diagnosis, is classified according to the WHO International Classification of Diseases for Oncology (ICD-O).²⁰ Commonly, surgical excision (or incision) biopsy removes the whole (or part) of the affected lymph

node for classification. A needle-core or endoscopic biopsy is reserved in cases where the location of the affected lymph node increases the risk of harm to the patient: in these cases, affected lymph nodes are likely to be surrounding organs (extranodal).

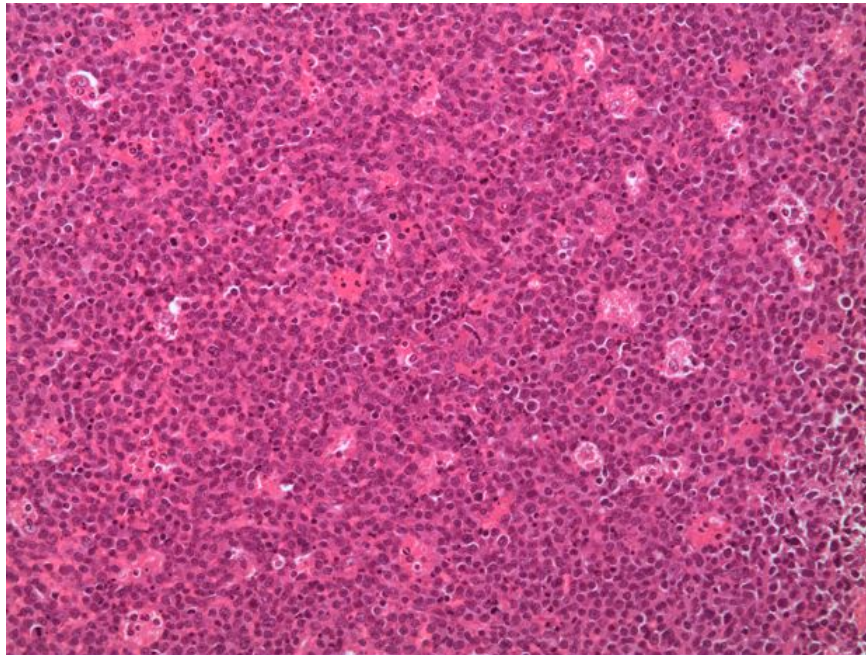


Figure 3: Microscopic image of a lymph node biopsy showing abnormal, large cells spread diffusely (characteristic of DLBCL)

Tumour grade

Tumour grade, otherwise called cancer differentiation, is an important prognostic value that describes the pace at which the cancer may develop and is useful for identifying the optimal treatment. Cancer cells that are well-differentiated closely resemble the original, normal cell; these cancer cells, for example follicular lymphoma, are slower-growing, or are in their early development, and usually have a better prognosis. On the other hand, poorly-differentiated cancer cells resemble the original, normal cell to a lesser extent; these cells, for example DLBCL, are fast-growing, and usually have a worse prognosis. Under the European Society for Medical Oncology guidelines, the morphological diagnosis for each subtype is confirmed with immunophenotypic investigations.⁷

Stage of disease

The stage of non-Hodgkin lymphoma at diagnosis, that is the extent to which the cancer has spread throughout the body, is indicative of treatment allocation and highly suggestive of prognostic outcomes. The Ann-Arbor staging system, applied initially to Hodgkin's lymphoma, provides a description of the extent of the cancer. There are four possible stages with modifiers appended to the stage description (table 3); for example, a patient who presents with affected lymph nodes on both sides of the diaphragm and has also spread outside of the lymphatic system (extranodal) is called stage IIIE.

Table 3: Ann-Arbor classification of stage at diagnosis for non-Hodgkin lymphoma

Stage	Location of cancer
I	Single region, usually one lymph node and the surrounding area
II	Two separate regions, both confined to one side of the diaphragm
III	Both sides of the diaphragm, including one organ or near the lymph nodes
IV	Disseminated involvement of one or more extralymphatic organs

Modifiers	Definition
A or B	A: absence of B-type symptoms, B: the presence of symptoms
S	Spleen involvement
E	Extranodal (not in lymph nodes)
X	Largest deposit (biopsy) is >10cm

1.2.3 Incidence patterns

Incidence worldwide

Non-Hodgkin lymphoma is one of the most common malignancies worldwide, ranked as high as the 5th most common in some countries; it is estimated that over half a million cases of, and a quarter of a million deaths due to, non-Hodgkin lymphoma were observed in 2018 alone.²¹ Due to its causes such as certain viruses, non-Hodgkin lymphoma is more prevalent than other cancers in north-east Africa.²¹

Global variations

According to recent data, although likely to be an artefact of changes in disease classification, countries with a very high human development index (HDI) have at least a two-fold higher incidence of NHL compared to countries with a lower HDI, with an age-standardised rate of over 10.3 males and 7.6 females per 100,000 being diagnosed in North America, Canada, north and western Europe, and Oceania.²² However, roughly 80% of the world's population, particularly countries with lower HDI, are not covered by cancer registry systems and incidence within such countries may be unreliable due to misdiagnosis and underenumeration.^{23,24} Underenumeration is likely to occur in countries with lower HDI because of the lack of facilities in underdeveloped healthcare systems to capture the fast-progressing lymphomas. Misdiagnosis can occur, even in countries with a high HDI and a well-developed healthcare system, because of vague or non-apparent symptoms for some lymphomas.^{25,26} Furthermore, haematological disease classifications have been redefined on multiple occasions over the past 20 years as more advanced diagnostic tools are produced.^{27,24} Cancer registries often report the diagnosis as a single overarching classifi-

cation (i.e., NHL 'not otherwise specified') because, unlike other cancers, lymphomas are diagnosed with a composition of histology, cytology, immunophenotyping, cytogenetics, imaging and clinical data, which is difficult for population-based epidemiological research to access systematically.²⁴

Time trends

Incidence rates were increasing at 3-4% per year in the 1970s and 1980s; in parts of Europe, North America, and Oceania, incidence rates were amongst the highest and generally increasing until the 1990s where they stabilised but still increased at a rate of 1-2% per year.^{28,29} Over the past 30 years, the age-standardised incidence rate has increased by 39% in the United Kingdom, males experience consistently higher rates than females. Although rates in the younger population has remained stable, the increasing trend is observed amongst those older than 50 years; the incidence rate of those aged over 80 years has increased by 67% since the 1990s.³⁰

National variations

Clinical commissioning groups (CCG) are National Health Service (NHS) organisations that represent a geographical region of England; they are responsible for the planning and commissioning of healthcare services within their region. There were 211 CCGs when they replaced Primary Care Trusts (PCT) in 2013; since then, the merging of CCGs reduced this number to 209 in 2015, and 191 in 2019.

Figure 4 shows the proportion of NHL diagnoses per 100,000 people within a boundary defined by a CCG (made using 'maps' R package, and linking CCG boundaries dataset to NHL dataset). This graph represents all patients diagnosed from 2005 to 2013, using boundaries of CCG defined as of 2016. The proportion of diagnoses changes slightly for each year, but is very similar. The boundaries for the CCG also differed for each year as CCGs were either split, merged or added; the coordinates for the CCG boundaries are available from the Office for National Statistics³¹ and the earliest data on the boundaries of each CCG are from 2016.

1.2.4 Risk factors

Age

Reasons for increasing incidence rates amongst older patients are not well defined. The concept of 'aging' could be viewed as biological or chronological (i.e., chronological age can be differentiated from age-related diseases).^{32,33} The occurrence of a mutation in a cell's genome that leads to cancer development becomes more likely over a longer period of time as the body is exposed to more carcinogens or a random error occurs in DNA replication.³⁴ An increasing age is associated with a reduction in the functional capacity of the immune

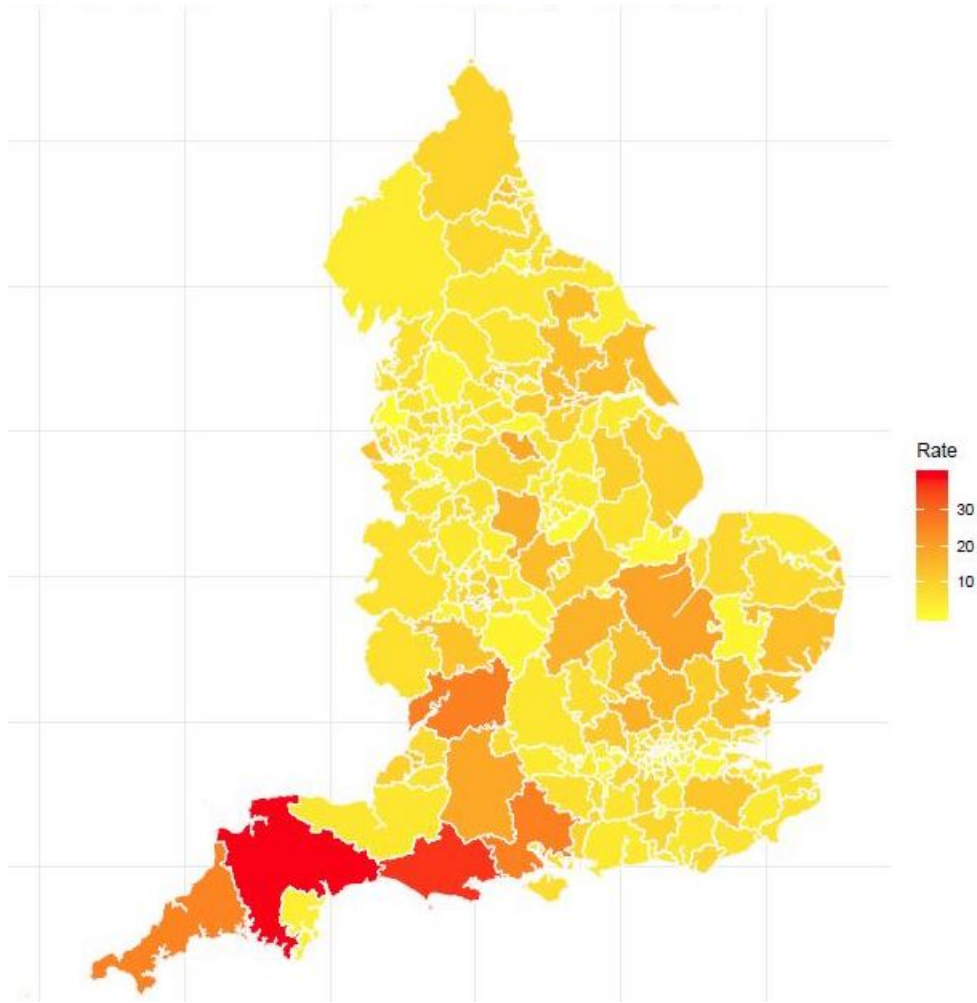


Figure 4: Age-standardised rate of NHL diagnosis (per 100,000 people) between 2005 and 2013 in England, bounded by Clinical Commissioning Groups.

system, specifically the output of T-cells due to thymic involution, leading to a decrease in the rate of detection of neoantigens produced by cancer cells.³⁵ Lymphoblastic lymphoma, a rare type of NHL, develops from mutated T-cells in the chest lymph nodes or thymus gland. However, lymphoblastic lymphoma is usually found amongst those under the age of 35. Thus, although biological aging is associated with chronological aging, there are subtle differences that must be taken into account.

Efficient DNA repair is essential to prevent cancer-causing mutations. Along the DNA damage repair (DDR) pathway, age-related changes occur in several DNA repair mechanisms: mismatch repair (MMR), base excision repair (BER), nucleotide excision repair (NER) and double-strand break (DSB).³⁶ Mutations that are not repaired by these mechanisms advance the aging process and may contribute to cancer cell development, if the mutation occurs within tumour-suppressor genes such as *p53*. *p53* genes maintain the genome stability, determine cell fate, and stimulate genes that reduce oxidative stress, the accumulation of which leads to prolonged longevity.³⁷ Although *p53* is central to cancer suppression, the gene is activated in latter phases of the DDR pathway, by which point

age-related changes may have sufficiently reduced, or mutated DNA within the p53 gene itself then reduces, the efficiency of the gene to suppress tumour malignancy.³⁷

Sex

Overall, non-Hodgkin lymphomas are more common, and age-standardised rates are higher, amongst males compared to females.²⁴ There is little difference in the sex rate ratio (males/females) of those with follicular or extranodal marginal zone lymphomas; whereas, there is a far higher incidence amongst males for other B-cell subtypes: particularly for Burkitt lymphoma.^{24,38} The association, and underlying biological mechanism, for the male-dominant incidence of NHL is as yet not fully understood.³⁸

Genetics

Due to the vast heterogeneity of NHL subtypes, the most common genes found to be mutated in NHL cases depends more on the subtype. Follicular lymphoma and diffuse large B-cell lymphoma are two of the most common subtypes of NHL. In follicular lymphoma, the overexpression of the antiapoptosis gene, *B-cell lymphoma 2 (BCL2)*, due to translocation is present in 85-90% of cases; other common genomic alterations include that of the histone modification gene *KMT2D* (80-90%) and mutations of Igh-epitopes in B-cell receptor signalling (85-95%).³⁹ In diffuse large B-cell lymphoma, the most common subtype of NHL, *BCL2* is the most common mutated expressed gene.^{40,41} The prevalent mutated genes for other forms of NHL depends on the subtype.⁴² Familial predisposition to NHL is estimated to be 1.7 times higher amongst first-degree relatives.⁴³

Lifestyle

Although the majority of studies did not find evidence of an association between smoking and the odds of overall lymphoma development,⁴⁴ subtype-specific studies found that those who smoked over a longer duration had an increased risk, with evidence of a dose-response effect, of follicular lymphoma.⁴⁵ As yet, there is no plausible biological explanation for a link between smoking and risk of follicular lymphoma. Similarly, the consumption of alcohol is not known to be a risk factor for NHL; contrary to other cancers, studies have shown a protective association, particularly with red wine.⁴⁶ This association could be explained by the anti-inflammatory properties of resveratrol (a phytochemical) within red wine. Obesity, and a higher body mass index, has been shown to increase the risk of specific NHL subtypes, such as diffuse large B-cell lymphoma,^{47,48} most likely due to the association between obesity and the hormone leptin, which exacerbates inflammatory reactions and influences T-lymphocyte function.

Other lifestyle risk factors have shown mixed conclusions on their association with NHL. Hair dyes manufactured before the 1980s have increased the risk of follicular lymphoma and chronic lymphocytic leukemia,^{49,50,51} Ultraviolet radiation increases the risk of NHL in England and Wales,⁵² which may explain the higher incidence in Oceania; however,

findings on these associations are inconsistent.⁵³ A meta-analysis showed the risk of NHL increased with each additional increase in consumption of dairy products.⁵⁴ However, this meta-analysis showed all case-control studies reported significant associations, and all cohort studies reported insignificant associations. There is unsubstantiated evidence of an increased risk of NHL from red meat items, but there was strong evidence of a reduced risk with increased intake of fruits and cruciferous vegetables.⁵⁵ The strongest evidence for occupational risk is amongst farmers or agricultural workers,^{56,44} which may be explained by exposure to chemicals such as solvents^{57,58} (benzene and trichloroethylene), pesticides⁵⁹ (organochlorine, organophosphates and carbamate insecticides), or polychlorinated biphenyls⁶⁰. Generally, health services, such as the NHS, suggest that modifiable risk factors for the risk of NHL are unclear and need further research.^{61,62}

Autoimmune disease

The cellular process in an autoimmune disease is described as the body's immune system to incorrectly identify the body's own tissues and organs as foreign cells; the immune system initiates phagocytosis, damaging healthy cells. The hyperactivity of the immune system leads to an increase in the proliferation of lymphocytes, which increases the risk of a cancerous lymph cell formation. Some autoimmune diseases have been linked to an increased risk of NHL, such as rheumatoid arthritis⁶³, systemic lupus erythematosus⁶⁴, and Sjogren's syndrome⁶⁵.

Infectious Diseases

Human immunodeficiency virus (HIV) weakens the immune system by attacking T-cells; these cells are responsible for removing waste cells and material, including infections. Weakened immune systems are more at risk of developing certain types of NHL, such as DLBCL.⁶⁶ In cases where the patient has HIV, a diagnosis of NHL is termed 'AIDS-defining NHL'. However, it is not clear whether HIV directly affects the DNA of lymphocytes and causes NHL, or if the virus substantially reduces the effectiveness of the immune system to a point where the lymphocytes are increasingly susceptible to structural changes of the cell's DNA.

Some viruses, such as Epstein-Barr virus (EBV), are known to cause NHL, for example Burkitt's lymphoma.⁶⁷ In countries where there is a high prevalence of EBV, for example African countries, Burkitt's lymphoma is the most common type of NHL.⁶⁸ Plasmodium falciparum malaria is also a risk factor for Burkitt's lymphoma because the presence of which exacerbates the amount of EBV-infected cells within lymph nodes. In Japan, during the 1970's, there was an unusual cluster of adult T-cell lymphoma (ATLL) cases that, after an epidemiological investigation, was shown to be caused by Human T-cell leukemia/lymphoma virus type 1 (HTLV-1).⁶⁹ Despite the normal stomach having no lymphoid tissue, gastric lymphomas occur because Helicobacter pylori infection causes the accumulation of lymphoid tissue with mucosa-associated lymphoid tissue (MALT) characteristics.⁷⁰

The association between some viruses, such as *Coxiella burnetti*, and B-cell non-Hodgkin lymphoma remains unclear.⁷¹

1.2.5 Treatment

Systemic treatments

Depending on the type, grade and stage of the lymphoma, the patient may have one of many differing treatments. Unlike most solid cancers where surgery is frequent, the treatment strategy for non-Hodgkin lymphoma is either watchful-waiting or first-line treatments such as radiotherapy, or immunochemotherapy.⁷² Watchful-waiting, generally advised for early-stage indolent (low-grade) lymphomas, is an approach in which the patient is observed for a period of time before introducing a medical intervention; for these patients, radiotherapy is recommended for non-metastatic bulky disease cases where the lymphoma can be precisely targeted. For radiotherapy, the aim is to control the lymphoma for as long as possible and eventually cure it. Radiotherapy may be too toxic for the organs when treating lymphomas with involvement of the lung or liver, particularly in cases with high tumour burden; instead, therapy as indicated for advanced stages is advised.⁷³ Local radiotherapy is advised for patients with localised stage IIA (asymptomatic) follicular lymphoma, otherwise a 'watch and wait' approach is advised.⁷⁴

In advanced-stage asymptomatic follicular lymphoma, rituximab induction therapy is usually offered. For those with advanced-stage symptomatic follicular lymphoma, a combination of chemotherapy and immunotherapy (monoclonal antibody) is recommended.^{74,75} For example, the most commonly prescribed regimen R-CHOP is a combination of chemotherapy drugs (cyclophosphamide, doxorubicin and vincristine), a steroid (prednisolone), and an immunotherapy (rituximab). Instead of CHOP, bendamustine may be prescribed with rituximab.⁷⁶ Amongst patients at risk or susceptible of cardiotoxic effects, doxorubicin is advised against by the clinician, and the patient may receive the less intense therapy of R-CVP (R-CHOP without doxorubicin).

The standard treatment for high-grade (aggressive) lymphomas, such as DLBCL, is R-CHOP; the dose-intensity of which depends on the stage, grade, and underlying health status.⁷⁷ The cellular composition of aggressive lymphomas are poorly-differentiated and often respond better to treatment compared to indolent lymphomas because immunochemotherapies target rapidly-growing cells. High-grade (aggressive) lymphomas are treated similarly to advanced low-grade lymphomas. For otherwise healthy patients, every twenty-one days six cycles of CHOP in combination with six doses of rituximab is administered. Concurrently to R-CHOP, patients with bulky disease have been shown to respond well to radiotherapy.⁷⁸ Evidence has shown that the presence of human immunodeficiency virus (HIV) does not worsen the outcome compared to HIV-negative patients, thus standard R-CHOP treatment is recommended in association with anti-viral therapy.⁷⁹ Indolent lymphomas,

such as low-grade follicular lymphoma, can go through histological transformation into high-grade forms. The reported incidence of histological transformation is broad and the treatment is individualised, but dose-intensification and consolidation is shown to improve the overall survival for these patients.⁸⁰

Emerging treatments

An immunotherapy known as Chimeric Antigen Receptor T-cell (CAR-T) therapy comes close to offering the patient a personalised treatment regimen. The process of CAR-T therapy is to: sample the patient's T-cells from their blood, genetically engineer the T-cell to present a certain type of antigen on the cell surface becoming a CAR-T cell, allow this cell to multiply, and finally, drip (intravenously) the CAR-T cells back into the patient's bloodstream. CAR-T cells travel through the bloodstream locating, binding, and destroying cancer cells. In January 2019, the use of CAR-T therapy was approved by the National Institute for Health and Care Excellence (NICE) for patients diagnosed with DLBCL in England.⁸¹ Since this approval, the NHS is providing CAR-T therapy for relapsed or refractory DLBCL patients who have experienced two or more systemic therapies.⁸²

1.2.6 Clinical management

Clinical Management

Earlier diagnosis of NHL is known to have better outcomes. Researchers and policy makers are interested in the characteristics of the patients who take certain routes to diagnosis (RTD).⁸³ The RTD, classified according to Elliss-Brookes et al (2012), are given in Table 4. Each route is a category representing the initial contact the patient had with the healthcare system that led to the diagnosis of the cancer.

It is often that patients recorded as 'death certificate only' are not diagnosed on the date of their death but at some time after their death, such as during a pathological investigation (an autopsy). These patients are likely to have been living with cancer prior to their death and for an unknown period of time: indicating that the diagnosis date (from pathological intervention) is an unknown-period-of-time later than the actual onset of the cancer. In survival analyses, if these patients were included in the analysis, then these patients would contribute a survival time of zero: giving an underestimate of the true survival time. Therefore, these patients are usually removed from survival analyses to avoid underestimation of survival time.

There are currently no widely recommended screening tests for NHL,^{84,85} thus there are no patients recorded as being diagnosed via a 'screen-detected' route. An 'unknown' route to diagnosis is recorded when there is no available information on the patients interaction with the healthcare system within clinical records 6-months prior to diagnosis. In

many cases. the patient’s first interaction is with a general practitioner (GP) referral to a haematologist-oncologist; if the patient presents with severe symptoms, or an advanced stage, the GP may refer the patient to see a haematologist-oncologist within two weeks: instigating a two-week-wait route to diagnosis. In severe cases, the lymphoma may initiate unexplained acute symptoms causing patients to present via Accident and Emergency. Upon consultation with a haematologist-oncologist, the patient will usually receive an excision biopsy and is encouraged to have a Positron Emission Tomography - Computed Tomography (PET/CT) scan, which is a combined medical imaging process that together gives radiologists a comprehensive view of the anatomical extent to which the cancer has spread. It can also be used as a guide to more accurately target affected extranodal lymph nodes during a biopsy.

Table 4: The eight routes to diagnosis defined by Elliss-Brookes *et al.* (2012)

Route	Description
Screen-Detected ¹	Detected via a screening programme
GP referral	Routine referrals made by a general practitioner
Emergency presentation	Emergency route via A&E, or other emergency referral
TWW ²	Urgent GP referral with a suspicion of cancer
Inpatient Elective	Admission from a waiting list, booked or planned
Other Outpatient	An outpatient appointment: self or unknown referral
DCO	No data available before the death certificate diagnosis
Unknown	No data available from any of the known routes

¹ Currently, there are no effective screening tests for non-Hodgkin lymphoma.

² Two-week wait.

1.2.7 Prognostic factors

The treatment allocated to the patient strongly determines the prognosis of NHL, but this prognosis also depends on an ensemble of tumour and patient characteristics, which largely defines the possible treatments. Both types of characteristics are important to consider and, for example, are included in the International non-Hodgkin Lymphoma Prognostic Index (IPI), which was devised to improve the post-treatment prognosis.⁸⁶

Grade

Patients with high-grade (aggressive), particular subtypes, bulky or distant stage of NHL exhibit higher mortality rates and lower survival probabilities.^{3,87,88} Of these characteristics, the lymphoma’s grade is a key indicator of the patient’s prognosis as high-grade, compared to low-grade, lymphomas are faster-growing and may hasten the patient’s date of death. On the other hand, high-grade lymphomas are more responsive to treatment and have a higher chance of cure.

Stage

Classified by the Ann Arbor system, and together with the respective subtype-specific International Prognostic Index, the stage of the lymphoma is another tumour-related prognostic indicator. Localised lymphomas are characteristic of an early stage at diagnosis where there is a single affected lymph node. Although staging is a key indicator of the performance of cancer diagnosis, it is rarely an independent prognostic factor when considering treatment regimens. Patients with a later stage at diagnosis (stages III and IV) have significantly lower survival probabilities than early stages.⁸⁸ Patients with primary lymph node lymphomas are more commonly diagnosed at a later stage (56% at stage III or IV) compared to early stage; whereas, primary extranodal lymphomas are more commonly diagnosed at an earlier stage (65% and 74% at stage I and II, respectively) compared to late stage.⁸⁹

Anatomical location

Prognosis of similarly-graded lymphomas, although having similar treatments, can vary due to anatomical location; a topographical examination may indicate secondary disease occurrence and provide a more accurate prediction of prognosis. Extranodal lymphomas afflicting the surrounding organs infer a worse prognosis, which also varies by location of specific anatomical sites. The distribution of extranodal involvement in NHL cases is uneven and they more commonly occur in the head and neck (Waldeyer's tonsillar ring), central nervous system, lung, skin, bone and gastrointestinal tract. Amongst DLBCL cases, the involvement of kidneys, lungs or reproductive organs exacerbates unfavourable outcomes; whereas, craniofacial, bone or thyroid involvement implies less severe ramifications. This dissimilarity in prognosis may be explained by the susceptibility of these anatomical sites in relation to a higher risk of central nervous system recurrence, and the capacity for the lymphoma's response to standard immunochemotherapy.⁹⁰ In studies investigating the cause of death, the most common was due to infection, mainly granulocytopenia, of an essential organ secondary to NHL.⁹¹ This association may be explained by the patient's susceptibility to infections after splenectomy or combination chemotherapy. For example, infections may be more severe to the patient if there is a presence of extranodal involvement in the lungs (or kidneys) where, due to the effect of chemotherapy, the lungs (or kidneys) are now less effective in removing the infection.

Patient characteristics

Along with stage at diagnosis, the IPI consisted of four other equally predictive factors: age older than 60 years, a performance status greater than 2 (with a maximum score of 5: indicating death), high lactate dehydrogenase levels, and the involvement of more than two extranodal sites. An IPI of 5 indicates a worst post-treatment prognosis because all available treatments cannot be delivered due to poor health conditions.⁹²

Lactate dehydrogenase (LDH) is a naturally occurring enzyme found in cells; its function is to catalyse the conversion of pyruvate into lactate and vice-versa. In normal cells, glucose is converted into pyruvate through the process of glycolysis, and LDH converts pyruvate into lactate in the absence, or low levels, of oxygen. At high concentrations of lactate, LDH displays feedback inhibition. However, in cancer cells this feedback ceases to function, leading to a higher uptake of glucose and resulting in abnormally high levels of lactate (lactic acid): this phenomenon is known as the Warburg effect.⁹³ This effect is utilised by fluorodeoxyglucose positron emission tomography (PET-CT) scans to locate cancer cells. Levels of LDH are also assessed from a patient's blood sample and high levels (greater than 250 units per litre of blood amongst adults) indicate tissue damage, although this test does not specify the type of tissue that is damaged.

Elevated blood LDH levels are known to be prevalent amongst cancer patients;^{94,95,96,97} recently, research has shown LDH levels to be a diagnostic marker of cancer existence up to 3 years prior to diagnosis, and indicative of overall survival for several cancers.⁹⁸ For NHL, better survival is observed for patients with normal LDH levels, independent of histological subtype and clinical stage.⁹⁹ In certain subtypes, such as DLBCL, a 1.5-fold increase in LDH over a period of 3 months is associated with an increased likelihood of relapse.¹⁰⁰ Although, elevated LDH levels are not always indicative of cancer progression. Hypothyroidism may cause increased LDH levels, and in cases where patients have previously been treated for NHL this has resulted in an unnecessary tumour hunt possibly due to the asymptomatic presentation of hypothyroidism in its early stages.¹⁰¹

An increasing chronological age is associated with lymphomas presenting with more aggressive biological features, such as DLBCL and grade 3b follicular lymphoma.¹⁰² Amongst older patients, the incidence of high-grade NHL is almost double that of low-grade NHL.¹⁰³ Unlike other cancers, such as breast cancer, younger age is not known to be associated with histological subtypes.

Another key prognostic indicator is the patient's performance score; a higher score indicates a sub-optimal response to treatment. Performance score, most commonly defined by the Eastern Co-operative Oncology Group (ECOG) scale,¹⁰⁴ consists of five categories ranging from 0 (asymptomatic and fully functional) to 4 (bedbound and completely disabled); the sixth category (a score of 5) indicates death and, for obvious reasons, is removed when assessing the potential for response to treatment. Generally, patient- and oncologist-allocated performance scores coincide, however, interobserver scores are not always perfectly correlated and are vulnerable to observer bias: the patient's performance score may also be time-dependent.¹⁰⁵ Both patient- and oncologist-allocated PS have been shown to be reliable prognostic markers.¹⁰⁵ Compared to other measures, the ECOG score shows greater prognostic prediction.¹⁰⁶

Similar to performance score, comorbidity status describes the underlying health condition of a cancer patient, and is suggestive of treatment and prognosis. The frequent unavoidable

dose-reductions of immunochemotherapy may explain this association with worse prognosis.¹⁰⁷ Additionally, high-intensity treatment is less frequent amongst older patients with high-impact comorbid conditions, which likely affects the survival within the immediate months after diagnosis.¹⁰⁸ Amongst patients with NHL, all-cause mortality is shown to be higher for patients with congestive heart failure, diabetes, or dementia.¹⁰⁹ Even though the prevalence of these comorbidities are higher amongst patients with NHL, the pathophysiological links between certain comorbid conditions and the development of NHL are not clear.^{110,111,112}

Previous research indicates that follicular lymphoma incidence is similar between males and females, however, women are less likely to develop more aggressive lymphoma subtypes compared to men; this lower incidence may be associated with greater exposure to female reproductive hormones.¹¹³ Some studies exhibit similar prognoses between females and males,⁸⁶ while others suggest males experience an inferior survival.^{114,115} Considering the plausibility of the inferior association is not well justified in literature, there is a suggestion that (amongst patients with DLBCL) females have a lower intrinsic clearance for higher doses of immunochemotherapy compared to males.¹¹⁶

1.3 Cancer survival

1.3.1 Measures of disease burden

Incidence, mortality and survival are three key estimates used to measure the burden of disease (e.g. cancer) in a population. Incidence describes the period-specific occurrence of the disease and is used to study the etiology of a disease and its outcome. In 2018, the estimated number of new cancer cases worldwide was roughly 18 million. Men had a 23% higher diagnosis rate than women, with rates of 315 and 238 per 100,000, respectively, in North America, north and west Europe and Oceania.²¹ The number of new cases is expected to increase to 30 million annually by 2040, with 35% of cancer deaths occurring in high-income countries.¹¹⁷

Mortality is a measure of the current impact of the disease on the population and the healthcare services in which they reside. In 2018, the age standardised mortality rate of NHL per 100,000 people worldwide was approximately 3.3 and 2.0 amongst males and females, respectively;²¹ in 2014, in England, the rates were 12.1 and 7.9, respectively.¹¹⁸ By the year 2035, the mortality rate of NHL in England is expected to decrease by 22% but the number of deaths is expected to increase by 32%:¹¹⁸ the opposing directions of change being explained by the increase in size and age of the population.

Survival is a measure of the length of time from the point of diagnosis through to the date of death. From a patient's or clinician's perspective, survival indicates the longevity of a treatment, but can also document the overall efficacy of policy plans, accessibility of

available treatment resources and inform the National Health Service framework. Relative survival, as opposed to cause-specific survival, is used when the cause of death of patients in a population is unreliable or unavailable. The relative survival setting, which uses life tables, provides measures of survival that is independent of the competing risks within the population. The measures of survival are independent from the information on cause of death, which is particularly useful when the cause of death is not unavailable or unreliable. Net survival, which can be estimated within the relative survival setting, is a measure of survival that is particularly useful for comparing cancer survival estimates between different populations and across time periods: thus, it is the approach used in this thesis.

1.3.2 Non-Hodgkin lymphoma survival

Temporal trends

Before 1990, 5-year survival for NHL patients in England was less than 50% with a slight improvement into the late 1990's.¹¹⁹ In contrast, between 2000 and 2011, 5-year survival steadily increased from 50% to 70% most likely due to improvements in diagnosis, clearer staging with PET-CT scans and more successful treatments such as immunotherapies.¹²⁰ Similarly, during 2010-2014, 5-year survival was above 70% amongst Scandinavian and Western European countries and Australia.¹²¹

Conditional survival

Compared to other developed countries, England has amongst the worst 5-year survival combining all cancers. For three common cancers (lung, colorectum and prostate) short-term survival is indicative of long-term survival. For NHL patients specifically, studies that separate 1-year survival from between 1- and 5-year survival show statistically significant detriment only for the former of the outcomes,¹²² which may be explained by differences in early diagnosis and/or initial management of advanced disease. Moreover, differences in long-term survival estimates between the UK and other European countries may not be explained by short-term survival.

1.4 Health inequality

1.4.1 Public health policies

Acknowledging the differential survival estimates, the National Health Service Cancer Plan,¹²³ devised in 2000, was the first comprehensive strategy attempt made by the National Health Service (NHS) to increase cancer survival of patients in England to compare with the best in Europe. Initial analyses found survival inequalities in England between populations and social groups: providing evidence against the fundamental precept that

access to healthcare was equitable. This precept assumes that timely diagnosis and availability and accessibility of treatments does not depend on biologically unrelated aspects, such as patient or healthcare system characteristics.

Possible reasons for the inequitable outcomes were delays in diagnosis and treatment for some patients, which led to a number of initiatives, for example the National Awareness and Early Diagnosis Initiative. The initiatives aimed to find out more about these patients who had a late diagnosis and emergency route to diagnosis, which explored the variation in survival due to patient characteristics. One of the main commitments of the NHS cancer plan was to reduce the gap in survival between socioeconomic groups. Thus, inequalities in cancer survival were investigated between patient characteristic such as age, deprivation, ethnicity, and lifestyle. Furthermore, it was suggested that access to high quality services varied across the NHS, such as insufficient access to radiotherapy services.^{124,125}

The Cancer Reform Strategy (CRS),¹²⁶ published in 2007, aimed to build on the developments toward improved survival initiated by the NHS Cancer Plan. The aims were adapted to improve services such that cancer survival is comparable to the best in the world. The CRS at this time recognised a number of factors that lead to inequalities in cancer survival, and suggested actions to combat the inequalities. Patients with a disability were recognised to be susceptible to a reduced survival of cancer, including haematological malignancies.

One of the goals of Cancer Research UK (CRUK) and CRS is that, by 2020, two-thirds of those with common cancers will survive for at least 5 years. The National Cancer Equality Initiative (NCEI) was set up to address this challenge and investigate the inequalities in cancer survival.¹²⁷ Some patient characteristics (such as age, gender, deprivation, and ethnicity) partly explained the inequality in survival, yet variations in survival remained. NCEI suggested that comorbidities could partly explain the variation in survival between deprivation groups because the prevalence of comorbidities was higher amongst those living in more deprived areas. This motivation applies in the context of this thesis in terms of area-level deprivation and multiple categorisations of comorbidity status.

1.4.2 Classifying social groups

Comparing survival estimates to other European countries requires the classification of sociodemographic measures to be consistent, accurate, and reliable across time periods, age groups, and countries. As opposed to simple classifications of population groups, such as gender or nationality, defining and measuring socioeconomic status is a multifaceted classification process.

Terms such as social class, social stratification, social or socioeconomic status (SES) have differing theoretical bases and provide different measurements given the context of the research field. Early definitions of SES arise from social (and philosophical) research con-

ducted by Karl Marx and Max Weber. According to Marx, an individual's SES was exogenous and structural, this viewpoint was largely due to the current capitalist economic system at the time. Conversely, Weber defined SES as a multidimensional classification system whereby individuals could trade and integrate their skills and abilities within society so as to improve their socioeconomic position. Weber's multiple dimensions for SES are domains such as income, education, and occupation, which are also dependent on the available opportunities given to an individual in society. The combination of Marx's structural and Weber's societal views helps to understand the relationship between SES and health within society. For example, an individual's occupation is obtained by employment opportunities within society and by familial welfare. Thus, individual-level explanations of deprivation cannot account for the entirety of the causes of deprivation.

In this thesis we refer to socioeconomic status as the socially derived economic factors that influence what positions individuals or groups hold within the multiple-stratified structure of a society.¹²⁸ The main purposes in measuring socioeconomic status are to describe and monitor diseases across societal groups or geographical regions, explain causal mechanisms, or statistically adjust for socioeconomic circumstances.¹²⁹ The object of examining inequalities in survival of NHL patients is the health policies and provision of health services delivered to geographical areas. In this context, area-level socioeconomic indicators are more appropriate as they aim to include all factors that ultimately shape health outcomes.¹³⁰

Area-level indicators can be obtained by aggregating individual-level measures of SES but this is often unfeasible. Instead, composite measures such as Indices of Multiple Deprivation, which comprises several domains, can be calculated from more easily obtainable data sets, such as local authority districts. More modern approaches to classify SES reflect the combinatorial effect of multiple factors (such as geographic, income, occupation or education) from different domains into a score or index. Census-based scores, such as Carstairs¹³¹ and Townsend¹³², and the more recent Indices of Multiple Deprivation,^{133,134} are examples of area-level scores that have been used to examine the relationship between deprivation and health outcomes within areas of certain characteristics. These scores have also been used to reflect individual-level deprivation, however, associations could be biased in either direction.^{135,136,137}

Areas that are marked as deprived may contain large numbers of people who are not deprived, and vice versa.¹³⁸ Smaller, rather than larger, areas are preferred for two reasons that do not undermine the small-area approach to defining deprivation: firstly, the majority of deprived people do not live in deprived areas;^{139,140} secondly, it is less cost-effective than general antideprivation policies that target deprived individuals wherever they live.¹⁴¹ Census-based data was previously used for small-area coverage, however, this data used problematic proxies (such as 'no access to a car') for income deprivation, and the collection was decennial. Since 2000, administrative data sources have provided analysis of depriva-

tion measures at intercensal periods.¹⁴² Since 2000, small-areas have been defined by the Office for National Statistics as super output areas (SOA); the lower-SOAs (LSOA) are adjacent output areas representing a median of 1500 individuals assigned to their LSOA based on postcodes look-up tables.

A combination of domain deprivation measures are used to index the LSOAs. An initial score is determined for each of the seven domains (income, employment, health and disability, education, barriers to housing and services, living environment, and crime) and the final Index of Multiple Deprivation (IMD) is a weighted combination of the seven domains. Adding to validity and reliability, LSOAs are consistent across time-periods, domains are given strict criteria to avoid any overlap in classification with other domains and to ensure information on a domain is available for the whole country. However, further work on the selection of the domain weights is ongoing and the use of subjective wellbeing has been proposed but not explored further.¹⁴³

A key purpose of measuring SES is to statistically adjust for socioeconomic circumstances, particularly when another factor (e.g. comorbidity - discussed below) is of primary interest. In this context, this thesis investigates the compositional effect of deprivation (encompassing multiple measures of deprivation) on survival of patients with non-Hodgkin lymphoma. Another key purpose, not of primary interest in this thesis, is the causal effect of deprivation on survival, which could argue for the examination of specific indicators (e.g. education) and their timing of exposure.^{144,145}

1.4.3 Classifying comorbidity status

Based on the IMD, inequalities in survival between deprivation groups have narrowed but still remain. Possible explanations, given by the latest cancer plan, are that comorbidities are more prevalent amongst those living in deprived areas; suggesting that comorbid conditions play a role towards the deprivation-gap in survival. Similarly to socioeconomic status, the task of classifying an individual's comorbidity status is more complex.

In the context of cancer epidemiology, comorbidities are defined as the coexistence of disorders, in addition to a primary disease of interest, which are causally unrelated to the primary disease (e.g. cancer or non-Hodgkin lymphoma).^{146,147} Reviews of literature on measures of comorbidity in cancer epidemiology concede there exists no gold standard approach, and the choice of the measure is at the discretion of the researcher based on the study question, population and data available.¹⁴⁸

In cancer epidemiology, the Charlson comorbidity index is most commonly used to assess the impact of comorbid conditions on cancer patient health outcomes for several reasons. The index is used extensively amongst cancer patient populations, most of the relevant conditions are included, there is strong evidence to support concurrent and predictive

validity, some evidence for moderate level of reliability, and is cheap and easy to use with routinely collected administrative data.¹⁴⁸ The Charlson index is a weighted index with the weights being equivalent to the adjusted relative risks for 1-year mortality for each comorbid condition. The weighted indices, such as Charlson, vary in terms of the number, and type, of comorbid conditions included in the index. For example, conditions such as cardiac, respiratory, liver and renal are included universally; however, obesity, amongst other conditions, are not included in the Charlson index but is a prognostic factor of survival for NHL patients.

The performance of the Charlson index in predicting short-mortality is similar to other indices, such as the Elixhauser index¹⁴⁹, and both indices are valid prognostic indicators across updated versions of the International Classification of Diseases.¹⁵⁰ Although having adequate predictive capability, there is disagreement in the validity of the Charlson index's discriminative ability for longer follow-up times.^{151,150,152,153}

1.5 Inequality and non-Hodgkin lymphoma

International differences in survival

Recent comparisons showed that the European age-standardised 5-year cancer survival of non-Hodgkin lymphoma patients was approximately 60% in 2007.¹⁵⁴ Except from Wales, England (56.7% survival) had lower survival compared to other countries in the United Kingdom and European areas: Northern (63.3%), Central (62.5%), Southern (58.7%). Only Eastern Europe (49.7%) showed a lower average survival, Wider international comparisons showed that, up to 2014, 5-year cancer survival improved to 64.9% in the United Kingdom but was still trailing behind other developed countries: Australia (71.2%), Canada (68.6%), United States (68.1%), Denmark (70.9%) and France (69.6%).¹⁵⁵

Survival at 1-year since diagnosis in England was significantly lower compared to the European average, but was no longer evident at either 5-years or 5-years given 1-year survival. Suggesting that the time lived amongst patients in England after diagnosis, but before 5-year survival, is far less than other European countries, on average. As is synonymous, the amount of life-years lost is greater in England than in European countries even though 5-year survival, conditional on 1-year survival, is similar.¹²²

Recent advances in statistical methods (i.e. relative survival and age-standardisation), data storage and collection, development of life tables and the increased rigor of adherence to follow-up have contributed to a more reliable and accurate comparison of survival estimates between countries. Yet, it is possible that some of the variation in survival estimates are due to differences in statistical methods, population coverage, diagnostic activity or over-diagnosis from the detection of less aggressive tumours that would not have reduced the patient's lifetime. Furthermore, the variation in survival between patients with the same

sociodemographic characteristics and the same morphology may be due to differences in diagnostic accuracy rate between countries.¹²⁰

Socioeconomic differences in survival

Survival of those living in deprived areas is substantially lower than those living in least deprived areas. Prior to the NHS Cancer Plan (2000), the deprivation gap in 5-year relative survival for males increased from 4.4% in 1986 to 7.3% in 1999, for females the deprivation gap remained around 5.4%.¹¹⁹ Most of the socioeconomic deficits in survival occur shortly after diagnosis, and tended to attenuate or disappear with time since diagnosis, which may also explain the international inequalities in survival with other European countries. After the NHS Cancer Plan, during initialisation and implementation, the deprivation gap did not change for males but widened for females by 2.0%.¹⁵⁶ Implying the cancer plan failed to target patients in more deprived areas and reduce the deprivation gap.

The proportion of avoidable deaths (i.e. when socioeconomic inequalities in excess mortality does not exist) has remained at around 17% since the mid-1990s. The stabilised avoidable deaths from 1996 to 2006 was partly due to the opposing changes in avoidable deaths between genders. Amongst males, from 1996 to 2006, this proportion increased from 12.6% to 14.3%; amongst females, this proportion decreased from 21.7% to 17.9%.¹⁵⁷ The impact of deprivation on the proportion of avoidable deaths not only differs between genders but has opposing effects. The higher proportion of avoidable deaths amongst women may be explained by the unemployment rate or the lower average income compared to males.

Possible explanations for the persistent inequalities in survival are the tumour, patient and healthcare system factors, particularly inequalities in access to the healthcare system, which have been reported in universal-access healthcare systems.¹⁵⁸ Inequalities in survival may also be due to underlying health conditions (comorbidities), such as congestive heart failure. For these patients a less intensive treatment toxicity is recommended. The prevalence and severity of comorbidities is unlikely to be uniformly distributed between countries, therefore treatment allocation and tumour management may explain the survival inequalities.

1.6 Study rationale

Inequalities in survival of patients with non-Hodgkin lymphoma persist, even after successive cancer plans aiming to improve survival for those living in more deprived areas. This thesis builds upon the theme of variation and inequality in survival through an examination of more recent data and assessment of the current framework of the healthcare system to reduce inequalities.

References

- [1] Terese Winslow LLC. Non-Hodgkin Lymphoma, Adult, Stage IV, 2009. URL <https://www.teresewinslow.com/>.
- [2] Cancer Research UK. Non-Hodgkin Lymphoma, 2018. URL <https://www.cancerresearchuk.org/about-cancer/non-hodgkin-lymphoma/about>.
- [3] Kate R Shankland, James O Armitage, and Barry W Hancock. Non-Hodgkin lymphoma. *The Lancet*, 380(9844):848–857, 9 2012. ISSN 0140-6736. doi: 10.1016/S0140-6736(12)60605-9. URL [https://doi.org/10.1016/S0140-6736\(12\)60605-9](https://doi.org/10.1016/S0140-6736(12)60605-9).
- [4] Steven H Swerdlow, Elias Campo, Stefano A Pileri, Nancy Lee Harris, Harald Stein, Reiner Siebert, Ranjana Advani, Michele Ghilmini, Gilles A Salles, Andrew D Zelenetz, and Elaine S Jaffe. The 2016 revision of the World Health Organization classification of lymphoid neoplasms. *Blood*, 127(20):2375–2390, 2016. doi: 10.1182/blood-2016-01-643569. URL <http://www.bloodjournal.org/content/bloodjournal/127/20/2375.full.pdf>.
- [5] NHS Digital. Non-Hodgkin Lymphoma: Symptoms, 2018. URL <https://www.nhs.uk/conditions/non-hodgkin-lymphoma/symptoms/>.
- [6] Mayo Clinic. Non-Hodgkin Lymphoma: Symptoms, 2018. URL <https://www.mayoclinic.org/diseases-conditions/non-hodgkins-lymphoma/symptoms-causes/syc-20375680>.
- [7] European Society for Medical Oncology. European Clinical Practice Guidelines: Haematological Malignancies, 2019. URL <https://www.esmo.org/Guidelines/Haematological-Malignancies>.
- [8] A Smith, S Crouch, S Lax, J Li, D Painter, D Howell, R Patmore, A Jack, and E Roman. Lymphoma incidence, survival and prevalence 2004–2014: sub-type analyses from the UK’s Haematological Malignancy Research Network. *Br J Cancer*, 112(9):1575–1584, 2015. doi: 10.1038/bjc.2015.94. URL <http://dx.doi.org/10.1038/bjc.2015.94><https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4453686/pdf/bjc201594a.pdf>.
- [9] Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. B Cells and Antibodies - Molecular Biology of the Cell, 2002. URL <https://www.ncbi.nlm.nih.gov/books/NBK26884/>.
- [10] Julie M Vose. Mantle cell lymphoma: 2017 update on diagnosis, risk-stratification, and clinical management. *American Journal of Hematology*, 92(8):806–813, 8 2017. ISSN 0361-8609. doi: <https://doi.org/10.1002/ajh.24797>. URL <https://doi.org/10.1002/ajh.24797>.

- [11] Thomas J Kipps, Freda K Stevenson, Catherine J Wu, Carlo M Croce, Graham Packham, William G Wierda, Susan O'Brien, John Gribben, and Kanti Rai. Chronic lymphocytic leukaemia. *Nature Reviews Disease Primers*, 3(1):16096, 2017. ISSN 2056-676X. doi: 10.1038/nrdp.2016.96. URL <https://doi.org/10.1038/nrdp.2016.96>.
- [12] Denis Burkitt. A sarcoma involving the jaws in african children. *British Journal of Surgery*, 46(197):218–223, 11 1958. ISSN 0007-1323. doi: 10.1002/bjs.18004619704. URL <https://doi.org/10.1002/bjs.18004619704>.
- [13] Georg Lenz, George W Wright, N C Tolga Emre, Holger Kohlhammer, Sandeep S Dave, R Eric Davis, Shannon Carty, Lloyd T Lam, A L Shaffer, Wenming Xiao, John Powell, Andreas Rosenwald, German Ott, Hans Konrad Muller-Hermelink, Randy D Gascoyne, Joseph M Connors, Elias Campo, Elaine S Jaffe, Jan Delabie, Erlend B Smeland, Lisa M Rimsza, Richard I Fisher, Dennis D Weisenburger, Wing C Chan, and Louis M Staudt. Molecular subtypes of diffuse large B-cell lymphoma arise by distinct genetic pathways. *Proceedings of the National Academy of Sciences*, 105(36):13520–13525, 2008. doi: 10.1073/pnas.0804295105. URL <http://www.pnas.org/content/105/36/13520.abstract><http://www.pnas.org/content/105/36/13520.full.pdf><http://www.pnas.org/content/pnas/105/36/13520.full.pdf>.
- [14] Luc Xerri, Stephan Dirnhofer, Leticia Quintanilla-Martinez, Birgitta Sander, John K C Chan, Elias Campo, Steven H Swerdlow, and German Ott. The heterogeneity of follicular lymphomas: from early development to transformation. *Virchows Archiv*, 468(2):127–139, 2016. ISSN 1432-2307. doi: 10.1007/s00428-015-1864-y. URL <https://doi.org/10.1007/s00428-015-1864-y>.
- [15] Dominique Bron, Nathalie Meuleman, on behalf of Eurobloodnet for rare diseases Hematology', and E H A S W G 'Aging and. Marginal zone lymphomas: second most common lymphomas in older patients. *Current Opinion in Oncology*, 31(5), 2019. ISSN 1040-8746. URL https://journals.lww.com/co-oncology/Fulltext/2019/09000/Marginal_zone_lymphomas__second_most_common.5.aspx.
- [16] Marc S Raab, Klaus Podar, Iris Breitkreutz, Paul G Richardson, and Kenneth C Anderson. Multiple myeloma. *The Lancet*, 374(9686):324–339, 7 2009. ISSN 0140-6736. doi: 10.1016/S0140-6736(09)60221-X. URL [https://doi.org/10.1016/S0140-6736\(09\)60221-X](https://doi.org/10.1016/S0140-6736(09)60221-X).
- [17] J A N Waldenström. Incipient myelomatosis or 'essential' hyperglobulinemia with fibrinogenopenia — a new syndrome? *Acta Medica Scandinavica*, 117(3-4):216–247, 1 1944. ISSN 0001-6101. doi: <https://doi.org/10.1111/j.0954-6820.1944.tb03955.x>. URL <https://doi.org/10.1111/j.0954-6820.1944.tb03955.x>.
- [18] Fernando Cabanillas and Bijal Shah. Advances in Diagnosis and Management of Diffuse Large B-cell Lymphoma. *Clinical Lymphoma, Myeloma and Leukemia*, 17

- (12):783–796, 12 2017. ISSN 2152-2650. doi: 10.1016/j.clml.2017.10.007. URL <https://doi.org/10.1016/j.clml.2017.10.007>.
- [19] H Tilly, M Gomes da Silva, U Vitolo, A Jack, M Meignan, A Lopez-Guillermo, J Walewski, M André, P W Johnson, M Pfreundschuh, and M Ladetto. Diffuse large B-cell lymphoma (DLBCL): ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Annals of Oncology*, 26(suppl 5):v116–v125, 2015. doi: 10.1093/annonc/mdv304. URL http://annonc.oxfordjournals.org/content/26/suppl_5/v116.shorthttp://annonc.oxfordjournals.org/content/26/suppl_5/v116.full.pdfhttps://watermark.silverchair.com/mdv304.pdf?token=AQECAHi208BE490oan9kKhW_Ercy7Dm3ZL_9Cf3qfKAc485ysgAAAd8wggHbBgkqhkiG9w0BBwa.
- [20] April Fritz, Constance Percy, Andrew Jack, Kanagaratnam Shanmugaratnam, Leslie H. Sobin, D. Maxwell Parkin, and Sharon L. Whelan. *International Classification of Diseases for Oncology*. World Health Organisation, 3rd editio edition, 2000. URL <https://apps.who.int/iris/handle/10665/42344>.
- [21] Freddie Bray, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L Siegel, Lindsey A Torre, and Ahmedin Jemal. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 68(6):394–424, 11 2018. ISSN 0007-9235. doi: 10.3322/caac.21492. URL <https://doi.org/10.3322/caac.21492>.
- [22] World Health Organisation. International Agency for Research on Cancer: cancer fact sheets, 2019. URL <https://gco.iarc.fr/today/fact-sheets-cancers>.
- [23] Donald M Parkin. The evolution of the population-based cancer registry. *Nature Reviews Cancer*, 6(8):603–612, 2006. ISSN 1474-1768. doi: 10.1038/nrc1948. URL <https://doi.org/10.1038/nrc1948>.
- [24] E Roman and A G Smith. Epidemiology of lymphomas. *Histopathology*, 58(1):4–14, 2011. doi: 10.1111/j.1365-2559.2010.03696.x. URL <https://www.scopus.com/inward/record.uri?eid=2-s2.0-78851470141&doi=10.1111%2Fj.1365-2559.2010.03696.x&partnerID=40&md5=7fb25aeb6c6dadfeb0d487e9ed2d6a79><http://onlinelibrary.wiley.com/doi/10.1111/j.1365-2559.2010.03696.x/abstract>.
- [25] D A Howell, A G Smith, and E Roman. Lymphoma: variations in time to diagnosis and treatment. *European Journal of Cancer Care*, 15(3):272–278, 7 2006. ISSN 0961-5423. doi: 10.1111/j.1365-2354.2006.00651.x. URL <https://doi.org/10.1111/j.1365-2354.2006.00651.x>.
- [26] D A Howell, A G Smith, and E Roman. Help-seeking behaviour in patients with lymphoma. *European Journal of Cancer Care*, 17(4):394–403, 7 2008. ISSN 0961-5423. doi: 10.1111/j.1365-2354.2007.00897.x. URL <https://doi.org/10.1111/j.1365-2354.2007.00897.x>.

- [27] S H Swerdlow. *WHO Classification of Tumours of Haematopoietic and Lymphoid Tissues; WHO Classification of Tumours, Volume 2*. International Agency for Research on Cancer, World Health Organisation, revised 4t edition, 2008.
- [28] Adalberto Miranda-Filho, Marion Piñeros, Ariana Znaor, Rafael Marcos-Gragera, Eva Steliarova-Foucher, and Freddie Bray. Global patterns and trends in the incidence of non-Hodgkin lymphoma. *Cancer Causes & Control*, 30(5):489–499, 2019. ISSN 1573-7225. doi: 10.1007/s10552-019-01155-5. URL <https://doi.org/10.1007/s10552-019-01155-5>.
- [29] Antonia M S Müller, Gabriele Ihorst, Roland Mertelsmann, and Monika Engelhardt. Epidemiology of non-Hodgkin’s lymphoma (NHL): trends, geographic distribution, and etiology. *Annals of Hematology*, 84(1):1–12, 2005. ISSN 1432-0584. doi: 10.1007/s00277-004-0939-7. URL <https://doi.org/10.1007/s00277-004-0939-7>.
- [30] Cancer Research UK. Cancer Research UK: non-Hodgkin lymphoma statistics, 2017. URL <https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/non-hodgkin-lymphoma>.
- [31] Office for National Statistics. Open Geography Portal, 2019. URL <https://geoportal.statistics.gov.uk>.
- [32] Mary C White, Dawn M Holman, Jennifer E Boehm, Lucy A Peipins, Melissa Grossman, and S Jane Henley. Age and cancer risk: a potentially modifiable relationship. *American journal of preventive medicine*, 46(3 Suppl 1):S7–S15, 3 2014. ISSN 1873-2607. doi: 10.1016/j.amepre.2013.10.029. URL <https://pubmed.ncbi.nlm.nih.gov/24512933https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4544764/>.
- [33] Nicholas Pavlidis, Giorgio Stanta, and Riccardo A Audisio. Cancer prevalence and mortality in centenarians: A systematic review. *Critical Reviews in Oncology/Hematology*, 83(1):145–152, 2012. ISSN 1040-8428. doi: <https://doi.org/10.1016/j.critrevonc.2011.09.007>. URL <http://www.sciencedirect.com/science/article/pii/S1040842811002307>.
- [34] Cristian Tomasetti, Lu Li, and Bert Vogelstein. Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention. *Science*, 355(6331):1330 LP – 1334, 3 2017. doi: 10.1126/science.aaf9011. URL <http://science.sciencemag.org/content/355/6331/1330.abstract>.
- [35] Sam Palmer, Luca Albergante, Clare C Blackburn, and T J Newman. Thymic involution and rising disease incidence with age. *Proceedings of the National Academy of Sciences*, 115(8):1883 LP – 1888, 2 2018. doi: 10.1073/pnas.1714478115. URL <http://www.pnas.org/content/115/8/1883.abstract>.
- [36] Vera Gorbunova, Andrei Seluanov, Zhiyong Mao, and Christopher Hine. Changes in DNA repair during aging. *Nucleic acids research*, 35(22):7466–7474, 2007. ISSN

- 1362-4962. doi: 10.1093/nar/gkm756. URL <https://pubmed.ncbi.nlm.nih.gov/17913742https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2190694/>.
- [37] Francis Rodier, Judith Campisi, and Dipa Bhaumik. Two faces of p53: aging and tumor suppression. *Nucleic acids research*, 35(22):7475–7484, 2007. ISSN 1362-4962. doi: 10.1093/nar/gkm744. URL <https://pubmed.ncbi.nlm.nih.gov/17942417https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2190721/>.
- [38] Nurit Horesh and Netanel A Horowitz. Does gender matter in non-hodgkin lymphoma? Differences in epidemiology, clinical behavior, and therapy. *Rambam Maimonides medical journal*, 5(4):e0038–e0038, 2014. doi: 10.5041/RMMJ.10172. URL <https://www.ncbi.nlm.nih.gov/pubmed/25386354https://www.ncbi.nlm.nih.gov/pmc/PMC4222427/https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4222427/pdf/rmmj-5-4-e0038.pdf>.
- [39] Randy D Gascoyne, Bertrand Nadel, Laura Pasqualucci, Jude Fitzgibbon, Jacqueline E Payton, Ari Melnick, Oliver Weigert, Karin Tarte, John G Gribben, Jonathan W Friedberg, John F Seymour, Franco Cavalli, and Emanuele Zucca. Follicular lymphoma: State-of-the-art ICML workshop in Lugano 2015. *Hematological Oncology*, 35(4):397–407, 12 2017. ISSN 0278-0232. doi: 10.1002/hon.2411. URL <https://doi.org/10.1002/hon.2411>.
- [40] Ryan D Morin, Maria Mendez-Lago, Andrew J Mungall, Rodrigo Goya, and Karen L Mungall. Frequent mutation of histone-modifying genes in non-Hodgkin lymphoma. *Nature*, 476(7360):298–303, 2011. ISSN 1476-4687. doi: 10.1038/nature10351. URL <https://doi.org/10.1038/nature10351>.
- [41] J M Schuetz, N A Johnson, R D Morin, D W Scott, K Tan, S Ben-Nierah, M Boyle, G W Slack, M A Marra, J M Connors, A R Brooks-Wilson, and R D Gascoyne. BCL2 mutations in diffuse large B-cell lymphoma. *Leukemia*, 26(6):1383–1390, 2012. ISSN 1476-5551. doi: 10.1038/leu.2011.378. URL <https://doi.org/10.1038/leu.2011.378>.
- [42] José P Vaqué, Nerea Martínez, Ana Batlle-López, Cristina Pérez, Santiago Montes-Moreno, Margarita Sánchez-Beato, and Miguel A Piris. B-cell lymphoma mutations: improving diagnostics and enabling targeted therapies. *Haematologica*, 99(2):222–231, 2 2014. ISSN 1592-8721. doi: 10.3324/haematol.2013.096248. URL <https://pubmed.ncbi.nlm.nih.gov/24497559https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3912951/>.
- [43] James R Cerhan and Susan L Slager. Familial predisposition and genetic risk factors for lymphoma. *Blood*, 126(20):2265–2273, 11 2015. ISSN 1528-0020. doi: 10.1182/blood-2015-04-537498. URL <https://pubmed.ncbi.nlm.nih.gov/26405224https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4643002/>.

- [44] Bryan A Bassig, Qing Lan, Nathaniel Rothman, Yawei Zhang, and Tongzhang Zheng. Current understanding of lifestyle and environmental factors and risk of non-hodgkin lymphoma: an epidemiological update. *Journal of cancer epidemiology*, 2012:978930, 2012. ISSN 1687-8566. doi: 10.1155/2012/978930. URL <https://pubmed.ncbi.nlm.nih.gov/23008714><https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3447374/>.
- [45] Lindsay M Morton, Patricia Hartge, Theodore R Holford, Elizabeth A Holly, and Tongzhang Zheng. Cigarette Smoking and Risk of Non-Hodgkin Lymphoma: A Pooled Analysis from the International Lymphoma Epidemiology Consortium (InterLymph). *Cancer Epidemiology Biomarkers & Prevention*, 14(4):925 LP – 933, 4 2005. doi: 10.1158/1055-9965.EPI-04-0693. URL <http://cebp.aacrjournals.org/content/14/4/925.abstract>.
- [46] Nathaniel C Briggs, Robert S Levine, Linda D Bobo, William P Haliburton, Edward A Brann, and Charles H Hennekens. Wine Drinking and Risk of Non-Hodgkin’s Lymphoma among Men in the United States: A Population-based Case-Control Study. *American Journal of Epidemiology*, 156(5):454–462, 9 2002. ISSN 0002-9262. doi: 10.1093/aje/kwf058. URL <https://doi.org/10.1093/aje/kwf058>.
- [47] Susanna C Larsson and Alicja Wolk. Obesity and risk of non-Hodgkin’s lymphoma: A meta-analysis. *International Journal of Cancer*, 121(7):1564–1570, 10 2007. ISSN 0020-7136. doi: 10.1002/ijc.22762. URL <https://doi.org/10.1002/ijc.22762>.
- [48] Susanna C Larsson and Alicja Wolk. Body mass index and risk of non-Hodgkin’s and Hodgkin’s lymphoma: A meta-analysis of prospective studies. *European Journal of Cancer*, 47(16):2422–2430, 11 2011. ISSN 0959-8049. doi: 10.1016/j.ejca.2011.06.029. URL <https://doi.org/10.1016/j.ejca.2011.06.029>.
- [49] Yawei Zhang, Silvia De Sanjose, Paige M Bracci, Lindsay M Morton, and Tongzhang Zheng. Personal use of hair dye and the risk of certain subtypes of non-Hodgkin lymphoma. *American journal of epidemiology*, 167(11):1321–1331, 6 2008. ISSN 1476-6256. doi: 10.1093/aje/kwn058. URL <https://pubmed.ncbi.nlm.nih.gov/18408225><https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4025953/>.
- [50] Silvia de Sanjosé, Yolanda Benavente, Alexandra Nieters, Lenka Foretova, Marc Maynadié, Pier Luigi Cocco, Anthony Staines, Martine Vornanen, Paolo Boffetta, Nikolaus Becker, Tomas Alvaro, and Paul Brennan. Association between Personal Use of Hair Dyes and Lymphoid Neoplasms in Europe. *American Journal of Epidemiology*, 164(1):47–55, 5 2006. ISSN 0002-9262. doi: 10.1093/aje/kwj187. URL <https://doi.org/10.1093/aje/kwj187>.
- [51] Francine Grodstein, Charles H Hennekens, Graham A Colditz, David J Hunter, and Meir J Stampfer. A Prospective Study of Permanent Hair Dye Use and Hematopoietic Cancer. *JNCI: Journal of the National Cancer Institute*, 86(19):1466–1470, 10 1994.

ISSN 0027-8874. doi: 10.1093/jnci/86.19.1466. URL <https://doi.org/10.1093/jnci/86.19.1466>.

- [52] G Bentham. Association between incidence of non-Hodgkin's lymphoma and solar ultraviolet radiation in England and Wales. *BMJ (Clinical research ed.)*, 312(7039):1128–1131, 5 1996. ISSN 0959-8138. doi: 10.1136/bmj.312.7039.1128. URL <https://pubmed.ncbi.nlm.nih.gov/8620128><https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2350672/>.
- [53] Patricia Hartge, Susan S Devesa, Dan Grauman, Thomas R Fears, and Joseph F Fraumeni Jr. Non-Hodgkin's Lymphoma and Sunlight. *JNCI: Journal of the National Cancer Institute*, 88(5):298–300, 3 1996. ISSN 0027-8874. doi: 10.1093/jnci/88.5.298. URL <https://doi.org/10.1093/jnci/88.5.298>.
- [54] Jia Wang, Xutong Li, and Dongfeng Zhang. Dairy Product Consumption and Risk of Non-Hodgkin Lymphoma: A Meta-Analysis. *Nutrients*, 8(3):120, 2 2016. ISSN 2072-6643. doi: 10.3390/nu8030120. URL <https://www.ncbi.nlm.nih.gov/pubmed/26927171><https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4808850/>.
- [55] Christine F Skibola. Obesity, diet and risk of non-Hodgkin lymphoma. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology*, 16(3):392–395, 3 2007. ISSN 1055-9965. doi: 10.1158/1055-9965.EPI-06-1081. URL <https://pubmed.ncbi.nlm.nih.gov/17337642><https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2819159/>.
- [56] Paolo Boffetta and Frank de Vocht. Occupation and the Risk of Non-Hodgkin Lymphoma. *Cancer Epidemiology Biomarkers & Prevention*, 16(3):369 LP – 372, 3 2007. doi: 10.1158/1055-9965.EPI-06-1055. URL <http://cebp.aacrjournals.org/content/16/3/369.abstract>.
- [57] P Cocco, A T'Mannetje, D Fadda, M Melis, N Becker, S de Sanjosé, L Foretova, J Mareckova, A Staines, S Kleefeld, M Maynadié, A Nieters, P Brennan, and P Boffetta. Occupational exposure to solvents and risk of lymphoma subtypes: results from the Epilymph case-control study. *Occupational and Environmental Medicine*, 67(5):341 LP – 347, 5 2010. doi: 10.1136/oem.2009.046839. URL <http://oem.bmj.com/content/67/5/341.abstract>.
- [58] Dominik D Alexander and Meghan E Wagner. Benzene Exposure and Non-Hodgkin Lymphoma: A Meta-Analysis of Epidemiologic Studies. *Journal of Occupational and Environmental Medicine*, 52(2), 2010. ISSN 1076-2752. URL https://journals.lww.com/joem/Fulltext/2010/02000/Benzene_Exposure_and_Non_Hodgkin_Lymphoma__A.9.aspx.
- [59] John J Spinelli, Carmen H Ng, Jean-Philippe Weber, Joseph M Connors, Randy D Gascoyne, Agnes S Lai, Angela R Brooks-Wilson, Nhu D Le, Brian R Berry, and

- Richard P Gallagher. Organochlorines and risk of non-Hodgkin lymphoma. *International Journal of Cancer*, 121(12):2767–2775, 12 2007. ISSN 0020-7136. doi: 10.1002/ijc.23005. URL <https://doi.org/10.1002/ijc.23005>.
- [60] Lawrence S Engel, Francine Laden, Aage Andersen, Paul T Strickland, and Nathaniel Rothman. Polychlorinated Biphenyl Levels in Peripheral Blood and Non-Hodgkins Lymphoma: A Report from Three Cohorts. *Cancer Research*, 67(11):5545 LP – 5552, 6 2007. doi: 10.1158/0008-5472.CAN-06-3906. URL <http://cancerres.aacrjournals.org/content/67/11/5545.abstract>.
- [61] NHS Digital. Non-Hodgkin Lymphoma: Causes, 2019. URL <https://www.nhs.uk/conditions/non-hodgkin-lymphoma/causes/>.
- [62] American Cancer Society. Non-Hodgkin Lymphoma Risk Factors, 2019. URL <https://www.cancer.org/cancer/non-hodgkin-lymphoma/causes-risks-prevention/risk-factors.html>.
- [63] Alina Klein, Aaron Polliack, and Anat Gafter-Gvili. Rheumatoid arthritis and lymphoma: Incidence, pathogenesis, biology, and outcome. *Hematological Oncology*, 36(5):733–739, 12 2018. ISSN 0278-0232. doi: 10.1002/hon.2525. URL <https://doi.org/10.1002/hon.2525>.
- [64] Sasha Bernatsky, Rosalind Ramsey-Goldman, Jeremy Labrecque, Lawrence Joseph, and Ann E Clarke. Cancer risk in systemic lupus: an updated international multi-centre cohort study. *Journal of autoimmunity*, 42:130–135, 5 2013. ISSN 1095-9157. doi: 10.1016/j.jaut.2012.12.009. URL <https://www.ncbi.nlm.nih.gov/pubmed/23410586https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3646904/>.
- [65] Elias Zintzaras, Michael Voulgarelis, and Haralampos M Moutsopoulos. The Risk of Lymphoma Development in Autoimmune Diseases: A Meta-analysis. *JAMA Internal Medicine*, 165(20):2337–2344, 11 2005. ISSN 2168-6106. doi: 10.1001/archinte.165.20.2337. URL <https://doi.org/10.1001/archinte.165.20.2337>.
- [66] David Aboulafia. Non-Hodgkin lymphoma in people with HIV. *The Lancet HIV*, 6, 2 2019. doi: 10.1016/S2352-3018(19)30039-6.
- [67] G Brady, G J MacArthur, and P J Farrell. Epstein-Barr virus and Burkitt lymphoma. *Journal of clinical pathology*, 60(12):1397–1402, 12 2007. ISSN 1472-4146. doi: 10.1136/jcp.2007.047977. URL <https://pubmed.ncbi.nlm.nih.gov/18042696https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2095571/>.
- [68] Jackson Orem, Edward Katongole Mbidde, Bo Lambert, Silvia de Sanjose, and Elisabete Weiderpass. Burkitt’s lymphoma in Africa, a review of the epidemiology and etiology. *African health sciences*, 7(3):166–175, 9 2007. ISSN 1729-0503. doi: 10.5555/afhs.2007.7.3.166. URL <https://www.ncbi.nlm.nih.gov/pubmed/18052871https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2269718/>.

- [69] B J Poiesz, F W Ruscetti, A F Gazdar, P A Bunn, J D Minna, and R C Gallo. Detection and isolation of type C retrovirus particles from fresh and cultured lymphocytes of a patient with cutaneous T-cell lymphoma. *Proceedings of the National Academy of Sciences of the United States of America*, 77(12):7415–7419, 12 1980. ISSN 0027-8424. doi: 10.1073/pnas.77.12.7415. URL <https://pubmed.ncbi.nlm.nih.gov/6261256><https://www.ncbi.nlm.nih.gov/pmc/articles/PMC350514/>.
- [70] A C Wotherspoon. Gastric MALT lymphoma and Helicobacter pylori. *The Yale journal of biology and medicine*, 69(1):61–68, 1996. ISSN 0044-0086. URL <https://pubmed.ncbi.nlm.nih.gov/9041690><https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2588981/>.
- [71] Sonja E van Roeden, Mirjam H A Hermans, Peet T G A Nooijen, Alexandra Herbers, Chantal P Bleeker-Rovers, Andy I M Hoepelman, Jan Jelrik Oosterheert, and Peter C Wever. Coxiella burnetii in non-Hodgkin lymphoma tissue samples: Innocent until proven otherwise? *Immunobiology*, 224(2):254–261, 2019. ISSN 0171-2985. doi: <https://doi.org/10.1016/j.imbio.2018.11.012>. URL <http://www.sciencedirect.com/science/article/pii/S017129851830189X>.
- [72] National Guideline Alliance (UK). *Non-Hodgkin’s Lymphoma: Diagnosis and Management*. National Institute for Health and Care Excellence, London, UK, no. 52 edition, 2016. URL <https://www.ncbi.nlm.nih.gov/books/NBK374283/>.
- [73] Jonathan W Friedberg, Michelle Byrtek, Brian K Link, Christopher Flowers, Michael Taylor, John Hainsworth, James R Cerhan, Andrew D Zelenetz, Jamie Hirata, and Thomas P Miller. Effectiveness of first-line management strategies for stage I follicular lymphoma: analysis of the National LymphoCare Study. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 30(27):3368–3375, 9 2012. ISSN 1527-7755. doi: 10.1200/JCO.2011.40.6546. URL <https://pubmed.ncbi.nlm.nih.gov/22915662><https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3675665/>.
- [74] National Institute for Health and Care Excellence. Non-Hodgkin lymphoma: diagnosis and management, 2016.
- [75] Christopher McNamara, Silvia Montoto, Toby A Eyre, Kirit Ardeshta, Cathy Burton, Tim Illidge, Kim Linton, Simon Rule, William Townsend, Wai L Wong, and Pam McKay. The investigation and management of follicular lymphoma. *British Journal of Haematology*, 191(3):363–381, 11 2020. ISSN 0007-1048. doi: <https://doi.org/10.1111/bjh.16872>. URL <https://doi.org/10.1111/bjh.16872>.
- [76] Robert Marcus, Andrew Davies, Kiyoshi Ando, Wolfram Klapper, Stephen Opat, Carolyn Owen, Elizabeth Phillips, Randeep Sangha, Rudolf Schlag, John F Seymour, William Townsend, Marek Trněný, Michael Wenger, Günter Fingerle-Rowson,

- Kaspar Rufibach, Tom Moore, Michael Herold, and Wolfgang Hiddemann. Obinutuzumab for the First-Line Treatment of Follicular Lymphoma. *New England Journal of Medicine*, 377(14):1331–1344, 10 2017. ISSN 0028-4793. doi: 10.1056/NEJMoa1614598. URL <https://doi.org/10.1056/NEJMoa1614598>.
- [77] H Tilly, M Gomes da Silva, U Vitolo, A Jack, M Meignan, A Lopez-Guillermo, J Walewski, M André, P W Johnson, M Pfreundschuh, and M Ladetto. Diffuse large B-cell lymphoma (DLBCL): ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Annals of Oncology*, 26(suppl 5):v116–v125, 2015. doi: 10.1093/annonc/mdv304. URL http://annonc.oxfordjournals.org/content/26/suppl_5/v116.short
http://annonc.oxfordjournals.org/content/26/suppl_5/v116.full.pdf
http://watermark.silverchair.com/mdv304.pdf?token=AQECAHi208BE490oan9kKhW_Ercy7Dm3ZL_9Cf3qfKAc485ysgAAAd8wggHbBgkqhkiG9w0BBwa.
- [78] Pfreundschuh M, Kuhnt E, and Trumper L. CHOP-like chemotherapy with or without rituximab in young patients with good-prognosis diffuse large-B-cell lymphoma: 6-year results of an open-label randomised study of the MabThera International Trial (MInT) Group. *Lancet Oncology*, 12:1013–1022, 2011.
- [79] Rita Coutinho, Alessia D Pria, and Shreyans Gandhi. HIV status does not impair the outcome of patients diagnosed with diffuse large B-cell lymphoma treated with R-CHOP in the cART era. *AIDS*, 28(5), 2014. ISSN 0269-9370. URL https://journals.lww.com/aidsonline/Fulltext/2014/03130/HIV_status_does_not_impair_the_outcome_of_patients.6.aspx.
- [80] Carla Casulo, W Richard Burack, and Jonathan W Friedberg. Transformed follicular non-Hodgkin lymphoma. *Blood*, 125(1):40–47, 1 2015. ISSN 0006-4971. doi: 10.1182/blood-2014-04-516815. URL <https://doi.org/10.1182/blood-2014-04-516815>.
- [81] National Institute for Health and Care Excellence. NICE recommends another revolutionary CAR T-cell therapy for adults with lymphoma, 2019. URL <https://www.nice.org.uk/news/article/nice-recommends-another-revolutionary-car-t-cell-therapy-for-adults-with-lymphoma>.
- [82] NHS Digital. CAR-T Therapy, 2019. URL <https://www.england.nhs.uk/cancer/cdf/car-t-therapy/>.
- [83] L Elliss-Brookes, S McPhail, A Ives, M Greenslade, J Shelton, S Hiom, and M Richards. Routes to diagnosis for cancer - determining the patient journey using multiple routine data sets. *Br J Cancer*, 107(8):1220–1226, 2012. doi: 10.1038/bjc.2012.408. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3494426/pdf/bjc2012408a.pdf>.

- [84] American Cancer Society. Can non-Hodgkin lymphoma be found early?, 2019. URL <https://www.cancer.org/cancer/non-hodgkin-lymphoma/detection-diagnosis-staging/detection.html>.
- [85] Cancer Research UK. Non-Hodgkin Lymphoma: Screening, 2019. URL <https://www.cancerresearchuk.org/about-cancer/non-hodgkin-lymphoma/getting-diagnosed/screening>.
- [86] The International Non-Hodgkin’s Lymphoma Prognostic Factors Project. A predictive model for aggressive non-Hodgkin’s lymphoma. *N Engl J Med*, 329(14):987–994, 1993. doi: 10.1056/nejm199309303291402.
- [87] Haematological Malignancy Research Network. Survival of non-Hodgkin lymphoma, 2016. URL <https://www.hmrn.org/statistics/survival>.
- [88] Saskia A M van de Schans, Liza N van Steenberghe, Jan Willem W Coebergh, Maryska L G Janssen-Heijnen, and Dick Johan van Spronsen. Actual prognosis during follow-up of survivors of B-cell non-Hodgkin lymphoma in the Netherlands. *Haematologica*, 99(2):339–345, 2 2014. ISSN 1592-8721. doi: 10.3324/haematol.2012.081885. URL <https://pubmed.ncbi.nlm.nih.gov/24038025https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3912965/>.
- [89] Andrew G Glass, Lucy H Karnell, and Herman R Menck. The national cancer data base report on non-hodgkin’s lymphoma. *Cancer*, 80(12):2311–2320, 12 1997. ISSN 0008-543X. doi: 10.1002/(SICI)1097-0142(19971215)80:12<2311::AID-CNCR13>3.0.CO;2-X. URL [https://doi.org/10.1002/\(SICI\)1097-0142\(19971215\)80:12%3C2311::AID-CNCR13%3E3.0.COhttp://2-x](https://doi.org/10.1002/(SICI)1097-0142(19971215)80:12%3C2311::AID-CNCR13%3E3.0.COhttp://2-x).
- [90] Thomas A Ollila and Adam J Olszewski. Extranodal Diffuse Large B Cell Lymphoma: Molecular Features, Prognosis, and Risk of Central Nervous System Recurrence. *Current treatment options in oncology*, 19(8):38, 6 2018. ISSN 1534-6277. doi: 10.1007/s11864-018-0555-8. URL <https://pubmed.ncbi.nlm.nih.gov/29931605https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6294323/>.
- [91] Stanley Ostrow, Charles H Diggs, John Sutherland, and Peter H Wiernik. Causes of death in patients with non-Hodgkin’s LYMPHOMA. *Cancer*, 48(3):779–782, 8 1981. ISSN 0008-543X. doi: 10.1002/1097-0142(19810801)48:3<779::AID-CNCR2820480320>3.0.CO;2-3. URL [https://doi.org/10.1002/1097-0142\(19810801\)48:3%3C779::AID-CNCR2820480320%3E3.0.COhttp://2-3](https://doi.org/10.1002/1097-0142(19810801)48:3%3C779::AID-CNCR2820480320%3E3.0.COhttp://2-3).
- [92] James O Armitage. Staging Non-Hodgkin Lymphoma. *CA: A Cancer Journal for Clinicians*, 55(6):368–376, 11 2005. ISSN 0007-9235. doi: 10.3322/canjclin.55.6.368. URL <https://doi.org/10.3322/canjclin.55.6.368>.
- [93] Sidney Weinhouse, Otto Warburg, Dean Burk, and Arthur L Schade. On Respiratory Impairment in Cancer Cells. *Science*, 124(3215):267 LP – 272, 8 1956. doi: 10.1126/

science.124.3215.267. URL <http://science.sciencemag.org/content/124/3215/267.abstract>.

- [94] Charles M Balch, Seng-Jaw Soong, Michael B Atkins, Antonio C Buzaid, Natale Cascinelli, Daniel G Coit, Irvin D Fleming, Jeffrey E Gershenwald, Alan Houghton Jr., John M Kirkwood, Kelly M McMasters, Martin F Mihm, Donald L Morton, Douglas S Reintgen, Merrick I Ross, Arthur Sober, John A Thompson, and John F Thompson. An Evidence-based Staging System for Cutaneous Melanoma. *CA: A Cancer Journal for Clinicians*, 54(3):131–149, 5 2004. ISSN 0007-9235. doi: 10.3322/canjclin.54.3.131. URL <https://doi.org/10.3322/canjclin.54.3.131>.
- [95] LaMont J Barlow, Gina M Badalato, and James M McKiernan. Serum tumor markers in the evaluation of male germ cell tumors. *Nature Reviews Urology*, 7(11):610–617, 2010. ISSN 1759-4820. doi: 10.1038/nrurol.2010.166. URL <https://doi.org/10.1038/nrurol.2010.166>.
- [96] Andrew J Armstrong, Daniel J George, and Susan Halabi. Serum Lactate Dehydrogenase Predicts for Overall Survival Benefit in Patients With Metastatic Renal Cell Carcinoma Treated With Inhibition of Mammalian Target of Rapamycin. *Journal of Clinical Oncology*, 30(27):3402–3407, 8 2012. ISSN 0732-183X. doi: 10.1200/JCO.2011.40.9631. URL <https://doi.org/10.1200/JCO.2011.40.9631>.
- [97] Sarah J Nagle, Kaitlin Woo, Stephen J Schuster, Sunita D Nasta, Edward Stadtmauer, Rosemarie Mick, and Jakub Svoboda. Outcomes of patients with relapsed/refractory diffuse large B-cell lymphoma with progression of lymphoma after autologous stem cell transplantation in the rituximab era. *American Journal of Hematology*, 88(10):890–894, 10 2013. ISSN 0361-8609. doi: 10.1002/ajh.23524. URL <https://doi.org/10.1002/ajh.23524>.
- [98] Wahyu Wulaningsih, Lars Holmberg, Hans Garmo, Håkan Malmstrom, Mats Lambe, Niklas Hammar, Göran Walldius, Ingmar Jungner, Tony Ng, and Mieke Van Hemelrijck. Serum lactate dehydrogenase and survival following cancer diagnosis. *British Journal of Cancer*, 113(9):1389–1396, 2015. ISSN 1532-1827. doi: 10.1038/bjc.2015.361. URL <https://doi.org/10.1038/bjc.2015.361>.
- [99] Luigi Endrizzi, Mario V Fiorentino, Luigi Salvagno, Romana Segati, Giovanni L Pappagallo, and Vinicio Fossier. Serum lactate dehydrogenase (LDH) as a prognostic index for non-Hodgkin’s lymphoma. *European Journal of Cancer and Clinical Oncology*, 18(10):945–949, 1982. ISSN 0277-5379. doi: [https://doi.org/10.1016/0277-5379\(82\)90242-5](https://doi.org/10.1016/0277-5379(82)90242-5). URL <http://www.sciencedirect.com/science/article/pii/0277537982902425>.
- [100] Basem Magdy William, Navneeth Rao Bongu, Martin Bast, Robert Gregory Bociek, Philip Jay Bierman, Julie Marie Vose, and James Olen Armitage. The utility of lactate dehydrogenase in the follow up of patients with diffuse large B-cell lymphoma.

- Revista brasileira de hematologia e hemoterapia*, 35(3):189–191, 2013. ISSN 1516-8484. doi: 10.5581/1516-8484.20130055. URL <https://pubmed.ncbi.nlm.nih.gov/23904809https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3728132/>.
- [101] David P Steensma and Thomas E Witzig. Elevated serum LDH in patients with non-Hodgkin’s lymphoma: not always an ominous sign. *British Journal of Haematology*, 107(2):463–464, 11 1999. ISSN 0007-1048. doi: 10.1046/j.1365-2141.1999.01786.x. URL <https://doi.org/10.1046/j.1365-2141.1999.01786.x>.
- [102] Ulrike Paul, Julia Richter, Christiane Stuhlmann-Laiesz, Markus Kreuz, and Wolfram Klapper. Advanced patient age at diagnosis of diffuse large B-cell lymphoma is associated with molecular characteristics including ABC-subtype and high expression of MYC. *Leukemia & Lymphoma*, 59(5):1213–1221, 5 2018. ISSN 1042-8194. doi: 10.1080/10428194.2017.1365851. URL <https://doi.org/10.1080/10428194.2017.1365851>.
- [103] Mary L Varterasian, John J Graff, Richard K Severson, Linda Weiss, Ayad M Aikatib, and Gregory P Kalemkerian. Non-Hodgkin’s Lymphoma: An analysis of the Metropolitan Detroit SEER Database. *Cancer Investigation*, 18(4):303–308, 1 2000. ISSN 0735-7907. doi: 10.3109/07357900009012172. URL <https://doi.org/10.3109/07357900009012172>.
- [104] Martin M Oken, Richard H Creech, Douglass C Tormey, John Horton, Thomas E Davis, Eleanor T McFadden, and Paul P Carbone. Toxicity and response criteria of the Eastern Cooperative Oncology Group. *American Journal of Clinical Oncology*, 5(6), 1982. ISSN 0277-3732. URL https://journals.lww.com/amjclinicaloncology/Fulltext/1982/12000/Toxicity_and_response_criteria_of_the_Eastern.14.aspx.
- [105] S P Blagden, S C Charman, L D Sharples, L R A Magee, and D Gilligan. Performance status score: do patients and their oncologists agree? *British journal of cancer*, 89(6):1022–1027, 9 2003. ISSN 0007-0920. doi: 10.1038/sj.bjc.6601231. URL <https://pubmed.ncbi.nlm.nih.gov/12966419https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2376959/>.
- [106] G Buccheri, D Ferrigno, and M Tamburini. Karnofsky and ECOG performance status scoring in lung cancer: A prospective, longitudinal study of 536 patients from a single institution. *European Journal of Cancer*, 32(7):1135–1141, 6 1996. ISSN 0959-8049. doi: 10.1016/0959-8049(95)00664-8. URL [https://doi.org/10.1016/0959-8049\(95\)00664-8](https://doi.org/10.1016/0959-8049(95)00664-8).
- [107] M L Janssen-Heijnen, D J van Spronsen, V E Lemmens, S Houterman, K D Verheij, and J W Coebergh. A population-based study of severity of comorbidity among patients with non-Hodgkin’s lymphoma: prognostic impact independent of Inter-

- national Prognostic Index. *Br J Haematol*, 129(5):597–606, 2005. doi: 10.1111/j.1365-2141.2005.05508.x. URL <https://www.ncbi.nlm.nih.gov/pubmed/15916681>.
- [108] D J van Spronsen, M L Janssen-Heijnen, W P Breed, and J W Coebergh. Prevalence of co-morbidity and its relationship to treatment among unselected patients with Hodgkin’s disease and non-Hodgkin’s lymphoma, 1993-1996. *Ann Hematol*, 78(7):315–319, 1999.
- [109] Laura Hester, Steven I Park, and Jennifer Leigh Lund. Patterns of comorbidity among older U.S. patients with non-Hodgkin lymphoma. *Journal of Clinical Oncology*, 34(7_suppl):304, 3 2016. ISSN 0732-183X. doi: 10.1200/jco.2016.34.7{_suppl.304. URL https://doi.org/10.1200/jco.2016.34.7_suppl.304.
- [110] Joanna Mitri, Jorge Castillo, and Anastassios G Pittas. Diabetes and risk of Non-Hodgkin’s lymphoma: a meta-analysis of observational studies. *Diabetes care*, 31(12):2391–2397, 12 2008. ISSN 1935-5548. doi: 10.2337/dc08-1034. URL <https://pubmed.ncbi.nlm.nih.gov/19033419https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2584201/>.
- [111] Jorge J Castillo, Nikhil Mull, John L Reagan, Saed Nemr, and Joanna Mitri. Increased incidence of non-Hodgkin lymphoma, leukemia, and myeloma in patients with diabetes mellitus type 2: a meta-analysis of observational studies. *Blood*, 119(21):4845–4850, 5 2012. ISSN 1528-0020. doi: 10.1182/blood-2011-06-362830. URL <https://pubmed.ncbi.nlm.nih.gov/22496152https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3367891/>.
- [112] Aneire E Khan, Valentina Gallo, Jakob Linseisen, Rudolf Kaaks, and Elio Riboli. Diabetes and the risk of non-Hodgkin’s lymphoma and multiple myeloma in the European Prospective Investigation into Cancer and Nutrition. *Haematologica*, 93(6):842 LP – 850, 6 2008. doi: 10.3324/haematol.12297. URL <http://www.haematologica.org/content/93/6/842.abstract>.
- [113] Jennifer S Lee, Paige M Bracci, and Elizabeth A Holly. Non-Hodgkin lymphoma in women: reproductive factors and exogenous hormone use. *American journal of epidemiology*, 168(3):278–288, 8 2008. ISSN 1476-6256. doi: 10.1093/aje/kwn119. URL <https://pubmed.ncbi.nlm.nih.gov/18550561https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2727261/>.
- [114] Sverker Hasselblom, Börje Ridell, Herman Nilsson-Ehle, and Per-Ola Andersson. The impact of gender, age and patient selection on prognosis and outcome in diffuse large B-cell lymphoma—a population-based study. *Leukemia & Lymphoma*, 48(4):736–745, 1 2007. ISSN 1042-8194. doi: 10.1080/10428190601187703. URL <https://doi.org/10.1080/10428190601187703>.
- [115] Sari Riihijärvi, Minna Taskinen, Mats Jerkeman, and Sirpa Leppä. Male gender is an adverse prognostic factor in B-cell lymphoma patients treated with im-

- munochemotherapy*. *European Journal of Haematology*, 86(2):124–128, 2 2011. ISSN 0902-4441. doi: 10.1111/j.1600-0609.2010.01541.x. URL <https://doi.org/10.1111/j.1600-0609.2010.01541.x>.
- [116] Michael Pfreundschuh, Niels Murawski, Samira Zeynalova, Viola Poeschel, and Norbert Schmitz. Male Sex Is Associated with Lower Rituximab Trough Serum Levels and Evolves as a Significant Prognostic Factor in Elderly Patients with DLBCL Treated with R-CHOP: Results From 4 Prospective Trials of the German High-Grade Non-Hodgkin-Lymphoma Study Group. *Blood*, 114(22):3715, 11 2009. ISSN 0006-4971. doi: 10.1182/blood.V114.22.3715.3715. URL <https://doi.org/10.1182/blood.V114.22.3715.3715>.
- [117] World Health Organisation. International Agency for Research on Cancer, 2018. URL <http://gco.iarc.fr/today/home>.
- [118] C R Smittenaar, K A Petersen, K Stewart, and N Moitt. Cancer incidence and mortality projections in the UK until 2035. *British journal of cancer*, 115(9):1147–1155, 10 2016. ISSN 1532-1827. doi: 10.1038/bjc.2016.304. URL <https://www.ncbi.nlm.nih.gov/pubmed/27727232><https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5117795/>.
- [119] B Rachet, E Mitry, A Shah, N Cooper, and M P Coleman. Survival from non-Hodgkin lymphoma in England and Wales up to 2001. *British Journal of Cancer*, 99(Suppl 1):S104–S106, 2008. doi: 10.1038/sj.bjc.6604605. URL <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2557528/><http://www.nature.com/bjc/journal/v99/n1s/pdf/6604605a.pdf><https://www.nature.com/articles/6604605.pdf>.
- [120] National Institute for Health and Care Excellence. Non-Hodgkin’s lymphoma: diagnosis and management. Technical report, National Institute for Health and Care Excellence, 2016. URL <https://www.nice.org.uk/guidance/ng52/evidence/full-guideline-2551524594>.
- [121] Claudia Allemani, Tomohiro Matsuda, Veronica Di Carlo, Rhea Harewood, Melissa Matz, Maja Nikšić, and et al. Global surveillance of trends in cancer survival 2000–14 (CONCORD-3): analysis of individual records for 37,513,025 patients diagnosed with one of 18 cancers from 322 population-based registries in 71 countries. *The Lancet*, 391(10125):1023–1075, 2018. doi: 10.1016/s0140-6736(17)33326-3.
- [122] C S Thomson and D Forman. Cancer survival in England and the influence of early diagnosis: what can we learn from recent EURO CARE results? *British Journal of Cancer*, 101(2):S102–S109, 2009. ISSN 1532-1827. doi: 10.1038/sj.bjc.6605399. URL <https://doi.org/10.1038/sj.bjc.6605399>.
- [123] Department of Health. The NHS Plan: a plan for action, a plan for reform. Technical report, Department of Health, 2000. URL www.doh.gov.uk/nhsplan.

- [124] C E Round, M V Williams, T Mee, N F Kirkby, T Cooper, P Hoskin, and R Jena. Radiotherapy Demand and Activity in England 2006-2020. *Clinical Oncology*, 25(9): 522–530, 9 2013. ISSN 0936-6555. doi: 10.1016/j.clon.2013.05.005. URL <https://doi.org/10.1016/j.clon.2013.05.005>.
- [125] M V Williams and K J Drinkwater. Radiotherapy in England in 2007: modelled demand and audited activity. *Clin Oncol (R Coll Radiol)*, 21(8):575–590, 2009. doi: 10.1016/j.clon.2009.07.003.
- [126] Department of Health and Social Care. NHS Cancer Reform Strategy. Technical report, London, 2007. URL <https://www.nhs.uk/NHSEngland/NSF/Documents/CancerReformStrategy.pdf>.
- [127] National Cancer Equality Initiative. Cancer Equalities. Technical report, National Cancer Intelligence Network, 2015. URL http://www.ncin.org.uk/cancer_type_and_topic_specific_work/topic_specific_work/equality.
- [128] J Lynch, G Kaplan, L F Berkman, and I Kawachi. *Social Epidemiology*. Oxford University Press, Oxford, 1st edition, 2000. URL <https://books.google.co.uk/books?id=DgzVAwAAQBAJ&lr=>.
- [129] Bruna Galobardes, John Lynch, and George Davey Smith. Measuring socioeconomic position in health research. *British Medical Bulletin*, 81-82(1):21–37, 1 2007. ISSN 0007-1420. doi: 10.1093/bmb/ldm001. URL <https://doi.org/10.1093/bmb/ldm001>.
- [130] H V Z Tunstall, M Shaw, and D Dorling. Places and health. *Journal of epidemiology and community health*, 58(1):6–10, 1 2004. ISSN 0143-005X. doi: 10.1136/jech.58.1.6. URL <https://pubmed.ncbi.nlm.nih.gov/14684719https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1757034/>.
- [131] Vera Carstairs and Russell Morris. Deprivation and mortality: an alternative to social class? *Journal of Public Health*, 11(3):210–219, 8 1989. ISSN 1741-3842. doi: 10.1093/oxfordjournals.pubmed.a042469. URL <https://doi.org/10.1093/oxfordjournals.pubmed.a042469>.
- [132] Peter Townsend, Peter Phillimore, and Alastair Beattie. *Health and deprivation: inequality and the North*. Routledge, 1988. ISBN 0709943512.
- [133] Department of the Environment Transport and the Regions. Measuring multiple deprivation at the small area level: the indices of deprivation 2000. Technical report, Department of the Environment, Transport and the Regions, London, UK, 2000.
- [134] Neighbourhood Renewal Unit - Office for the Deputy Prime Minister. The English indices of deprivation 2004 (revised). Technical report, Neighbourhood Renewal Unit, Office for the Deputy Prime Minister, London, UK, 2004.

- [135] S V Subramanian, J T Chen, D H Rehkopf, P D Waterman, and N Krieger. Subramanian et al. Respond to “Think Conceptually, Act Cautiously”. *American Journal of Epidemiology*, 164(9):841–844, 11 2006. ISSN 0002-9262. doi: 10.1093/aje/kwj315. URL <https://doi.org/10.1093/aje/kwj315>.
- [136] Arline T Geronimus. Invited Commentary: Using Area-based Socioeconomic Measures—Think Conceptually, Act Cautiously. *American Journal of Epidemiology*, 164(9):835–840, 11 2006. ISSN 0002-9262. doi: 10.1093/aje/kwj314. URL <https://doi.org/10.1093/aje/kwj314>.
- [137] Fiona C Ingleby, Aurélien Belot, Iain Atherton, Matthew Baker, Lucy Elliss-Brookes, and Laura M Woods. Assessment of the concordance between individual-level and area-level measures of socio-economic deprivation in a cancer patient cohort in England and Wales. *BMJ Open*, 10(11):e041714, 11 2020. doi: 10.1136/bmjopen-2020-041714. URL <http://bmjopen.bmj.com/content/10/11/e041714.abstract>.
- [138] Gillian Smith. Area-based initiatives: the rationale and options for area targeting. *LSE STICERD Research Paper No. Case025*, 1999.
- [139] Mark Kleinman. There goes the neighbourhood: area policies and social exclusion. *New Economy*, 6(4):188–192, 1999. ISSN 1070-3535.
- [140] Laura Woods, B Rachet, and M Coleman. Choice of geographic unit influences socioeconomic inequalities in breast cancer survival. *British journal of cancer*, 92:1279–1282, 5 2005. doi: 10.1038/sj.bjc.6602506.
- [141] PIU. Reaching Out: The Role of Central Government at Local and Regional Level, 2000.
- [142] Michael Noble, Gemma Wright, George Smith, and Chris Dibben. Measuring multiple deprivation at the small-area level. *Environment and Planning A*, 38:169–185, 2 2006. doi: 10.1068/a37168.
- [143] Geeta Gandhi Kingdon and John Knight. Subjective well-being poverty vs. income poverty and capabilities poverty? *The Journal of Development Studies*, 42(7):1199–1224, 2006. ISSN 0022-0388.
- [144] S O Dalton, M Düring, L Ross, K Carlsen, P B Mortensen, J Lynch, and C Johansen. The relation between socioeconomic and demographic factors and tumour stage in women diagnosed with breast cancer in Denmark, 1983-1999. *British journal of cancer*, 95(5):653–659, 9 2006. ISSN 0007-0920. doi: 10.1038/sj.bjc.6603294. URL <https://pubmed.ncbi.nlm.nih.gov/16909141https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2360690/>.
- [145] C L Hart and G Davey Smith. Relation between number of siblings and adult mortality and stroke risk: 25 year follow up of men in the Collaborative study.

- Journal of Epidemiology and Community Health*, 57(5):385 LP – 391, 5 2003. doi: 10.1136/jech.57.5.385. URL <http://jech.bmj.com/content/57/5/385.abstract>.
- [146] Miquel Porta. *A Dictionary of Epidemiology*. Oxford University Press, 2014. ISBN 9780195314496. URL <http://www.oxfordreference.com/view/10.1093/acref/9780195314496.001.0001/acref-9780195314496>.
- [147] A R Feinstein. The pre-therapeutic classification of co-morbidity in chronic disease. *J Chronic Dis*, 23(7):455–468, 1970.
- [148] Diana Sarfati. Review of methods used to measure comorbidity in cancer populations: No gold standard exists. *Journal of Clinical Epidemiology*, 65(9):924–933, 9 2012. ISSN 0895-4356. doi: 10.1016/j.jclinepi.2012.02.017. URL <https://doi.org/10.1016/j.jclinepi.2012.02.017>.
- [149] Anne Elixhauser, Claudia Steiner, D Robert Harris, and Rosanna M Coffey. Comorbidity Measures for Use with Administrative Data. *Medical Care*, 36(1):8–27, 1998. URL https://journals.lww.com/lww-medicalcare/Fulltext/1998/01000/Comorbidity_Measures_for_Use_with_Administrative.4.aspx.
- [150] Bing Li, Dewey Evans, Peter Faris, Stafford Dean, and Hude Quan. Risk adjustment performance of Charlson and Elixhauser comorbidities in ICD-9 and ICD-10 administrative databases. *BMC health services research*, 8:12, 1 2008. ISSN 1472-6963. doi: 10.1186/1472-6963-8-12. URL <https://pubmed.ncbi.nlm.nih.gov/18194561https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2267188/>.
- [151] Juho Pylväläinen, Kirsi Talala, Teemu Murtola, Kimmo Taari, Jani Raitanen, Teuvo L Tammela, and Anssi Auvinen. Charlson Comorbidity Index Based On Hospital Episode Statistics Performs Adequately In Predicting Mortality, But Its Discriminative Ability Diminishes Over Time. *Clinical epidemiology*, 11:923–932, 10 2019. ISSN 1179-1349. doi: 10.2147/CLEP.S218697. URL <https://pubmed.ncbi.nlm.nih.gov/31695505https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6805117/>.
- [152] Danielle A Southern, Hude Quan, and William A Ghali. Comparison of the Elixhauser and Charlson/Deyo Methods of Comorbidity Measurement in Administrative Data. *Medical Care*, 42(4), 2004. ISSN 0025-7079. URL https://journals.lww.com/lww-medicalcare/Fulltext/2004/04000/Comparison_of_the_Elixhauser_and_Charlson_Deyo.8.aspx.
- [153] Nils Gutacker, Karen Bloor, and Richard Cookson. Comparing the performance of the Charlson/Deyo and Elixhauser comorbidity measures across five European countries and three conditions. *European Journal of Public Health*, 25(suppl_1):15–20, 2 2015. ISSN 1101-1262. doi: 10.1093/eurpub/cku221. URL <https://doi.org/10.1093/eurpub/cku221>.

- [154] Roberta De Angelis, Milena Sant, Michel P Coleman, Silvia Francisci, Paolo Baili, Daniela Pierannunzio, Annalisa Trama, Otto Visser, Hermann Brenner, Eva Ardanz, Magdalena Bielska-Lasota, Gerda Engholm, Alice Nennecke, Sabine Siesling, Franco Berrino, and Riccardo Capocaccia. Cancer survival in Europe 1999–2007 by country and age: results of EURO CARE-5—a population-based study. *The Lancet Oncology*, 15(1):23–34, 2014. doi: 10.1016/s1470-2045(13)70546-1.
- [155] Claudia Allemani, Tomohiro Matsuda, Veronica Di Carlo, Rhea Harewood, Melissa Matz, Maja Nikšić, and et al. Global surveillance of trends in cancer survival 2000–14 (CONCORD-3): analysis of individual records for 37,513,025 patients diagnosed with one of 18 cancers from 322 population-based registries in 71 countries. *The Lancet*, 391(10125):1023–1075, 2018. doi: 10.1016/s0140-6736(17)33326-3.
- [156] B Rachet, L Ellis, C Maringe, T Chu, U Nur, M Quaresma, A Shah, S Walters, L Woods, D Forman, and M P Coleman. Socioeconomic inequalities in cancer survival in England after the NHS cancer plan. *Br J Cancer*, 103(4):446–453, 2010. doi: <http://www.nature.com/bjc/journal/v103/n4/supinfo/6605752s1.html>. URL <http://dx.doi.org/10.1038/sj.bjc.6605752><https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2939774/pdf/6605752a.pdf>.
- [157] Libby Ellis, Michel P Coleman, and Bernard Rachet. How many deaths would be avoidable if socioeconomic inequalities in cancer survival in England were eliminated? A national population-based study, 1996–2006. *European Journal of Cancer*, 48(2):270–278, 1 2012. ISSN 0959-8049. doi: 10.1016/j.ejca.2011.10.008. URL <https://doi.org/10.1016/j.ejca.2011.10.008>.
- [158] E Kane, D Howell, A Smith, S Crouch, C Burton, E Roman, and R Patmore. Emergency admission and survival from aggressive non-Hodgkin lymphoma: A report from the UK’s population-based Haematological Malignancy Research Network. *European Journal of Cancer*, 78:53–60, 2017. doi: 10.1016/j.ejca.2017.03.013. URL <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85017365106&doi=10.1016%2Fj.ejca.2017.03.013&partnerID=40&md5=a7e66328fab209dca4231def72852249><https://www.sciencedirect.com/science/article/pii/S0959804917308316?via%3Dihub>.

2 Material

This thesis required the examination of routinely collected data from multiple sources in England including cancer registries, public health bodies and government organisations (Table 5). This chapter describes the data, and their source, and details the cleaning and preparation.

Table 5: Patient data linkage between sources of data sets

Data	Source	Description
Patient and tumour ^{1,2}	NCRAS	Registrations of lymphomas diagnosed in 2005-2013 in England followed up to 2015.
Deprivation ¹	gov.uk	Scores of seven domains that together make up the IMD score for each LSOA.
Comorbidity ²	NHS England	Admissions, A&E attendences and outpatient appointments at NHS hospitals prior to cancer diagnosis.
Population	ONS	Counts of residents in England for each year between 2005 and 2015 by age group, gender and deprivation.
Deaths	ONS	Counts of deaths in England for each year between 2005 and 2013 by age group and gender.
Life tables	ONS/CSG/ICON	Validated, age-specific mortality rates in 2005 to 2013 by age group, gender and deprivation.

¹ Linked on postcode of individual cancer patient and location of LSOA

² Linked on pseudonymised patient and tumour identification numbers

NCRAS: National Cancer Registry and Analysis Service. **ONS**: Office for National Statistics. **NHS**: National Health Service. **IMD**: Index of Multiple Deprivation. **LSOA**: Lower Super Output Area. **A&E**: Accident and Emergency. **CSG**: Cancer Survival Group. **ICON**: Inequalities in Cancer Outcomes Network.

2.1 Tumour data

National cancer registries systematically collect information on tumours from all patients diagnosed with any cancer in England. Records (tumour data) are stored and managed by the National Cancer Registry and Analysis Service (NCRAS). Data is sent from multiple sources (histopathology and haematology services, medical records, radiotherapy departments, hospices, independent hospitals, screenings services, death certificates, general practices, and other UK cancer registries) to the NCRAS and merged together. The data set comprised of all patients diagnosed with non-Hodgkin lymphoma in England during

2005 to 2013 with follow up until the end of 2015. In total, there were 84,504 non-Hodgkin lymphoma registry records (patients). The cancer registry data is then linked to data containing information on patient’s sociodemographic characteristics and prior admission to hospital (Table 6).

Table 6: Variables in each of the data sets

Variable	CR	IMD	HES	LT
Unique patient identification number	X		X	
Unique tumour identification number	X		X	
Gender	X		X	X
Age (years)	X		X	X
Calendar year of death	X			X
Country/region of diagnosis	X		X	
Ethnicity	X		X	
Full date of diagnosis	X		X	
Day, month and year of death (if death)	X			
Cause of death	X			
Country/region at death	X			X
Four-digit ICD-10 code	X			
Five-digit ICD-O code	X			
Detailed tumour characteristics	X			
Type of healthcare organisation at diagnosis	X		X	
Measure of deprivation		X		X
Lower super output area code	X	X	X	
Previous admissions or appointments to hospital	X		X	
Four digit ICD-10 code of previously diagnosed diseases (co-morbidity)			X	

CR: Cancer registry, **IMD:** Index of Multiple Deprivation, **HES:** Hospital Episode Statistics, **LT:** Population mortality life tables, **ICD-(O):** International Classification of Diseases (for Oncology)

2.1.1 Tumour data cleaning and manipulation

Ineligible records

Ineligible records were those with any of: incomplete or invalid data, resident outside of England, in situ neoplasm, benign or uncertain behaviour, or metastatic.¹

Eligible records excluded from analysis

Of those that were eligible, records were excluded from the analysis if they were any of: aged under 15 or over 100 years at diagnosis, a cancer recurrence, invalid date of death, death certificate only (DCO) registration, synchronous tumour, or multiple records of the same primary tumours.

Patients who are diagnosed via a DCO route to diagnosis are those patients where a previous admission to hospital that is related to the cancer cannot be found within six months prior to their date of death. In other words, the time between diagnosis of the cancer and the date of death is unknown: thus, the follow-up time is unknown. These patients are not included survival estimates as their follow-up time is either null or unreliable and will bias the population survival estimates downwards. The downward bias arises because these patients are likely to have been living with NHL for an unknown period of time before their death.

Grouping of histological codes

Non-Hodgkin lymphoma is a heterogeneous group of malignancies characterised as either low grade (indolent and slower growing) or high grade (aggressive and faster growing). The morphological codes were categorised into eight subgroups (subtypes) based on ICD-O classification system (Table 7).²

2.1.2 Tumour data description

Of the patients diagnosed with NHL in England during 2005-2013, DLBCL and Follicular lymphomas were most common and 19.3% were without a specified grade (Table 7). The proportion of unspecified lymphoma decreased year-on-year (Figure 5) and there was a harmonious, opposing relationship with the number of DLBCL diagnoses. For most NHL subtypes, females were, on average, slightly older than males (Figure 6). For most subtypes, patients were diagnosed around 70 years old, except for Burkitt lymphoma that was around 55 years old.

Information on the country of diagnosis was available for all patients: all were diagnosed

Table 7: Distribution of non-Hodgkin lymphoma subtypes for patients in England diagnosed from 2005-2013, with respective morphology and topography ICD-O-3 codes

Index	Site group (subtype)	Grade	Topography	Morphology	n	%
1	CLL/SLL	Indolent	C82.0-C85.9	9670, 9823	4,043	4.8
2	Waldenstrom macroglobulinemia	Indolent	C82.0-C85.9	9761	2,453	2.9
3	Mantle cell	Indolent	C82.0-C85.9	9673	3,549	4.2
4	Diffuse large B-cell	Aggressive	C82.0-C85.9	9680, 9688, 9737-9738	30,750	36.4
5	Burkitt	Aggressive	C82.0-C85.9	9687, 9826	1,077	1.3
6	Follicular	Indolent	C82.0-C85.9	9690-9691, 9695, 9698	15,624	18.5
7	Mature T-cell	Aggressive	C82.0-C85.9	9702	6,066	7.2
8	Marginal zone B-cell	Indolent	C82.0-C85.9	9689, 9699, 9760, 9764, 9699	4,615	5.5
9	Not otherwise specified	N/A	C82.0-C85.9	9591, 9675, 9735	10,308	12.2
10	Other ²	N/A	C82.0-C85.9	9591, 9675, 9735	6,019	7.1
Total					84,504	100.00 ¹

NA: Not applicable (there was not subtype information), **CLL/SLL**: Chronic lymphocytic leukemia / Small-cell lymphocytic lymphoma.

¹ Percentages may not equate to 100.0% due to rounding

² The morphology code specifies these patients are diagnosed with NHL. However, the description states 'other'; these patients are classified similarly to 'Not Otherwise Specified'.

with England. Ethnicity records were missing on 29.7% of the patients. Of the observed ethnicity records 94.2% were White (or any variation thereof). Information on the cause of death was complete for 82.4% of patients who died in the period 2005-2015. The proportion of complete information on cause of death (amongst those who died) was substantially lower amongst two cancer registries: North West & Mersey and Trent. Information on the clinical commissioning group, and lower super output area (thus deprivation level), was observed for 100 per cent of patients. The majority of patients (34.8%) were diagnosed through a general practitioner referral, but 25.0% of patients were diagnosed through A&E; the proportion of patients with missing information for route to diagnosis was 6.8%.

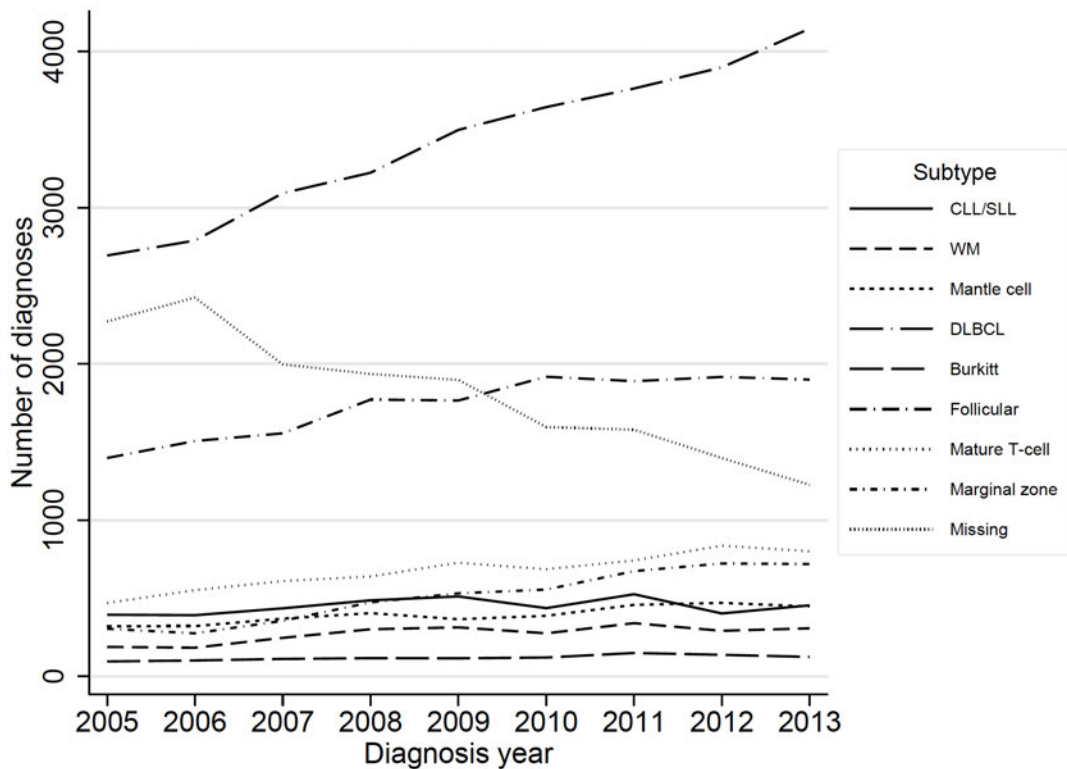


Figure 5: Number of patients diagnosed for each category of NHL subtype in England, 2005-2013.

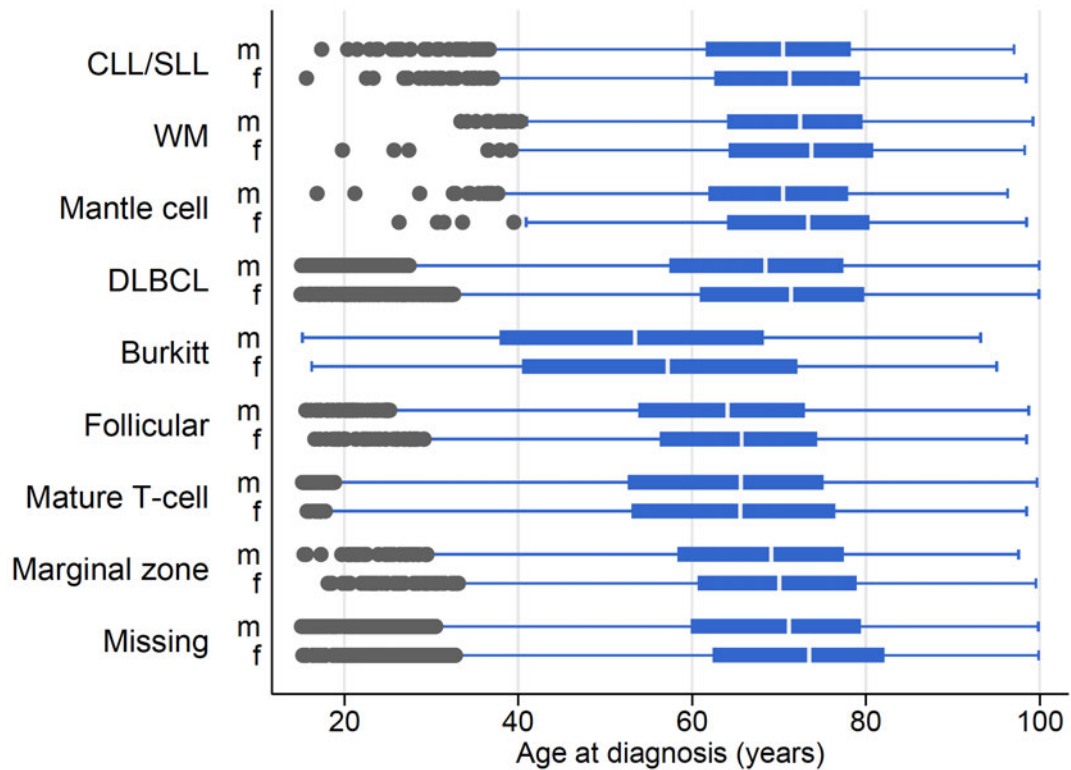


Figure 6: Box plot of age at diagnosis by gender (m: male, f: female) for each category of NHL subtype in England, 2005-2013.

2.2 Deprivation data

When individual-level information is not available, the socioeconomic status of a cancer patient is measured at the ecological-level where the deprivation level of an area is defined by a small area-based score. This thesis applies the approach of an ecological measure of deprivation because none of the routine data sets used in this thesis contains individual-level information on deprivation.

It is known that survival estimates vary by deprivation level, and it is important to first consider whether the variation in survival is partly due to the variation of the underlying measure of deprivation. Information on the geographical boundaries of LSOAs are collected from a decennial census, initially in 2001 and repeated in 2011; there were only minor changes in the number, and population characteristics, of LSOAs.³ The deprivation score (routinely available from administrative data) that is applied to each LSOA was generated initially in 2001 and then updated in 2004, 2007, 2010, 2015 and 2019. For each update, the same approach structure and methodology were applied, which allows these relative deprivation score to be compared over time periods. Research has shown that the choice of the geographic unit (e.g. LSOA or higher layer output areas) influences the socioeconomic inequalities in cancer survival, but not the choice of the index.⁴ Therefore, the stability of deprivation levels for each LSOA across time periods leads us to assume that the updates in deprivation scores of the LSOAs explain little of the deprivation-gap in survival.

Derivation of deprivation score

As deprivation is understood to be multi-dimensional, it is measured by the composition of seven distinct domains: income, employment, education, health, crime, barriers to housing & services, and living environment.⁵ Each domain consists of an accumulation of indicators, which are chosen to be specific to that domain. For example, if a house was without central heating this would indicate a more deprived living environment and not a barrier to housing since the house is already occupied. Information on these indicators is recorded within local authority districts, and combined to make up the singular score for each of the domains. This process is repeated for each LSOA giving seven scores (representing the seven domains) for each LSOA.

To combine the scores into an overall IMD score requires standardisation then transformation (exponential transformation of the ranked domain score) to reduce cancellation effects.⁶ For example, low scores in a domain is not cancelled out by high scores in another domain. The standardised and transformed domain scores are combined into an overall index via weighting. Weighting each domain results in a greater influence for some domains over others. Since the first Indices of Multiple Deprivation, same application of weights have been applied for each subsequent update (Table 8). Income and Employment domains have a higher weight due to the perception that these domains have a greater direct impact on the overall deprivation experience of the area.⁷ Changes to the values of the weights have been assessed using empirical methodologies but all suitable alternatives showed consistent results to the initial weights, and altering the weights in such a way had little impact on the ranks of the LSOAs.⁸

Table 8: Weights of each domain that comprises the Index of Multiple Deprivation

Domain	Weight (%)
Income	22.5
Employment	22.5
Health and disability	13.5
Education, Skills and Training	13.5
Barriers to Housing and Services	9.3
Crime	9.3
Living Environment	9.3

Linking deprivation data to tumour records

The deprivation quintile (where 1 is least deprived, 2, 3, 4, and 5 is most deprived) was linked to patient records in the cancer registry dataset by LSOA codes at the time of cancer diagnosis; all patients were successfully linked to their respective LSOA at the time

of diagnosis. Patients were assigned deprivation quintiles from IMD 2007 if they were diagnosed in 2005-2006, IMD 2010 for diagnoses in 2007-2009, and IMD 2015 for diagnoses in 2010-2013.

Description of deprivation data

The distribution of deprivation quintiles amongst patients diagnosed with NHL was similar across IMD updates (Table 9). For each time period (2007, 2010, or 2015), there was a lower proportion of diagnoses amongst patients in more deprived LSOAs.

Table 9: Proportion of cancer patients diagnosed with NHL for each IMD update by deprivation quintile

Deprivation quintile	IMD		
	2007 (%)	2010 (%)	2015 (%)
Least deprived	22.2	21.8	23.0
2	22.3	22.5	22.2
3	21.0	21.1	21.0
4	18.1	18.5	18.2
Most deprived	16.4	16.1	15.7

Generally, there is consistency in the allocated deprivation quintile to LSOAs for each IMD update.⁵ Cancer patients who were in a certain deprivation quintile at the time of their diagnosis would have been in a similar deprivation quintile had they been diagnosed at a different time period and thus a different IMD update (Table 10). Patients who were in a certain quintile during a certain IMD update would have been at most three quintiles away from the quintile in which they were originally measured: never four or more. Of the patients diagnosed between 2005-2006, and whose deprivation was measured by IMD 2007, 2522 were in the most deprived quintile according to IMD 2007 and IMD 2004 (Table 10a); 217 patients would have been in quintile 4 in IMD 2004, and no patients would have been in less deprived quintiles. Similarly, there were 224 patients who were in quintile 4 in IMD 2007 who would have been in the most deprived quintile in IMD 2004.

Geographically, by cancer registry (Table 11), the highest proportion of diagnoses amongst most deprived patients were in North Western (25.0%), Northern and Yorkshire (22.2%), and West Midlands (20.0%), and lowest amongst those in Oxford (5.1%), Eastern (6.5%) and South and West (6.8%). The difference is partly due to the higher rates of deprivation in Northern counties compared to Southern counties.

Table 10: Number of cancer patients having an agreeable deprivation quintile between the measures directly before or after their diagnosis. 1 is least deprived, 5 is most deprived. Numbers shaded in gray indicate agreement in deprivation level between different time periods of IMD measures.

(a) Patients diagnosed in 2005-2006

		IMD 2004				
		1	2	3	4	5
IMD 2007	1	3274	437	1	0	0
	2	473	2750	488	3	0
	3	3	497	2605	392	0
	4	0	3	382	2419	224
	5	0	0	0	217	2522

(b) Patients diagnosed in 2007-2009

		IMD 2007				
		1	2	3	4	5
IMD 2010	1	5449	622	6	0	0
	2	715	4774	765	4	0
	3	2	794	4548	541	0
	4	0	0	569	4242	338
	5	0	0	0	370	4125

(c) Patients diagnosed in 2010-2013

		IMD 2010				
		1	2	3	4	5
IMD 2015	1	7863	1287	28	1	0
	2	1264	6207	1363	15	2
	3	19	1578	5843	954	10
	4	0	12	1253	5327	662
	5	0	0	7	820	5435

2.3 Comorbidity data

Classifying comorbidity status

The NHS Cancer Plan recognised the inequalities in survival and suggested that, since they are more prevalent amongst more deprived areas, comorbidities may partly explain

Table 11: Proportion of patients diagnosed within each cancer registry by deprivation quintile

Cancer Registry	Deprivation quintile					Total (%)
	1 (%)	2 (%)	3 (%)	4 (%)	5 (%)	
Northern and Yorkshire	17.3	20.2	19.5	20.9	22.2	100.0
North Western	19.9	18.2	17.8	19.1	25.0	100.0
Trent	17.6	21.4	22.7	20.6	17.7	100.0
West Midlands	19.1	23.2	19.1	18.7	20.0	100.0
Eastern	25.5	25.0	24.4	18.6	6.5	100.0
Oxford	43.4	22.2	15.6	13.8	5.1	100.0
South and West	23.2	26.4	25.6	18.1	6.8	100.0
Thames	19.9	19.6	20.2	22.1	18.2	100.0

the deprivation gap in survival. Comorbidities are defined as the coexistence of disorders, in addition to a primary disease of interest, which are causally unrelated to the primary disease (i.e. cancer).^{9,10} For example, a patient may have previously been diagnosed with diabetes before their cancer diagnosis, and since diabetes is chronic disease it will coexist but is considered unrelated (or not a cause of) the cancer. Information on an individual's previous diagnosis of a disorder is collected and stored within a population-based database called the Hospital Episode Statistics database. This database holds administrative and clinical records of all the admissions, A&E attendances and outpatient appointments to NHS hospitals by individuals in England. Healthcare providers collect this information on predefined periodic dates during each year. As HES data is collected from Clinical Commissioning Groups, the records also contain information on private patients treated in NHS hospitals.

Deriving comorbidity score

The Charlson comorbidity index¹¹ (CCI) is the most commonly used in epidemiological studies for several reasons; the index is: a valid prognostic indicator across updated versions of the International Classification of Diseases, applicable to different diseases, was devised in a hospital context, and is sufficient when only administrative (not also clinical) data is available. The index is a weighted score of the number of comorbid conditions, where the weights of the score is the severity of the comorbid condition. The higher the score the more severe the effect of the underlying comorbid conditions on the patient's health outcome.

Although the approach of Charlson *et al.* was to develop a score that described the impact of comorbidities on health outcomes, it did not, however, take account of the patient's likely treatment. For example, treatment or surgery is more risky for particular comorbidities. Thus, the Royal College of Surgeons' (RCS) adaptation of the CCI may more accurately

classify the severity of comorbidities in this setting.¹² The RCS-CCI criteria reduces the number of effectual comorbidities to fourteen; however, the interest of this thesis is in patients with a first diagnosis of cancer and two categories were removed from the RCS-CCI index: any malignancy and metastatic solid tumour. (Data cleaning removed any patients with a previous malignancy or tumour; therefore, there would be no patients within this criteria).

Table 12: Royal College of Surgeons Charlson Score indicating International Classification of Disease tenth revision code for 14 categories

Index	Disease category	ICD-10 code	N (%)
1	Myocardial infarction	I21, I22, I23, I252	719 (0.9%)
2	Congestive cardiac failure	I11, I13, I255, I42, I50, I517	873 (1.0%)
3	Peripheral vascular disease	I70-73, I770, I771, K551, K558, K559, R01, Z958, 959	615 (0.7%)
4	Cerebrovascular disease	G45, G46, I60-69	830 (1.0%)
5	Dementia	A810, F00-03, F051, G30, G31	275 (0.3%)
6	Chronic pulmonary disease	I26, I27, J40-45, J46, J47, J60-67, J684, J701, J703	2777 (3.3%)
7	Rheumatological disease	M05, M06, M09, M120, M315, M32-36	869 (1.0%)
8	Liver disease	B18, I85, I864, I982, K70, K71, K721, K729, K76, R162, Z944	307 (0.4%)
9	Diabetes mellitus	E10-14	2182 (2.6%)
10	Hemiplegia or paraplegia	G114, G81-83	165 (0.2%)
11	Renal disease	I12, I13, N01, N03, N05, N07, N08, N171, N172, N18, N19, N25, Z49, Z940, Z992	996 (1.2%)
12	AIDS/HIV	B20-24	49 (0.1%)
13*	Any malignancy	C00-C26, C30-34, C37-41, C43, C45-58, C60-76, C80-85, C88, C90-97	N/A
14*	Metastatic solid tumour	C77-C79	N/A
Total			10,657 (12.6%)

*Index 13 and 14 were removed

Previous studies have shown a lack of consistency, validity and reproducibility when deriving an individual's comorbidity score.¹³ Maringe *et al.* (2017)¹⁴ developed a robust algorithm to capture an internally, and externally, valid comorbidity score based on crucial assumptions that were not included in previous studies. The algorithm specifically attempts to minimise selection bias by defining an optimal *time-window* for a comorbidity to occur. For example, by allocating the same amount of person-time at risk of having

a comorbidity for any patient, which gives each patient the same probability of being diagnosed with a comorbidity. Furthermore, an optimal *restriction-window* is defined and applied to avoid collider bias arising due to the cancer causing the comorbidity. Maringe *et al.* show that the optimal time- and restriction-window are 6 years and 6 months, respectively. The restriction-window of 6 months is applied in this thesis. The time-window was set to 24 months because HES data was not available before 2003 and the earliest cancer diagnosis of the patients in the cancer registry data was 2005. However, the validity of the a patient's comorbidity status remains high because a high proportion of comorbidity information is collected within 2 years prior to diagnosis; in other words, a high proportion of comorbidities are diagnosed within 2 years prior to cancer diagnosis.^{14,15}

The total number of comorbid conditions recorded amongst NHL patients between 2003 and 2013 was 10,657 (12.6%) (Table 12). The most common comorbidity was chronic pulmonary disease (2777; 3.3%) followed by diabetes 2182 (2.6%). Figure 7 shows the proportion of patients by comorbidity score for all twelve comorbidities. For all comorbidities, except AIDS/HIV, patients tended to be aged between 60 and 90 years old, and the most common age was around 80 years old. Comorbidities were rarely observed amongst those younger than 60 years, except for those with AIDS/HIV, rheumatological or liver disease. Patients with dementia, hemi/paraplegia, renal disease and AIDS/HIV are scored at least two or more and would not have a comorbidity score as low as one due to their severity of impact on the patient's overall health. There was an increasing trend in the prevalence of comorbidity amongst patients living in deprived areas (Figure 8). The prevalence of a comorbidity score of one was similar across deprivation levels; however, there were opposing trends by deprivation levels comparing a comorbidity score of zero to two or more.

2.4 Population and mortality data

Life table estimates (derived from population statistics data) is required for the calculation of cancer survival within the relative survival setting. Annual mid-year population estimates of England are obtained from the Office for National Statistics for England and Wales.¹⁶ Estimates are defined according to a standard demographic method: the cohort component method. The estimates, obtained from census data, are highly validated and are extensively used in public health research. The components of the method incorporate natural change (births, deaths and aging), migration and special populations. Deaths data are obtained from the Civil Registration System (administered by the ONS), which captures information on all deaths in England, including deaths of those usually resident outside England.

Life tables, used for estimating net survival, can be constructed from raw counts of deaths in the population stratified by age, gender, deprivation and calendar period. Occasionally, the estimated background mortality may not match the true background mortality of the

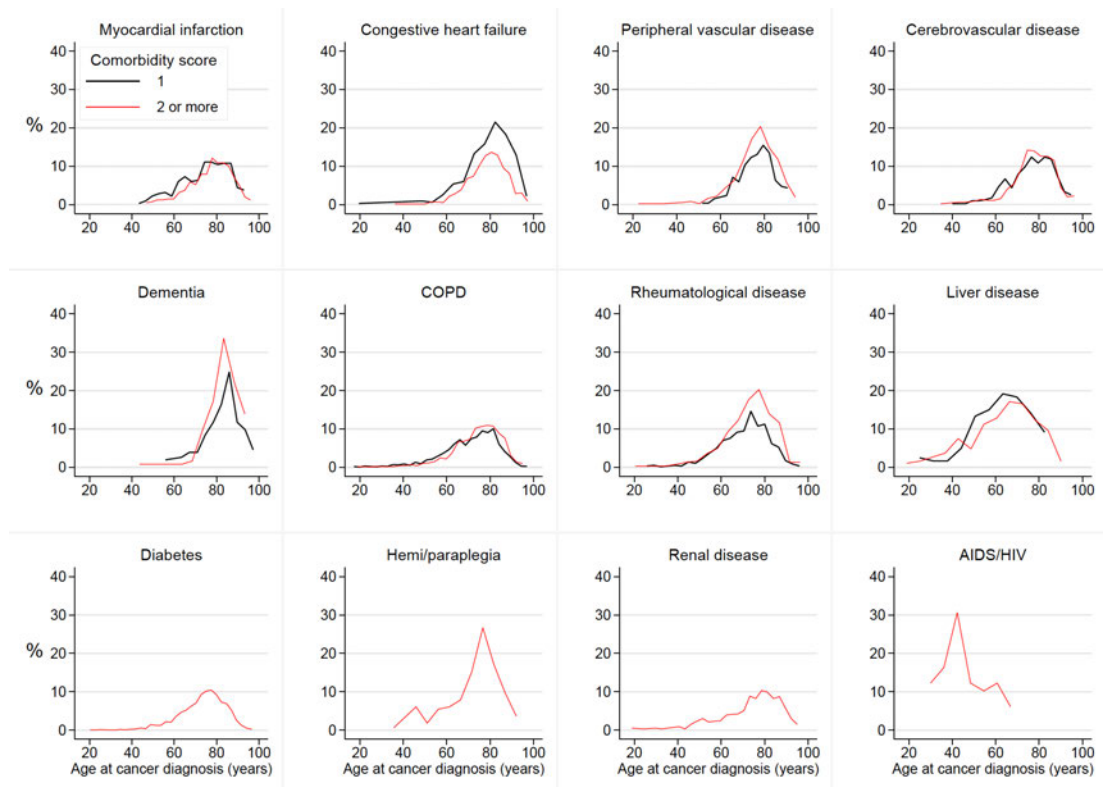


Figure 7: Prevalence of comorbidity score amongst non-Hodgkin lymphoma patients in England diagnosed 2005-2013. Diabetes, hemi/paraplegia, renal disease and AIDS/HIV are automatically scored as 2 or more. COPD: Chronic obstructive pulmonary disease.

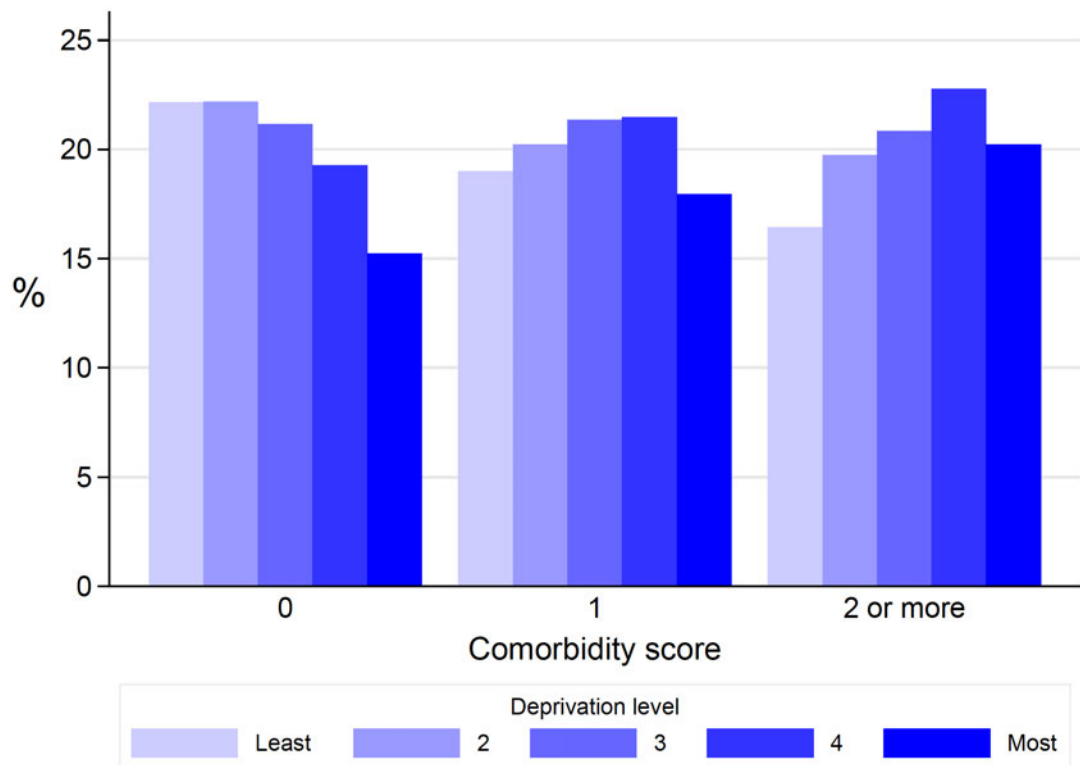


Figure 8: Probability of comorbidity score by deprivation level amongst non-Hodgkin lymphoma patients in England diagnosed 2005-2013.

cancer patient with a particular set of stratified variables. For example, the cancer patient may have a particular characteristic that separates them from their matched population mortality estimate; this can lead to biased estimates of excess mortality. Not including a particular characteristic in the life table (e.g., smoking-adjusted life tables) has a small impact on survival estimates.¹⁷ However, recent advances in methodology have proposed two parametric corrections in the excess hazard regression model to account for possible mismatches in the life table and thus misspecification of the background mortality rate. The two approaches are to include (i) a single-parameter or (ii) a random effect (frailty).¹⁸

Raw mortality rates are vulnerable to scarcity of events in stratified groups (i.e., low number of cancer patient deaths within the cohort data for subcategories of the variables used to stratify the life tables). For example, at the time a cancer patient is diagnosed, there could be low numbers of deaths amongst male patients of a certain age who are living in more deprived areas; therefore, the expected mortality for the matched cancer patient (with the aforementioned characteristics) will be weighted by a low number of observations in the population. Stability is introduced by smoothing the raw mortality rates of life tables.^{19,20}

References

- [1] Ruoran Li, Louise Abela, Jonathan Moore, Laura M. Woods, Ula Nur, Bernard Rachet, Claudia Allemani, and Michel P. Coleman. Control of data quality for population-based cancer survival analysis. *Cancer Epidemiology*, 38(3):314–320, 6 2014. ISSN 1877783X. doi: 10.1016/j.canep.2014.02.013.
- [2] S H Swerdlow. *WHO Classification of Tumours of Haematopoietic and Lymphoid Tissues; WHO Classification of Tumours, Volume 2*. International Agency for Research on Cancer, World Health Organisation, revised 4t edition, 2008.
- [3] Office for National Statistics. Changes in lower super output area codes between 2001 and 2011 in England, 2018.
- [4] Laura Woods, B Rachet, and M Coleman. Choice of geographic unit influences socioeconomic inequalities in breast cancer survival. *British journal of cancer*, 92:1279–1282, 5 2005. doi: 10.1038/sj.bjc.6602506.
- [5] Office for National Statistics. The English Indices of Deprivation 2019 (IoD2019). Technical report, Ministry of Housing, Communities & Local Government, London, UK, 2019. URL <https://www.gov.uk/government/statistics/english-indices-of-deprivation-2019>.
- [6] David McLennan, Stefan Noble, Michael Noble, Emma Plunkett, Gemma Wright, and Nils Gutacker. The English Indices of Deprivation 2019: Technical Report. Technical report, Ministry of Housing, Communities and Local Government, London, UK, 2019. URL <https://www.gov.uk/government/publications/english-indices-of-deprivation-2019-technical-report>.
- [7] Peter Townsend. Deprivation. *Journal of Social Policy*, 16(2):125–146, 4 1987. ISSN 0047-2794. doi: 10.1017/S0047279400020341. URL https://www.cambridge.org/core/product/identifier/S0047279400020341/type/journal_article.
- [8] Verity Watson, Chris Dibben, Matt Cox, Iain Atherton, Matt Sutton, and Mandy Ryan. Testing the Expert Based Weights Used in the UK’s Index of Multiple Deprivation (IMD) Against Three Preference-Based Methods. *Social Indicators Research*, 144(3):1055–1074, 2019. ISSN 1573-0921. doi: 10.1007/s11205-018-02054-z. URL <https://doi.org/10.1007/s11205-018-02054-z>.
- [9] Miquel Porta. *A Dictionary of Epidemiology*. Oxford University Press, 2014. ISBN 9780195314496. URL <http://www.oxfordreference.com/view/10.1093/acref/9780195314496.001.0001/acref-9780195314496>.
- [10] A R Feinstein. The pre-therapeutic classification of co-morbidity in chronic disease. *J Chronic Dis*, 23(7):455–468, 1970.

- [11] Mary E Charlson, Peter Pompei, Kathy L Ales, and C Ronald MacKenzie. A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation. *Journal of Chronic Diseases*, 40(5):373–383, 1987. doi: [https://doi.org/10.1016/0021-9681\(87\)90171-8](https://doi.org/10.1016/0021-9681(87)90171-8). URL <http://www.sciencedirect.com/science/article/pii/0021968187901718><https://www.sciencedirect.com/science/article/pii/0021968187901718?via%3Dihub>.
- [12] J N Armitage and J H van der Meulen. Identifying co-morbidity in surgical patients using administrative data with the Royal College of Surgeons Charlson Score. *Br J Surg*, 97(5):772–781, 2010. doi: 10.1002/bjs.6930.
- [13] Diana Sarfati. Review of methods used to measure comorbidity in cancer populations: No gold standard exists. *Journal of Clinical Epidemiology*, 65(9):924–933, 9 2012. ISSN 0895-4356. doi: 10.1016/j.jclinepi.2012.02.017. URL <https://doi.org/10.1016/j.jclinepi.2012.02.017>.
- [14] Camille Maringe, Helen Fowler, Bernard Rachet, and Miguel Angel Luque-Fernandez. Reproducibility, reliability and validity of population-based administrative health data for the assessment of cancer non-related comorbidities. *PLoS ONE*, 12(3):e0172814, 2017. doi: 10.1371/journal.pone.0172814. URL <https://doi.org/10.1371/journal.pone.0172814><https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5338773/pdf/pone.0172814.pdf>.
- [15] Helen Fowler, Aurelien Belot, Libby Ellis, Camille Maringe, Miguel Angel Luque-Fernandez, Edmund Njeru Njagi, Neal Navani, Diana Sarfati, and Bernard Rachet. Comorbidity prevalence among cancer patients: a population-based cohort study of four cancers. *BMC Cancer*, 20(1):2, 2020. ISSN 1471-2407. doi: 10.1186/s12885-019-6472-9. URL <https://doi.org/10.1186/s12885-019-6472-9>.
- [16] Office for National Statistics. Population estimates for the UK, England and Wales, Scotland and Northern Ireland. Technical report, London, UK, 2020.
- [17] L. Ellis, M. P. Coleman, and B. Rachet. The impact of life tables adjusted for smoking on the socio-economic difference in net survival for laryngeal and lung cancer. *British Journal of Cancer*, 111(1):195–202, 5 2014. ISSN 15321827. doi: 10.1038/bjc.2014.217. URL www.bjcancer.com.
- [18] Francisco J. Rubio, Bernard Rachet, Roch Giorgi, Camille Maringe, and Aurélien Belot. On models for the estimation of the excess mortality hazard in case of insufficiently stratified life tables. *Biostatistics (Oxford, England)*, 22(1):51–67, 1 2021. ISSN 14684357. doi: 10.1093/biostatistics/kxz017. URL <https://academic.oup.com/biostatistics/article/22/1/51/5499194>.
- [19] Bernard Rachet, Camille Maringe, Laura M Woods, Libby Ellis, Devon Spika, and Claudia Allemani. Multivariable flexible modelling for estimating complete, smoothed

life tables for sub-national populations. *BMC Public Health*, 15(1):1240, 2015. ISSN 1471-2458. doi: 10.1186/s12889-015-2534-3. URL <https://doi.org/10.1186/s12889-015-2534-3>.

- [20] Inequalities in Cancer Outcomes Network. Life tables for England and Wales by sex, calendar period, region and deprivation., 2013.

3 Methods

3.1 Survival analyses

3.1.1 Data settings

Survival is an indicator of the efficiency of a health care system to manage cancer, and highlights the burden of cancer in a given population. In epidemiology, survival is used to identify differences in health care management between population groups or trends observed over time periods. In the overall survival setting, the survival probability of a population of patients is the survival where any cause of death contributes to the event of interest (i.e., death). This estimator of survival is influenced by the mortality hazard from causes other than the cancer of interest.

When the interest is in the impact of a particular type of cancer, the survival estimator would need to provide an estimate that was not influenced by death from other causes. In this case, death from other causes are considered as competing risks. Therefore, it is impossible to measure the time until the patient's death that is due to cancer for any patient who dies from a competing risk. Thus, the net survival probability is an attractive estimator because it is not influenced by other causes of death, and is defined as the survival if cancer were the only cause of death.^{1,2}

Cancer registry data holds information on patient's date, and cause, of death recorded by routinely collected death certificates. However, even if the cause of death were available (cause-specific setting) the records may be inaccurate or unreliable. Assigning an exact cause of death is a complex process. For example, the exact cause may not be discernible by the doctor, or the registrar coding the cause may have numerous, various fields (primary, secondary, underlying, etc). The recording of death certificates is a legal requirement in England, along with the recording of internationally recognised codes (ICD-10) for the cause of death. However, there is variability in temporal, and geographical, routine registration of the cause of death; thus, causes of death may not be comparable over time and between geographies. Population-based analyses of cancer survival are challenging to construct due to the uncertainty of the cause of death.

Furthermore, within the cause-specific setting, the assumption of independent censoring may not hold and can lead to biased survival estimates. For example, patients who survive their cancer are generally younger, physically active, have less severe (or fewer) comorbidities and earlier-staged tumours. The information on patients who die from competing risks are not only associated with the probability of censoring but also the survival; therefore, censoring in this paradigm is informative. The covariable, for example the patient's age, influences the cause-specific mortality hazard and the other-cause mortality hazard.

The relative survival setting addresses this often invalid assumption by assuming that the competing causes of death are approximated using population mortality estimates (stratified by age, gender, deprivation and calendar year of diagnosis) based on the general population from which the cancer patients are drawn. Two common approaches have been developed to account for the bias arising from the influence of other causes of death: the relative survival ratio and the hazard modelling approach.

The relative survival (RS) ratio is a comparison of the overall survival to the survival that is expected in a similar population that was not diagnosed. This approach matches cancer patients to patients in the life tables. However, patients with the lowest survival exit the study before those with highest survival; thus, the sample begins to represent a different population (e.g., a younger population who are unlikely to experience competing causes of death). Therefore, the relative survival ratio is a biased estimator of net survival.

The hazard modelling approach, the second approach, assumes that the observed mortality hazard can be decomposed into the cancer-specific mortality hazard (i.e., the excess mortality hazard) and the general population mortality hazard. Unlike in the cause-specific setting where the two mortality processes are assumed independent, in the hazard modelling approach the excess mortality hazard is not assumed to be the same for all patients. Indeed, in the hazard modelling approach, the assumption of independent censoring is assumed to hold conditional on accounting for the information on patients' sociodemographic characteristics contained within the life tables (e.g., age at cancer diagnosis, gender, deprivation, and calendar year at diagnosis).

3.1.2 Net survival measure

Survival estimates derived within the relative survival setting are valid under certain assumptions. Firstly, non-informative administrative censoring argues that patients are only censored alive at the end of the follow-up of the study. This assumption would be invalid if patients were censored alive before the end of follow-up, which is more likely for healthier patients who attend a consult shortly after their diagnosis but, due to their healthier status, do not return for another consult until after the end of the study. Secondly, it is assumed that the time to death due to cancer is conditionally independent of the time to death due to other causes, given the demographic variables used to construct life tables. Thirdly, the estimation of population background mortality is assumed to be sufficient to capture the mortality due to other causes. Lastly, it is assumed that a patient's survival time is independent of the survival time of another patient.

Net survival (NS) is the percentage of patients that are alive at a certain time point, if cancer were the only cause of death. This method isolates the mortality due to cancer that is in excess of the mortality due to other causes.³ This measure removes the bias arising from competing risks (which is approximated by the mortality from life tables). In

England, life tables are stratified by age, gender, deprivation and calendar year, and are used to estimate the expected general population mortality. Life tables are used to adjust the observed (overall, λ_O) mortality in the cancer patient cohort from the mortality due to other causes (population, λ_P) in a similar cohort without cancer. This assumes that the observed mortality is the sum of the mortality due to cancer (excess, λ_E) and the mortality due to other causes. More formally, this is called the additive model:

$$\lambda_O(t) = \lambda_E(t) + \lambda_P(t).$$

Under the aforementioned assumptions, λ_E is called the excess mortality hazard (EMH), which is given by

$$\lambda_E(t) = \lim_{\delta t \rightarrow 0} \frac{P(t < T_E \leq t + \delta t | T_E > t)}{\delta t}.$$

This is the limit, as dt approaches 0 (a small interval), of the probability that the event of interest occurs between time t and time $t + dt$ given that the event of interest has not yet occurred up to time t . Given a population all alive now, the hazard is the proportion of the population that will die in the next short unit of time, divided by the length of the short time unit, thus giving a rate. As in classical survival analysis, the relationship that links the excess hazard to the net survival is

$$S_E(t) = \exp\left\{-\int_0^t \lambda_E(u) du\right\}.$$

This relationship is made possible under the assumption that the excess hazard and population hazard are conditionally independent given the set of covariates defined in the life table (age, gender, deprivation, and calendar year at diagnosis).

Estimating net survival: Non-parametric approach

Of the approaches used to estimate net survival, the Pohar Perme approach provides a consistent estimator of population excess hazard (and, therefore, of population net survival).³ The principle of this approach is, at each event time, to estimate the net survival for patient i using their observed mortality (from the cancer data set) weighted by their expected survival probability at that time (measured using life tables).

The weighting is necessary to account for informative censoring due to competing cause of death. The Pohar Perme approach accounts for informative censoring by using the inverse probability of censoring as weights. In other words, an individual's survival is weighted by the inverse of their expected survival probability, S_{P_i} , which is derived from population life tables.

To account for informative censoring, the components of the additive hazard model must be weighted by the individual's inverse probability of censoring. The additive hazard model assumes that

$$\lambda_{Ei}(t) = \lambda_{Oi}(t) - \lambda_{Pi}(t),$$

where

$$\lambda_{Oi}(t) = \frac{dN(t)}{Y(t)},$$

which is the ratio of the number of events in an interval dt , over the number of patients at risk at the start of the interval. To account for informative censoring, the components of $\lambda_{Oi}(t)$ is weighted such that

$$\lambda_{Oi}(t) = \frac{dN^W(t)}{Y^W(t)},$$

where

$$dN^W(t) = \sum_{i=1}^N \frac{dN_i(t)}{S_{Pi}(t)}$$

and

$$Y^W(t) = \sum_{i=1}^N \frac{Y_i(t)}{S_{Pi}(t)}.$$

The second component of the additive hazard model, the population hazard $\lambda_{Pi}(t)$, estimated from population life tables, is corrected for informative censoring by:

$$\lambda_P(t) = \frac{\sum_{i=1}^N Y_i^W(t) \lambda_{Pi}(t) dt}{Y^W(t)}.$$

Putting the weighted hazards into the additive hazard model gives an estimate of the cumulative excess hazard:

$$\hat{\Lambda}_E(t) = \int_0^t \frac{dN^W(u)}{Y^W(u)} du - \int_0^t \frac{\sum_{i=1}^N Y_i^W(u) d\Lambda_{Pi}(u)}{Y^W(u)} du.$$

Thus, the population net survival is

$$\hat{S}_E(t) = \exp(-\hat{\Lambda}_E(t)).$$

Net survival estimates obtained from this approach are independent of the other causes of death. The interpretation of a survival estimate derived in this way is the average of the survival ratios for each individual. Net survival estimates can be obtained for specific characteristics (such as age at diagnosis, gender, socioeconomic status, ethnicity, etc).

Estimating net survival: Regression models

Regression models are used to estimate the association between a group of variables and the excess hazard, or are used to predict the net survival for a particular group of patients sharing homogeneous characteristics. This differs from the non-parametric Pohar Perme approach that uses inverse probability weighting; whereas, excess hazard models account for informative censoring by adjusting for the covariable defined in the life table. Modelling predicts the net survival of a given time t , for each individual (including those who died, or were censored, before time t). The average of the individual net survival estimates gives the population net survival at time t . The survival estimates are unbiased if the model correctly incorporates (via modelling) the effect of the covariables on the risk of censoring due to competing causes of death.

The additive hazard model incorporates the covariables such that

$$\lambda_{O_i}(t, a, \mathbf{x}, \mathbf{z}) = \lambda_{E_i}(t, a, \mathbf{x}, \mathbf{z}) + \lambda_{P_i}(a + t, \mathbf{z})$$

where t is the time since cancer diagnosis, a is the age at cancer diagnosis, \mathbf{x} the vector of covariables, and \mathbf{z} the vector of the variables in the life table that remove the bias arising from censoring.

The most common functional form of a multivariable excess hazard model is multiplicative, such that

$$\lambda_{E_i}(t, a, \mathbf{x}, \mathbf{z}) = \lambda_{0,E_i}(t) * \exp(f(t_i, \beta, \mathbf{x}_i, \mathbf{z}_i))$$

where $\lambda_{0,E_i}(t)$ denotes the individual baseline excess hazard. The function $f()$ can be a flexible function of time t , incorporating non-linear and time-dependent effects of the vector of covariates \mathbf{x} and \mathbf{z} and their corresponding parameters, β .

The net survival of patient i is the survival derived from the individual excess mortality hazard $\lambda_{E_i}(t)$, such that

$$S_{N_i}(t) = \exp\left(-\int_0^t \lambda_{E_i}(u) du\right).$$

Thus, the marginal net survival is the average of individual net survival functions

$$S_N(t) = \frac{1}{n} \sum_{i=1}^n S_{N_i}(t).$$

3.1.3 Strategy of survival estimation

Due to its flexibility, the excess hazard model can include a series of functional forms to improve the fit of the model. Before advances in methodology, the step function allowed the use of more flexible functions.⁴ More recently, and after advances in computing and methodology, common functional forms include:

1. Baseline hazard: standard distributions (e.g. exponential, Weibull, log-logistic, etc.), fractional polynomials,⁵ restricted cubic splines,^{6,7,8} B-splines,⁹ or penalised tensor splines.¹⁰
2. Non-linear effects (a non-linear relationship between the outcome and independent variables) and time-dependent effects (an interaction between the follow-up time and a variable considered to have a varying effect on the excess hazard over time).
3. Interactions (effect modification): such as the effect of a variable on the excess hazard differs within levels of another variable.

The algebraic expression for the excess hazard model that includes the above functional forms is

$$\lambda_E(t, \mathbf{x}) = \lambda_0(t) * \exp(\beta_1 x_1 + f(x_2) + g(t)x_3)$$

where $\lambda_0(t)$ is the baseline hazard, x_1 a variable with a linear effect, $f(x_2)$ a function for a variable with a non-linear effect, and $g(t)x_3$ a function for a variable with a time-dependent effect on the excess hazard. The need for non-linear and time-dependent effects can be validated using Akaike Information Criterion.

3.1.4 Multilevel excess hazard models

The study design may feature a hierarchical structure and the excess hazard model would need to account for the dependency between outcomes (i.e., mortality hazard).¹¹ By in-

corporating a random effect, w_d , into the excess hazard model to represent the correlation between outcomes:

$$\lambda_E(t, \mathbf{x}|w_d) = \lambda_0(t) * \exp(\beta_1 x_1 + f(x_2) + g(t)x_3 + w_d)$$

The term $w_d \sim N(0, \sigma^2)$ represents a random effect for clusters $d = 1, \dots, D$, and might be validated by comparing the likelihood ratio statistic to a mixture of chi-squares with 0 and 1 degree of freedom ($-2llr(\theta_0) \sim \chi_{0,1}^2$).

3.2 Causal inference

This section introduces the estimation of the causal effect of an intervention on a time-to-event for a single outcome. The effect is estimated via regression standardization using flexible parametric survival models. The methods detailed here explain how to estimate the standardized cumulative hazard function for a single survival outcome (e.g., cumulative hazard of death at 5 years since diagnosis) assuming independent censoring.

Potential outcomes framework

Randomised control trials are preferred over observational studies to estimate the causal effect between an exposure and an outcome. However, it may be unethical or unfeasible to allocate an exposure to a patient (e.g., to allocate an individual to smoke cigarettes or to allocate an individual to have diabetes). Often, research questions in health sciences are causal and are estimated with classical methods such as regression. However, in observational studies (contrary to randomised control studies), the measure of the exposure-outcome association using regression cannot be interpreted as causal if the individuals' characteristics differ between exposure groups and they are not adjusted for in the analysis. Otherwise, if multivariable regression is used, there is a causal interpretation (although sometimes conditional) if the assumptions hold. Moreover, several methods have been developed to measure the causal effect.

To estimate the causal effect of an exposure on an outcome, Jerzy Neyman introduced the potential outcomes framework specifically for randomised control trials, which was generalised by Donald Rubin: extending causal inference from randomised experiments to observational data.¹² To illustrate the framework, let the outcome, Y , be a binary indicator for death at a specific time point after cancer diagnosis, with a binary exposure (e.g., treatment), X , and measured confounders (\mathbf{C}). For a binary exposure, each patient in the sample has two potential outcomes (i.e., $Y(\mathbf{a})$), where $Y(1)$ denotes the potential outcome if they received the exposure, and $Y(0)$ denotes the potential outcome if they did not receive the exposure.¹² Since, only one exposure-outcome combination is observed (i.e., treated then survived, or treated then died), only one of the potential outcomes is observed. The causal effect of interest is the contrast between the *potential outcomes* under different exposure levels (i.e., the difference between $E[Y(1)] - E[Y(0)]$).¹³ In practice, the contrast cannot be observed, but can be estimated from the data under the following assumptions:

1. **Counterfactual consistency** holds if the potential outcome ($Y(1)$) for an individual (had they been exposed) is the same as the observed outcome of the individual when exposed, and likewise for unexposed individuals. In other words, the definition of the exposure and outcome is consistent for all individuals, and that there are no differing versions of the interventions.

2. **Conditional exchangeability** holds if the measured and unmeasured confounders of the exposure-outcome relationship are equally distributed between the exposed and the unexposed groups. In randomised studies, conditional exchangeability holds because the exposed individuals, had they not been exposed, would have had the same potential outcome as the unexposed, and vice versa.
3. **Positivity** holds if the probability of being exposed (and similarly for all other predictors) is greater than zero: and therefore, less than one. When this assumption is violated, it is typically because the target population is poorly defined (trying to estimate the effect of a treatment on people who would never receive it anyway).
4. **Noninterference** holds if the potential outcome of one individual was not influenced by the exposure of another individual.
5. **Independent censoring** holds if, within any subgroup of interest, individuals censored at time t are representative of all of the individuals in that subgroup who remain at risk at time t : in other words, censoring is independent provided the censoring occurs randomly within any subgroup.¹⁴
6. **Absence of time-dependent confounding** holds if the *values* of any independent variables are constant over time or that the *effect* of an independent variable is constant over time (i.e., absence of an interaction between the constant-value confounder and time).¹⁵

Regression standardisation

In survival analysis, when the research question focuses on the causal effect of an exposure (e.g., a treatment) on an outcome (e.g., time-to-death or the hazard of death) in a population, the effect is often measured using a parametric model (classical regression). A parametric model is a model for which assumptions are made about the relationship between the outcome and the predictors. The parametric model usually involves adjusting for covariates and obtaining estimates of the exposure-outcome relationship for each level of the exposure: giving estimates that are interpreted as conditional on the covariates. Adjusting for categorical covariates, such as the patient's gender, assumes that the effect of interest is constant across levels of the confounders (when not including interaction terms in the model). However, a model often includes interaction terms, along with non-linear and time-dependent effects, which complicates the interpretation of the main effect of interest. Instead, the effect of interest (e.g., the cumulative hazard) can be standardised to give an average cumulative hazard for a level of the exposure over the distribution of the confounders in the sample: avoiding the aforementioned difficulty of interpretation.^{16,17} Also, the standardised model can incorporate more complex structures without adding to the complexity of the interpretation.

Regression standardisation is obtained by:

1. Fitting a regression model to predict the function for the cumulative hazard of death, separately by each level of the exposure, generating a prediction for each level of the exposure.
2. The predictions are averaged over a 'standard' confounder distribution in the sample to produce a standardised cumulative hazard for each level of the exposure.
3. These standardised cumulative hazards are then contrasted to provide standardised measures of association. Assuming all confounding is controlled for, the standardised measures of association are interpreted as average causal effects.

The contrast of the standardised cumulative hazards (to give the average causal effect) is obtained by calculating the difference in the average of the predicted hazards (the number of predictions is equal to the number of observations, N , in the sample).^{18,19} As an example, for the time-to-death, t , of a sample of n individuals with a binary exposure, X , and measured covariates Z , the 'standard' distribution is the marginal distribution of Z in the sample. Meaning that each level of the exposure is given the same covariate distribution. A standardised cumulative hazard curve is generated for each level of the binary exposure X , averaged over the predictions of each individual, and contrasted, such that

$$\frac{1}{N} \sum_{i=1}^N H(t|X = 1, \mathbf{Z}_i) - \frac{1}{N} \sum_{i=1}^N H(t|X = 0, \mathbf{Z}_i).$$

This contrast is possible under the assumptions outlined in section 3.2: counterfactual consistency, conditional exchangeability, positivity, noninterference, independent censoring, and absence of time-dependent confounding.

Flexible parametric survival models

Flexible parametric survival models are used to parametrically estimate the proportional hazard of an event, such as the probability of death in a time interval amongst a sample of cancer patients.^{20,21,7} The parametric model can be derived from the relationship between the survival function and the hazard function:

$$S(t) = \exp(-\lambda t)$$

Then, transforming to the log cumulative hazard scale

$$\ln[H(t)] = \ln[-\ln(S(t))]$$

$$\ln[H(t)] = \ln(\lambda) + \ln(t)$$

and including covariates gives

$$\ln[H(t|\mathbf{x}_i)] = \ln(\lambda) + \ln(t) + \mathbf{x}_i\beta$$

This is a parametric proportional hazards model (parametric model) and, compared to the widely-used semi-parametric Cox model, it has additional flexibility. For example, the Cox model has no widely-accepted approach for non-proportional hazards.⁷ There is further flexibility in the parametric model such that, assuming proportional hazards, and when modelling on the (log) cumulative hazard scale, parameters (β) can still be interpreted as hazard ratios.

To overcome the inflexible shape of the baseline hazard that arises from the assumption of linearity in log time ($\ln(t)$), the parametric model can incorporate restricted cubic splines to provide smoothed cumulative hazard curves and aid predictions. Restricted cubic splines are outlined in section 3.1.3 and are applied in the same way here.

3.3 Dependent discrete data

Binary outcomes are often modelled using logistic regression models, which are part of the generalised linear model family. Examples of binary outcomes in healthcare settings are: route to diagnosis (e.g., admission via emergency department vs other) and vital status (i.e., deceased vs alive). In population-based studies, at an individual level, a patient's health outcome is viewed as independent from another patient's outcome; however, due to the universal healthcare system, a patient's outcome is more likely to be similar to another patient if both patients shared a characteristic of the cluster in which they reside (e.g. lower super output area, hospital attended at diagnosis, or Clinical Commissioning Groups). To obtain reliable inferences, this correlation between outcomes of patients from the same cluster must be accounted for. The section defines the logistic regression model for binary outcomes and outlines how correlation is incorporated into the analysis using multilevel logistic regression models (i.e., generalised linear mixed models).

Let Y_j be the binary outcome for patient j . We assume that Y_j follows a Bernoulli distribution with success probability π_{ij} , i.e.,

$$Y_j \sim \text{Bernoulli}(\pi_j).$$

A crude model (including a single continuous variable) is defined as

$$\text{logit}(\pi_j) = \beta_0 + \beta_1 A_j$$

where A is the continuous variable age at diagnosis. Categorical variables are included as

$$\text{logit}(\pi_j) = \beta_0 + \sum_{k=1}^K \beta_k C_{jk}$$

where C_k for $k = 1, 2, \dots, K - 1$ are dummy variables for a categorical variable with K levels.

Not accounting for clustering, the multivariable logistic regression model is

$$\text{logit}(\pi_j) = \beta_0 + \beta_1 A_j + \beta_2 G_j + \beta_3 E_j + \sum_{k=1}^4 \beta_{4k} D_{jk} + \sum_{k=1}^2 \beta_{5k} C_{jk} + \beta_6 R_j$$

where patient characteristics are the continuous variable for age (A : 18 - 99 years), binary variables for gender (G : 0 - male, 1 - female), ethnicity (E : 0 - white, 1 - other), route to diagnosis (R : 0 - elective, 1 - emergency), and categorical variables for deprivation (D : 1

- least deprived, 2, 3, 4, 5 - most deprived), comorbidity status (C: 0 - no comorbidity, 1 - comorbidity, 2 - multimorbidity). Route was used as an explanatory variable when analysing survival probability and was used as an outcomes when analysing the probability of expedited diagnostic route.

To account for the correlation between patient outcomes within a cluster (as outlined above), a random intercept is included in the models above. Here, in the univariable generalised linear mixed effects model (GLMM), the model for the outcome of patient j from cluster i is given by

$$Y_{ij}|b_i \sim \text{Bernoulli}(\pi_{ij})$$

where $b_i \sim N(0, \sigma_b^2)$. Incorporating this into the crude model (above) gives

$$\text{logit}(\pi_{ij}) = \beta_0 + b_i + \beta_1 A_{ij}$$

and categorical variables were included as

$$\text{logit}(\pi_{ij}) = \beta_0 + b_i + \sum_{k=1}^K \beta_k C_{ijk}.$$

Thus, the multivariable GLMM is defined as

$$\text{logit}(\pi_{ij}) = \beta_0 + b_i + \beta_1 A_j + \beta_2 G_j + \beta_3 E_j + \sum_{k=1}^4 \beta_{4k} D_{jk} + \sum_{k=1}^2 \beta_{5k} C_{jk} + \beta_6 R_j.$$

The levels of the explanatory variables are as defined earlier. The need for the random effect was tested using a mixture of chi-square tests with 0 and 1 degrees of freedom.

The estimate of the random effect, obtained upon model fitting, are called empirical Bayes (EB) estimates. The EB estimates are usually graphed to identify clusters which are outlying in terms of the success probability of the binary outcome. The EB estimates can be grouped, for example, by identifying clusters (e.g. Clinical Commissioning Groups) that have outlying probabilities of emergency route to cancer diagnosis. The estimates can also be grouped by characteristics that were not included in the model (e.g. population density within that cluster) could identify a common characteristic of certain clusters that has an impact on the patient's outcome.

GLMM can be implemented using the commands *gllamm* and *xtlogit* in Stata, or using the *glmer* function of the *lme4* package in R software.

3.4 Missing data analysis

Missing data patterns

Patterns of missing data are useful to identify information that could be overlooked. There are two types of missing data patterns:

1. **Monotone.** Missing data follows a monotone pattern if the (fully- and partially-observed) variables can be reordered such that, for every unit i and variable j (where $j = 1, 2, \dots, p$), if unit i is *observed* on variable j , it is observed on all variables $j' < j$.
2. **Non-monotone.** Fully- and partially-observed variables in a data set that cannot be ordered in a monotone pattern are considered non-monotone missing data pattern.

Monotone missing data patterns often occur in longitudinal analyses. A monotone missing data pattern would occur when, for example, a patient regularly attends a clinic and their health status (e.g. blood pressure, heart rate, vital status) is recorded but for a particular reason does not attend subsequent appointments. Non-monotone missing data patterns would occur in a longitudinal study if a patient attends at least one appointment after having missed at least one previous appointment. Monotone missing data patterns require less complex methods to handle the missing data in comparison to data with a non-monotone missing data pattern. Monotone patterns can also occur in cross-sectional studies. For example, for a given set of patients with fully-observed age and gender records, data can be rearranged to create a monotone missing data pattern (Table 13). In this example, ethnicity is missing when route and stage is missing (1% of cases), it is also missing when stage is missing (4% of cases), and occasionally missing when all other variables are observed (5% of cases).

Table 13: Monotone missing data pattern

Pattern	Variable					No.	% of total
	Age	Gender	Route	Stage	Ethnicity		
1	X	X	X	X	X	900	90%
2	X	X	X	X	.	50	5%
3	X	X	X	.	.	40	4%
4	X	X	.	.	.	10	1%

However, in cross-sectional studies, particular care is required in clarifying the pattern. Unlike in longitudinal studies, cross-sectional studies are not temporal and the rearrange-

ment of variables are not constrained by the time point at which a patient attends a clinic. Continuing with the example above, another partially-observed variable *ethnicity* is introduced. However, the behaviour of missing data within this variable has resulted in a non-monotone missing data pattern (Table 14). The difference here is that, in pattern 3, ethnicity was fully-observed for some patients who had a missing stage at diagnosis, and these patterns cannot be rearranged into a monotone missing data pattern.

Table 14: Non-monotone missing data pattern

Pattern	Variable					No.	% of total
	Age	Gender	Route	Stage	Ethnicity		
1	X	X	X	X	X	900	90%
2	X	X	X	X	.	50	5%
3	X	X	X	.	X	40	4%
4	X	X	.	.	.	10	1%

3.4.1 Missing data mechanisms

Decisions on how to handle missing data are directed by the assumptions of how the missing data occurs. Making the wrong assumption about the mechanism of the missing data can lead to bias in inferences. This is in addition to the potential loss of efficiency in standard errors. "The extent of information loss is not directly linked to the proportion of incomplete records. Instead it is intrinsically linked to the analysis question".^{22,23} Data can be missing due to three different mechanisms:

1. Missing completely at random (MCAR)
2. Missing at random (MAR)
3. Missing not at random (MNAR)

A complete case analysis is an analysis including only those patients whose variables are observed (i.e., patients are not included if their observations on variables are missing). A complete case analysis assumes that data are MCAR. MCAR assumes that the probability that a value of a variable is missing is neither dependent on the unobserved value itself, nor on the observed data in the other variables for that patient. While a complete case analysis would be unbiased if the data are MCAR, it would nevertheless be potentially inefficient.

An analysis under MAR assumes that the probability that a value is missing depends on the observed data in the patient's other variables, and, given these observed data, the probability does not additionally depend on the unobserved value itself. An analysis under MNAR, even given the observed data in the patient's other variables (i.e., even conditioning on these), the probability of a missing value depends on the unobserved value itself.

More formally, for $Y_{i,j}$ where $j = 1, 2, \dots, p$ variables for individual $i = 1, 2, \dots, n$, let $R_{i,j} = 1$ if the value of $Y_{i,j}$ is observed, and $R_{i,j} = 0$ if the value of $Y_{i,j}$ is missing. The missing data mechanism is defined as

$$P(\mathbf{R}_i | \mathbf{Y}_i).$$

This is the probability of observing individual i 's data, given the set of variables \mathbf{Y}_i .

Missing completely at random

Data is MCAR if the probability of a value being missing is independent of the observed and missing data of that individual. More formally,

$$P(\mathbf{R}_i | \mathbf{Y}_i) = P(\mathbf{R}_i).$$

If the data is MCAR, the data is representative of the underlying population from which the sample is taken, and inferences are unbiased. However, due to the loss of data, the standard errors, and consequently the confidence intervals, are potentially wider (i.e., loss of efficiency).

Missing at random

Data is MAR if the probability distribution of \mathbf{R}_i is independent of the missing data, given the observed data for that individual. Observed data on an individual is defined as $\mathbf{Y}_{i,O}$, and missing data as $\mathbf{Y}_{i,M}$. More formally,

$$P(\mathbf{R}_i | \mathbf{Y}_i) = P(\mathbf{R}_i | \mathbf{Y}_{i,O}).$$

Under this assumption, the probability of a value being missing for an individual is independent of the missing value itself given the observed data. The dependency of the probability distribution on the missing value is removed conditional on the observed data for that individual. The MAR assumption can be further split into two scenarios: covariate-dependent missing at random (CMAR), and outcome-dependent missing at random (OMAR).

Covariate-dependent missing at random

Under covariate-dependent missing at random (CMAR), the probability of a value being missing is independent of the missing value itself given the observed data of the *covariates*. That is to say the outcome variable holds no further information as to the probability that a value is missing. More formally, assume there is a single outcome variable, let $\mathbf{Y}_{i,O} = (\mathbf{Y}_{i,O}^C, Y_{i,O}^Y)$, where $\mathbf{Y}_{i,O}^C$ is the vector of covariate-only variables, and $Y_{i,O}^Y$ is outcome variable, then

$$P(\mathbf{R}_i|\mathbf{Y}_i) = P(\mathbf{R}_i|\mathbf{Y}_{i,O}^C).$$

Inferences drawn from variables that are CMAR are unbiased in a complete-case analysis, but are potentially inefficient.

Outcome-dependent missing at random

Under outcome-dependent missing at random (OMAR), the probability of a value being missing is independent of the missing value itself given the observed covariates *and* the outcome variable. More formally,

$$P(\mathbf{R}_i|\mathbf{Y}_i) = P(\mathbf{R}_i|\mathbf{Y}_{i,O}^C, Y_{i,O}^Y) = P(\mathbf{R}_i|\mathbf{Y}_{i,O}).$$

Inferences drawn from variables that are OMAR are biased in a complete-case analysis and potentially inefficient.

Missing not at random

MNAR assumes that the probability of a value being missing depends on the value of the missing observation itself, and the dependence remains even after conditioning on $\mathbf{Y}_{i,O}$. More formally,

$$P(\mathbf{R}_i|\mathbf{Y}_i) \neq P(\mathbf{R}_i|\mathbf{Y}_{i,O}).$$

Inference under MNAR assumption requires explicit specification of the joint distribution of \mathbf{Y}_i and \mathbf{R}_i

$$P(\mathbf{R}_i|\mathbf{Y}_i)P(\mathbf{Y}_i) = P(\mathbf{R}_i, \mathbf{Y}_i) = P(\mathbf{Y}_i|\mathbf{R}_i)P(\mathbf{R}_i).$$

The left-hand side is the selection model (also known as the selection model factorisation of the joint distribution), and the right-hand side is the pattern mixture model (also known as the pattern mixture factorisation of the joint distribution). Both models can be used to infer the other, and to specify the MNAR mechanism. Inferences drawn from variables

that are MNAR after using certain methods to deal with missing values (i.e. multiple imputation conducted under MAR) may lead to biased results.²⁴

Exploring the missing data mechanism

The missingness mechanism is, at best, an assumption. However, the observed data can be used to form assumptions by (i) exploring the missing data pattern, and (ii) applying logistic regression analyses of \mathbf{R}_i on observed and near-fully observed variables. Investigating the probability of missing data will give an indication, and highlight evidence, for or against a possible missing data mechanism.

Focusing on the case of missing data in explanatory variables, the framework to explore the probability of a missing value for a set of partially-observed variables, \mathbf{V} , is in general defined as follows. For each partially-observed variable, define an indicator variable, R , which takes the value 1 if the value of that variable is missing, and 0 otherwise. For any given individual, the probability of a missing value within a partially-observed variable V is given by

$$\text{logit}[P(R_{V_i} = 1)] = \beta_0 + \beta_1 Y_i + \mathbf{X}_i^* \boldsymbol{\beta}^*$$

where Y_i is the outcome variable of the substantive analysis, and \mathbf{X}_i^* is the vector of fully and near-fully observed covariables (excluding variable V). In survival analysis settings, it is customary to include, in this logistic regression model, the follow-up time, T , and the vital status, δ .

3.4.2 Multiple imputation

Imputation model

The imputation model should contain all the variable in the substantive model, including the outcome variable. It should also reflect/take into account any non-linear, time-dependent and interaction effects, and any auxiliary variables satisfying the following conditions:²²

1. To reduce the bias from complete case analysis include variables predictive of the probability of missing values and the underlying missing value itself
2. To improve efficiency include variables predictive of only the underlying missing value itself

In the survival analysis settings, the structure of the imputation model includes the outcome defined by the Nelson-Aalen estimate of the cumulative hazard, $H(T)$, where $T \in \mathbb{R}$

$0 \leq T \leq 1$; and the vital status indicator, δ , taking values 1 if the patient has died and 0 otherwise.²⁵

As with any model with several binary or a categorical variables, perfect prediction issues may arise. Perfect predictions results in unstable parameter estimates and inflated standard errors.²² To overcome this issue, a good starting point is to seek if possible a sensible reduction in the number of categories through regrouping.

General outline

For simplicity, let $\mathbf{Y}_{i,j} = (\mathbf{Y}_{i,1}, \mathbf{Y}_{i,2})$, that is two continuous variables, where \mathbf{Y}_2 (a covariate) is MAR conditional on \mathbf{Y}_1 (the outcome). The substantive model is the regression of \mathbf{Y}_1 on \mathbf{Y}_2 , which will be biased under a complete case analysis. Since Y_2 is MAR conditional on Y_1 , then a regression of Y_2 on Y_1 (i.e., where Y_2 is used as the outcome) using the complete records is valid.

The imputation procedure is to:

1. Fit the regression of \mathbf{Y}_2 on \mathbf{Y}_1 to the complete data:

$$\mathbf{Y}_{i,2} = \alpha_0 + \beta_1 \mathbf{Y}_{i,1} + e_i, \quad e_i \stackrel{i.i.d.}{\sim} N(0, \sigma_{(2|1)}^2),$$

obtaining $\hat{\alpha}_0$, $\hat{\beta}_1$, and $\hat{\sigma}_{(2|1)}^2$. The regression above is termed an imputation model.

2. Draw multiple times from the distribution of the missing data given the observed, for $k = 1, 2, \dots, K$, taking account of the uncertainty, giving K imputed data sets.
3. Fit the substantive model to the K imputed data sets, obtaining K estimates of $\alpha_{0,k}$, $\hat{\beta}_{1,k}$, and $\hat{\sigma}_{(2|1),k}^2$.
4. Combine the k estimates for inference according to Rubin's rules.²⁶

Rubin's rules to combine the estimates are:

for the estimate of β

$$\hat{\beta}_{MI} = \frac{1}{K} \sum_{k=1}^K \hat{\beta}_k$$

and estimate of the variance is

$$\hat{V}_{MI} = \hat{W} + \left(1 + \frac{1}{K}\right) \hat{B}$$

where

$$\hat{W} = \frac{1}{K} \sum_{k=1}^K \hat{\sigma}_k^2, \quad \text{and} \quad \hat{B} = \frac{1}{K-1} \sum_{k=1}^K (\hat{\beta}_k - \hat{\beta}_{MI})^2.$$

\hat{W} is the within-imputation variance, and \hat{B} is the between-imputation variance.

Imputation procedures

Sequential imputation is an approach to multiple imputation when the missing data pattern is monotone; however, in our case there is a non-monotone pattern and the joint modelling approach can be used with the multivariate normal distribution. The multivariate normal imputation treats discrete variables as continuous: implying the distribution of other variables is conditioned on a linear function of the discrete variables, if these discrete variables are fully observed. In the latent normal approach, the discrete variables are modelled using latent normal variables. These are then jointly modelled together with the continuous variables using the multivariate normal model.²² The latent normal approach extends naturally to the clustered/hierarchical data setting. In this section, multiple imputation is explained, followed by the latent normal approach, and finally, software packages used to impute the missing values.

Joint modelling

Joint modelling makes no assumptions about the missing data pattern but the missingness mechanism is assumed to be MAR. In this framework, the imputation model is the multivariate normal model:

$$\mathbf{Y} \sim N(\boldsymbol{\beta}, \boldsymbol{\Omega}),$$

for a matrix of variables and parameters, respectively,

$$\mathbf{Y} = \begin{pmatrix} Y_{i,1} \\ Y_{i,2} \\ \vdots \\ Y_{i,p} \end{pmatrix} \text{ and } \boldsymbol{\beta} = \begin{pmatrix} \beta_{0,1} \\ \beta_{0,2} \\ \vdots \\ \beta_{0,p} \end{pmatrix},$$

where $\boldsymbol{\Omega}$ is the unstructured covariate matrix. Imputation then proceeds via the Gibbs sampler.²²

There are methods for imputing missing data in continuous variables; however, in this thesis the variables with missing data are all categorical (nominal or ordinal) variables. While categorical variables can be imputed under the multivariate normal model treating them as continuous and then applying various rounding off approaches, an alternative approach is to utilise the latent normals for these categorical variables. Therefore, multiple imputation of binary and ordinal data requires an adaptation of the multivariate normal approach: the latent normal approach, which is described in the next section.

Latent normal joint modelling approach

Considering probit regression of a binary variable Y_i on a constant, then a latent normal variable is defined as $Z_i \sim N(\beta, 1)$, Z_i equivalent to $Y_i = 1$, where β is the coefficient of the probit regression. The latent normal formulation is equivalent to the probit model,

which links in naturally with the multivariate normal imputation model. The latent normal variables can then be jointly modelled with the continuous ones in the multivariate normal model.²²

Multilevel multiple imputation

Often in observational data, patient health outcomes are similar not only based on the patient-level characteristics but on the characteristics of the area in which they reside. This is known as hierarchical studies, and the variables representing the characteristics of how the patients are grouped are known as cluster-level variables. If cluster-level variables are included in a substantive model, then the cluster-level variables should also be taken into account in the imputation model. By including a random intercept u_j (where $j = 1, \dots, J$ clusters), the substantive model could be a multilevel logistic regression model or a multilevel access hazard model (described in previous sections). Following on from the joint modelling approach, random intercepts can be incorporated into the imputation model to form a joint random intercept imputation model.²²

3.4.3 Number of imputations

It has been shown that multiple imputation is highly efficient even for as small as 5 imputations.^{22,27} However, this result, which applies to estimation of the substantive model parameters, does not extend to estimation of p-values. It has therefore been noted that if one wishes the error in estimating the p-values to be small, at least 100 imputations will be required.²² Therefore, apart from large and computationally intensive problems, it is advisable to choose a larger rather than smaller number of imputations. For large and computationally intensive problems, it is sensible to pause after a small number of imputations and check whether inferences are clear cut.

3.4.4 Hypothesis testing after multiple imputation

When the number of imputations is large, inference for β can be conducted using Wald tests. However, this approach requires the number of imputations to be large enough for the normal approximation to hold. For a small number of imputations there are several other approaches:

1. t-test;
2. Approximate F-test; and
3. Likelihood Ratio test.

Firstly, one approach is to use t-tests (an adaptation of the Wald test) that includes an

additional term in the variance. The additional term accounts for the uncertainty due to finite imputations: as the number of imputations tends to infinity, the additional variance becomes negligible.^{26,28} For inferences regarding a vector of parameters in the substantive model, an F-type test is available.^{29,30} Thirdly, an approach was developed as an extension of the multiple imputation combination rules to likelihood ratio statistics so that likelihood ratio tests for parameters in the substantive model could be conducted.³¹ In practicality, if computational time is not a burden, it is advocated to increase the number of imputations such that approximate normal assumptions hold.

3.4.5 Pitfalls

Incompatibility

It is crucial that the imputation model contains all of the variables in the substantive model, and reflects interactions terms, non-linear effects and time-dependent effects contained in the substantive model: failure to account for these can lead to biased estimates and invalid inferences. For example, suppose the substantive model is the linear regression of Y (the dependent variable) on X_1 (partially-observed variable) with fully-observed covariate X_2 . If the substantive model also included an interaction $X_1 * X_2$ then the imputation model should also reflect this interaction. For example, imputing as if no interaction is present, and then computing the interaction after imputation (passive imputation) has been shown to result in attenuated associations. On the other hand, simply computing the interaction and including it in the imputation model (the so-called "just another variable" approach) has been shown to generally give invalid estimates under MAR.²²

One way of accounting for the interaction, if X_2 is categorical, is to impute X_1 separately for each level of X_2 , then for each imputation append the data sets for each level of X_2 before proceeding with analysis of the imputed data sets and combination using Rubin's rules.²²

Another approach is to use so-called substantive model compatible multiple imputation methods. These methods have been proposed recently and they are still in development.^{32,33,34} In these methods, the substantive model is recognised at the imputation stage. Substantive model compatible multiple imputation approaches are available for various substantive models under the full conditional specification approach^{32,33} and the latent normal joint modelling approach,³⁴ but work is still needed to avail these for complex substantive models such as multilevel excess hazard models. Particularly in the excess hazard model context, there is need for methodological and software development work to enable imputation compatibly with, for example, time-dependent effects of incompletely observed variables such as stage.

Sensitivity analysis

Multiple imputation is most commonly performed assuming the missing data is *missing at random*. However, this assumption is untestable, and whether or not this assumption is valid cannot be confirmed. Multiple imputation can be performed under a missing not at random assumption. Results can then be compared to those under MAR to assess the stability of the results, or lack thereof, under the different assumptions about the missing data mechanism. This is referred to as sensitivity analysis.²²

3.4.6 Available software packages

Joint modelling using the *jomo* package does not explicitly distinguish between ordered and unordered categorical variables, which may result in loss of efficiency. The *jomo* package uses the generic latent normal categorical algorithm for the imputation of unordered categorical data, and does not explicitly reflect the order within a categorical variable. However, the general algorithm for imputing ordinal data using the unordered latent normal imputation model results in negligible loss of efficiency.³⁵ From this, results show there is no reason to suggest that using the generic latent normal categorical algorithm will not work well even if the data are truly ordinal.

3.5 Conclusion

This chapter has outlined the various methods used in this thesis to answer each of the aims. The methods described here are used in the following chapters, which summarise the findings from each of the research aims. As will be seen, these methods are incorporated with missing data analysis to provide less biased and potentially more efficient estimates.

References

- [1] Robin Schaffar, Elisabetta Rapiti, Bernard Rachet, and Laura Woods. Accuracy of cause of death data routinely recorded in a population-based cancer registry: impact on cause-specific survival and validation using the Geneva cancer registry. *BMC Cancer*, 13(1):609, 2013. ISSN 1471-2407. doi: 10.1186/1471-2407-13-609. URL <https://doi.org/10.1186/1471-2407-13-609>.
- [2] Aurélien Belot, Aminata Ndiaye, Miguel-Angel Luque-Fernandez, Dimitra-Kleio Kipourou, Camille Maringe, Francisco Javier Rubio, and Bernard Rachet. Summarizing and communicating on survival data according to the audience: a tutorial on different measures illustrated with population-based cancer registry data. *Clinical epidemiology*, 11:53–65, 1 2019. ISSN 1179-1349. doi: 10.2147/CLEP.S173523. URL <https://pubmed.ncbi.nlm.nih.gov/30655705><https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6322561/>.
- [3] Maja Pohar-Perme, Stare Janez, and Estève Jacques. On Estimation in Relative Survival. *Biometrics*, 68(1):113–120, 2012. doi: doi:10.1111/j.1541-0420.2011.01640.x. URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1541-0420.2011.01640.x>.
- [4] J Estève, E Benhamou, M Croasdale, and L Raymond. Relative survival and the estimation of net survival: Elements for further discussion. *Stat Med*, 9(5):529–538, 1990. doi: 10.1002/sim.4780090506. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.4780090506>.
- [5] P. Royston, G. Ambler, and W. Sauerbrei. The use of fractional polynomials to model continuous risk variables in epidemiology. *International Journal of Epidemiology*, 28(5):964–974, 10 1999. ISSN 0300-5771. doi: 10.1093/ije/28.5.964. URL <https://academic.oup.com/ije/article-lookup/doi/10.1093/ije/28.5.964>.
- [6] Patrick Royston. Flexible Parametric Alternatives to the Cox Model, and more. *The Stata Journal: Promoting communications on statistics and Stata*, 1(1):1–28, 11 2001. ISSN 1536-867X. doi: 10.1177/1536867X0100100101. URL <http://journals.sagepub.com/doi/10.1177/1536867X0100100101>.
- [7] Patrick Royston and Mahesh K B Parmar. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics in Medicine*, 21(15): 2175–2197, 8 2002. ISSN 0277-6715. doi: <https://doi.org/10.1002/sim.1203>. URL <https://doi.org/10.1002/sim.1203>.
- [8] Patrick Royston. Flexible Parametric Alternatives to the Cox Model: Update. *The Stata Journal: Promoting communications on statistics and Stata*, 4(1):98–101, 3 2004. ISSN 1536-867X. doi: 10.1177/1536867X0100400112. URL <http://journals.sagepub.com/doi/10.1177/1536867X0100400112>.

- [9] R Giorgi, M Abrahamowicz, C Quantin, P Bolard, J Esteve, J Gouvernet, and J Faivre. A relative survival regression model using B-spline functions to model non-proportional hazards. *Stat Med*, 22(17):2767–2784, 2003. doi: 10.1002/sim.1484.
- [10] Laurent Remontet, Zoé Uhry, Nadine Bossard, Jean Iwaz, Aurélien Belot, Coraline Danieli, Hadrien Charvat, and Laurent Roche. Flexible and structured survival model for a simultaneous estimation of non-linear and non-proportional effects and complex interactions between continuous variables: Performance of this multidimensional penalized spline approach in net survival trend analysis. *Statistical Methods in Medical Research*, 28(8):2368–2384, 8 2019. ISSN 14770334. doi: 10.1177/0962280218779408. URL <http://journals.sagepub.com/doi/10.1177/0962280218779408>.
- [11] H Charvat, L Remontet, N Bossard, L Roche, O Dejardin, B Rachet, G Launoy, and A Belot. A multilevel excess hazard model to estimate net survival on hierarchical data allowing for non-linear and non-proportional effects of covariates. *Stat Med*, 35(18):3066–3084, 2016. doi: 10.1002/sim.6881.
- [12] D.B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974.
- [13] Donald B Rubin. Causal Inference Using Potential Outcomes. *Journal of the American Statistical Association*, 100(469):322–331, 2005. URL <http://dx.doi.org/10.1198/016214504000001880>.
- [14] D G Kleinbaum. Survival Analysis, a Self-Learning Text. *Biometrical Journal*, 40(1):107–108, 4 1998. ISSN 0323-3847. doi: [https://doi.org/10.1002/\(SICI\)1521-4036\(199804\)40:1<107::AID-BIMJ107>3.0.CO;2-9](https://doi.org/10.1002/(SICI)1521-4036(199804)40:1<107::AID-BIMJ107>3.0.CO;2-9). URL [https://doi.org/10.1002/\(SICI\)1521-4036\(199804\)40:1%3C107::AID-BIMJ107%3E3.0.COhttp://2-9](https://doi.org/10.1002/(SICI)1521-4036(199804)40:1%3C107::AID-BIMJ107%3E3.0.COhttp://2-9).
- [15] R M Daniel, S N Cousens, B L De Stavola, M G Kenward, and J A C Sterne. Methods for dealing with time-dependent confounding. *Statistics in Medicine*, 32(9):1584–1618, 4 2013. ISSN 0277-6715. doi: <https://doi.org/10.1002/sim.5686>. URL <https://doi.org/10.1002/sim.5686>.
- [16] Arvid Sjölander. Regression standardization with the R package stdReg. *European Journal of Epidemiology*, 31(6):563–574, 2016. ISSN 1573-7284. doi: 10.1007/s10654-016-0157-3. URL <https://doi.org/10.1007/s10654-016-0157-3>.
- [17] K Rothman, S Greenland, and T L Lash. *Modern Epidemiology*. Lippincott, Williams & Wilkins, 3rd edition, 2008.
- [18] Paul C Lambert, Sally R Wilkes, and Michael J Crowther. Flexible parametric modelling of the cause-specific cumulative incidence function. *Statistics in Medicine*, 36(9):1429–1446, 4 2017. ISSN 0277-6715. doi: <https://doi.org/10.1002/sim.7208>. URL <https://doi.org/10.1002/sim.7208>.

- [19] Paul Lambert. STPM2_STANDSURV: Stata module to obtain standardized survival curves after fitting an stpm2 survival model, 6 2018. URL <https://econpapers.repec.org/RePEc:boc:bocode:s458289>.
- [20] Michal Abrahamowicz, Todd Mackenzie, and John M Esdaile. Time-Dependent Hazard Ratio: Modeling and Hypothesis Testing with Application in Lupus Nephritis. *Journal of the American Statistical Association*, 91(436):1432–1439, 12 1996. ISSN 0162-1459. doi: 10.1080/01621459.1996.10476711. URL <https://www.tandfonline.com/doi/abs/10.1080/01621459.1996.10476711>.
- [21] Michal Abrahamowicz, Roxane du Berger, and Steven A Graver. Flexible Modeling of the Effects of Serum Cholesterol on Coronary Heart Disease Mortality. *American Journal of Epidemiology*, 145(8):714–729, 4 1997. ISSN 0002-9262. doi: 10.1093/aje/145.8.714. URL <https://doi.org/10.1093/aje/145.8.714>.
- [22] James R. Carpenter and Michael G. Kenward. *Multiple Imputation and Its Application*. John Wiley & Sons, Ltd, 1st edition, 1 2013.
- [23] Paul Madley-Dowd, Rachael Hughes, Kate Tilling, and Jon Heron. The proportion of missing data should not be used to guide decisions on multiple imputation. *Journal of Clinical Epidemiology*, 110:63–73, 2019. ISSN 0895-4356. doi: <https://doi.org/10.1016/j.jclinepi.2019.02.016>. URL <http://www.sciencedirect.com/science/article/pii/S0895435618308710>.
- [24] Rachael A Hughes, Jon Heron, Jonathan A C Sterne, and Kate Tilling. Accounting for missing data in statistical analyses: multiple imputation is not always the answer. *International Journal of Epidemiology*, 48(4):1294–1304, 3 2019. ISSN 0300-5771. doi: 10.1093/ije/dyz032. URL <https://doi.org/10.1093/ije/dyz032>.
- [25] Ian R White and Patrick Royston. Imputing missing covariate values for the Cox model. *Statistics in medicine*, 28(15):1982–1998, 7 2009. ISSN 1097-0258. doi: 10.1002/sim.3618. URL <https://pubmed.ncbi.nlm.nih.gov/19452569https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2998703/>.
- [26] Donald B. Rubin. *Multiple imputation for nonresponse in surveys*. Wiley, New York, 1987. ISBN 978-0-471-65574-9.
- [27] Geert Molenberghs and Geert Verbeke. *Models for Discrete Longitudinal Data*. Springer-Verlag New York, New York, 1 edition, 2005.
- [28] J Barnard and D B Rubin. Miscellanea. Small-sample degrees of freedom with multiple imputation. *Biometrika*, 86(4):948–955, 12 1999. ISSN 0006-3444. doi: 10.1093/biomet/86.4.948. URL <https://doi.org/10.1093/biomet/86.4.948>.
- [29] K H Li, T E Raghunathan, and D B Rubin. Large-Sample Significance Levels from Multiply Imputed Data Using Moment-Based Statistics and an F Refer-

- ence Distribution. *Journal of the American Statistical Association*, 86(416):1065–1073, 12 1991. ISSN 0162-1459. doi: 10.1080/01621459.1991.10475152. URL <https://doi.org/10.1080/01621459.1991.10475152>.
- [30] Jerome P Reiter. Small-sample degrees of freedom for multi-component significance tests with multiple imputation for missing data. *Biometrika*, 94(2):502–508, 6 2007. ISSN 0006-3444. doi: 10.1093/biomet/asm028. URL <https://doi.org/10.1093/biomet/asm028>.
- [31] Xiao-Li Meng and Donald B Rubin. Performing Likelihood Ratio Tests with Multiply-Imputed Data Sets. *Biometrika*, 79(1):103–111, 1 1992. ISSN 00063444. doi: 10.2307/2337151. URL <http://www.jstor.org/stable/2337151>.
- [32] Jonathan W. Bartlett, Shaun R. Seaman, Ian R. White, and James R. Carpenter. Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model. *Statistical Methods in Medical Research*, 24(4):462–487, 8 2015. ISSN 14770334. doi: 10.1177/0962280214521348. URL <http://journals.sagepub.com/doi/10.1177/0962280214521348>.
- [33] Jonathan W. Bartlett and Tim P. Morris. Multiple imputation of covariates by substantive-model compatible fully conditional specification. *Stata Journal*, 15(2):437–456, 6 2015. ISSN 15368734. doi: 10.1177/1536867x1501500206.
- [34] Matteo Quartagno and James R Carpenter. *Multilevel Multiple Imputation in presence of interactions, non-linearities and random slopes*. PhD thesis, London School of Hygiene and Tropical Medicine, 2018.
- [35] Matteo Quartagno and James R Carpenter. Multiple imputation for discrete data: Evaluation of the joint latent normal model. *Biometrical journal. Biometrische Zeitschrift*, 61(4):1003–1019, 7 2019. ISSN 1521-4036. doi: 10.1002/bimj.201800222. URL <https://www.ncbi.nlm.nih.gov/pubmed/30868652><https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6618333/>.

4 Patient characteristics and survival

This chapter, and the following chapters (i.e., 4.2, 5.1, 5.2), answer the aims and objectives of the thesis using, and applying, the material and methods outlined in Chapters 2 and 3. Firstly, a description of survival amongst patient characteristics is provided (Chapter 4.1), followed by the association between survival and patient characteristics (Chapter 4.2), then an evaluation of the comorbidity, and deprivation, gaps in short-term survival (Chapter 5.1), and finally an investigation into the variation in access to the healthcare system between comorbidity status (Chapter 5.2).

4.1 Differences in survival by patient characteristics

4.1.1 Introduction

The first aim of the thesis was to contribute to research around the description in survival of patients with non-Hodgkin lymphoma (NHL). The objectives were to (i) estimate the survival of patients with diffuse large B-cell lymphoma or follicular lymphoma by patient characteristics (i.e., comorbidity status, deprivation level, and gender), and (ii) compare 5-year survival estimates between patient characteristics with a focus on comorbidity status and deprivation level.

Here, the scene is set to measure the current performance of the healthcare system by calculating crude age-standardised 5-year net survival. The structure of this subsection first provides a brief overview of the background and methods used, followed by the results, and finally discusses the main findings along with further research. See Appendix A.5.1 for explicit details on the findings of the original research.

4.1.2 Overview of paper

Background

To assess the performance of a healthcare system in managing cancer patient outcomes within a country, comparisons of cancer survival can be made to other countries based on socioeconomic and demographic characteristics. In 2000, the National Health Service (NHS) Cancer Plan,¹ intended, at the time, to compare current cancer survival estimates to other universal healthcare systems in Europe: survival in England was below average. Furthermore, the sociodemographic gap in cancer survival was wider in England and more efforts were made to reduce the gaps. The Cancer Reform Strategy (2007),² hypothesised that the gaps could be explained by the prevalence of comorbid conditions (such as chronic obstructive pulmonary disease, or congestive heart failure). Targets were created

by Cancer Research UK (CRUK), such as two-thirds survival at 5-years since diagnosis;³ also, the National Cancer Equality Initiative was constructed to investigate and improve the inequalities in cancer survival.⁴ Of the most commonly diagnosed cancers, survival of patients with non-Hodgkin lymphoma are lower in England and the deprivation gap in survival persists.

Non-Hodgkin lymphoma (NHL) is a heterogeneous group of malignancies with varying morphology and topography.^{5,6} Diffuse large B-cell lymphoma (DLBCL) and follicular lymphoma (FL) are the two most common subtypes of NHL.⁷ In 2018, the age-standardised incidence rate of NHL in the United Kingdom (UK) was the second highest in the world (12 cases per 100,000 individuals),⁸ This is expected to increase as the world, and England, is experiencing an increase in life expectancy. Moreover, survival of patients differ by the behaviour of the subtypes,⁹ and the survival amongst patients with shared characteristics, with the same subtype, is unclear. An aging population is associated with a greater incidence of comorbid conditions, which is indicative of worse health outcomes. Previous research has shown the socioeconomic-gap in survival in England widened amongst certain patient characteristics, yet little is known regarding the impact of comorbid conditions on cancer survival.^{5,10,11,12}

Recent population-based cancer registry data linked to various population-based health records can provide an in-depth perception of the performance of healthcare systems, measured by 5-year cancer survival estimates, including for patients with NHL subtypes who have underlying comorbid conditions. The aim of this section is to provide a description of cancer survival by patient characteristics and to describe deprivation, and comorbidity, gaps in survival.

Methods

As detailed in Chapter 2, the data contains information on all patients diagnosed with diffuse large B-cell lymphoma (DLBCL) or follicular lymphoma (FL) in England between 2005 and 2013 (with follow up to 2015), which was linked to population-based health records. Overall, 30,274 and 15,583 patients were diagnosed with DLBCL and FL, respectively. The Pohar Perme method of net survival (Chapter 3.1.2) was used to estimate 5-year net survival for patients with shared characteristics.¹³ The cohort approach was used for patients diagnosed between 2005-2010, and the hybrid approach¹⁴ was used for survival in 2014-2015: the hybrid approach uses survival information from patients diagnosed in 2012-2013. Survival estimates were standardised to the International Cancer Survival Standard weights for group 1.¹⁵ Survival was estimated for gender, socioeconomic status and comorbidity status.

Results

Overall, the average age at diagnosis was 67 and 64 years for DLBCL and FL, respectively.

The proportion of patients with comorbidity was 11% and 8% amongst DLBCL and FL, respectively. For both subtypes, there was a decreasing trend in the proportion of cancer diagnoses for each increase in deprivation level. The prevalence of comorbidity increased for each increase in deprivation level.

Net survival is one of the key measures of the performance of a healthcare system. Here, the non-parametric estimates of net survival give us some ideas about the current performance of the National Health Service in managing patients with cancer.

In particular, it shows that those living in more deprived areas had worse 5-year survival compared to less deprived areas and that, over time, this deprivation gap in survival did not change for patients diagnosed with DLBCL but slightly narrowed amongst those with FL (from 6.6% to 5.2%). Similarly, the gap in survival between comorbidity groups narrowed more among patients with FL than among those with DLBCL. For both DLBCL and FL, survival was also generally higher amongst females and, over time, females had a greater improvement in survival compared to males: the gender-gap in survival widened, with females having greater survival. This could be explained by the fact that females with comorbidities had slightly greater improvements in survival than males with comorbidities.

As hypothesised, for both DLBCL and FL, patients living in more deprived areas had worse survival if they also had any underlying comorbid conditions. For both DLBCL and FL, there were large deprivation gaps in survival for those with none or severe comorbidities, and a negligible deprivation gap amongst those with mild comorbidity status.

Discussion

Age-standardised 5-year net survival for patients with DLBCL or FL has improved over time, and improved more for females, those living in more deprived areas, or any form of comorbidities. The socioeconomic gap in survival narrowed for patients with FL but did not narrow for patients with DLBCL. Patients with comorbidities and living in more deprived areas experienced worse survival compared to others. For DLBCL, the comorbidity-gap in survival narrowed for patients in least deprived areas but not for those in most deprived areas; for FL, the comorbidity-gap in survival narrows for all patients, regardless of socioeconomic status.

This paper showed net survival of patients diagnosed with DLBCL between 2005-2010 was around 59%, and in comparison to other European countries, between 2006-2008, net survival in Northern Europe at 64% was better than England.¹⁶ During 2012-2013, this paper showed survival had greatly increased in England to around 72% but the inequalities in survival between socioeconomic groups and comorbidity status did not narrow. The increase over time is likely due to improved effectiveness of treatments, such as rituximab; thus, any increase in England is also expected in other European countries. The inequality gaps in survival are a measure of the healthcare system's ability to improve outcomes for

patients; however, since these gaps have not narrowed, it is expected that the survival of patients with NHL is still not comparable to the best in Europe and greater focus is needed to reduce the socioeconomic inequalities in survival in England.

4.1.3 Conclusion

This section has provided a glimpse of the current inequalities in survival of patients with diffuse large B-cell lymphoma or follicular lymphoma. The survival probability estimates are only part of the story to locate, measure and reduce the inequalities. Not only is there a deprivation gap in survival, but also a comorbidity gap. The next challenge, in section 4.2, will be to assess the association between patient characteristics, the healthcare pathway, and survival of these patients by using an excess hazard model; thus, this procedure to estimate net survival (via the excess hazard) will incorporate missing data analysis.

4.2 Association between patient characteristics and survival

4.2.1 Introduction

The second aim was to quantify the association between patient characteristics and survival from non-Hodgkin lymphoma in England. The objectives were to (i) build an excess mortality hazard model adjusting for patient characteristics, (ii) incorporate parameters to estimate non-linear and time-dependent effects of patient characteristics, and (iii) expand the model to incorporate correlation in outcomes between patients. In this section an alternative method was used to measure net survival and estimate the association between patient characteristics and survival, whilst accounting for healthcare-level characteristics. A brief overview is provided of the background, methods, main results, and a discussion of the main findings along with further research: the work outlined resulted in a paper currently under review (see Appendix A.5.2).

4.2.2 Overview of paper

Background

Patients living in more deprived areas or with co/multimorbidities are expected to have lower chances of 5-year survival as shown in 4.1, and previous research.^{10,7} Also, patients with a combination of deprivation and comorbidities were expected to have worse survival. Over time, survival generally improved, particularly for those expected to have worse survival, and the socioeconomic gap in survival narrowed for some but still remained.

In 4.1, stratification methods were used so it was not possible to investigate survival for the combination of more than two variables, nor was it possible to estimate associations. A natural approach is to build a parametric model that adjusts for multiple variables, which can include complex associations (e.g., non-linear, time-dependent, and interaction terms) and hierarchical effects.

Previous studies have investigated the association for other cancers and found that comorbidity explains little of the socioeconomic-gap in cancer survival, even after accounting for the healthcare pathway.¹⁷ Although, these results cannot be generalisable to NHL as each cancer has a unique diagnostic process in which inequalities can occur. Furthermore, the healthcare pathway is a crucial step during the cancer diagnostic investigation to achieve optimal prognosis. Previous research has shown that healthcare access could partly explain the inequalities.¹⁸ The healthcare pathway (route to, and stage at, diagnosis) will need to be accounted for. To date, it is unclear what association there is between each patient characteristics, the healthcare pathway and survival from NHL.

The survival of patients within certain clusters, defined by lower super output areas

(LSOA), is expected to be more similar to patients from clusters with similar characteristics. Since socioeconomic status is derived from cluster-level variables there may also be elements of the cluster (in which patients reside) that are not taken into account when measuring deprivation indices.¹⁹ If the Indices of Multiple Deprivation (IMD) accurately measures socioeconomic status, then the cluster-level variable of the cancer patient should not show variation in health outcomes. Therefore, this study incorporates the hierarchical nature of the association between patient health outcomes (i.e., chance of death).

The aim of this study is estimate the association between patient characteristics, the healthcare pathway, and survival of patients with non-Hodgkin lymphoma in England, and to provide population-based evidence of explanations in survival inequalities. The objectives are to model the excess hazard of death due to NHL, accounting for complex associations (e.g., non-linear, time-dependent, and effect modification) and hierarchical effects, and to estimate the socioeconomic inequalities in 5-year net survival by comorbidity status.

Methods

As detailed in Chapter 2, this study contains information on patients diagnosed with DLBCL or FL in England between 2005 and 2013 with information on their follow-up to 2015. Overall, 29,898 and 15,516 patients were diagnosed with DLBCL or FL, respectively, between the ages of 18 and 94. Patients aged 95 or older were removed because 5-year survival estimates are not compatible with life tables that give expected mortality up to 99 years old. It is possible to generate life tables up to age 105, however net survival assumes that the patient dies only from the studied cancer. This assumption is not meaningful at the oldest ages as, in the population, the probability to die from causes other than the cancer is high.

A multilevel excess hazard regression model (Chapter 3.1.4) was used to estimate 5-year net survival for patients with shared characteristics, accounting for clustering due to lower-super output area (LSOA). This multilevel model formed the base of the substantive model used to combine estimates after multiple imputation of variables with missing data (Chapter 3.4). The imputation model included all variables in the substantive model and auxiliary variables (vital status and Nelson-Aalen estimate of the cumulative hazard²⁰) to increase information on variables with missing data. I imputed 10 data sets, estimated the parameters using the substantive model for each imputed data set and combined results using Rubin's rules.^{21,22} Tests for the overall effect of age were done using the F-based procedure for the test of multiple parameter: the code for this is supplied in the Appendix A.6.²³

Results

By building an excess hazard regression model, the association between patient character-

istics and the excess mortality hazard, adjusted for other factors, could be obtained along with the excess mortality hazard ratio (EMHR).

As expected from conclusions in 4.1, amongst DLBCL and FL, respectively, those with comorbidities had 1.23 (95% Confidence Interval -CI-: 1.14-1.32) and 1.52 (95% CI 1.25-1.84) time higher excess mortality hazard compared to those without comorbidities. Patients in most deprived areas showed 1.22 (95% CI 1.18-1.27) and 1.45 (95% CI 1.30-1.62) times higher excess mortality hazard compared to those in least deprived areas. For both DLBCL and FL, those diagnosed through an emergency route had approximately three times higher excess mortality compared to those diagnosed through a non-emergency route.

Although the excess hazard fluctuates by age and time since diagnosis, short-term excess mortality was indicative of long-term survival. Non-linear and time-dependent effects were measured in the excess hazard model and the EMHR was shown to be higher immediately, and at 5-years, after cancer diagnosis for all ages. Within the first 6 months after diagnosis, the EMHR of older and younger patients was markedly different compared to those 70 years old, and is more similar at 2 years since diagnosis before the EMHR then differs again over a longer follow-up time (i.e., at 5-years). Regardless of the comorbidity status, the deprivation-gap in survival was apparent from approximately 3 months after diagnosis.

Discussion

After adjusting for comorbidity status, there remains a higher excess mortality hazard (EMH) amongst patients living in more deprived areas in England. There was a higher EMH amongst patients with comorbidities or multimorbidities, adjusting for age, deprivation level, ethnicity, route to diagnosis and accounting for patient's area of residence. Also, for all ages, there was an increasing EMH within the first year since diagnosis and was highest amongst those of older ages. This suggests opportunities to reduce socioeconomic inequalities that are not accounted for by comorbidity status, particularly within the first year after diagnosis.

The inequalities in survival that remain amongst patients with DLBCL may be partly explained by treatment allocation and cardiotoxicity. As recommended by national guidelines, the first-line treatment for DLBCL are immunochemotherapies often used as a chemotherapy cocktail that includes doxorubicin. An intensive treatment plan of doxorubicin is known to increase the risk of cardiotoxicity leading to cardiac-specific adverse events (e.g., congestive heart failure). Thus, clinicians of patients with underlying cardiac-related comorbid conditions may encourage patients to undergo a less intensive treatment plan. Therefore, this highlights the need for better monitoring of cardiac-related comorbid conditions to reduce the impact on the choice of treatment allocation.

Although socioeconomic status could be considered dynamic, such that individuals can increase or decrease their deprivation level, comparisons of area-level measures of deprivation

over time show that the deprivation level of a particular area changes only slightly, possibly due to the measure being relative rather than absolute. Therefore, given that underlying deprivation levels are not expected to change, there is an opportunity for the healthcare system to adapt and more accurately target areas of more severe deprivation.

This study also showed that expedited diagnostic route (i.e., two-week-wait) for patients suspected of advanced cancer is at least as beneficial as a standard diagnostic route (i.e., general practitioner referral) for patients with unexplained symptoms possibly related to cancer. The two-week-wait referral pathway is recommended for patients suspected to have advanced cancer and worse survival compared to those with milder symptoms; thus, the results suggests that the TWW pathway is performing better than general practitioner referral because patients referred via TWW pathway are more likely to have more severe symptoms.

4.2.3 Conclusion

The results showed that socioeconomic inequalities in survival persist, after adjusting for patient and healthcare pathway characteristics. Comorbidity status (i.e., comorbidity and multimorbidity) is associated with a higher excess mortality hazard compared to those without comorbidity. Even though comorbidity status only partly explains the socioeconomic inequalities in survival, further research into the effect of comorbidity status on survival is important for clinical reasons; for example, the effect of comorbidity on diagnostic delay, treatment allocation, and short- and long-term survival.

It has been shown that long-term survival estimates, if conditional on short-term survival, are comparable between England and other European countries. In Chapter 5.1, the socioeconomic inequalities in survival within the first year since diagnosis are assessed and the combination of comorbidity and deprivation on the hazard of death is measured. Thus, any differences found in short-term mortality between England and European countries may explain the differences in long-term outcomes (i.e., 5-year survival).

References

- [1] Department of Health. The NHS Plan: a plan for action, a plan for reform. Technical report, Department of Health, 2000. URL www.doh.gov.uk/nhsplan.
- [2] Department of Health and Social Care. NHS Cancer Reform Strategy. Technical report, London, 2007. URL <https://www.nhs.uk/NHSEngland/NSF/Documents/CancerReformStrategy.pdf>.
- [3] Cancer Research UK. Our strategy to beat cancer sooner. Technical report, 2014. URL <https://www.cancerresearchuk.org/about-us/our-organisation/our-strategy-to-beat-cancer-sooner>.
- [4] National Cancer Equality Initiative. Cancer Equalities. Technical report, National Cancer Intelligence Network, 2015. URL http://www.ncin.org.uk/cancer_type_and_topic_specific_work/topic_specific_work/equality.
- [5] B Rachet, E Mitry, A Shah, N Cooper, and M P Coleman. Survival from non-Hodgkin lymphoma in England and Wales up to 2001. *British Journal of Cancer*, 99(Suppl 1):S104–S106, 2008. doi: 10.1038/sj.bjc.6604605. URL <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2557528/http://www.nature.com/bjc/journal/v99/n1s/pdf/6604605a.pdf><https://www.nature.com/articles/6604605.pdf>.
- [6] Manuela Quaresma, Michel P Coleman, and Bernard Rachet. 40-year trends in an index of survival for all cancers combined and survival adjusted for age and sex for each cancer in England and Wales, 1971–2011: a population-based study. *The Lancet*, 385(9974):1206–1218, 2015. doi: 10.1016/s0140-6736(14)61396-9.
- [7] A Smith, S Crouch, S Lax, J Li, D Painter, D Howell, R Patmore, A Jack, and E Roman. Lymphoma incidence, survival and prevalence 2004–2014: sub-type analyses from the UK’s Haematological Malignancy Research Network. *Br J Cancer*, 112(9):1575–1584, 2015. doi: 10.1038/bjc.2015.94. URL <http://dx.doi.org/10.1038/bjc.2015.94><https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4453686/pdf/bjc201594a.pdf>.
- [8] World Health Organisation. International Agency for Research on Cancer, 2018. URL <http://gco.iarc.fr/today/home>.
- [9] D J van Spronsen, M L Janssen-Heijnen, V E Lemmens, W G Peters, and J W Coebergh. Independent prognostic effect of co-morbidity in lymphoma patients: results of the population-based Eindhoven Cancer Registry. *Eur J Cancer*, 41(7):1051–1057, 2005. doi: 10.1016/j.ejca.2005.01.010. URL <https://www.ncbi.nlm.nih.gov/pubmed/15862755>.
- [10] A Smith, S Crouch, D Howell, C Burton, R Patmore, and E Roman. Impact of age and socioeconomic status on treatment and survival from aggressive

- lymphoma: a UK population-based study of diffuse large B-cell lymphoma. *Cancer Epidemiol*, 39(6):1103–1112, 2015. doi: 10.1016/j.canep.2015.08.015. URL http://ac.els-cdn.com/S1877782115001794/1-s2.0-S1877782115001794-main.pdf?_tid=f59fbca0-b191-11e6-9962-00000aacb360&acdnat=1479915293_d20c14fe03600c6e72b7975c7adc06c3https://www.sciencedirect.com/science/article/pii/S1877782115001794?via%3Dihub.
- [11] M Sogaard, R W Thomsen, K S Bossen, H T Sorensen, and M Norgaard. The impact of comorbidity on cancer survival: a review. *Clin Epidemiol*, 5(Suppl 1):3–29, 2013. doi: 10.2147/clip.S47150. URL <https://www.dovepress.com/getfile.php?fileID=18023>.
- [12] B Rachet, L Ellis, C Maringe, T Chu, U Nur, M Quaresma, A Shah, S Walters, L Woods, D Forman, and M P Coleman. Socioeconomic inequalities in cancer survival in England after the NHS cancer plan. *Br J Cancer*, 103(4):446–453, 2010. doi: <http://www.nature.com/bjc/journal/v103/n4/supinfo/6605752s1.html>. URL <http://dx.doi.org/10.1038/sj.bjc.6605752https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2939774/pdf/6605752a.pdf>.
- [13] Maja Pohar-Perme, Stare Janez, and Estève Jacques. On Estimation in Relative Survival. *Biometrics*, 68(1):113–120, 2012. doi: doi:10.1111/j.1541-0420.2011.01640.x. URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1541-0420.2011.01640.x>.
- [14] Hermann Brenner and Bernard Rachet. Hybrid analysis for up-to-date long-term survival rates in cancer registries with delayed recording of incident cases. *European Journal of Cancer*, 40(16):2494–2501, 2004. doi: <https://doi.org/10.1016/j.ejca.2004.07.022>. URL <http://www.sciencedirect.com/science/article/pii/S0959804904006082>.
- [15] Isabella Corazziari, Mike Quinn, and Riccardo Capocaccia. Standard cancer patient population for age standardising survival ratios. *European Journal of Cancer*, 40(15):2307–2316, 2004. doi: <https://doi.org/10.1016/j.ejca.2004.07.002>. URL <http://www.sciencedirect.com/science/article/pii/S0959804904005283>.
- [16] M Sant, P Minicozzi, M Mounier, L A Anderson, H Brenner, B Holleccek, R Marcos-Gragera, M Maynadie, A Monnereau, G Osca-Gelis, O Visser, and R De Angelis. Survival for haematological malignancies in Europe between 1997 and 2008 by region and age: results of EURO CARE-5, a population-based study. *Lancet Oncol*, 15(9):931–942, 2014. doi: 10.1016/s1470-2045(14)70282-7. URL http://ac.els-cdn.com/S1470204514702827/1-s2.0-S1470204514702827-main.pdf?_tid=e148ae24-b191-11e6-a7f8-00000aacb35e&acdnat=1479915259_bdcf6be905e2ee773a0aa167c2452264https://www.sciencedirect.com/science/article/pii/S1470204514702827?via%3Dihub.

- [17] Cristina Renzi, Georgios Lyratzopoulos, Willie Hamilton, Camille Maringe, and Bernard Rachet. Contrasting effects of comorbidities on emergency colon cancer diagnosis: a longitudinal data-linkage study in England. *BMC Health Services Research*, 19(1):311, 2019. ISSN 1472-6963. doi: 10.1186/s12913-019-4075-4. URL <https://doi.org/10.1186/s12913-019-4075-4>.
- [18] Alberto Quaglia, Marina Vercelli, Roberto Lillini, Eugenio Mugno, Jan Willem Coebergh, Mike Quinn, Carmen Martinez-Garcia, Riccardo Capocaccia, and Andrea Micheli. Socio-economic factors and health care system characteristics related to cancer survival in the elderly: A population-based analysis in 16 European countries (ELDCARE project). *Critical Reviews in Oncology/Hematology*, 54(2):117–128, 2005. ISSN 1040-8428. doi: <https://doi.org/10.1016/j.critrevonc.2004.12.001>. URL <http://www.sciencedirect.com/science/article/pii/S1040842804002197>.
- [19] L M Woods, B Rachet, and M P Coleman. Origins of socio-economic inequalities in cancer survival: a review. *Annals of Oncology*, 17(1):5–19, 9 2005. ISSN 0923-7534. doi: 10.1093/annonc/mdj007. URL <https://doi.org/10.1093/annonc/mdj007>.
- [20] Milena Falcaro, Ula Nur, Bernard Rachet, and James R. Carpenter. Estimating excess hazard ratios and net survival when covariate data are missing strategies for multiple imputation. *Epidemiology*, 26(3):421–428, 5 2015. ISSN 15315487. doi: 10.1097/eDe.0000000000000283. URL https://journals.lww.com/epidem/Fulltext/2015/05000/Estimating_Excess_Hazard_Ratios_and_Net_Survival.19.aspx.
- [21] R. J. Little and D. B. Rubin. *Statistical Analysis with Missing Data*. John Wiley & Sons, Inc., New York, 1987.
- [22] Donald B. Rubin. *Multiple imputation for nonresponse in surveys*. Wiley, New York, 1987. ISBN 978-0-471-65574-9.
- [23] James R. Carpenter and Michael G. Kenward. *Multiple Imputation and Its Application*. John Wiley & Sons, Ltd, 1st edition, 1 2013.

5 The role of comorbidity on survival

5.1 The role of comorbidity in explaining cancer survival differences

5.1.1 Introduction

The focus of this section answers the third aim: to evaluate the inequalities in short-term mortality amongst patients with non-Hodgkin lymphoma in England. The objectives were to (i) develop a model for the short-term mortality risk standardised to the distribution of patient characteristics, and (ii) predict and compare the cumulative mortality hazard between comorbidity status and deprivation levels. The work here resulted in a paper currently under review (see Appendix A.5.3), and follows on from section 4.2 by shedding light on socioeconomic, and comorbidity, inequalities in survival within a short time interval after cancer diagnosis. The methods used are related to another paper (see Appendix A.5.4) written concurrent to this thesis and published in *Statistics in Medicine* (2021). Here, short-term mortality outcomes are contrasted between levels of socioeconomic status, and comorbidity status, using direct standardisation methods. Note that the results are obtained via causal inference methods; however, they do not have a causal interpretation. A brief overview of the background is provided, followed by the methods used, main results obtained, and a discussion of the main findings along with further research.

5.1.2 Overview of paper

Background

In sections 4.1 and 4.2, where the outcome of interest was 5-year survival, the results showed there were not only socioeconomic inequalities in survival but also inequalities between comorbidity status. However, it was not clear when the inequalities are conceived along the patient and healthcare pathway. For other cancers, research suggests that the differences in 5-year survival estimates between England and other European countries is smaller if estimates were conditional on survival after 1 year since diagnosis; the excess mortality hazard is greater within 1 year since diagnosis in England.^{1,2,3,4} This suggests that sources of long-term survival inequality originate closer to the time at diagnosis. Furthermore, the presence of underlying comorbid conditions will heavily influence the treatment allocation and management of a cancer patient after diagnosis. Evaluating the inequalities in short-term mortality by comorbidity status, and the pathways through which they occur, could provide further insight into how to improve long-term health outcomes.

In section 4.2, an excess hazard model was used as long-term measures of survival are more likely to be affected by competing risks and informative censoring. Short-term measures are less likely to be biased by informative censoring as there is less time for competing

risks to occur; therefore, a patient's mortality is more likely to be due to the cancer, or its treatment, than other causes. Furthermore, the interpretation of parameters from a model that adjusts for other variables can be unclear. For example, a model that estimates the association between age and excess hazard of death, and adjusts for a categorical variable (e.g., gender), would be interpreted as the association for an average gender. Instead, standardisation would provide predictions of the survival estimate for each level of the categorical variables, which can be contrasted to find the difference in the average survival time. Details of this approach are explained in a paper written concurrent to this thesis (see Appendix A.5.4).

This section aims to estimate the inequalities in short-term mortality (of patients diagnosed with NHL in England) by comorbidity status and socioeconomic status. The objectives are to build a flexible parametric survival model to estimate the cumulative hazard of death for each patient, then to standardise the cumulative hazards by patient characteristics, and finally to contrast the cumulative hazard comparing comorbidity and socioeconomic statuses.

Methods

As detailed in Chapter 2, this study contains information on patients diagnosed with DLBCL or FL in England between 2005 and 2013 with information on their follow-up to 2015. Overall, 27,379 and 14,043 patients were diagnosed with DLBCL or FL, respectively, between the ages of 45 and 99. Short-term mortality was first described using 1-year cumulative hazard calculated using the non-parametric Nelson-Aalen estimator.⁵ Flexible parametric survival models (Chapter 3.2) were used to model the non-linear mortality risk with age. The cumulative incidence of death at 1-year since diagnosis was derived for each comorbidity status and socioeconomic status and standardised (Chapter 3.2) to the empirical distribution of the confounders. The standardised cumulative incidence was graphically illustrated for socioeconomic status and comorbidity status.

Results

For DLBCL, 33% of patients died within 1 year after diagnosis. Patients with multimorbidity, had 1.4 times the mortality hazard of those without comorbidity at 1 year since diagnosis (HR: 1.44; CI: 1.34 - 1.55, respectively), standardised to the distribution of gender, deprivation and ethnicity. For FL (where 8% died within 1 year after diagnosis), those with multimorbidity had twice the mortality hazard (HR: 2.17; CI: 1.78 - 2.64, respectively) compared to those without comorbidities. The difference in mortality hazard between DLBCL (HR: 2.2) and FL (HR: 1.4) amongst those with multimorbidities could be because the baseline hazard amongst patients with DLBCL (i.e., those without comorbidities) is higher than the baseline hazard of FL patients.

For both DLBCL and FL, there was a clear increase in mortality hazard amongst more

deprived areas, which was not explained by comorbidity status. Amongst those with DLBCL, the combined effect of multimorbidity and deprivation on mortality hazard at 1 year since diagnosis was 1.9 (CI 1.70 - 2.07) times higher compared to those without comorbidities and living in least deprived areas. For FL, and for the same comparison, the mortality hazard was 3.3 (CI 2.48 - 4.28) times higher at 1 year.

Discussion

This study showed that multimorbidity and deprivation, and their combination, are strong independent predictors short-term mortality amongst patients with DLBCL or FL in England. Also, there was evidence of an increasing trend in short-term mortality with an increase in the deprivation of an area. This suggests that although there are inequalities in survival between patient and healthcare pathway characteristics, the source originates around the pre-, peri-, or immediately post-diagnostic periods. The increase in short-term mortality amongst patients with comorbidities may be explained by the nature and presentation of underlying pre-diagnostic symptoms. Comorbid conditions presenting with symptoms similar to that of DLBCL or FL may hide the underlying cancer and delay the diagnosis, whereas dissimilar symptoms may hasten the diagnosis; thus, this highlights the need to investigate the chances of certain diagnostic routes amongst patients with comorbid conditions. In addition, the comorbid inequalities in health outcomes may be explained by an increased demand on the healthcare service, particularly amongst more populated areas.

5.1.3 Conclusion

Thus far, sections 4.1, 4.2 and this section (5.1) have shown that socioeconomic inequalities in long-term survival are still apparent, they are only partly explained by comorbidity status in combination with other patient and healthcare pathway characteristics, and they originate before or immediately after diagnosis. The following section (5.2) explores the inequalities around the time of diagnosis and investigates the inequalities in diagnostic delay amongst patients with DLBCL or FL in England.

5.2 Impact of comorbidity on patient's access to the healthcare system

5.2.1 Introduction

The focus of this section answers the fourth aim: to investigate the variation in access to the healthcare system amongst patients with non-Hodgkin lymphoma in England. The objectives were to (i) assess the association between diagnostic delay and patient characteristics, and (ii) describe patterns in diagnostic delay by population density. Following on from section 5.1, detailed here is an overview of the probability of cancer diagnosis delay, which resulted in a paper published in the *British Journal of Cancer* (see Appendix A.5.5).⁶ A brief overview of the background is provided, along with a description of the methods used, followed by the main results, and finishing with a discussion of the main findings coupled with further research.

5.2.2 Overview of paper

Background

In sections 4.1, 4.2, and 5.1, it is shown that patients living in more deprived areas, with more numerous or severe comorbidities, are more likely to experience worse survival. Socioeconomic inequalities in long- and short-term survival were observed but not fully explained by patient characteristics, and the conception of inequalities was concluded to have occurred pre-, peri-, or (immediately) post-diagnosis.

Patients experiencing diagnostic delay are more likely, via more intensive treatment plans, to experience worse short- or long-term survival.⁷ Not unexpectedly, for other cancers, research has shown that patients experiencing diagnostic delay are those living in more deprived areas and with comorbidities.^{8,9,10} A universal healthcare system, such as the National Health Service (NHS), commits to providing equal access to healthcare regardless of the patient's characteristics. While this access may be equitable, the act of receiving the appropriate and necessary care for the, as yet, unknown cancer is complicated by the presence of comorbidities. The mechanism for this could be explained by comorbidities expressing symptoms similar or dissimilar to cancer may delay or hasten the diagnosis, respectively. For example, a swollen abdomen and fatigue in diabetes,¹¹ and chest pain in congestive heart failure,¹² are symptoms of diseases prevalent amongst cancer patients, which could explain misdiagnosis and diagnostic delay.¹³ As yet, the socioeconomic inequalities in diagnostic delay amongst patients with non-Hodgkin lymphoma has not been fully explored.

Clinical commissioning groups (CCGs) are responsible for the provision of healthcare services within a geographical region for a diverse population of patients.¹⁴ All CCGs are provided with guidelines on how to care for patients, however the delivery of care may dif-

fer between CCGs to match the needs of the population in which it represents; therefore, patients within a CCG are more likely to have similar health outcomes.^{15,16} In addition, a CCG with more dense populations may experience more demand on its services, thus patients may 'compete' for healthcare appointments, such as general practitioners. The hypothesis is that patients in more densely populated CCGs are more likely to experience diagnostic delay.

This section aims to explain the variation in diagnostic delay between socioeconomic groups and provide alternative explanations and recommendations to enhance expedited diagnostic route. The objectives of this section are to estimate the odds of diagnostic delay in socioeconomic groups accounting for comorbidity status and other patient characteristics, to evaluate the variation in diagnostic delay between CCGs and describe the influence of population density in CCGs.

Methods

As detailed in Chapter 2, this study contains information on patients diagnosed with DLBCL or FL in England between 2005 and 2013 with information on their follow-up to 2015. Overall, 30,078 and 15,551 patients were diagnosed with DLBCL or FL, respectively, between the ages of 18 and 99. A multivariable generalised linear mixed-effect model (Chapter 3.3) was used to estimate the cluster-specific odds of emergency route to, and late stage at, cancer diagnosis, accounting for patient characteristics and the clustering due to CCGs.^{17,18,19,20} Empirical Bayes estimation of the random effect was used to graphically illustrate the contribution to the variance from each CCG. The random effects for each CCG were extracted and graphical markers were highlighted and resized based on the population density within each CCG. From this graphical display it was possible to identify patterns in the effect of each CCG.

Results

Amongst those with DLBCL, the odds of emergency diagnostic route was significantly higher amongst patients living in more deprived areas, those with comorbidities or multimorbidities, and those of ethnic minorities but the odds of emergency route to diagnosis was similar between males and females. However, amongst those with FL, only the presence of multimorbidity (not comorbidity) and living in more deprived areas increased the odds of emergency route to diagnosis. Also, females were significantly more likely to be diagnosed through other routes. For both DLBCL and FL, the odds of emergency route to diagnosis increased with age.

As observed through graphical illustration (see paper in Appendix A.5.5), the empirical Bayes estimates of the CCG random effect for route to diagnosis showed that, amongst patients with DLBCL, CCGs with greater population densities tended to have patients who were more likely to be diagnosed through emergency route. This was not apparent

for patients with FL.

Discussion

This study showed that, for both DLBCL and FL, deprivation level was a strong independent predictor of diagnostic delay after adjusting for comorbidity status; however, accounting for clustering due to CCGs appeared to increase the strength of the association. This suggests that patients are more likely to have similar diagnostic routes if they reside within the same CCG. This could be explained by the shared allocation of resources such as general practitioner availability and locations, specialist clinicians, specialist hospitals, and diagnostic facilities (e.g., PET-CT scans) that follow CCG-specific guidelines and procedures. Previous studies have recommended providing less variability in the number of pre-diagnosis general practitioner appointments,²¹ clearly defined symptoms and guidelines,²² and expedited contact with lymph node diagnostic clinics.²³ This highlights more precise targeting of resource allocation is needed in more deprived areas, and promotes further investigations into assessing the interaction between patients with prediagnostic symptoms and the healthcare system.

Patients diagnosed with DLBCL or FL are almost twice as likely to have a diagnostic delay if they have underlying comorbid conditions, this increased risk was not explained by their socioeconomic status or other patient characteristics. A possible explanation of this could be that patients who are diagnosed via emergency route may have similar general practitioner consultations in the year prior to cancer diagnosis.²⁴ These patients may also have cancers that are associated with less obvious symptoms. Patients with cancers that have less obvious symptoms, who also have underlying comorbid conditions, may not be referred for further investigation until the underlying comorbid condition is being well managed or further investigations may be considered to risky if they are invasive. For example, patients with less obvious cancer symptoms, and with cardiac-related comorbid conditions, may not be referred to have an ultrasound guided core biopsy (an excision of a lymph node deep in the body) if the investigative procedure posed greater risk at exacerbating the underlying cardiac-related comorbid condition.

Furthermore, there was a pattern between higher population density and diagnostic delay amongst patients with DLBCL and FL. There is a lack in research of the relationship between population density and diagnostic delay, however other studies have found that physician supply and primary care physician density are associated with lower incidence of late stage cancer.²⁵ As DLBCL is an aggressive faster-growing form of NHL, patients may have less time, and therefore greater difficulty, to secure an appointment with a general practitioner or a specialist. In addition, the demand in access to general practitioners may be higher in more densely populated communities, which is also associated with more deprived areas of the population. This increases pressure on the healthcare system and exacerbates the difficulty in securing a general practitioner appointment. As FL is an indolent slower-growing form, patients may have more time to present with symptoms and

the patterns observed for aggressive cancers may not be apparent for indolent cancers.

5.2.3 Conclusion

This study showed that there were inequalities in diagnostic delay between socioeconomic groups, which was not explained by comorbidity status. Patients were more likely to experience diagnostic delay if they have underlying comorbid conditions. Further research could focus on whether the socioeconomic inequalities in diagnostic delay are explained by availability of specific diagnostic facilities (i.e., PET-CT scans), other research could investigate whether there are common pre-diagnostic symptoms amongst patients who are diagnosed through emergency route to diagnosis.

References

- [1] C S Thomson and D Forman. Cancer survival in England and the influence of early diagnosis: what can we learn from recent EURO CARE results? *British Journal of Cancer*, 101(2):S102–S109, 2009. ISSN 1532-1827. doi: 10.1038/sj.bjc.6605399. URL <https://doi.org/10.1038/sj.bjc.6605399>.
- [2] Gerda Engholm, Anne Mette T. Kejs, David H. Brewster, Maria Gaard, Lars Holmberg, Roger Hartley, Robert Iddenden, Henrik Møller, Risto Sankila, Catherine S. Thomson, and Hans H. Storm. Colorectal cancer survival in the Nordic countries and the United Kingdom: Excess mortality risk analysis of 5 year relative period survival in the period 1999 to 2000. *International Journal of Cancer*, 121(5):1115–1122, 9 2007. ISSN 00207136. doi: 10.1002/ijc.22737. URL <http://doi.wiley.com/10.1002/ijc.22737>.
- [3] Lars Holmberg, Fredrik Sandin, Freddie Bray, Mike Richards, James Spicer, Mats Lambe, Åsa Klint, Mick Peake, Trond Eirik Strand, Karen Linklater, David Robinson, and Henrik Møller. National comparisons of lung cancer survival in England, Norway and Sweden 2001-2004: Differences occur early in follow-up. *Thorax*, 65(5):436–441, 5 2010. ISSN 14683296. doi: 10.1136/thx.2009.124222. URL <http://thorax.bmj.com/>.
- [4] Henrik Moller, Fredrik Sandin, Freddie Bray, Åsa Klint, Karen M. Linklater, Arnie Purushotham, David Robinson, and Lars Holmberg. Breast cancer survival in England, Norway and Sweden: a population-based comparison. *International Journal of Cancer*, 127(11):2630–2638, 2 2010. ISSN 00207136. doi: 10.1002/ijc.25264. URL <http://doi.wiley.com/10.1002/ijc.25264>.
- [5] P Armitage, G Berry, and JNS Matthews. *Statistical methods in medical research*. Blackwell Science, Oxford, 4th edition, 2002.
- [6] Matthew J Smith, Miguel Angel Luque Fernandez, Aurélien Belot, Matteo Quartagno, Audrey Bonaventure, Sara Benitez Majano, Bernard Rachet, and Edmund Njeru Njagi. Investigating the inequalities in route to diagnosis amongst patients with diffuse large B-cell or follicular lymphoma in England. *British Journal of Cancer*, 2021. ISSN 1532-1827. doi: 10.1038/s41416-021-01523-6. URL <https://doi.org/10.1038/s41416-021-01523-6>.
- [7] E Kane, D Howell, A Smith, S Crouch, C Burton, E Roman, and R Patmore. Emergency admission and survival from aggressive non-Hodgkin lymphoma: A report from the UK’s population-based Haematological Malignancy Research Network. *European Journal of Cancer*, 78:53–60, 2017. doi: 10.1016/j.ejca.2017.03.013. URL <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85017365106&doi=10.1016%2Fj.ejca.2017.03.013&partnerID=40&md5=a7e66328fab209dca4231def72852249><https://www.sciencedirect.com/science/article/pii/S0959804917308316?via%3Dihub>.

- [8] Jason Gurney, Diana Sarfati, and James Stanley. The impact of patient comorbidity on cancer stage at diagnosis. *British Journal of Cancer*, 113(9):1375–1380, 2015. ISSN 1532-1827. doi: 10.1038/bjc.2015.355. URL <https://doi.org/10.1038/bjc.2015.355>.
- [9] Diana Sarfati, Bogda Koczwara, and Christopher Jackson. The impact of comorbidity on cancer and its treatment. *CA: A Cancer Journal for Clinicians*, 66(4):337–350, 7 2016. ISSN 0007-9235. doi: 10.3322/caac.21342. URL <https://doi.org/10.3322/caac.21342>.
- [10] Theodosia Salika, Georgios Lyratzopoulos, Katriina L Whitaker, Jo Waller, and Cristina Renzi. Do comorbidities influence help-seeking for cancer alarm symptoms? A population-based survey in England. *Journal of Public Health*, 40(2): 340–349, 6 2017. ISSN 1741-3842. doi: 10.1093/pubmed/fox072. URL <https://doi.org/10.1093/pubmed/fox072>.
- [11] Joanna Mitri, Jorge Castillo, and Anastassios G Pittas. Diabetes and risk of Non-Hodgkin’s lymphoma: a meta-analysis of observational studies. *Diabetes care*, 31(12):2391–2397, 12 2008. ISSN 1935-5548. doi: 10.2337/dc08-1034. URL <https://pubmed.ncbi.nlm.nih.gov/19033419><https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2584201/>.
- [12] Ramesh M Gowda and Ijaz A Khan. Clinical Perspectives of Primary Cardiac Lymphoma. *Angiology*, 54(5):599–604, 9 2003. ISSN 0003-3197. doi: 10.1177/000331970305400510. URL <https://doi.org/10.1177/000331970305400510>.
- [13] Helen Fowler, Aurelien Belot, Libby Ellis, Camille Maringe, Miguel Angel Luque-Fernandez, Edmund Njeru Njagi, Neal Navani, Diana Sarfati, and Bernard Rchet. Comorbidity prevalence among cancer patients: a population-based cohort study of four cancers. *BMC Cancer*, 20(1):2, 2020. ISSN 1471-2407. doi: 10.1186/s12885-019-6472-9. URL <https://doi.org/10.1186/s12885-019-6472-9>.
- [14] NHS England. Clinical Commissioning Groups (CCGs), 2021.
- [15] Office for National Statistics. Index of cancer survival for Clinical Commissioning Groups in England: adults diagnosed 2001 to 2016 and followed up to 2017. Technical report, Office for National Statistics, London, UK, 2019. URL <https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/bulletins/indexofcancersurvivalforclinicalcommissioninggroupsinengland/adultsdiagnosed2001to2016andfollowedupto2017>.
- [16] London School of Hygiene and Tropical Medicine. Expert comment on ONS cancer survival bulletins, 0. URL https://www.lshtm.ac.uk/newsevents/news/2014/comment_cancer_survival.html.

- [17] Geert Molenberghs and Geert Verbeke. *Models for Discrete Longitudinal Data*. Springer-Verlag New York, New York, 1 edition, 2005.
- [18] Alan Agresti. *Categorical Data Analysis*. John Wiley & Sons, Inc., Online, 2 edition, 2002.
- [19] Garrett M Fitzmaurice, N.M. Laird, and J.H. Ware. *Applied Longitudinal Analysis*. John Wiley & Sons, Inc., New York, 2 edition, 2011.
- [20] Sophia Rabe-Hesketh and Anders Skrondal. *Multilevel and Longitudinal Modelling Using Stata, Volume II: Categorical Responses, Counts, and Survival*. Stata Press, 3 edition, 2012.
- [21] D A Howell, A G Smith, and E Roman. Lymphoma: variations in time to diagnosis and treatment. *European Journal of Cancer Care*, 15(3):272–278, 7 2006. ISSN 0961-5423. doi: 10.1111/j.1365-2354.2006.00651.x. URL <https://doi.org/10.1111/j.1365-2354.2006.00651.x>.
- [22] D A Howell, A G Smith, and E Roman. Help-seeking behaviour in patients with lymphoma. *European Journal of Cancer Care*, 17(4):394–403, 7 2008. ISSN 0961-5423. doi: 10.1111/j.1365-2354.2007.00897.x. URL <https://doi.org/10.1111/j.1365-2354.2007.00897.x>.
- [23] I Chau, M T Kelleher, D Cunningham, A R Norman, A Wotherspoon, P Trott, P Rhys-Evans, G Querci Della Rovere, G Brown, M Allen, J S Waters, S Haque, T Murray, and L Bishop. Rapid access multidisciplinary lymph node diagnostic clinic: analysis of 550 patients. *British journal of cancer*, 88(3):354–361, 2 2003. ISSN 0007-0920. doi: 10.1038/sj.bjc.6600738. URL <https://pubmed.ncbi.nlm.nih.gov/12569376https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2747551/>.
- [24] C Renzi, G Lyratzopoulos, T Card, T P C Chu, U Macleod, and B Rachet. Do colorectal cancer patients diagnosed as an emergency differ from non-emergency patients in their consultation patterns and symptoms? A longitudinal data-linkage study in England. *British Journal of Cancer*, 115:866, 2016. doi: 10.1038/bjc.2016.250<https://www.nature.com/articles/bjc2016250>{\#}supplementary-information. URL <https://doi.org/10.1038/bjc.2016.250>.
- [25] Ashwin N Ananthakrishnan, Raymond G Hoffmann, and Kia Saeian. Higher Physician Density is Associated with Lower Incidence of Late-stage Colorectal Cancer. *Journal of General Internal Medicine*, 25(11):1164–1171, 2010. ISSN 1525-1497. doi: 10.1007/s11606-010-1457-z. URL <https://doi.org/10.1007/s11606-010-1457-z>.

6 Discussion

6.1 Summary

This thesis investigated the inequalities in health outcomes, particularly survival, of patients with non-Hodgkin lymphoma in England. Survival, a patient health outcome, is a key metric to assess the performance of public health systems: other key metrics are route to, and stage at, diagnosis. Reducing the socioeconomic inequalities, which is known to be a cornerstone for the variation of patient health outcomes, is a high priority target for public health policies. Previously, the perspective was that patient characteristics explain these inequalities; however, this thesis shows that this perspective is only partly true.

There are wide differences in cancer survival by deprivation level, some of which are explained by individual factors, but the proportion of inequalities explained by these factors is only little. For example, when a comparison is made between outcomes of patients with different deprivation categories but with the same comorbidity, and when the differences in survival are more or less the same overall, this suggests that comorbidity explains none of the inequalities. In other words, reducing the impact of comorbidity amongst the more deprived will reduce the differences in cancer survival (since comorbidities are more prevalent in deprived patients and is associated with higher cancer mortality), but will not reduce the inequalities. Furthermore, patients in more deprived areas receive suboptimal care in comparison to less deprived patients (assuming that reliable and valid comorbidity records are obtainable and all confounders are measured). To explain the remaining variation in patient health outcomes, I examined possible reasons for the inequalities beyond the conventional perspective, and challenged the concept that comorbidities and healthcare system factors at least partly account for the remaining variation in patient health outcomes.

By definition, a universal healthcare service, such as the National Health Service, offers equitable access to care and health facilities regardless of patient characteristics. However, in England, inequalities in patient health outcomes, such as survival differences between socioeconomic groups, are evident. There are two possible conjectures:

1. The first conjecture is that patient characteristics influence the likelihood of accessing the healthcare system, which then impacts on patient health outcomes.
2. On the other hand, the second conjecture is that aspects of the healthcare system make it difficult for either all patients or those with certain characteristics to receive optimal care.

In this thesis, I disentangled these conjectures incrementally to describe the overall picture of sources of inequality. To explore the first conjecture, I identified important risk factors of survival, including the association between comorbidity status and survival. Noticing

that population short-term mortality is an indicator of longer-term outcomes, I assessed the association between these risk factors and health outcomes immediately after the patient's cancer diagnosis. To explore the second conjecture, I examined healthcare system characteristics and the likelihood of certain key metrics for the performance of the healthcare system, such as the relationship between population density and a patient having an emergency route to diagnosis: indicating unequal competition for healthcare resources. The ideas brought forward in this thesis provide a novel perspective of what, and by how much, some patient and healthcare system characteristics influence the health outcomes of patients in a population. The ideas brought forward in thesis aids policy recommendations to more efficiently target areas of improvement of the healthcare system.

6.2 Interpretations

6.2.1 Socioeconomic status

In 4.1, I focused on describing age-standardised 5-year net survival by patient characteristics (i.e., gender and socioeconomic status). Patients living in more deprived areas experienced a lower probability of 5-year net survival compared to those in less deprived areas. These results were expected because National Health Service cancer plans recognised, and aimed to reduce, the inequality in survival between socioeconomic groups. However, even though this thesis realises the inequalities were expected, the results did not show a reduction in the inequality. For example, the survival estimates of those living in least deprived areas improved at apparently the same rate as those in the most deprived groups. Suggesting that the survival of the population has increased collectively but the inequality observed from previous time periods still remains.

There are possible explanations for the static socioeconomic gap in survival. The NHS Cancer Plan (2015) noted that the prevalence of comorbidities were higher in more deprived areas. This thesis has shown that comorbidity status is associated with higher excess mortality hazard after adjusting for deprivation and other patient characteristics. However, the presence of a comorbidity is not thought of as acting as a competing risk of cancer mortality (i.e., the direct effect of comorbidity on mortality) because the competing risks are partially taken into account by the deprivation-specific life tables. Instead, the effect of comorbidity is thought to complicate the diagnosis (e.g., higher chance of diagnostic delay) and limit the therapeutic options or complicate the treatment (e.g., less intensive treatments or closer expert monitoring).

6.2.2 Comorbidity status

Descriptive statistics, in Chapter 4.1, showed that the prevalence of comorbidities was higher amongst patients living in more deprived areas. In Chapter 4.2 I modelled the excess mortality hazard, adjusted for patient characteristics and healthcare pathway factors, to assess the association between deprivation, comorbidity status, and net survival. The results, as expected, showed that comorbidity status only partly explains the socioeconomic inequalities in survival of patients with non-Hodgkin lymphoma. Generally, amongst patients from similarly deprived areas, the survival of patients with comorbidities was much lower than those without comorbidities. For those in more deprived areas, patients experienced a worse survival if they had any underlying comorbidity or multimorbidities.

Patients with comorbidities or multimorbidities experienced substantially lower survival in comparison to those without, and did not fully explain the socioeconomic gap in survival. A possible explanation is that patients with at least one comorbidity are at an increased risk of having a delayed diagnosis, since patients with an early cancer diagnosis are more likely to have a better prognosis. This suggests that there are elements of the healthcare system that are missing the targets for early diagnosis, particularly amongst those patients with more complicated medical histories. Indeed, the care and management of patients with comorbidities is systematically different in comparison to those without comorbidities, but further research could explore whether the differences in management are consistent within differing deprivation levels. Furthermore, even after accounting for cancer stage diagnosis in the excess hazard model (4.2), and in line with the hypothesis, there remains a gap in survival between those with or without comorbidity even shortly after their diagnosis. Therefore, it was of particular interest whether, and where, the inequalities in survival arose within the short time period after diagnosis.

In Chapter 5.1, I investigated the association between socioeconomic status and multimorbidity with the probability of short-term mortality. This time frame was considered important because population estimates of long-term survival, such as 5 years, may be explained by short-term mortality, such as 1-year. Even at 1 year after diagnosis there were inequalities in survival between socioeconomic groups and between patients with differing comorbidity statuses. A possible explanation is that the inequalities in survival are conceived before, during or immediately after the cancer diagnosis. The elusive explanation for the inequalities may be hidden in the interactions between the patients, with certain characteristics, and the healthcare system pathway.

6.2.3 Healthcare pathway

It is well known that diagnostic delay (i.e., late stage at, or emergency route to, diagnosis) is indicative of a poorer prognosis; thus, it may be that certain patient characteristics

are more likely to experience diagnostic delay. Results from 5.2 show that patients from more deprived areas, and those with comorbidities, had a greater chance of diagnostic delay. However, I also found that the chances of diagnostic delay was associated with the greater population density of the area in which the patient is represented by a clinical commissioning group. This could suggest that certain characteristics of the healthcare system are contributing to the inequalities in survival. For example, high deprivation is accounted for in the allocation of GP resources but not for diagnostic testing and secondary care resources. Densely populated CCGs, which are correlated with more deprived areas, normally may experience more demand for services and diagnostic tools than CCGs representing less densely populated areas. Moreover, there could be other characteristics of CCGs, not yet realised, that are associated to diagnostic delay and, therefore, poorer health outcomes.

6.3 Implications

6.3.1 Health policies

It was already known that socioeconomic inequalities in survival were present but this thesis has begun the exploration into why the inequalities persist. Despite successive cancer plans and health policies devised to reduce the socioeconomic inequalities in cancer survival,^{1,2,3} the deprivation-gap and inequalities in survival remain.⁴ Reasons for the persisting disparities were thought to be due to patient characteristics, such as comorbidity; however, this thesis challenges that theory and suggests comorbidity only partly explains the disparities and the healthcare pathway may have more influence than previously expected. Furthermore, this thesis highlights the importance of accounting for comorbidity status when assessing the socioeconomic inequalities in health outcomes of cancer patients.

There is currently a primitive understanding of the interplay between patient characteristics and the healthcare pathway, particularly for patients with NHL. Thus far, the interplay has been investigated for common malignancies (i.e., such as colorectal, lung, and breast cancer), which have found that patients with comorbidities have greater risk of diagnostic delay even though they have a higher frequency of contact with the healthcare system.⁵ This thesis has found results in concordance of this theory, which implies that public health policies must do more to clearly disentangle disease-specific symptoms (e.g., separating cancer-related symptoms from underlying comorbidity symptoms) and provide interventions to support the diagnostic process. On the other hand, disentangling comorbidity- from cancer-related symptoms are more difficult for patients with comorbidities than those without; thus, the challenge could be what comorbidity-gap is acceptable and how to narrow the gap.

Consistent with previous studies,⁶ survival after general practitioner referral (non-emergency)

diagnosis is significantly better compared to other routes to diagnosis. However, this thesis also found that patients diagnosed through two-week wait (TWW) referral pathway showed no significant difference in survival compared to GP referral. This is a surprising finding since those patients who were diagnosed via TWW are expected to have worse symptoms and, therefore, worse survival.

There are two possible reasons for the absence of a difference survival between TWW and GP referral. Firstly, GPs could advocate for a prompt referral (mirroring the waiting time of TWW) even though the patient is not on the TWW pathway. In effect, creating a proxy TWW pathway, which results in GP-referred patients experiencing similar access to healthcare facilities as those on the TWW pathway. Secondly, on the other hand, patients referred through the TWW pathway have more severe symptoms and are expected to have lower chance of survival. This thesis showed no difference in health outcomes: implying that the TWW pathway prevents patients (who have more severe symptoms) from having a worse survival compared to the GP-referred patients. This suggests that the performance of the TWW pathway is at least as beneficial to a patient's survival as GP referral. In brief, the two possible reasons are the effectiveness of GP referrals in acquiring healthcare resources or the effectiveness of the TWW pathway in treating more severe cancers. Throughout England, there is an increasing trend in TWW referrals,⁷ possibly due to redefined and clearer criteria, such as an 'unexplained lymphadenopathy or splenomegaly',⁸ or an increase in availability of consultants or specialists.

An incidental finding from this thesis is that, after accounting for patient characteristics, population density is associated with the probability of diagnostic delay amongst aggressive lymphomas: presenting what could be called competing demand for healthcare services. Research into this effect in England is sparse but population density is correlated with cancer mortality, due to higher incidence, in the Western World.⁹ However, as shown in this thesis, competing demand was not observed for indolent lymphomas. This could be because, since indolent lymphomas are slower-growing, the cancer will spend more time in each stage before becoming more severe: there is more time available to seek out a specialist before an emergency diagnostic route is required. This implies that access to the healthcare services, facilities, and appointments amongst patients with faster-growing aggressive lymphomas are not equitably distributed, nor potentially available in the areas where they are needed. Moreover, since there is a correlation between population density and deprivation, CCGs that represent more deprived areas of the population usually have additional resources in primary care. However, further research is needed to understand the resource allocation for secondary care, specialised services and diagnostic testing facilities. Population density may represent unmeasured factors of the healthcare system (e.g., number or availability of haematologists or oncologists) or the environment in which the CCG presides (e.g, dense populations are coterminous with worsening environment).

Another possible explanation is that the competing demand effect could be explained by

the prevalence of comorbidity amongst non-cancer patients within a CCG-specific area. For example, since patients with comorbidities are more likely to interact with the healthcare system more often than patients without comorbidities, the higher the prevalence of comorbidity in an area the more demand on the healthcare service. Implying that the availability of the healthcare service would need to adapt to the demographic changes in comorbid conditions in specific areas of England.

6.4 Limitations

6.4.1 Data

Generalisability

The results found in this thesis are relevant to patients diagnosed with non-Hodgkin lymphoma in England but may also be generalisable to developed countries with a similar structure for their universal healthcare system. The results may also be generalisable to patients diagnosed up to December 2019. This is because there is broad similarity in mechanisms of diagnosis, treatment allocation, accessibility of the healthcare system, and the management of care. These results would not be generalisable to patients diagnosed or treated after January 2020 because the effect of the coronavirus pandemic systematically changed the aforementioned practices of the healthcare system in England. Furthermore, cancer registries provide a national coverage of cancer patients and encapsulate all diagnoses made within the population.

Subtype classification

Cancer diagnoses were made according to the most recent version of the World Health Organisation's International Classification of Diseases for Oncology, third edition. For patients diagnosed with non-Hodgkin lymphoma between 2005 and 2010 the second edition was used, for diagnoses made after 2010 the third edition was used. For both editions, there is large granularity of non-Hodgkin lymphoma subtypes, for example there are at least two different types of diffuse large B-cell lymphoma (i.e., germinal center B-cell like DLBCL, activated B-cell like DLBCL, and peripheral mediastinal B-cell like DLBCL). The data available did not differ between specific subtypes of diffuse large B-cell lymphoma, nor for other subtypes. The different forms of diffuse large B-cell lymphoma can have varying prognoses, therefore grouping the different forms would have the effect of averaging the outcomes of patients who have different forms of diffuse large B-cell lymphoma. If this thesis used the more granular forms of diffuse large B-cell lymphoma then the methods used throughout this thesis may not be feasible due to data sparsity: it is a balance between a clinical interpretation and a statistical interpretation, in other words it is a balance between having precise statistics and clinically-relevant interpretations.

Occasionally, non-Hodgkin lymphoma cases are classified as 'not otherwise specified' (NOS), which is a category of a classification system where the subtype of the lymphoma was either not investigated or identified after hematopathological investigation but where the patient presents with symptoms in concordance with non-Hodgkin lymphoma. Overall, 20% of cases were NOS. If those cases were, according to ICD, unclassifiable subtypes (or classifiable subtypes other than DLBCL or FL), then the results of this thesis would not be biased. However, if they were in fact DLBCL or FL cases, then this could bias the results. For example, if the cases were predominantly patients with severe cases of DLBCL or FL, then the survival estimates of this thesis would be biased upwards (i.e., an overestimate of the survival due to sampling healthier patients).

The occurrence of NOS cases could be explained by the diagnostic procedure. Most lymphomas are diagnosed via surgical excision biopsy (SEB). A systematic review showed that SEB has a higher diagnostic yield in comparison to core needle biopsy (97.5% vs 91.4%) because it provides more tissue for architectural analysis, even though core needle biopsy (CNB) usually provides enough material for diagnosis.¹⁰ The choice of biopsy needle depends on the size and anatomical location of the lesions, and the patient's performance status and likelihood of tolerating the SEB. It is possible that the lymphoma was not classified because of the biopsy needle used (e.g., a SEB for a patient with poor performance status).¹¹ Since most lymphoma subtypes, particularly DLBCL and FL, are not known to be associated with specific anatomical locations and would not be more likely to have a CNB (with a lower diagnostic yield), this suggests that NOS lymphoma cases are non-differentially misclassified. Moreover, there are certain subtypes, such as primary central nervous system lymphoma, that would require an extraction of cerebrospinal fluid or a brain biopsy to diagnose, which has its own estimated diagnostic yield.¹² Nevertheless, in this thesis, since the proportion of patients with DLBCL or FL is similar to what is expected in the general population as shown by previous research,¹³ and that other subtypes are slightly below what is expected, it may be possible that the NOS cases are more likely to be subtypes other than DLBCL or FL that were more difficult to classify based on the morphology or due to difficulties obtaining the biopsy.

Comorbidity status

Defining a patient's comorbidity status is a delicate process and can lead to misclassification or selection bias if not carefully scrutinised; determining the independence of the comorbidity from the cancer is crucial to avoid differential misclassification. To avoid selection bias, each patient must have the same amount of person-time at risk for a comorbidity to be registered, which allows equal probability for all patients to be diagnosed with a comorbidity. I used a robust algorithm to determine the comorbidity status for each patient.¹⁴ The same study showed that the optimal time window for the assessment of comorbidities was 6 years prior to cancer diagnosis (with a right truncation of 6 months prior to cancer diagnosis). Although the optimal time window was not possible in this

thesis due to data availability, misclassification and selection bias was minimised by using the algorithm.

Indices used to derive a patient's comorbidity status may not be valid for several reasons. Firstly, the Charlson index is valid for USA and Canada health databases, which are likely to differ in coding practices from those in England, and records may not concord with other data settings (e.g., Hospital Episode Statistics database). Secondly, improvement in treatments since the development of the Charlson index (in 1987) would reduce the estimated impact of the comorbidity on the patient's survival. Thirdly, comorbidity status may not be cancer-specific or comorbidities may be more prevalent amongst certain cancers. For example, given two patients with similar characteristics except from their cancer, their comorbidities may contribute differential risk to their health outcomes simply due to the cancer. The Royal College of Surgeons' adaptation¹⁵ of the Charlson score¹⁶ used in this thesis, reassesses the risk of comorbidities on health outcomes and avoids misclassifying medical complications as a comorbidity. Furthermore, while the Charlson score was developed using United States' hospital records, the Royal College of Surgeons' Charlson score was developed using Hospital Episode Statistics data and would be valid for coding practices of data used in this thesis. This thesis may contain misclassification bias due to undeveloped cancer-specific comorbidity indices, and specific comorbidities could have been considered, however the focus of this thesis was on the overall pressure of the presence of comorbidities on patient health outcomes and the healthcare system in England.

Hospital Episode Statistics (HES) data contains information on all patients admitted to a hospital (secondary care) in England. It is possible that some comorbidities were not observed because they were diagnosed, and treated, in primary care settings (e.g., diabetes diagnosed during a general practitioner consultation). However, the Royal College of Surgeons' comorbidity index, amongst other indices, are constructed based on the impact of the comorbidity on the risk of mortality; in other words, severe comorbidities that require hospitalisation. Severe comorbid conditions, which are likely to affect care decisions, are more likely to be recorded in hospitals. Comorbidities of the RCS comorbidity index are those that often require hospitalisation, leading to a record within HES data. Previous research has shown that combining primary care records to secondary care data identifies a greater proportion of comorbidity within the population;¹⁷ however, Crooks *et al* show the inclusion of comorbidities identified from primary care records does not have a large effect on predicted cancer survival beyond results obtained using secondary care data.

Socioeconomic status

The Index of Multiple Deprivation (IMD), an area-level measurement for a patient's socioeconomic status, may not be as valid as an individual-level measurement, such as income. Ingleby *et al* have shown that an ecological-level measure of a patient's income has low concordance with a patient-level measurement: in fact, it is only slightly better than a coin toss.¹⁸ This suggests that misclassification of a patient's socioeconomic status may be

more likely to occur when only the income domain of IMD was used.

However, there is better concordance between ecological- and patient-level measures for other domains such as education and occupation. Socioeconomic status, as described in section 1.4.2, encompasses not only the individual's income but also the characteristics of the environment in which they live that can contribute to deprivation. Using all seven domains of the IMD may reduce the risk in misclassifying socioeconomic status in comparison to using only the income domain. On the other hand, one of the seven IMD domains is 'health deprivation and disability', which measures the risk of premature death and the impairment of quality of life through poor physical or mental health. This domain could be autocorrelated with a patient's comorbidity status or the health outcome of interest during this thesis, which could result in biased estimates of the outcome of interest (e.g., cancer survival). This domain has a weighting of 13.5% within the IMD, which, if there is autocorrelation, may not have a large influence on the conclusions of the study. Therefore, even though IMD quintiles are closely correlated to income domain quintiles, further research could utilise the full IMD quintiles except from the 'health deprivation and disability' domain. As stated in Ingleby *et al* (2020), an ecological measure of socioeconomic status only partially captures the relationship between deprivation and patient health outcomes,¹⁸ and, if available, patient- and ecological-measures of socioeconomic status are needed in studies exploring the relationship between deprivation and health outcomes.

6.4.2 Unmeasured variables

Non-surgical treatment data was not available from Public Health England during 2005 through to 2013, this information may partly explain the socioeconomic or comorbid inequalities in survival. Amongst patients with early-stage low-grade lymphomas, there is a high survival even at 5 years since diagnosis. Not knowing whether the patients were placed on 'watch and wait' may have influenced the conclusions of this thesis since treatments for these patients are emerging and being considered more regularly.¹⁹ For example, even before 5 years since diagnosis, low-grade follicular lymphoma can go through histological transformation (becoming a higher-grade follicular lymphoma) and start to show more obvious, serious symptoms. For these higher-grade cases, patients are offered radiotherapy or, in more severe cases, immunochemotherapy.^{20,21} These patients are expected to have a worse survival due to the advanced nature of the lymphoma but, additionally, since data was unavailable, it is not known whether the patients, or those with certain characteristics (i.e., less deprived), received treatment after they were placed on 'watch and wait'. Therefore, access to treatment for patients with follicular lymphoma might partly explain socioeconomic inequalities in survival. Since this information was unavailable, the conclusions of this thesis are based on the observed patient, and tumour, characteristics at the time of cancer diagnosis.

High grade, and advanced-stage low-grade, lymphomas are commonly treated with rituximab in combination with cyclophosphamide, vincristine, doxorubicin, and prednisone.²² Doxorubicin is associated with a higher risk of cardiotoxicity, which may reduce the left ventricular ejection fraction.²³ Doctors and patients may be less inclined to use doxorubicin if the patient had an underlying multimorbidity that included a cardiac disease. The lack of the recommended treatment (doxorubicin) would result in the patient receiving a less intensive treatment regime, which may lead to an increased risk of mortality amongst patients with multimorbidities. This suggests that the risk of mortality due to cancer may be overestimated amongst patients with co/multimorbidity.

Information on a patient's performance score²⁴ (i.e., the patient's ability to carry out everyday tasks) was not available from the data but, if measured, could provide information on the therapeutic options that were available to the patients. In addition, performance status is crucial for indicating the patient's ability to tolerate their treatment.

6.4.3 Missing data

The complexity of diagnosing subtypes of non-Hodgkin lymphoma (NHL) can cause a lack in granular information within data. Some subtypes of NHL cannot be identified from the tools at the disposal of the healthcare system and, as a consequence, some patients will not have a recorded subtype. In this thesis, 20% of patients with NHL did not have a record for their specific subtype (i.e., leading to a missing record for the patient's subtype). Selecting only those patients with an observed subtype may lead to selection bias. If the subtype was considered to be a variable in the analysis, then this could also be termed collider bias. In other words, the sample we select is not due to chance (for example, through randomisation or, in this case, even selecting all the available patients) but is potentially driven by some factors. In this case, the factor is the missingness indicator for subtype. Thinking of a causal diagram, the missingness indicator of subtype would be a collider under the assumptions that the missing subtypes are *missing at random* or *missing not at random*. Therefore, conditioning on the collider will introduce bias in the association between comorbidity (the exposure) and survival (the outcome). If subtype is *missing completely at random*, then by only selecting the observed subtypes the results lose efficiency but the estimates would be unbiased. If subtype is *missing at random* given certain variables, then this may introduce selection bias. To avoid the selection bias, the missing data can either be imputed or observations can be weighted to represent the original population (for example, by using inverse probability of treatment weighting methods). The two approaches are detailed in the following paragraphs.

The first approach would be to impute the patient's NHL subtype by defining a missing data mechanism and imputation model based on the observed variables (i.e., age, gender, ethnicity, deprivation, stage at diagnosis, route to diagnosis, and comorbidity status). The

caveat to this approach is that there are possibly too many subtypes to impute, and the imputation model would not converge due to data sparsity. The issue with data sparsity could be avoided if the subtypes were grouped into aggressive or indolent lymphomas but this would mean losing granularity and clinically relevant information.

The second approach could be to focus on subtypes that are encapsulated within the data. Two very common types of NHL are diffuse large B-cell lymphoma and follicular lymphoma (used throughout this thesis). They are well-known subtypes of NHL and can be more easily diagnosed than other subtypes. In other words, the missing subtypes may more likely be less well-known subtypes. Assuming that only a small proportion of the missing subtypes would be either DLBCL or FL, then the amount of selection bias would also be small. This is a reasonable approach unless the proportion of missing subtypes being DLBCL or FL is large, meaning that there is a factor that drives DLBCL or FL to be missing. Factors that drive a cancer diagnosis to be missing are often the factors that contribute to the complexity of making the diagnosis: for DLBCL and FL this is not the case since they are more easily diagnosed than other subtypes. Thus, assuming the missing subtypes are missing completely at random, selecting only the patients with observed DLBCL and FL is justified because the proportion of patients with DLBCL and FL in the cancer data set resembles the proportions expected in the population. This suggests that the data set captures almost all of the patients with DLBCL or FL, or enough that selection bias due to missing subtype records is negligible.

For the remaining variables with missing data (i.e. ethnicity, route to diagnosis, and stage at diagnosis) multiple imputation was performed, which has the potential to mitigate bias and loss of efficiency. Whether multiple imputation provides gains over a complete case analysis cannot be simply determined from the proportion of incomplete cases in a single variable.²⁵ Indeed, potential benefits from multiple imputation depend on factors such as whether missing data occur in the explanatory variable of interest or covariates, and interrelationships between the variables.²⁶ Lee and Carlin (2012)²⁶ and White and Carlin (2010)²⁷ highlight the need to conduct both a complete case analysis and multiple imputation, and to carefully compare results.

6.4.4 Methods

Net survival

Net survival is the survival probability derived from the cancer-specific hazard of dying. Net survival is the only measure allowing a proper comparison of different populations according to time, geography or other characteristics.²⁸ Several previously developed estimators (i.e., Ederer I, Hakulinen, and Ederer II)^{29,30} were thought of as estimating net survival but were instead estimating the relative survival ratio.^{28,31} These approaches inconsistently estimate survival because, while the bias in small samples is often difficult to extract due

to large variation, the bias in large samples still persists.²⁸ This inconsistency has been shown in practical settings.^{31,32}

However, there are disadvantages of using the Pohar Perme approach. Confidence intervals of the aforementioned approaches are similar in short-term survival (i.e., up to 5-years since diagnosis).³² However, for longer-term survival estimates the Pohar Perme approach is susceptible to larger variation because of the inclusion of very old patients.³¹ These patients carry little or no information on long-term net survival because the probability of competing risks of death is very high. For the other approaches, the variation remains comparatively small because the estimators assume that the information on the variation is provided by younger patients.^{31,32} This assumption may not be valid and further research is required in this area. To avoid this large variation, in this thesis survival estimates were calculated based on a sample of patients up to 94 years old. Therefore, 5-year survival estimates could be obtained for these patients since the weights were based on population mortality data with life tables containing ages up to 99 years (where still a large sample of mortality was observed in the population).

Alternatively, to maintain the purpose of comparability between time periods and countries, net survival can be estimated using a multivariable excess hazard model. Modelling is advantageous because the effects of prognostic variables often change with time since diagnosis, particularly for patients with aggressive, treatable cancers such as diffuse large B-cell lymphoma. Modelling the variation in hazard ratios over time is necessary to avoid biased estimates. In this thesis, a non-linear effect of age was included because the mortality hazard differs across age at diagnosis: often there is a lower hazard for younger ages and a higher hazard for older ages. In addition, a time-varying effect of age was included because the mortality hazard amongst patients of a certain age is expected to be different at alternative times since diagnosis. For example, amongst patients of 70 years old, the mortality hazard is expected to be higher at the time of cancer diagnosis, lower between 1 and 4 years since diagnosis, then higher again after 5 years since diagnosis. Including non-linear and time-varying effects can complicate the interpretations of the model's fixed effect parameters for age and other covariates, but including the effects is often necessary if the assumptions of linear effects and proportional hazards are not justified, which could be assessed using Akaike Information Criterion.^{33,34}

Non-linear and time-varying effects of covariates in an excess hazard model are often modelled using splines.³³ Splines introduce flexibility to more accurately model the data without specifying *a priori* assumptions about the time-varying effect. Such *a priori* restrictions occur when using piecewise proportional hazard models³⁵, which often include several prespecified segments that does not result in a smoothed curve. Restricted cubic splines present an alternative approach to produce smoothed curve. However, to avoid numerical problems, the splines are usually categorised, which complicates the clinical relevance of the results.³⁶ In addition, restricted cubic splines are constrained to have linear tails, which

limits their flexibility.^{36,37} Since their development, B-splines have been used in excess hazard models because they have additional flexibility compared to previous approaches and remain clinically relevant with complex models.³⁶ Using a relatively low number of B-splines ensures sufficient smoothness whilst maintaining a low number of parameters and reduces the risk of overfitting the data.

In this thesis, excess hazard models with different numbers, and locations, of knots were considered throughout. However, as also found in this thesis, previous studies have shown that a different number and location of knots are of limited importance and provide similar conclusions.^{38,36}

Furthermore, the excess hazard model can include a random effect that provides a way to handle the hierarchical structure within a study.³⁴ The hierarchical structure arises because patients living in the same geographical area (i.e., cluster) may share unobserved characteristics, such as primary care facilities, hospitals, or environmental characteristics.³⁴ These shared characteristics may induce a correlation between patient health outcomes, violating the assumption that patients have independent health outcomes. The random effect, once included, models the unobserved heterogeneity of patient health outcomes between clusters. Often, the Gamma distribution is used to model the random effect because the marginal distribution has a closed-form expression,^{34,39} but this is not possible within the excess hazard framework because of the decomposition of the overall hazard (i.e., into excess and population hazard). The random effect was assumed to be normally distributed (i.e., with mean 0 and standard deviation σ), which was plausible since the random effect can be interpreted as the sum of a large number of unobserved characteristics at the cluster level; also, by the central limit theorem, there was a sufficiently large number of clusters.³⁴

Causal inference

Causal inference methods (e.g., generalisation of standardisation) is advantageous over regression-based methods when the interest of the study is in the relationship between only one specific variable and the outcome and the relationships with the other variables (i.e., confounders and the outcome) are of negligible interest. In regression-based models, the interpretation of the relationship of interest is hampered when there is an interaction between two variables (i.e., effect modification). For example, the interpretation of the association between comorbidity and the risk of short-term mortality could be complexified by an interaction between sex and deprivation level. The interpretation of the association is a change in the risk of short-term mortality comparing those with comorbidity to those without comorbidity *for patients with baseline (i.e., reference) categories of the other variables*: this is an interpretation of a conditional association.⁴⁰ Instead, when the interest of the study is in the marginal interpretation of the association, standardisation assumes that the effect of comorbidity on short-term mortality can differ by other variables (e.g., sex or deprivation level). However, standardisation requires making additional assumptions,

which are discussed hereafter.

Propensity score adjustment

In section 4.2, I investigated the association between multimorbidity and short-term mortality by using the parametric g-formula (i.e., weighting using a propensity score). The weight attributed to a patient is the inverse of their probability to be in a particular exposure group. The propensity score model was built by adjusting for age at diagnosis, sex, deprivation level and ethnicity. This approach assumes that the propensity score model is correctly specified. Initially, it was assumed that the patient's age had a linear effect on the odds of mortality, but including non-linear effects provided a better prediction of the odds of mortality. Similarly, a non-linear effect of comorbidity status over time provided a better prediction of the odds of mortality. However, this approach is sensitive to studies where there are patients with low probabilities of being in a particular exposure group. The consequence of this can lead to inaccurate or unstable weights that heavily, or too lightly, weight certain patients. This is closely related to the positivity assumption, which assumes that each patient has a non-zero probability of being in any exposure group.

The disadvantage of having unstable weights can be overcome by trimming the sample to those patients who fall within a certain range of plausible weights (e.g., removing those who have weights beyond the 2.5 and 97.5 centiles).^{41,42,43} Trimming the weights reduces the variance but may introduce bias. Alternatively, the weights can be truncated, whereby all the values of the weights higher than a user-specified maximum value are replaced by the threshold value.^{41,42,43} In section 4.2, the sample of patients was reduced to those aged older than 45 years at cancer diagnosis because those who were younger were unlikely to have multimorbidity. Through checking overlap plots of the propensity score, it was not necessary to further trim or truncate the weights.

Conditional exchangeability

For the conditional exchangeability assumption to hold requires the counterfactual outcome for exposed patients to be the same for unexposed patients, if the unexposed patients were, possibly contrary to fact, exposed.^{44,45} In other words, the assumption of conditional exchangeability holds if the measured and unmeasured confounders of the exposure-outcome relationship are equally distributed between the exposed and the unexposed groups. This assumption is violated if there are unmeasured confounders that are not accounted for in the statistical model. In this thesis, information on baseline patient characteristics (i.e., age, sex, ethnicity, and socioeconomic status) was available, and the risk of short-term mortality was standardised to the distribution of these variables. On the other hand, information on the patient's lifestyle characteristics were not available. As outlined in chapter 1, certain risk factors, such as smoking and obesity, could increase the risk of having certain comorbidities and are known to increase the risk of mortality. These unmeasured confounders could explain some of the association between comorbidity and short-term

mortality.

Due to data availability, information on each cancer patient's treatment (e.g., R-CHOP) was not available. A patient's treatment allocation depends on the presence of underlying comorbidities,^{19,20,21} particularly a history of cardiovascular disease, and could be considered as a mediator between comorbidity status and risk of mortality. Thus, the assumption of conditional exchangeability could still hold without considering treatment as a confounder, but it may explain some of the effect of comorbidity on risk of mortality when behaving as a mediator. A model that adjusts for the patient's treatment (i.e., a variable that is affected by comorbidity and a risk factor for risk of mortality) will provide an unbiased estimate of the association between comorbidity and mortality, but a biased estimate of a causal effect.⁴⁶

Counterfactual consistency

Consistency is assumed to hold if the variants of the exposure (i.e., variants of comorbidity) do not have different effects on the outcome (i.e., mortality).^{46,47} In section 5.1 the exposure was comorbidity and it was the presence of an underlying health condition (independent of the patient's cancer) that increased the risk of mortality. There were variants of a comorbidity such as liver, cardiovascular, and peripheral vascular diseases. Patients were then categorised into three groups: no comorbidity, one comorbidity, or 'two or more' comorbidities (multimorbidity). Since it was possible for patients to have a singular, but differing, comorbidity (or a different combination of comorbidities that summed a patient's multimorbidity), then it was possible that patients placed in the same exposure group (i.e., comorbidity) could have a different variant of that exposure group. For example, patients within the 'one comorbidity' group are patients who have one of twelve different comorbidities. Armitage *et al* (2010), in their paper on the Royal College of Surgeon's adaptation of the Charlson comorbidity index, identified these twelve comorbidities because they increased the risk of mortality.¹⁵ But, since the increased risk of mortality was not the same for all twelve comorbidities, the consistency assumption in this thesis is not assumed to hold. A caveat is that the Charlson Comorbidity Index, and the Royal College of Surgeons' adaptation, include comorbidities that only *increased* the risk of mortality. Therefore, if there were any bias resulting from the violation of the consistency assumption, the presence of this bias would affect the magnitude of the effect of comorbidity on survival, but not the direction of the effect.

Noninterference

Noninterference is the assumption that the potential outcome of one individual was not influenced by the exposure of another individual. More specifically, in section 5.1, a patient having one or more of twelve comorbidities could not influence the risk of short-term mortality of another patient. For example, the risk of mortality in one patient is not influenced by another patient having cardiovascular disease. The noninterference assumption

is often debated in studies of communicable diseases. This relates to this thesis because one possible, although very unlikely, violation of the noninterference assumption could be a patient with human immunodeficiency virus infecting another patient: and increasing the risk of their mortality.

Dependent discrete data

In section 5.2, a multilevel model was used to measure the association between patient characteristics and diagnostic route, while accounting for the hierarchical nature arising from correlation between patients within the same cluster (i.e., clinical commissioning groups [section 1.2.3]). Without accounting for clustering, the standard errors of regression coefficients will be underestimated, which may artificially inflate the chances of obtaining statistically significant results (i.e., type I error). The use of multilevel models are also advantageous for identifying potentially outlying groups (i.e., clinical commissioning groups that have patients who are more likely to have an early or late stage at diagnosis), and can help to indicate whether there are any unmeasured cluster-level characteristics that contribute to the chances of a patient experience the outcome (i.e., diagnostic delay). Generally, multilevel modelling requires a sufficiently large number of clusters such that the central limit theorem can be expected to hold. In this thesis, slightly over 200 clinical commissioning groups represented the clusters, which is assumed to be large enough for the central limit theorem, and the random effect for each cluster was graphed to assess the plausibility of the normality assumption.

The multilevel model built in section 5.2 assumes that the probability of a patient experiencing diagnostic delay is independent from another patient, after accounting for clustering arising from clinical commissioning group. Consideration was given to available data that represented other clusters, such as lower super output area, primary care (i.e., local general practitioner), or secondary care (i.e., local hospital). As the outcome was diagnostic delay (i.e., a comparison between emergency diagnostic route and other routes), the use of primary or secondary care variables to represent clusters would not be appropriate because the outcome was the diagnostic route itself (i.e., possibility of autocorrelation). Clinical commissioning groups (CCG) represent a level higher than, and encompass multiple, primary and secondary care facilities. Since there are unmeasured characteristics of CCGs, such as population density, it provided further insight into the overall performance in relation to the patients they represent. Lastly, including the lower super output area as a random effect may not provide further information beyond that of socioeconomic status since both are closely related, and including an additional random effect may introduce unnecessary computational complexity and difficulties with interpretation.

Missing data analysis

Sensitivity analysis

Throughout this thesis missing data was handled by imputing under a missing at random (MAR) assumption, which is to assume that the probability of data being missing is independent of the missing observations conditional on the observed data.⁴⁸ However, if the probability of the data being missing depends on the value of the missing observations, even after conditioning on the observed data, then data are missing not at random (MNAR).⁴⁸ Neither of these assumptions can be confirmed from the observed data alone, hence the need for sensitivity analysis to assess robustness of inference.^{48,49} Sensitivity to the violation of the MAR assumption can be assessed by imputing missing values assuming the data is MNAR.^{48,50,51,52} Sensitivity analysis after multiple imputation can be carried out within two generic frameworks: pattern mixture models (PMM) or selection factorisation models (SFM).^{48,51,52}

Often, there are a number of patterns of missing observations, each with a different joint distribution; the overall density is then the average over the patterns, giving the term pattern mixture model.⁴⁸ The PMM framework assumes that the observations are stratified based on patterns of missing data, then distinct models are formulated to estimate parameters within each pattern.⁵³ In the PMM framework, there are two common approaches to multiple imputation (MI) under MNAR assumption: prior distributions and shift parameters (delta adjustment method). Previous research has shown consistency in inferences of point estimates and confidence intervals obtained from the two approaches.⁵³ The prior distribution approach may be a more natural method to use because of its simplicity to incorporate uncertainty about the missing data mechanism by using conjugate prior distributions, which are informed by the distribution of the variable under analysis. The delta adjustment method requires making a possibly difficult choice for appropriate sensitivity parameters (often based on expert clinicians' opinions).^{48,54} When prior knowledge is difficult to obtain, Gachau *et al* (2020) suggest using the delta adjustment method with tipping-point analysis as a possible alternative.^{53,55,56}

The selection factorisation framework (SFM) is based on modelling the probability that the observation is missing, given the observed data; in other words, this framework describes assumptions regarding the mechanisms leading to missing data.⁴⁸ In contrast to PMF, where data were imputed under a mixture of MNAR mechanisms, SFM entails imputing under the MAR assumption and then combining estimates by using a weighted average approach.⁵¹ An approximate approach, proposed by Carpenter *et al* (2007), combines estimates using Rubin's rules but with a weighted average of the imputation estimates that up-weight imputations more likely to occur under a MNAR mechanism.

In this thesis, there was missing data in the covariates (stage at diagnosis and ethnicity) and an outcome (route to diagnosis). Further research, beyond this thesis, could have explored the sensitivity to departures from the MAR assumption after multiple imputation. Due to computational complexity, and available software, pattern mixture models could be utilised instead of selection factorisation models. Reliable expert opinions on probability estimates

(to inform sensitivity parameters) may be difficult to obtain because the variability in sociodemographic characteristics across geographical regions would require a large number of clinicians to contribute to the collection. The two approaches in the PMM framework (i.e., prior distribution and delta adjustment) could be used, and their simplicity and logistical involvement compared in order to recommend the appropriate approach for the study design.

6.5 Recommendations

6.5.1 Inequalities

Socioeconomic inequalities in survival remained even after accounting for comorbidities; furthermore, these inequalities were apparent when measuring short-term survival. To address socioeconomic inequalities in survival, future studies should focus on the interaction between the patient and the healthcare pathway prior, during, and immediately after cancer diagnosis. For example, are there socioeconomic inequalities in the number of presentations to a general practitioner prior to cancer diagnosis? Are there socioeconomic inequalities in the time until a lymph node biopsy (or PET-CT scan) during cancer diagnosis? Are there socioeconomic inequalities in the treatment allocation (or treatment adaptation) after cancer diagnosis?

Further research is needed to establish an NHL-specific comorbidity score, given that there is an increased risk of adverse events, such as cardiotoxicity. The RCS Charlson comorbidity score was built for patients where the first-line treatment was surgical intervention. A score could be developed in the setting where the first-line treatment of the cancer are pharmaceutical drugs (e.g., immunochemotherapies). Thus, the cancer-specific score will have greater validity in predicting the risk of mortality amongst these patients with cancer. Furthermore, if data availability allows, studies should use an optimal time-window for a comorbidity to be recorded.¹⁴

Treatment allocation would not explain the inequalities in survival between comorbidity status because patients with comorbidities would be given less intensive treatment due to the underlying comorbidity. It is clinically reasonable to prescribe a different treatment based on a patient's underlying medical history but it is not clinically reasonable based on the patient's socioeconomic status. Therefore, treatment should be considered as a mediator between comorbidity and survival outcomes. Further research could focus on the probability of first-line treatment allocation comparing patients living in more deprived areas to those in less deprived areas.

Further research could explore the association between specific comorbid conditions and the survival of patients with those conditions. In this thesis, I used a score to determine the impact of the comorbid condition on the patient's risk of a health outcome. This score

equates some comorbid conditions to have the same impact as other comorbid conditions: this assumes that the effect of the current condition is the same between two different comorbid conditions. Patients with cardiac related comorbid conditions would have a low score similar to that of patients who have dementia. However, the treatment for NHL often includes doxorubicin and the risk of a detrimental health outcome to the patient would be higher if they had a cardiac related comorbidity compared to a patient who had dementia.

6.5.2 Methods

The latent normal joint modelling multiple imputation approach under a missing at random assumption was used throughout this thesis to account for the variables with missing data. This approach allows imputation of a mix of variable types, while accounting for multilevel structures arising from clustering of patients.^{48,57,58} As with all missing data problems, it is impossible to distinguish between a missing at random and a missing not at random mechanism based on the observed data.^{48,59,60,61} Follow-up work will therefore involve assessing sensitivity of the results to departures from the missing at random mechanism, by imputing under a missing not at random assumption.

Although the methods used in this thesis reduced the risk of bias and inefficiency, they are still lacking applicability in certain elements. For example, it is likely that the stage of a patient's cancer would change over time; thus, the substantive, and imputation, model would need to account for a time-varying variable. Imputation methods do not currently have a facility for imputing missing data in a categorical variable that is also included as a time dependent variable in the substantive model. A solution to this problem would be applicable to not only NHL but for other cancers because the stage of the cancer would also be expected to change over time. Research on conducting multiple imputation compatibly with analysis models containing interactions, non-linear effects, hierarchies and other complex structures is ongoing.^{62,63,64,65} Follow-up work could involve developing methods for imputing partially observed time-dependent variables, particularly in a multilevel excess hazard regression model.

6.5.3 Coronavirus

The impact of coronavirus on the functionality of healthcare systems has been extremely detrimental, particularly for patients with cancer. These patients have experienced increasing difficulty in accessing the healthcare system, delays in being referred to a specialist oncologists, delays in diagnosis, delays in accessing facilities such as MRI scans CT scan to diagnose and assess the extent of cancer, delays in access to treatments such as surgery or medicines, and reduced resources in the management and care after diagnosis. The implications of coronavirus On the healthcare system and consequently on cancer patient care

are as yet not fully understood. However, it is clear that the aforementioned implications of coronavirus will certainly cause a plateau in the increase of survival and may even decrease survival probabilities.^{66,67} Since long-term estimates are driven by short-term survival, it may be clear what will happen within the next five years given that it has been at least one year since coronavirus became a pandemic. In other words, healthcare systems are struggling to care for cancer patients along with the additional pressure of a pandemic,⁶⁷ and the inequalities in survival are expected to increase, particularly for patients living in deprived areas or with multimorbidities, or both.

The incidence of NHL has been increasing since the early 21st century and was expected to continue given the aging population of England. However, the coronavirus pandemic may be a competing risk for the incidence of NHL because it is known to more severely affect those of older ages or with preexisting conditions. This may cause a phenomenon where the incidence of NHL actually decreases over the next decade. Indeed, patterns are already forming amongst other cancers showing incidences are far lower than expected.^{68,69,70}

6.6 Conclusion

The overall aim was to investigate inequalities in survival of patients with non-Hodgkin lymphoma in England. Through this thesis, a retrospective approach to locate and measure inequalities was taken. Starting from measuring long-term outcomes, 5-year net survival was inequitable between patient characteristics (Section 4.1), then it was observed that the associations between patient characteristics and 5-year net survival were greater for certain characteristics, accounting for other factors (Section 4.2). Bringing the focus closer towards the time at diagnosis, the short-term survival was also inequitable between socioeconomic groups (Section 5.1). Finally, the locus on inequalities was pre-, peri-, or immediately post-diagnosis (Section 5.2).

In summary, the 5-year survival of patients with non-Hodgkin lymphoma has increased in comparison to previous years; however, inequalities remain for certain patient characteristics, such as socioeconomic status and comorbidity status. There is an increased risk of short- and long-term mortality amongst those of an older age, living in more deprived areas, or with at least one comorbidity. However, there is no evidence of a difference in the risk of mortality amongst those diagnosed through the two-week wait referral system in comparison to a general practitioner referral. The risk of emergency route to diagnosis, in comparison to other routes to diagnosis, is higher in these pairs of characteristics (i.e. socioeconomic status and comorbidity status), and the risk is higher within geographical areas with more densely populated healthcare systems.

The increased risk of short-term mortality could be reduced by adapting the current healthcare system to manage the more complicated diagnoses amongst those with comorbidities or multimorbidities. The increased risk of short-term mortality could also be reduced by

adapting the healthcare system within areas with more dense populations, which is also correlated with areas that are more deprived and higher prevalence of comorbidities.

References

- [1] Department of Health. The NHS Plan: a plan for action, a plan for reform. Technical report, Department of Health, 2000. URL www.doh.gov.uk/nhsplan.
- [2] Department of Health. Improving Outcomes: a strategy for cancer. Technical report, 2011. URL <https://www.gov.uk/government/publications/the-national-cancer-strategy>.
- [3] National Institute for Health and Care Excellence. Haematological cancers: improving outcomes. Technical report, Department of Health, London, UK, 2016. URL <https://www.nice.org.uk/guidance/ng47>.
- [4] Aimilia Exarchakou, Bernard Rachet, Aurélien Belot, Camille Maringe, and Michel P Coleman. Impact of national cancer policies on cancer survival trends and socioeconomic inequalities in England, 1996-2013: population based study. *BMJ (Clinical research ed.)*, 360:k764–k764, 3 2018. ISSN 1756-1833. doi: 10.1136/bmj.k764. URL <https://www.ncbi.nlm.nih.gov/pubmed/29540358><https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5850596/>.
- [5] Cristina Renzi, Georgios Lyratzopoulos, Willie Hamilton, Camille Maringe, and Bernard Rachet. Contrasting effects of comorbidities on emergency colon cancer diagnosis: a longitudinal data-linkage study in England. *BMC Health Services Research*, 19(1):311, 2019. ISSN 1472-6963. doi: 10.1186/s12913-019-4075-4. URL <https://doi.org/10.1186/s12913-019-4075-4>.
- [6] E Kane, D Howell, A Smith, S Crouch, C Burton, E Roman, and R Patmore. Emergency admission and survival from aggressive non-Hodgkin lymphoma: A report from the UK’s population-based Haematological Malignancy Research Network. *European Journal of Cancer*, 78:53–60, 2017. doi: 10.1016/j.ejca.2017.03.013. URL <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85017365106&doi=10.1016%2Fj.ejca.2017.03.013&partnerID=40&md5=a7e66328fab209dca4231def72852249><https://www.sciencedirect.com/science/article/pii/S0959804917308316?via%3Dihub>.
- [7] National Cancer Registration and Analysis Service. Routes to Diagnosis, 2020.
- [8] National Institute for Health and Care Excellence. Suspected Cancer: recognition and referral, 2021. URL <https://www.nice.org.uk/guidance/ng12/chapter/1-Recommendations-organised-by-site-of-cancer>.
- [9] C Pritchard and B Evans. Population density and cancer mortality by gender and age in England and Wales and the Western World 1963–93. *Public Health*, 111(4): 215–220, 7 1997. ISSN 00333506. doi: 10.1038/sj.ph.1900367. URL <https://pubmed.ncbi.nlm.nih.gov/9242033/>.

- [10] Dale Seviar, Mehreen Yousuff, Zoe Chia, Keith Ramesar, Joel Newman, and David C Howlett. Image-guided core needle biopsy as the first-line diagnostic approach in lymphoproliferative disorders—A review of the current literature. *European Journal of Haematology*, 106(2):139–147, 2 2021. ISSN 0902-4441. doi: <https://doi.org/10.1111/ejh.13532>. URL <https://doi.org/10.1111/ejh.13532>.
- [11] Alice Johl, Eva Lengfelder, Wolfgang Hiddemann, Wolfram Klapper, and the German Low-grade Lymphoma Study Group (GLSG). Core needle biopsies and surgical excision biopsies in the diagnosis of lymphoma—experience at the Lymph Node Registry Kiel. *Annals of Hematology*, 95(8):1281–1286, 2016. ISSN 1432-0584. doi: [10.1007/s00277-016-2704-0](https://doi.org/10.1007/s00277-016-2704-0). URL <https://doi.org/10.1007/s00277-016-2704-0>.
- [12] Alexis A Morell, Ashish H Shah, Claudio Cavallo, Daniel G Eichberg, Christopher A Sarkiss, Ronald Benveniste, Michael E Ivan, and Ricardo J Komotar. Diagnosis of primary central nervous system lymphoma: a systematic review of the utility of CSF screening and the role of early brain biopsy. *Neuro-oncology practice*, 6(6):415–423, 12 2019. ISSN 2054-2577. doi: [10.1093/nop/npz015](https://doi.org/10.1093/nop/npz015). URL <https://pubmed.ncbi.nlm.nih.gov/31832211https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6899047/>.
- [13] A Smith, S Crouch, S Lax, J Li, D Painter, D Howell, R Patmore, A Jack, and E Roman. Lymphoma incidence, survival and prevalence 2004–2014: sub-type analyses from the UK’s Haematological Malignancy Research Network. *Br J Cancer*, 112(9):1575–1584, 2015. doi: [10.1038/bjc.2015.94](https://doi.org/10.1038/bjc.2015.94). URL <http://dx.doi.org/10.1038/bjc.2015.94https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4453686/pdf/bjc201594a.pdf>.
- [14] Camille Maringe, Helen Fowler, Bernard Rchet, and Miguel Angel Luque-Fernandez. Reproducibility, reliability and validity of population-based administrative health data for the assessment of cancer non-related comorbidities. *PLoS ONE*, 12(3):e0172814, 2017. doi: [10.1371/journal.pone.0172814](https://doi.org/10.1371/journal.pone.0172814). URL <https://doi.org/10.1371/journal.pone.0172814https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5338773/pdf/pone.0172814.pdf>.
- [15] J N Armitage and J H van der Meulen. Identifying co-morbidity in surgical patients using administrative data with the Royal College of Surgeons Charlson Score. *Br J Surg*, 97(5):772–781, 2010. doi: [10.1002/bjs.6930](https://doi.org/10.1002/bjs.6930).
- [16] Mary E Charlson, Peter Pompei, Kathy L Ales, and C Ronald MacKenzie. A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation. *Journal of Chronic Diseases*, 40(5):373–383, 1987. doi: [https://doi.org/10.1016/0021-9681\(87\)90171-8](https://doi.org/10.1016/0021-9681(87)90171-8). URL <http://www.sciencedirect.com/science/article/pii/0021968187901718https://www.sciencedirect.com/science/article/pii/0021968187901718?via%3Dihub>.

- [17] C J Crooks, J West, and T R Card. A comparison of the recording of comorbidity in primary and secondary care by using the Charlson Index to predict short-term and long-term survival in a routine linked data cohort. *BMJ Open*, 5(6):e007974, 6 2015. ISSN 2044-6055. doi: 10.1136/BMJOPEN-2015-007974. URL <https://bmjopen.bmj.com/content/5/6/e007974><https://bmjopen.bmj.com/content/5/6/e007974.abstract>.
- [18] Fiona C Ingleby, Aurélien Belot, Iain Atherton, Matthew Baker, Lucy Elliss-Brookes, and Laura M Woods. Assessment of the concordance between individual-level and area-level measures of socio-economic deprivation in a cancer patient cohort in England and Wales. *BMJ Open*, 10(11):e041714, 11 2020. doi: 10.1136/bmjopen-2020-041714. URL <http://bmjopen.bmj.com/content/10/11/e041714.abstract>.
- [19] James O Armitage and Dan L Longo. Is watch and wait still acceptable for patients with low-grade follicular lymphoma? *Blood*, 127(23):2804–2808, 6 2016. ISSN 0006-4971. doi: 10.1182/blood-2015-11-632745. URL <https://doi.org/10.1182/blood-2015-11-632745>.
- [20] Christopher McNamara, Silvia Montoto, Toby A Eyre, Kirit Ardeshta, Cathy Burton, Tim Illidge, Kim Linton, Simon Rule, William Townsend, Wai L Wong, and Pam McKay. The investigation and management of follicular lymphoma. *British Journal of Haematology*, 191(3):363–381, 11 2020. ISSN 0007-1048. doi: <https://doi.org/10.1111/bjh.16872>. URL <https://doi.org/10.1111/bjh.16872>.
- [21] National Institute for Health and Care Excellence. Non-Hodgkin lymphoma: diagnosis and management, 2016.
- [22] H Tilly, M Gomes da Silva, U Vitolo, A Jack, M Meignan, A Lopez-Guillermo, J Walewski, M André, P W Johnson, M Pfreundschuh, and M Ladetto. Diffuse large B-cell lymphoma (DLBCL): ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Annals of Oncology*, 26(suppl 5):v116–v125, 2015. doi: 10.1093/annonc/mdv304. URL http://annonc.oxfordjournals.org/content/26/suppl_5/v116.shorthhttp://annonc.oxfordjournals.org/content/26/suppl_5/v116.full.pdfhttps://watermark.silverchair.com/mdv304.pdf?token=AQECAHi208BE490oan9kKhW_Ercy7Dm3ZL_9Cf3qfKAc485ysgAAAd8wggHbBgkqhkiG9w0BBwa.
- [23] G. Curigliano, D. Lenihan, M. Fradley, S. Ganatra, A. Barac, A. Blaes, J. Herrmann, C. Porter, A. R. Lyon, P. Lancellotti, A. Patel, J. DeCara, J. Mitchell, E. Harrison, J. Moslehi, R. Witteles, M. G. Calabro, R. Orecchia, E. de Azambuja, J. L. Zamorano, R. Krone, Z. Iakobishvili, J. Carver, S. Armenian, B. Ky, D. Cardinale, C. M. Cipolla, S. Dent, and K. Jordan. Management of cardiac disease in cancer patients throughout oncological treatment: ESMO consensus recommendations. *Annals of Oncology*, 31

- (2):171–190, 2 2020. ISSN 15698041. doi: 10.1016/j.annonc.2019.10.023. URL <https://doi.org/10.1016/j.annonc.2019.10.023>.
- [24] Martin M Oken, Richard H Creech, Douglass C Tormey, John Horton, Thomas E Davis, Eleanor T McFadden, and Paul P Carbone. Toxicity and response criteria of the Eastern Cooperative Oncology Group. *American Journal of Clinical Oncology*, 5(6), 1982. ISSN 0277-3732. URL https://journals.lww.com/amjclinicaloncology/Fulltext/1982/12000/Toxicity_and_response_criteria_of_the_Eastern.14.aspx.
- [25] Paul Madley-Dowd, Rachael Hughes, Kate Tilling, and Jon Heron. The proportion of missing data should not be used to guide decisions on multiple imputation. *Journal of Clinical Epidemiology*, 110:63–73, 2019. ISSN 0895-4356. doi: <https://doi.org/10.1016/j.jclinepi.2019.02.016>. URL <http://www.sciencedirect.com/science/article/pii/S0895435618308710>.
- [26] Katherine J Lee and John B Carlin. Recovery of information from multiple imputation: a simulation study. *Emerging Themes in Epidemiology*, 9(1):3, 2012. ISSN 1742-7622. doi: 10.1186/1742-7622-9-3. URL <https://doi.org/10.1186/1742-7622-9-3>.
- [27] Ian R White and John B Carlin. Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Statistics in Medicine*, 29(28):2920–2931, 12 2010. ISSN 0277-6715. doi: 10.1002/sim.3944. URL <https://doi.org/10.1002/sim.3944>.
- [28] Maja Pohar Perme, Janez Stare, and Jacques Estève. On Estimation in Relative Survival. *Biometrics*, 68(1):113–120, 3 2012. ISSN 0006-341X. doi: 10.1111/j.1541-0420.2011.01640.x. URL <https://doi.org/10.1111/j.1541-0420.2011.01640.x>.
- [29] F Ederer, L M Axtell, and S J Cutler. The relative survival rate: a statistical methodology. *Natl Cancer Inst Monogr*, 6:101–121, 1961.
- [30] T Hakulinen. Cancer survival corrected for heterogeneity in patient withdrawal. *Biometrics*, 38(4):933–942, 1982.
- [31] Maja Pohar Perme, Jacques Estève, and Bernard Rachtel. Analysing population-based cancer survival - settling the controversies. *BMC cancer*, 16(1):933, 12 2016. ISSN 1471-2407. doi: 10.1186/s12885-016-2967-9. URL <https://www.ncbi.nlm.nih.gov/pubmed/27912732https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5135814/>.
- [32] Karri Seppä, Timo Hakulinen, and Arun Pokhrel. Choosing the net survival method for cancer survival estimation. *European Journal of Cancer*, 51(9):1123–1129, 6 2015. ISSN 0959-8049. doi: 10.1016/j.ejca.2013.09.019. URL <https://doi.org/10.1016/j.ejca.2013.09.019>.

- [33] L Remontet, N Bossard, A Belot, J Estève, and the French network of cancer registries FRANCIM. An overall strategy based on regression models to estimate relative survival and model the effects of prognostic factors in cancer survival studies. *Statistics in Medicine*, 26(10):2214–2228, 5 2007. ISSN 0277-6715. doi: 10.1002/sim.2656. URL <https://doi.org/10.1002/sim.2656>.
- [34] H Charvat, L Remontet, N Bossard, L Roche, O Dejardin, B Rachet, G Launoy, and A Belot. A multilevel excess hazard model to estimate net survival on hierarchical data allowing for non-linear and non-proportional effects of covariates. *Stat Med*, 35(18):3066–3084, 2016. doi: 10.1002/sim.6881.
- [35] P Bolard, C Quantin, J Esteve, J Faivre, and M Abrahamowicz. Modelling time-dependent hazard ratios in relative survival: Application to colon cancer. *Journal of Clinical Epidemiology*, 54(10):986–996, 2001. ISSN 0895-4356. doi: [https://doi.org/10.1016/S0895-4356\(01\)00363-8](https://doi.org/10.1016/S0895-4356(01)00363-8). URL <https://www.sciencedirect.com/science/article/pii/S0895435601003638>.
- [36] R Giorgi, M Abrahamowicz, C Quantin, P Bolard, J Esteve, J Gouvernet, and J Faivre. A relative survival regression model using B-spline functions to model non-proportional hazards. *Stat Med*, 22(17):2767–2784, 2003. doi: 10.1002/sim.1484.
- [37] C J Stone and C Koo. Additive splines in statistics. In *Proceedings of the Statistical Computing Section*, pages 45–48, 1985.
- [38] H Charvat, N Bossard, L Daubisse, F Binder, A Belot, and L Remontet. Probabilities of dying from cancer and other causes in French cancer patients based on an unbiased estimator of net survival: A study of five common cancers. *Cancer Epidemiology*, 37(6):857–863, 2013. ISSN 1877-7821. doi: <https://doi.org/10.1016/j.canep.2013.08.006>. URL <https://www.sciencedirect.com/science/article/pii/S1877782113001215>.
- [39] L Duchateau and P Janssen. *The Frailty Model*. Springer-Verlag New York, New York, NY, 2008.
- [40] Ian Shrier, Annabelle Redelmeier, Mireille E Schnitzer, and Russell J Steele. Challenges in interpreting results from ‘multiple regression’ when there is interaction between covariates. *BMJ Evidence-Based Medicine*, 26(2):53 LP – 56, 4 2021. doi: 10.1136/bmjebm-2019-111225. URL <http://ebm.bmj.com/content/26/2/53.abstract>.
- [41] Til Stürmer, Kenneth J Rothman, Jerry Avorn, and Robert J Glynn. Treatment effects in the presence of unmeasured confounding: Dealing with observations in the tails of the propensity score distribution-A simulation study. *American Journal of Epidemiology*, 172(7):843–854, 2010. ISSN 14766256. doi: 10.1093/aje/kwq198. URL <https://pubmed.ncbi.nlm.nih.gov/20716704/>.

- [42] Stephen R Cole and Miguel A Hernán. Constructing inverse probability weights for marginal structural models, 2008. ISSN 00029262. URL <https://academic.oup.com/aje/article/168/6/656/88658>.
- [43] Yongling Xiao, Erica E.M. Moodie, and Michal Abrahamowicz. Comparison of approaches to weight truncation for marginal structural Cox models. *Epidemiologic Methods*, 2(1):1–20, 12 2013. ISSN 2161962X. doi: 10.1515/em-2012-0006.
- [44] Miguel Ángel Hernán and James M Robins. Estimating causal effects from epidemiological data. *Journal of Epidemiology and Community Health*, 60(7):578–586, 7 2006. ISSN 0143-005X. doi: 10.1136/JECH.2004.029496. URL <https://europepmc.org/articles/PMC2652882https://europepmc.org/article/PMC/2652882>.
- [45] Michael Schomaker. Regression and Causality. *arXiv:2006.11754*, 6 2020. URL <http://arxiv.org/abs/2006.11754>.
- [46] James M Robins, Miguel Ángel Hernán, and Babette Brumback. Marginal Structural Models and Causal Inference in Epidemiology. *Epidemiology*, 11(5), 2000. ISSN 1044-3983. URL https://journals.lww.com/epidem/Fulltext/2000/09000/Marginal_Structural_Models_and_Causal_Inference_in.11.aspx.
- [47] Stephen R Cole and Constantine E Frangakis. The Consistency Statement in Causal Inference: A Definition or an Assumption? *Epidemiology*, 20(1), 2009. ISSN 1044-3983. URL https://journals.lww.com/epidem/Fulltext/2009/01000/The_Consistency_Statement_in_Causal_Inference__A.3.aspx.
- [48] James R. Carpenter and Michael G. Kenward. *Multiple Imputation and Its Application*. John Wiley & Sons, Ltd, 1st edition, 1 2013.
- [49] Geert Molenberghs and Geert Verbeke. *Models for Discrete Longitudinal Data*. Springer-Verlag New York, New York, 1 edition, 2005.
- [50] Garrett M Fitzmaurice, N.M. Laird, and J.H. Ware. *Applied Longitudinal Analysis*. John Wiley & Sons, Inc., New York, 2 edition, 2011.
- [51] Vanina Héraud-Bousquet, Christine Larsen, James Carpenter, Jean-Claude Desenclos, and Yann Le Strat. Practical considerations for sensitivity analysis after multiple imputation applied to epidemiological studies with incomplete data. *BMC Medical Research Methodology*, 12(1):73, 2012. ISSN 1471-2288. doi: 10.1186/1471-2288-12-73. URL <https://doi.org/10.1186/1471-2288-12-73>.
- [52] Victoria Liublinska and Donald B Rubin. Sensitivity analysis for a partially missing binary outcome in a two-arm randomized clinical trial. *Statistics in medicine*, 33(24):4170–4185, 10 2014. ISSN 1097-0258. doi: 10.1002/sim.6197. URL <https://pubmed.ncbi.nlm.nih.gov/24845086https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4297215/>.

- [53] Susan Gachau, Matteo Quartagno, Edmund Njeru Njagi, Nelson Owuor, Mike English, and Philip Ayieko. Handling missing data in modelling quality of clinician-prescribed routine care: Sensitivity analysis of departure from Missing at Random (MAR) assumption. *Statistical methods in medical research*, 29(10):3076, 10 2020. doi: 10.1177/0962280220918279. URL [/pmc/articles/PMC7116368//pmc/articles/PMC7116368/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC7116368/](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7116368/).
- [54] Stef van Buuren. *Flexible imputation of missing data*. Chapman & Hall/CRC, New York, NY, 2018.
- [55] Daniel Mark Tompsett, Finbarr Leacy, Margarita Moreno-Betancur, Jon Heron, and Ian R White. On the use of the not-at-random fully conditional specification (NAR-FCS) procedure in practice. *Statistics in Medicine*, 37(15):2338–2353, 7 2018. ISSN 0277-6715. doi: <https://doi.org/10.1002/sim.7643>. URL <https://doi.org/10.1002/sim.7643>.
- [56] Finbarr P Leacy, Sian Floyd, Tom A Yates, and Ian R White. Analyses of Sensitivity to the Missing-at-Random Assumption Using Multiple Imputation With Delta Adjustment: Application to a Tuberculosis/HIV Prevalence Survey With Incomplete HIV-Status Data. *American Journal of Epidemiology*, 185(4):304–315, 2 2017. ISSN 0002-9262. doi: 10.1093/aje/kww107. URL <https://doi.org/10.1093/aje/kww107>.
- [57] James Carpenter, Harvey Goldstein, and Michael Kenward. REALCOM-IMPUTE software for multilevel multiple imputation with mixed response types. *Journal of Statistical Software*, 45, 12 2011. doi: 10.18637/jss.v045.i05.
- [58] Matteo Quartagno and James R Carpenter. Multiple imputation for discrete data: Evaluation of the joint latent normal model. *Biometrical journal. Biometrische Zeitschrift*, 61(4):1003–1019, 7 2019. ISSN 1521-4036. doi: 10.1002/bimj.201800222. URL <https://www.ncbi.nlm.nih.gov/pubmed/30868652https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6618333/>.
- [59] Geert Molenberghs, Caroline Beunckens, Cristina Sotito, and Michael G Kenward. Every Missingness Not at Random Model Has a Missingness at Random Counterpart with Equal Fit. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 70(2):371–388, 3 2008. ISSN 13697412, 14679868. URL <http://www.jstor.org/stable/20203830>.
- [60] Geert Verbeke and Geert Molenberghs. Arbitrariness of models for augmented and coarse data, with emphasis on incomplete data and random effects models. *Statistical Modelling*, 10, 11 2010. doi: 10.1177/1471082X0901000403.
- [61] Geert Molenberghs, Edmund Njagi, Michael Kenward, and Geert Verbeke. Enriched-Data Problems and Essential Non-Identifiability. *International Journal of Statistics in Medical Research*, pages 16–44, 1 2012. doi: 10.6000/1929-6029.2012.01.01.02.

- [62] Jonathan W Bartlett, Shaun R Seaman, Ian R White, and James R Carpenter. Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model. *Statistical Methods in Medical Research*, 24(4):462–487, 2 2014. ISSN 0962-2802. doi: 10.1177/0962280214521348. URL <https://doi.org/10.1177/0962280214521348>.
- [63] Tim P Morris and Jonathan W Bartlett. Multiple Imputation of Covariates by Substantive-model Compatible Fully Conditional Specification. *Stata Journal*, 15: 437–456, 1 2015. doi: 10.1177/1536867X1501500206.
- [64] Ruth H Keogh and Tim P Morris. Multiple imputation in Cox regression when there are time-varying effects of covariates. *Statistics in Medicine*, 37(25):3661–3678, 11 2018. ISSN 0277-6715. doi: 10.1002/sim.7842. URL <https://doi.org/10.1002/sim.7842>.
- [65] Matteo Quartagno and James R Carpenter. *Multilevel Multiple Imputation in presence of interactions, non-linearities and random slopes*. PhD thesis, London School of Hygiene and Tropical Medicine, 2018.
- [66] Camille Maringe, James Spicer, Melanie Morris, Arnie Purushotham, Ellen Nolte, Richard Sullivan, Bernard Rachet, and Ajay Aggarwal. The impact of the COVID-19 pandemic on cancer deaths due to delays in diagnosis in England, UK: a national, population-based, modelling study. *The Lancet. Oncology*, 21(8):1023–1034, 8 2020. ISSN 1474-5488. doi: 10.1016/S1470-2045(20)30388-0. URL <https://pubmed.ncbi.nlm.nih.gov/32702310><https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7417808/>.
- [67] Amit Sud, Bethany Torr, Michael E. Jones, John Broggio, Stephen Scott, Chey Loveday, Alice Garrett, Firza Gronthoud, David L. Nicol, Shaman Jhanji, Stephen A. Boyce, Matthew Williams, Elio Riboli, David C. Muller, Emma Kipps, James Larkin, Neal Navani, Charles Swanton, Georgios Lyratzopoulos, Ethna McFerran, Mark Lawler, Richard Houlston, and Clare Turnbull. Effect of delays in the 2-week-wait cancer referral pathway during the COVID-19 pandemic on cancer survival in the UK: a modelling study. *The Lancet Oncology*, 21(8):1035–1044, 8 2020. ISSN 14745488. doi: 10.1016/S1470-2045(20)30392-2. URL <https://pubmed.ncbi.nlm.nih.gov/32702311/>.
- [68] Ermengol Coma, Carolina Guiriguat, Nuria Mora, Mercè Marzo-Castillejo, Mencia Benítez, Leonardo Méndez-Boo, Francesc Fina, Mireia Fàbregas, Albert Mercadé, and Manuel Medina. Impact of the COVID-19 pandemic and related control measures on cancer diagnosis in Catalonia: a time-series analysis of primary care electronic health records covering about five million people. *BMJ Open*, 11(5):e047567, 5 2021. ISSN 2044-6055. doi: 10.1136/bmjopen-2020-047567. URL <https://bmjopen.bmj.com/lookup/doi/10.1136/bmjopen-2020-047567>.

- [69] Toral Gathani, Gill Clayton, Emma MacInnes, and Kieran Horgan. The COVID-19 pandemic and impact on breast cancer diagnoses: what happened in England in the first half of 2020. *British Journal of Cancer*, 124(4):710–712, 2 2021. ISSN 15321827. doi: 10.1038/s41416-020-01182-z. URL <https://pubmed.ncbi.nlm.nih.gov/33250510/>.
- [70] Eva J.A. Morris, Raphael Goldacre, Enti Spata, Marion Mafham, Paul J. Finan, Jon Shelton, Mike Richards, Katie Spencer, Jonathan Emberson, Sam Hollings, Paula Curnow, Dominic Gair, David Sebag-Montefiore, Chris Cunningham, Matthew D. Rutter, Brian D. Nicholson, Jem Rashbass, Martin Landray, Rory Collins, Barbara Casadei, and Colin Baigent. Impact of the COVID-19 pandemic on the detection and management of colorectal cancer in England: a population-based study. *The Lancet Gastroenterology and Hepatology*, 6(3):199–208, 3 2021. ISSN 24681253. doi: 10.1016/S2468-1253(21)00005-4. URL <https://pubmed.ncbi.nlm.nih.gov/33453763/>.

A Appendix

A.1 Details of ethics approvals obtained



Observational / Interventions Research Ethics Committee

Mr Matthew Smith
LSHTM

23 January 2018

Dear Mr Matthew Smith

Study Title: Investigating the disparities in survival of patients diagnosed with non-Hodgkin lymphoma: variation due to socio-economic status, age and comorbidities

LSHTM Ethics Ref: 14613

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	Matthew Smith CV 2017	01/08/2017	1
Protocol / Proposal	Project proposal of ALR for NHL data	10/11/2017	1

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,



Professor John DH Porter
Chair

ethics@lshtm.ac.uk
<http://www.lshtm.ac.uk/ethics/>

A.2 Details of copyright approvals



THESIS COPYRIGHT PERMISSION FORM

Title(s) of the Image(s): Terese Winslow LLC owns the copyright to the following image(s):

Title(s) of illustration(s): Lymphoma, Non-Hodgkin, Adult, Stage IV or Lymphoma, Adult, Stage IV

Description of the Work: Terese Winslow LLC hereby grants permission to reproduce the above image(s) for use in the work specified:

Thesis title: Survival of Non-Hodgkin Lymphoma: Investigating the Inequalities

If available: DOI (Digital Object Identifier), ISSN (International Standard Serial Number), or other publication identifier number

University: London School of Hygiene and Tropical Medicine, London UK.

License Granted: Terese Winslow LLC hereby grants limited, non-exclusive worldwide print and electronic rights only for use in the Work specified. Terese Winslow LLC grants such rights "AS IS" without representation or warranty of any kind and shall have no liability in connection with such license.

Restrictions: Reproduction for use in any other work, derivative works, or by any third party by manual or electronic methods is prohibited. Ownership of original artwork, copyright, and all rights not specifically transferred herein remain the exclusive property of Terese Winslow LLC. Additional license(s) are required for ancillary usage(s).

Credit must be placed adjacent to the image(s) in the following format:

For the National Cancer Institute © ^{2009 or 2019} ~~(copyright year)~~ Terese Winslow LLC, U.S. Govt. has certain rights

Permission granted to:

Author name: Matthew James Smith

Mailing address: Cancer Survival Group, Faculty of EPH, London School of Hygiene and Tropical Medicine, London, WC1E 7HT, United Kingdom.

Email address: matthew.smith1@lshtm.ac.uk

Phone number: work: +44 (0)20 7636 8636 ext. 2495 mobile: +44 (0)7900 915763

Signature _____ **Date** 16/04/2019

Signature _____ **Date** 16th April 2019

Digitally signed by TERESE WINSLOW
Date: 2019.04.16 13:04:36 -04'00'

Terese Winslow, CMI, Member

Terese Winslow LLC, Medical Illustration
714 South Fairfax Street, Alexandria, Virginia 22314
(703) 836-9121
terese@teresewinslow.com
www.teresewinslow.com

A.3 Comorbidity algorithm

The optimal window was 6 years according to Maringe *et al.* (2017). We ran the algorithm based on a 6 month restriction window and 72 month time window, calculated the comorbidity score, and plotted the probability of the comorbidity score over age at diagnosis (figure 9) and socioeconomic status (figure 10). The distribution of comorbidity score was very similar to a time window of 24 months.

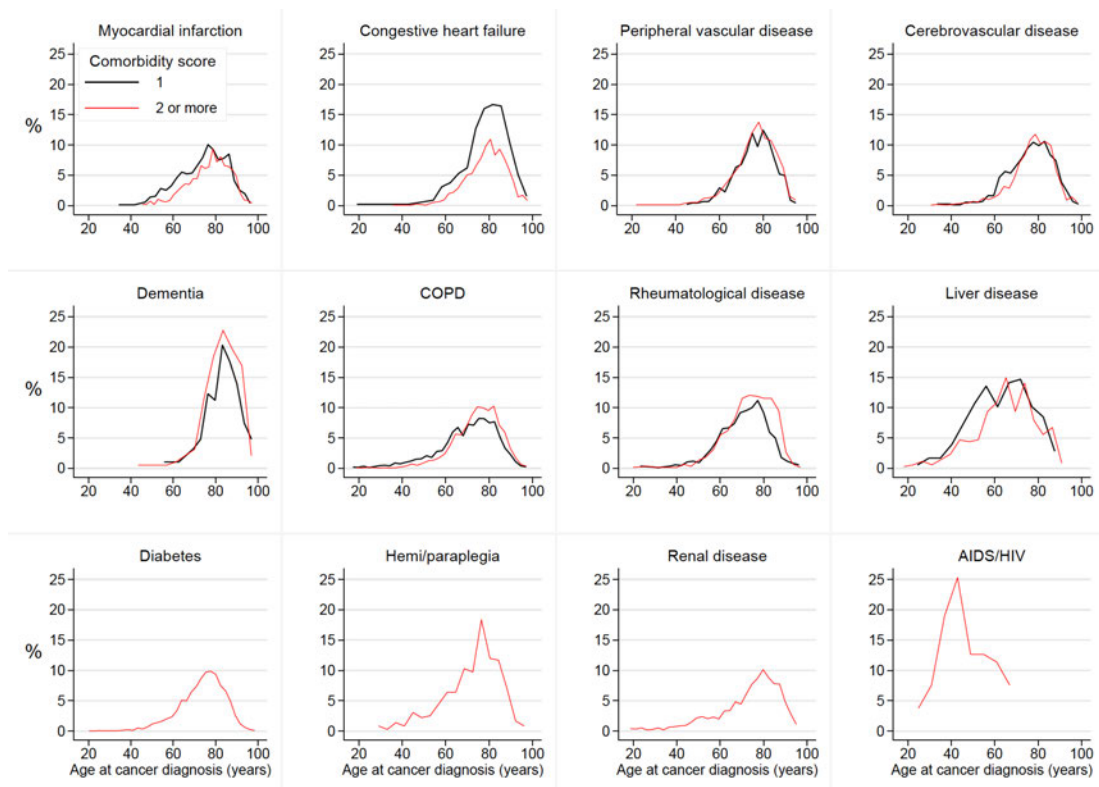


Figure 9: Probability of comorbidity score amongst non-Hodgkin lymphoma patients in England diagnosed 2009-2013. Diabetes, hemi/paraplegia, renal disease and AIDS/HIV are automatically scored as 2 or more. COPD: Chronic obstructive pulmonary disease. Note: using a 6-year optimal time window for comorbidities to be recorded.

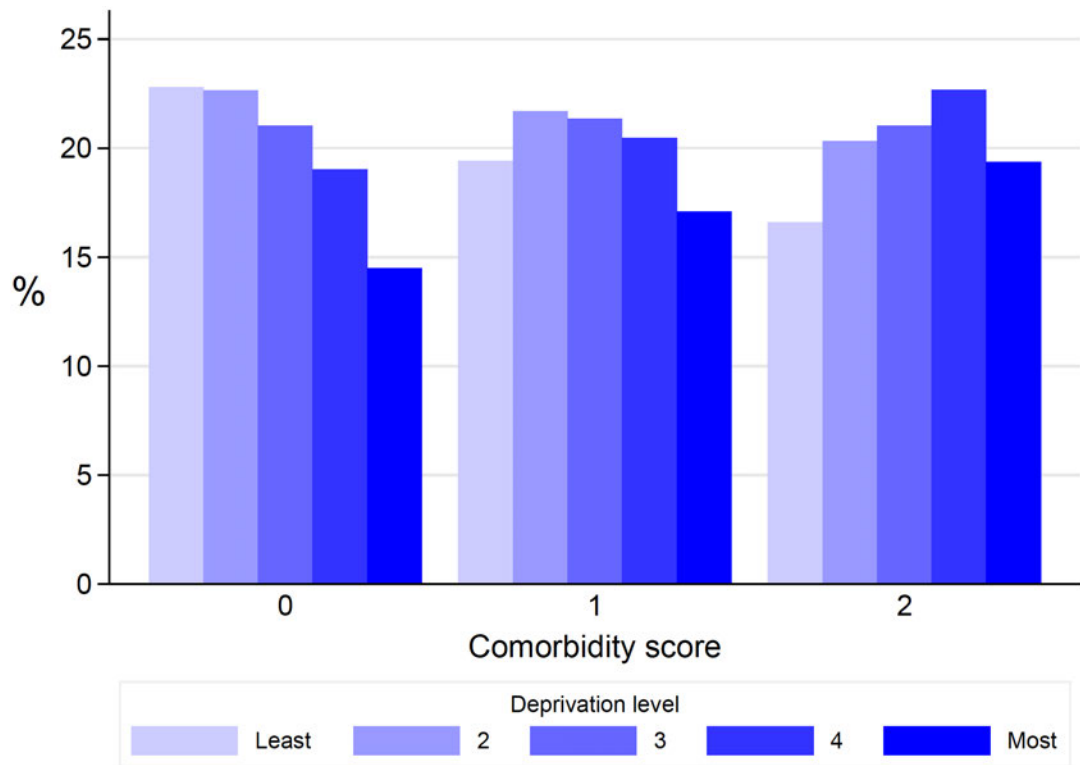


Figure 10: Probability of comorbidity score by deprivation level amongst non-Hodgkin lymphoma patients in England diagnosed 2005-2013. Note: using a 6-year optimal time window for comorbidities to be recorded.

A.4 Publications

The following publications are papers in which I am an author:

Maringe, Camille; Benitez Majano, Sara; Exarchakou, Aimilia; **Smith, Matthew**; Rachet, Bernard; Belot, Aurélien; Leyrat, Clémence; (2020) *Reflection on modern methods: trial emulation in the presence of immortal-time bias. Assessing the benefit of major surgery for elderly lung cancer patients using observational data.* Int J Epidemiol, 49 (5). pp. 1719-1729. ISSN 0300-5771 DOI: <https://doi.org/10.1093/ije/dyaa057>

Smith, M.J., Fernandez, M.A.L., Belot, A. et al. *Investigating the inequalities in route to diagnosis amongst patients with diffuse large B-cell or follicular lymphoma in England.* Br J Cancer (2021). <https://doi.org/10.1038/s41416-021-01523-6>

A.5 Research papers

A.5.1 Descriptive survival of patients with non-Hodgkin lymphoma

This research contains information on the description of survival of patients with non-Hodgkin lymphoma in England. The research presented here has not been submitted to a journal.

Title

Survival of patients diagnosed with diffuse large B-cell lymphomas or follicular lymphomas in England
between 2005 and 2013: a population-based descriptive analysis

Authors

Matthew J. Smith^{1*}, Aurélien Belot¹, Miguel Angel Luque Fernandez^{1,2,3}, Audrey Bonaventure⁴, Sara Benitez
Majano¹, Bernard Rachet¹, Edmund Njeru Njagi¹

Authors' affiliations

¹ Inequalities in Cancer Outcomes Network, Department of Non-Communicable Disease Epidemiology,
London School of Hygiene and Tropical Medicine, Keppel Street, London, WC1E 7HT, UK

² Noncommunicable Disease and Cancer Epidemiology Group, Instituto de Investigación Biosanitaria de
Granada, Ibs.GRANADA, Andalusian School of Public Health, Granada, Spain

³ Centro de Investigación Biomédica en Red de Epidemiología y Salud Pública (CIBER of Epidemiology and
Public Health, CIBERESP), Madrid, Spain

⁴ CRESS, Université de Paris, INSERM, UMR 1153, Epidemiology of Childhood and Adolescent Cancers
Team, Villejuif, France

Corresponding author*

Matthew J. Smith
LSHTM, Keppel Street, London, WC1E 7HT, UK
Email: matthew.smith1@lshtm.ac.uk

Word count: Abstract: 370; Text: 3291; Tables: 3; Figures: 7

Abstract

Background

Analysing patterns of survival amongst patients with non-Hodgkin's lymphoma (NHL) is a priority for the Department of Health and other healthcare systems (National Health Service). Within population-based data, patient characteristics, pathways of care, and geographical disparities can be indicative for the average length of survival after diagnosis. Differences in net survival time are thought to be due to these characteristics. However, further research is needed to explore differences in survival by patient's comorbidity status.

Methods

Information on patients diagnosed between 2005 and 2013 with non-Hodgkin lymphoma (NHL), including cancer diagnosis (topography, morphology), age at diagnosis, gender, deprivation level, and comorbidity status were collected from a linkage of databases in England. Comorbidity was defined by the Charlson comorbidity index and categorised into three groups of severity. Net survival (NS) is used to determine 5-year survival. The comorbidity gap in survival due to age, deprivation, and gender is estimated and trends in the deprivation gap are discussed.

Results

Out of 45,857 NHL patients, men were more likely to have diffuse large B-cell lymphoma (DLBCL), and women follicular lymphoma (FL). 10.7% of DLBCL patients, and 7.6% of FL patients had at least one comorbidity. Over time, the comorbidity-gap in survival narrowed by 7.1% for FL, and by 0.4% for DLBCL. Amongst those without a comorbidity, 5-year net survival was 74% for DLBCL and 89.2% for FL; patients with a severe comorbidity score experienced worse survival at 55.2% and 73.4%, respectively. Over time, the deprivation-gap in survival narrowed by 1.4% for FL; there was no change in the deprivation-gap for DLBCL (stable at 5.6%). For all patient characteristics, survival of NHL improved over time; however, the improvement in survival was heterogeneous.

Conclusion

We found that patients with a severe comorbidity score had a worse 5-year net survival compared to those without comorbidities; and that comorbidities-related survival differ by subtype. We also found the deprivation gap in survival remains unchanged for DLBCL, but is slightly narrower for FL. In conclusion, the results suggest that survival of NHL is improving over time and for all patient risk factors considered in this study. However, there are vast differences in the improvement of survival for comorbidity status, and deprivation; suggesting health care system factors benefit certain patient characteristics more than others.

Introduction

Non-Hodgkin lymphoma (NHL) is a heterogeneous group of malignancies. It arises when B- and T-lymphocytes, a type of immune cells, undergo uncontrolled proliferation. It is estimated that, in 2018, the age-standardised incidence rate of NHL (standardised to the world population) in the United Kingdom (UK) was 12 cases per 100,000 individuals,¹ which is the second highest rate of NHL diagnosis after the United States of America. NHL is more commonly diagnosed amongst the elderly (>65 years);² however, the distribution of subtypes (depending on the morphology of tumour cells) vary for different age groups.³ The average life expectancy in the UK is increasing; therefore, the total number of NHL cases is expected to increase.

The National Health Service (NHS) Cancer Plan,⁴ devised in 2000, was the first comprehensive strategy attempt made by the NHS to increase the survival of patients with cancer in England, aiming to reach survival comparable with the best in Europe; in 2000, cancer survival in the UK was found to be poorer. Possible reasons discussed then were socioeconomic inequalities and delays in diagnosis and treatment. Inequalities in cancer survival were investigated; the suggested sources were patient characteristics: age, deprivation, ethnicity, and lifestyle, and one of the main commitments of the NHS cancer plan was to reduce the gap in survival between socio-economic groups.

The Cancer Reform Strategy (2007) (CRS),⁵ on the back of the NHS Cancer Plan, aimed to improve services such that cancer survival is comparable to the best in the world. The CRS at that time recognised that, amongst all cancers including haematological malignancies, patients with a disability were susceptible to a reduced survival. One of the goals of Cancer Research UK (CRUK) and CRS is that, by 2020, two-thirds of those with common cancers will survive for at least 5 years.⁶ The National Cancer Equality Initiative (NCEI) was set up to address this challenge and investigate the inequalities in cancer survival.⁷ Factors recognised to be contributing to the inequality of cancer survival were: age, gender, deprivation, and ethnicity; but not comorbidity. While previous research has considered NHL as a single type of cancer, NHL is a group of malignancies with different clinical features and prognosis.^{8,9} Diffuse large B-cell lymphoma (DLBCL) and follicular lymphoma (FL) are the two most common subtypes of NHL; they represented approximately 57% of

all diagnoses in 2015.³ More recent studies have focused on separating the subtypes because survival of an aggressive NHL subtype (e.g. DLBCL) is expected to be lower than that for an indolent subtype (e.g. FL).¹⁰ However most studies have provided a summary of survival estimates for each subtype without regards to patient characteristics.

Survival for patients with NHL differs by age, gender, and ethnicity; however little is known regarding differences by comorbidity status.^{8,11} With an aging population, patients are more likely to develop a variety of illnesses, or comorbidities, other than cancer. In population-based studies, a patient's comorbid condition is most commonly classified according to the Charlson comorbidity index (CCI).¹² Previous research shows that comorbidity has a detrimental impact on certain outcomes for patients with NHL.¹³

Up to 2001, the deprivation gap in survival in the UK has appeared to increase; for women this deprivation gap, on average, increased significantly,¹⁴ which also indicated a significant deprivation gap in 1-year survival amongst NHL after the initialisation and during the implementation of the NHS Cancer Plan.

This study uses England population-based cancer registry data linked to various population-based health records with the aim of investigating the inequalities in survival of DLBCL and FL by patient's comorbidity status and deprivation level. We hypothesise that patients with a higher comorbidity status, or living in most deprived areas, have a worse survival. Diagnoses were recorded between 1st January 2005 and 31st December 2013, with follow-up to 31st December 2015.

Materials and Methods

Data source

Information on patients with non-Hodgkin lymphoma (NHL) are collected from the linkage of English cancer registry data, the Cancer Analysis System (CAS) and Hospital Episode Statistics (HES) datasets.^{15,16} The cancer dataset, CAS, contains information collected on patients diagnosed with NHL between 2005 and 2013, with follow up to 31st December 2015. Information on the patient's cancer, subtype (morphology), and date of diagnosis are collected from the core dataset: national cancer registry and analysis service (NCRAS). The NCRAS is linked to CAS and the HES dataset, which contains information on patient's admissions, accident and emergency presentations, and outpatient appointments at an NHS hospital.

Study population

Patients were included if they were diagnosed with NHL (either DLBCL or FL) between 1st January 2005 and 31st December 2013 and aged between 15 and 99 years at the time of diagnosis; with follow-up to 31st December 2015. DLBCL and FL are chosen because of their high prevalence in comparison to other less prevalent subtypes, and their morphological similarity to other aggressive and indolent subtypes. DLBCL and FL are diagnosed according to the 10th revision of the International Statistical Classification of Diseases and Related Problems (ICD) (**Supplementary Table 2**).¹⁷

Variable Definitions

From the linkage of the datasets, there is a more extensive collection of patient's records. Comorbidity index, after removing any patients with a previous malignancy, was classified according to the Royal College of Surgeons (RCS) Charlson Score (**Supplementary table 1**), and derived from the HES dataset for all patients, then the status was developed using a robust algorithm with an optimal time window of 6 to 24 months prior to cancer diagnosis.^{18,19} Socio-demographic characteristics are collected through the linkage of the datasets. Ethnicity (HES dataset) is recorded as either: white, black, Asian, or other. Area-level deprivation (HES dataset), classified into one of five quintiles, is determined by the Index of Multiple Deprivation (IMD), which is based on the Lower Super Output Area (LSOA) residence of the patient at the time of cancer diagnosis; LSOA

is a geographical location for a median of 1500 inhabitants.²⁰ Stage at diagnosis was defined according to the Ann Arbor classification.²¹ Route to diagnosis is determined using the CAS dataset.²² Since no official screening programme is established for NHL, this study does not include a reference to a diagnosis via screening.

Missing data

NHL subtype is missing in 20% of diagnoses and this results in a patient having a morphology classified as ‘not otherwise specified’ (NOS). We include all patients diagnosed with DLBCL or FL, the prevalence of which is consistent with estimates of the population.

Statistical Analysis

We tabulate patient characteristics amongst DLBCL and FL subtypes, then calculated the odds of missing subtype (i.e., NOS) with 95% confidence interval (CI) (**Supplementary table 3**).

We used the Pohar Perme method for net survival (NS), the survival of patients accounting for other competing risks of death, to estimate 5-year survival probability in the relative survival setting.²³ We use two approaches to estimate 5-year NS: the cohort approach for patients diagnosed between 2005 and 2010 (*interval 1*), and the hybrid approach for patients followed up to 2014 and 2015 (*interval 2*) (**Figure 1**).^{24,25} NS for a group of patients, with shared characteristics, is derived from a survival function predicted by the excess hazard (EH) of death due to cancer. The EH is found from a decomposition of the overall hazard of death λ_O such that $\lambda_O = \lambda_E + \lambda_P$. Where the overall hazard λ_O is assumed the sum of the hazard due to the event λ_E (i.e. hazard due to NHL) and the hazard due to the other causes in the population λ_P , estimated using life tables.²⁶ The relationship between the mortality hazard and the survival is $S_E(t) = \exp\left\{-\int_0^t \lambda_E(u)du\right\}$, under the assumption the hazard due to NHL and the hazard of death in the population are conditionally independent, given the set of covariates in the study. The net survival for patients with NHL is then estimated as a weighted average of the individual hazards

$$S_N(t) = \frac{1}{n} \sum_{i=1}^n \frac{S_{O_i}(t)}{S_{P_i}(t)}.$$

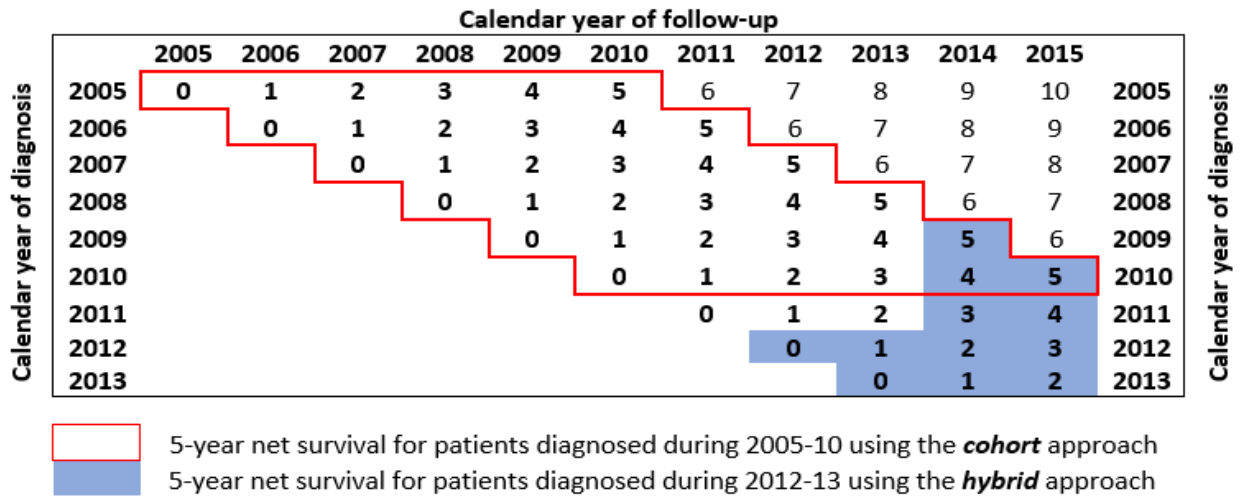


Figure 1: Cohort approach (red line) and hybrid approach (blue highlights) to estimate 5-year net survival for patients diagnosed with diffuse large B-cell lymphoma or follicular lymphoma in England between 2005 and 2013.

Age-standardised estimates of 5-year net survival (ASNS) are calculated using the International Cancer Survival Standard (ICSS) weights for group 1 due to the increased incidence of NHL with age.²⁷ Age groups were defined as 15-44, 45-54, 55-64, 65-74, and 75-99 years. The standard errors of the age-standardised survival are estimated using the Greenwood formula.²⁸ We present age-standardised 5-year net survival by comorbidity status, stratified by deprivation and gender and used the *stns* package in Stata statistical software for all analyses.²⁹

Results

Patient characteristics

Overall, 85,598 patients were diagnosed with NHL in England between 2005 and 2013. Patients with DLBCL (n=30,274) or FL (n=15,583) accounted for 54% of all patients diagnosed with NHL: distribution of patient characteristics are shown in **Table 1**. The percentage of patients who were 65 or older was 63% and 49.7% for DLBCL and FL, respectively. The percentage of patients experiencing mild comorbidity for DLBCL and FL was 10.7% and 7.6%, respectively: for severe comorbidity this was 5.9% and 3.9%, respectively. Males (54.1%) were more likely to have DLBCL than females (45.9%); whereas the reverse was observed for FL: males (47.2%) compared to females (52.8%). There was a higher prevalence of DLBCL diagnoses made amongst patients living in affluent areas (21.2%) compared to more deprived areas (16.0%); the same was observed for FL: affluent (22.8%) compared to deprived areas (14.4%).

A large majority of the DLBCL and FL patients were of white ethnicity (94.1% and 94.9%, respectively). Patients with DLBCL were more likely to be diagnosed through an emergency admission (A&E) compared to general practitioner (GP) referral: 33.0% compared to 27.8%. Whereas patients with FL were more likely to be diagnosed through GP than A&E: 41.4% compared to 12.4%. A large proportion of records for ethnicity were missing: DLBCL (23.0%) and FL (25.0%). For stage at diagnosis, patients were more likely to present with either stage I or IV for both DLBCL and FL. Stage at diagnosis is not considered for further analysis in this study due to the high proportion of missing data for DLBCL (78.1%) and FL (77.5%).

Table 1: Distribution of patient characteristics by NHL subtype for patients diagnosed with diffuse large B-cell lymphoma or follicular lymphoma in England between 2005 and 2013.

	NHL Subtype	
	DLBCL n = 30,274 (35.3%) * N (%)	FL n = 15,583 (18.2%) * N (%)
Age at diagnosis		
Mean (SD)	67.6 (15.0)	64.0 (13.6)
Age categories		
15-44	2,704 (8.9)	1,509 (9.7)
45-54	2,873 (9.5)	2,324 (14.9)
55-64	5,633 (18.6)	3,995 (25.6)
65-74	8,261 (27.3)	4,320 (27.7)
75+	10,803 (35.7)	3,435 (22.0)
Comorbidity**		
None	27,029 (89.3)	14,404 (92.4)
Mild	1,604 (5.3)	642 (4.1)
Severe	1,641 (5.4)	537 (3.5)
Gender		
Male	16,381 (54.1)	7,355 (47.2)
Female	13,893 (45.9)	8,228 (52.8)
Deprivation		
Affluent	6,404 (21.2)	3,554 (22.8)
2	6,737 (22.3)	3,532 (22.7)
3	6,326 (20.9)	3,309 (21.2)
4	5,938 (19.6)	3,934 (18.8)
Deprived	4,847 (16.0)	2,244 (14.4)
Missing	22 (0.07)	10 (0.06)
Ethnicity		
White	21,948 (72.5)	11,093 (71.2)
Black	325 (1.1)	118 (0.8)
Asian	795 (2.6)	351 (2.3)
Other	251 (0.8)	133 (0.9)
Missing	6,955 (23.0)	3,888 (25.0)
Route		
GP referral	8,243 (27.2)	6,313 (40.5)
Emergency	9,797 (32.4)	1,891 (12.1)
Inpatient elective	735 (2.4)	325 (2.1)
Other outpatient	3,021 (10.0)	1,906 (12.2)
TWW	6,956 (23.0)	3,920 (25.2)
Unknown	943 (3.1)	904 (5.8)
Missing	579 (1.9)	324 (2.1)
Stage		
I	1,831 (6.1)	924 (5.9)
II	1,316 (4.4)	519 (3.3)
III	1,120 (3.7)	901 (5.8)
IV	2,353 (7.8)	1,159 (7.4)
Missing	23,654 (78.1)	12,080 (77.5)

Percentages may not sum to 100.0% due to rounding. DLBCL: Diffuse large B-cell lymphoma. FL: Follicular lymphoma (all grades).

*Other subtypes accounted for 39,839 (46.5%) of all NHL cases diagnosed between 2005 and 2013

**CCI measured by the coded as: 0 – none, 1 – mild, 2 or more – severe

Distribution of age categories over time

The average age at diagnosis was higher amongst DLBCL than FL. Patients with DLBCL were more likely to be diagnosed after 75 years: FL between 65-74 years. The proportion of patients diagnosed within certain age groups differed over time (**figure 2**); more recently, since 2010, patients with DLBCL and FL were diagnosed at an older age compared to diagnoses made prior to 2010.

Distribution of deprivation amongst CCI scores

The distribution of CCI scores differed by deprivation groups and by NHL subtype (**figure 3**). For both subtypes, the proportion of diagnoses for patients with a higher CCI score increased with each level of deprivation. In other words, there was a higher percentage of patients with a mild or severe CCI score amongst those who were more deprived; this result was more pronounced amongst patients with DLBCL than FL.

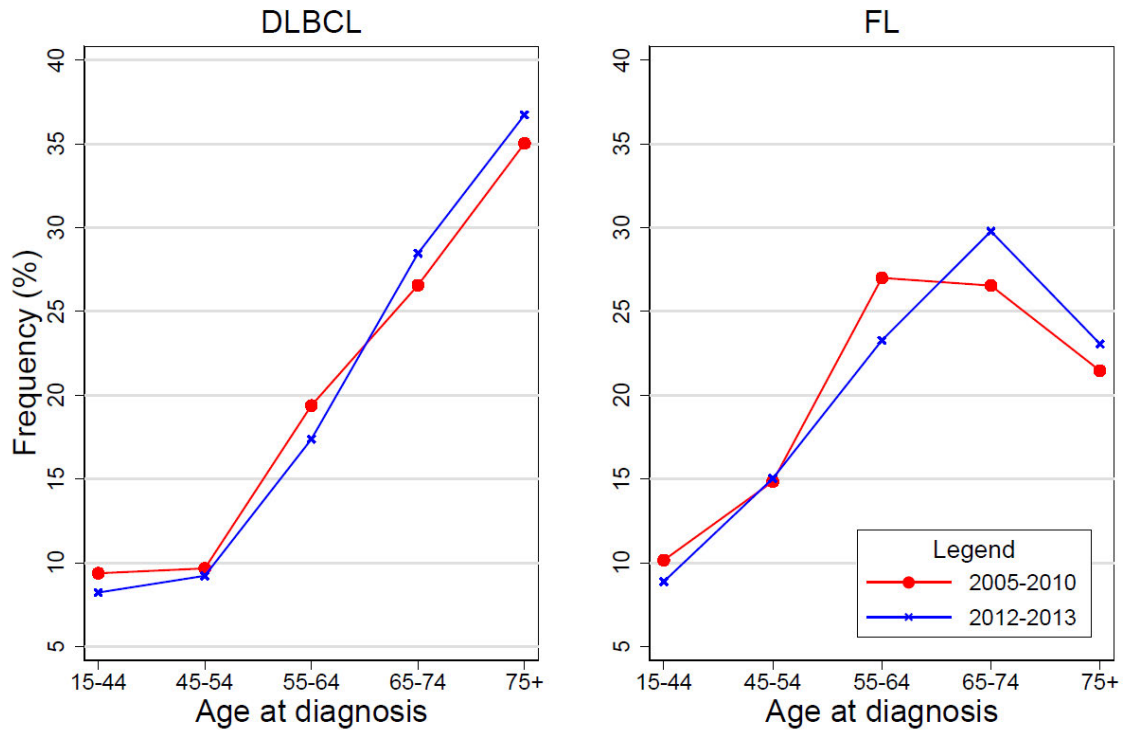


Figure 2. Distribution of age groups amongst patients diagnosed with diffuse large B-cell lymphoma (left) or follicular lymphoma (right) between 2005 and 2013.

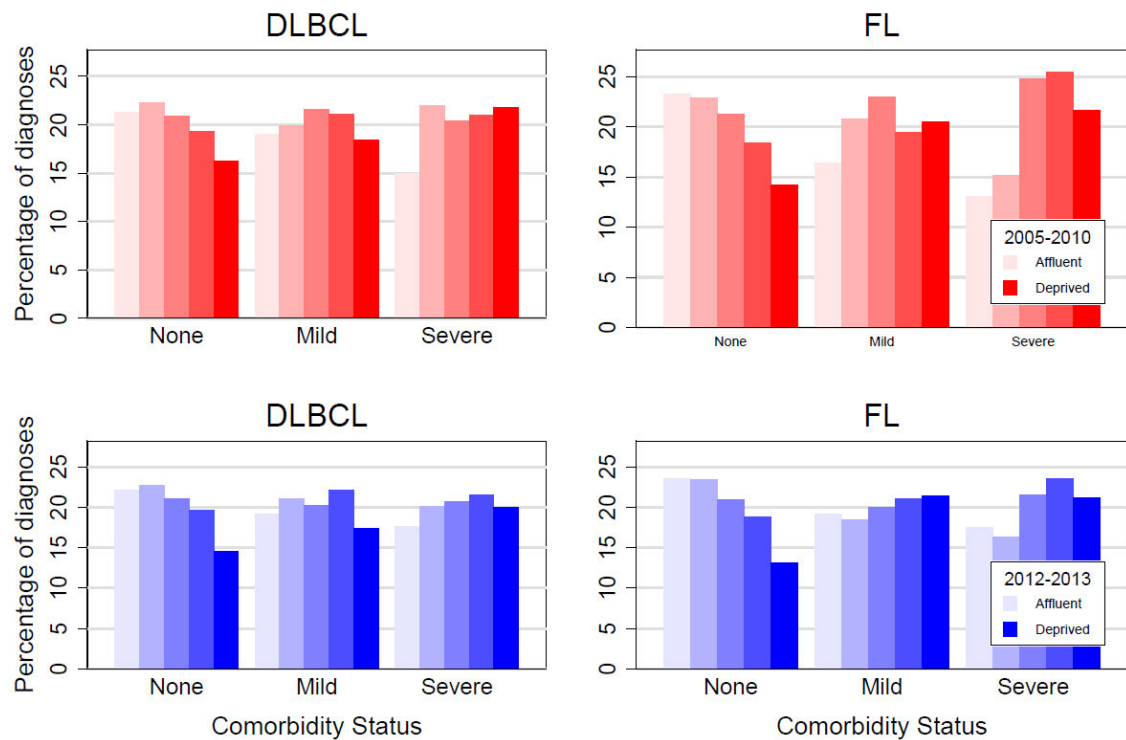


Figure 3. Distribution of socioeconomic status by comorbidity score amongst patients diagnosed with diffuse large B-cell lymphoma (left) or follicular lymphoma (right) between 2005 and 2013.

Net survival

Age-standardised 5-year survival (ASNS) was highest amongst those without a comorbidity, and generally lower amongst those with a mild or severe comorbidity score (**table 2** and **figure 4**). The gap in survival between those without comorbidities compared mild or severe comorbidities was 15.3% and 19.2%, respectively; whereas, using the hybrid approach, these survival gaps were 10.2% and 18.8%, respectively. For FL, those with a mild CCI score were more likely to survive to 5-years than those with a severe CCI score, the comorbidity gap between those without comorbidities compared to mild or severe comorbidities was 12.5% and 22.9%, respectively; whereas, using the hybrid approach, these survival gaps were 6.7% and 15.8%, respectively.

For patients with DLBCL, there was a higher survival comparing patients living in least deprived to most deprived areas: the improvement over time was comparable across deprivation groups. The socioeconomic gap in 5-year survival was similar and did not narrow in 2005-2010 compared to 2012-2013. Amongst patients with FL, there was a narrowing in socioeconomic gap from 6.6% in 2005-2010 to 5.2% in 2012-2013. Although survival improved for both genders, the gender gap in survival widened for both DLBCL and FL as females had a greater improvement than males (**table 2**). Amongst patients with 'not otherwise specified' subtypes, the improvement in survival was similar across comorbidity score, deprivation level and gender (**Supplementary table 4**).

Table 2: Age-standardised 5-year net survival (ASNS) estimates by socio-demographic characteristics for patients diagnosed with diffuse large B-cell lymphoma or follicular lymphomas in England during 2005-2010 and 2012-2013.

	DLBCL (5-year net survival %)			FL (5-year net survival %)		
	2005-2010	2012-2013	Difference	2005-2010	2012-2013	Difference
Comorbidity*						
<i>None</i>	60.9 (60.1-61.8)	74.0 (73.1-74.9)	+13.1	83.8 (82.6-85.1)	89.2 (88.5-89.9)	+5.4
<i>Mild</i>	45.6 (41.7-49.7)	63.8 (60.8-66.8)	+18.2	71.3 (66.0-77.0)	82.5 (79.0-86.1)	+11.3
<i>Severe</i>	41.7 (38.0-45.7)	55.2 (52.0-58.6)	+13.5	60.9 (54.8-67.7)	73.4 (69.3-77.8)	+12.5
Deprivation						
<i>Affluent</i>	60.2 (58.5-61.9)	74.7 (72.2-77.2)	+14.5	82.7 (80.4-85.1)	90.6 (89.3-91.9)	+7.9
2	62.0 (60.3-63.7)	74.4 (73.2-75.6)	+12.4	84.4 (81.9-86.9)	88.1 (86.7-89.6)	+3.7
3	59.9 (58.1-61.7)	72.8 (71.5-74.1)	+12.9	82.3 (79.8-84.9)	89.1 (87.7-90.6)	+6.8
4	58.5 (56.6-60.5)	70.0 (68.6-71.5)	+11.5	83.1 (80.2-86.1)	85.1 (83.4-86.9)	+2.0
<i>Deprived</i>	54.6 (52.6-57.0)	69.1 (67.4-70.8)	+14.5	76.1 (72.9-79.5)	85.4 (83.3-87.6)	+9.3
Gender						
<i>Male</i>	58.2 (57.1-59.4)	70.2 (68.7-71.7)	+12.0	81.4 (79.6-83.3)	86.9 (85.8-88.0)	+5.5
<i>Female</i>	60.5 (59.3-61.7)	74.8 (73.9-75.7)	+14.3	82.9 (81.4-84.5)	89.0 (88.1-89.9)	+6.1

AS net survival is presented by 5-year NS (CI)

NOS – not otherwise specified

* As measured by the CCI score

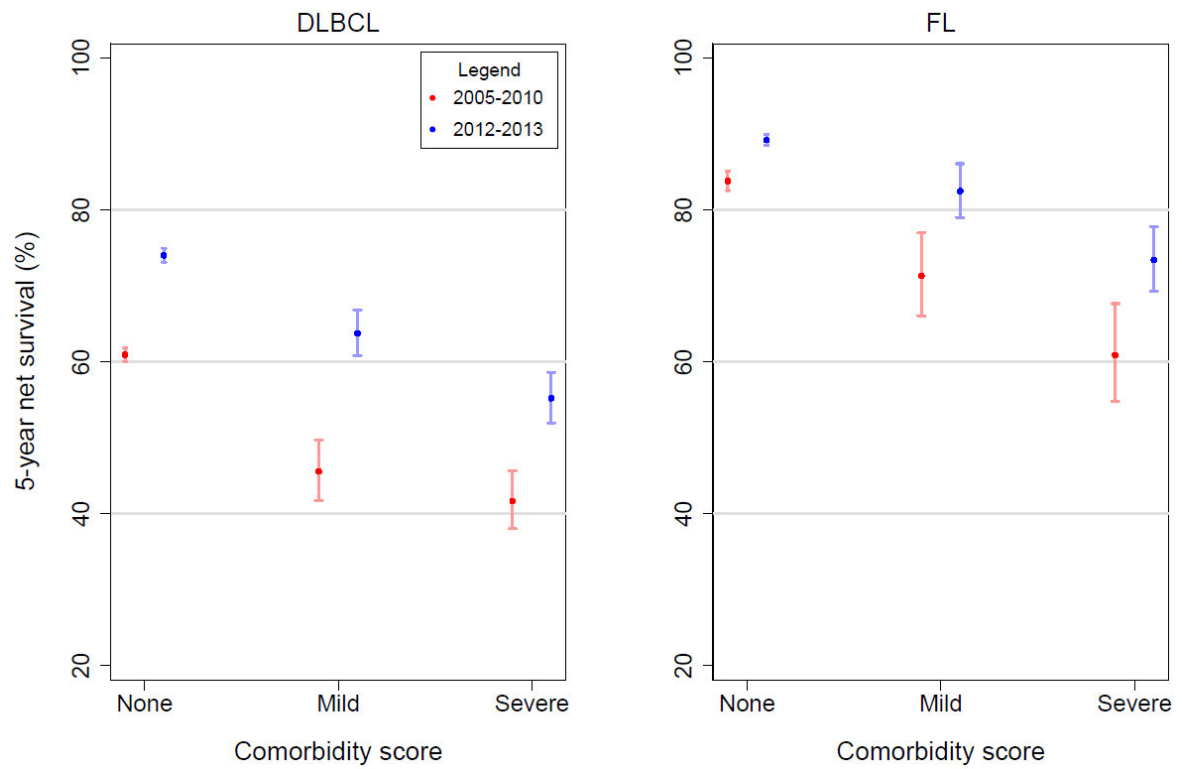


Figure 4. Age-standardised 5-year net survival estimates by comorbidity score amongst patients diagnosed with diffuse large B-cell lymphomas (left) or follicular lymphomas (right) in England between 2005 and 2013

Stratification

Table 3 and **figures 5, 6, and 7** show stratified age-standardised 5-year net survival estimates. For those with DLBCL, the socioeconomic gap in survival amongst males narrowed from 7.0% in 2005-2010 to 5.7% in 2012-2013; amongst females, the gap widened from 3.9% to 6.1% over time. For those with FL, the gap widened in males from 8.9% in 2005-2010 to 10.5% in 2012-2013; amongst females, the gap narrowed from 4.9% to 1.3% over time.

For those with DLBCL in 2005-2010, the socioeconomic gap in survival amongst those with no comorbidities vs severe comorbidity was 4.6% vs 5.2%; however, in 2012-2013, the socioeconomic gap widened to 5.1% vs 10.5%, for the same comparison. For those with FL in 2005-2010, the socioeconomic gap in survival amongst those with no comorbidities vs severe comorbidities was 5.2% vs -5.5% (meaning the most deprived patients had better survival compared to least deprived); however, in 2012-2013, the socioeconomic gap was unchanged for those with no comorbidities, and actually reversed for those with severe comorbidities (from -5.5% to 3.1%).

Table 3: Age-standardised (AS) net survival (NS) estimates for each combination of stratification between comorbidity, deprivation, and gender amongst patients diagnosed with diffuse large B-cell lymphoma or follicular lymphomas in England during 2005 to 2013

		DLBCL (5-year net survival %)			FL (5-year net survival %)		
		2005-10	2012-13	Difference	2005-10	2012-13	Difference
Comorbidity							
None	m	59.9 (58.7-61.1)	71.9 (70.3-73.6)	+12.0	83.0 (81.0-85.0)	88.0 (86.9-89.1)	+5.0
	f	62.0 (60.8-63.3)	76.1 (75.2-77.0)	+14.1	84.5 (82.9-86.2)	90.1 (89.2-91.1)	+5.6
Mild	m	43.6 (38.4-49.4)	61.5 (57.5-65.8)	+17.9	64.3 (56.4-73.3)	83.0 (77.8-88.6)	+18.7
	f	47.6 (42.3-53.5)	66.5 (62.4-70.8)	+18.9	71.3 (64.7-78.5)	82.2 (77.8-86.9)	+10.9
Severe	m	41.6 (37.1-46.7)	53.8 (49.7-58.1)	+12.2	59.2 (51.8-67.7)	72.1 (66.3-78.3)	+12.9
	f	41.6 (35.9-48.1)	57.3 (52.2-63.0)	+15.7	57.1 (48.6-66.9)	74.9 (69.3-81.0)	+17.8
Deprivation							
Affluent	m	59.7 (57.4-62.1)	72.4 (68.6-76.4)	+12.7	82.3 (78.8-85.9)	90.6 (88.7-92.6)	+8.3
	f	60.7 (58.2-63.2)	77.4 (75.6-79.2)	+16.7	82.9 (80.0-86.0)	90.5 (88.7-92.3)	+7.6
2	m	60.1 (57.8-62.4)	72.4 (70.7-74.2)	+12.3	84.2 (80.5-88.1)	87.4 (85.3-89.7)	+3.2
	f	64.1 (61.6-66.7)	76.7 (75.0-78.4)	+12.6	84.5 (81.4-87.7)	88.7 (86.8-90.6)	+4.2
3	m	58.6 (56.1-61.2)	69.8 (68.0-71.8)	+11.2	81.7 (77.9-85.7)	88.5 (86.3-90.7)	+6.8
	f	60.6 (58.0-63.3)	75.8 (73.9-77.6)	+15.2	82.6 (79.3-86.1)	89.7 (87.8-91.5)	+7.1
4	m	58.0 (55.3-60.8)	68.5 (66.5-70.6)	+10.5	81.3 (76.7-86.1)	82.7 (79.9-85.7)	+1.4
	f	59.3 (56.6-62.1)	71.4 (69.4-73.5)	+12.1	84.4 (80.8-88.1)	86.8 (84.6-89.0)	+2.4
Deprived	m	52.7 (49.7-55.9)	66.7 (64.4-69.2)	+14.0	73.4 (68.1-79.0)	80.1 (76.4-83.9)	+6.7
	f	56.8 (53.8-59.8)	71.3 (68.9-73.6)	+14.5	78.0 (74.0-82.2)	89.2 (86.7-91.8)	+11.2
Comorbidity							
None	Affluent	61.2 (59.5-63.0)	75.9 (73.4-78.5)	+14.7	83.3 (80.9-85.7)	91.7 (90.4-93.0)	+8.4
	2	63.8 (62.0-65.6)	75.9 (74.6-77.2)	+12.1	86.1 (83.6-88.7)	89.4 (88.0-90.9)	+3.3
	3	61.7 (59.7-63.6)	74.4 (73.0-75.8)	+12.7	83.7 (81.0-86.4)	90.2 (88.7-91.7)	+6.5

	4	59.9 (57.9-62.0)	71.5 (70.0-73.1)	+11.6	85.6 (82.5-88.8)	85.9 (84.0-87.7)	+0.3
	Deprived	56.6 (54.3-59.0)	70.8 (69.0-72.6)	+14.2	78.1 (74.5-81.7)	86.7 (84.4-89.0)	+8.6
Mild	Affluent	44.7 (36.6-54.5)	63.0 (56.6-70.2)	+18.3	59.7 (49.6-71.8)	84.0 (82.2-85.8)	+24.3
	2	48.3 (40.4-57.6)	65.3 (59.5-71.6)	+17.0	46.2 (38.6-55.2)	77.5 (70.4-85.2)	+31.3
	3	43.1 (36.3-51.2)	62.0 (56.0-68.6)	+18.9	58.7 (49.3-69.9)	83.3 (76.8-90.3)	+24.6
	4	40.6 (33.2-49.7)	65.2 (59.2-71.7)	+24.6	58.1 (47.1-71.5)	83.2 (75.1-92.1)	+25.1
	Deprived	45.1 (37.0-55.0)	64.2 (57.5-71.6)	+19.1	64.2 (52.8-78.2)	82.1 (74.6-90.4)	+17.9
Severe	Affluent	38.2 (30.3-48.1)	62.8 (56.0-70.3)	+24.6	48.4 (36.2-64.6)	77.8 (70.2-86.2)	+29.4
	2	40.4 (33.7-48.6)	58.6 (52.4-65.4)	+18.2	51.6 (39.4-67.4)	72.2 (63.4-82.2)	+20.6
	3	45.5 (38.0-54.5)	54.8 (48.0-62.6)	+9.3	41.4 (31.2-54.9)	73.2 (64.5-83.2)	+31.8
	4	46.8 (39.3-55.8)	49.5 (42.6-57.4)	+2.7	48.2 (37.0-62.7)	70.9 (62.1-81.0)	+22.7
	Deprived	32.7 (25.8-41.5)	52.3 (45.3-60.3)	+19.6	53.9 (43.9-66.1)	74.7 (66.4-83.9)	+20.8

m – male, f – female, DLBCL – diffuse large B-cell lymphoma, FL – follicular lymphoma

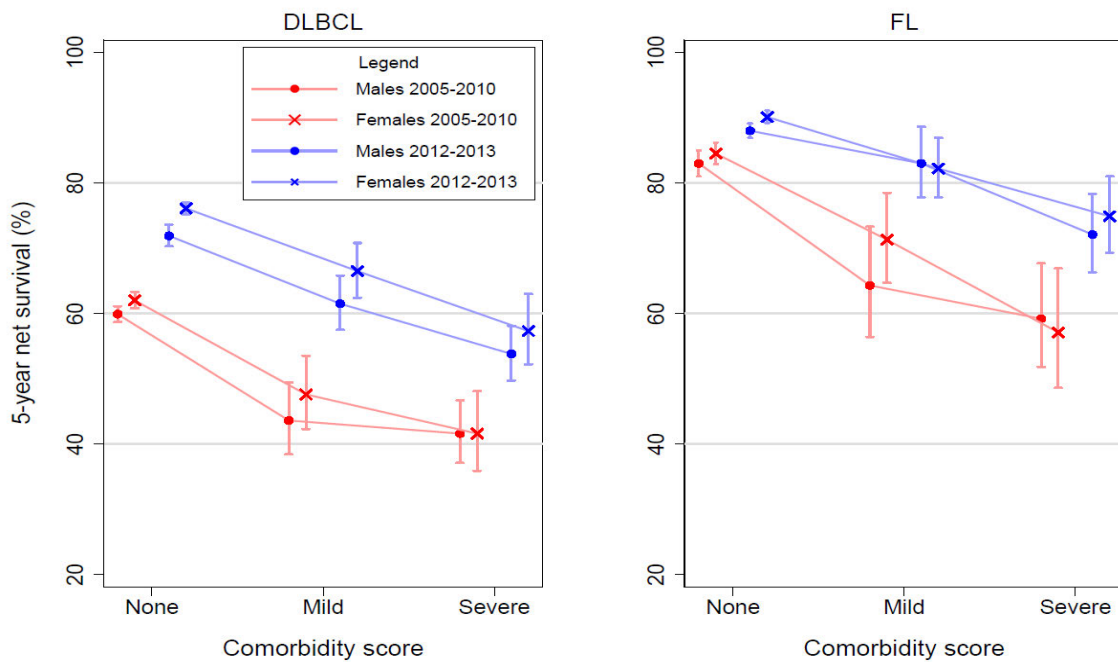


Figure 5. Stratified age-standardised 5-year net survival estimates by comorbidity and gender amongst patients diagnosed with diffused large B-cell lymphoma (left) or follicular lymphoma (right) in England between 2005 and 2013.

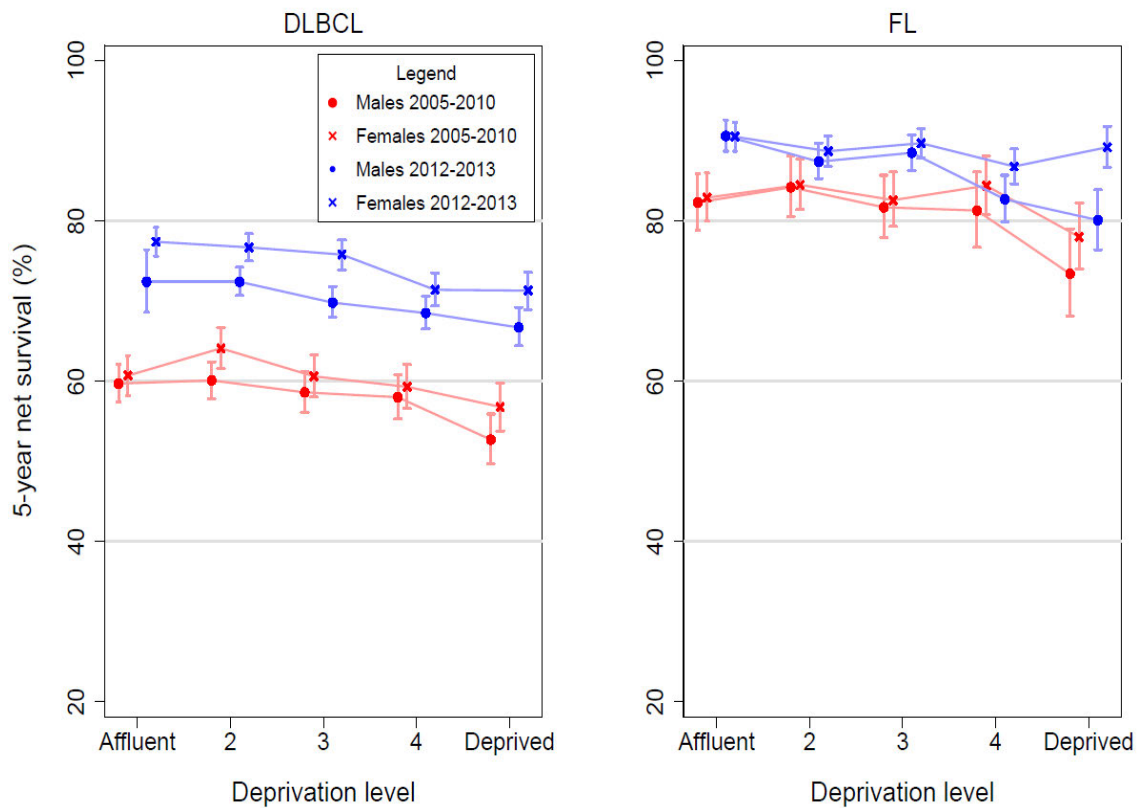


Figure 6. Stratified age-standardised 5-year net survival estimates by deprivation and gender amongst patients diagnosed with diffuse large B-cell lymphoma (left) or follicular lymphomas (right) in England between 2005 and 2013

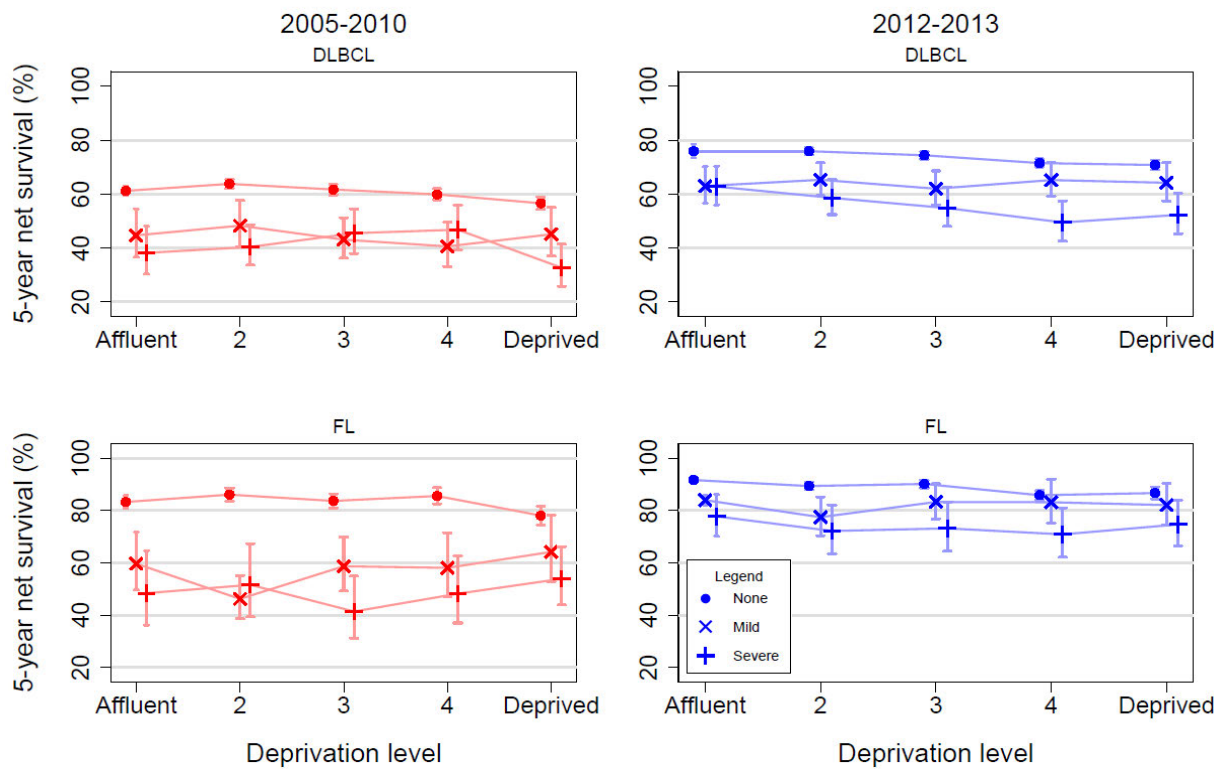


Figure 7. Stratified age-standardised 5-year net survival estimates by deprivation and comorbidity amongst patients diagnosed with diffuse large B-cell lymphoma (left) or follicular lymphoma (right) in England between 2005 and 2013

Discussion

We found that, overall, 5-year age-standardised net-survival (ASNS) for patients with Diffuse Large B-cell lymphoma (DLBCL) and follicular lymphoma (FL) improved over time. Survival improved more for females, and those living in more deprived areas, or with more severe comorbidities. However, socioeconomic inequalities in survival still remain and actually widened in certain stratifications of patient characteristics. In 2005-2010, for both DLBCL and FL, 5-year survival was worse for patients with any comorbidity compared to those without comorbidity. In 2012-2013, for DLBCL, this comorbidity-gap in survival narrows for those in the least deprived areas but remains apparent in the more deprived areas; however, for FL, this comorbidity-gap in survival narrows for all patients regardless of deprivation level.

Stratification analysis highlighted a potential interaction between deprivation and comorbidity for patients with DLBCL, gender and comorbidity for FL, and gender and deprivation for FL. Modelling the interaction, using parametric methods, could be used to estimate the synergistic effect of these interactions on a patient's survival.

The increase in survival may be explained by a number of factors. The approval of immunotherapy (rituximab) in 1997 for treatment of DLBCL and FL has been found to be an effective treatment for those of an advanced age, and could partly explain the improvement in older age-groups.³⁰ However, in the treatment of NHL, rituximab is often used in combination with anthracyclines, one of which is doxorubicin. An increase in dosage of doxorubicin is correlated with an increase in the incidence of adverse effects (cardiotoxicity), usually in the form of congestive heart failure.³¹ According to the National Institute for Health and Care Excellence, patient's treatment should be decided based on age and the International Prognostic Index, which accounts for age, the patient's performance status (reflecting general condition) and other clinical prognosis factors; and also suggests, patients with high IPI score or with cardiac dysfunction (at risk of cardiotoxicity) may be given a less-intensive treatment regimen.³² This less intensive treatment allocation may explain some of the inequality in survival between comorbidity scores observed in this study. Studies investigating treatment strategies for patients with differing comorbidity scores may provide further insight.

This study supports findings from that highlight the detrimental impact of comorbidities on a patient's survival, which has been well documented.¹³ Our results suggest that the socioeconomic- and comorbidity-gap in survival is narrowing over time for FL patients, but not for DLBCL patients, which would warrant further investigations. Suggestions to improve outcomes for patients with higher comorbidity scores include: novel treatment strategies, inclusion of elderly patients in clinical trials, and investigation of dose-allocation amongst those with higher comorbidity scores.³³ The improvement in survival has also shown a reduction of the socioeconomic-gap in survival. Patients in all levels of deprivation are experiencing an improvement in survival, however this study supports the paradigm that the gap in survival does not appear to be closing consistently.⁸

Regarding a patient's comorbid condition, we assume that Hospital Episode Statistics (HES) data are a valid source of comorbidity records. HES data is subjected to automatic data cleaning and derivation rules, contributing to high internal validity.³⁴ Beforehand, concerns lingered over the external validity of HES data due to the apparent lack of involvement from clinicians when documenting records. Records not entered by clinicians may give a suboptimal understanding of the patient's medical condition, requiring interpretations to be made, and potentially leading to a reduced quality of records, such as additional comorbidities remaining absent to the HES data.

Moreover, we include comorbidities if they were recorded within a defined time-window and produced a comorbidity score using a robust algorithm. We assumed that once a patient is diagnosed with a chronic comorbidity, the patient is burdened until the time of cancer diagnosis. As shown by Maringe *et al.* (2017), a paradoxical association can occur when the probability of a patient's comorbidity is associated with the patient's cancer survival, which can be avoided with a time-window restriction prior to the cancer diagnosis.¹⁹ This study embraces the restriction window and aims to avoid the paradoxical association.

A strength of this study is that all NHL cases were diagnosed and coded with the latest version ICD for oncology (third edition).¹⁷ A review was conducted in 2008 which provided a more accurate diagnosis tool for some haematopoietic and lymphoid tissues. Since the implementation of the review, NHL diagnoses may have

become more accurate; however, it is expected that any differences in survival estimates within this study between pre- and post-2010 would not be due to this review as all cases of DLBCL or FL were considered.

Furthermore, by removing patients with a previous malignancy from the criteria of a comorbidity, this study represents patients with NHL who are diagnosed with any cancer for the first time. The removal of these patients may provide a more accurate estimate of the survival for first-time NHL patients. We also excluded patients who were diagnosed via a death certificate only (DCO), because we do not know when they were diagnosed with cancer; therefore, we should not include them in survival analyses as this will remove immortal time bias.

A limitation is that from the total of 85,598 NHL cases identified between 2005 and 2013, 20% of these cases were identified as 'not otherwise specified' (NOS), meaning that the subtype was not identified for over 17,000 cases. However, the proportion of DLBCL and FL cases in this sample were similar to that expected in the population, therefore, selection bias is expected to be negligible.

Further studies investigating NHL survival are strongly advised to estimate survival not only by subtypes (as they behave differently according to morphology and prognosis) but also by deprivation and comorbidity score. Additionally, survival may differ for patients with certain combinations of characteristics, for example, it was not clear if there is effect modification between gender, deprivation levels, and comorbidity status. Therefore, more complex methods such as flexible parametric models may provide insight into interactions between patient's characteristics, such as deprivation and comorbidity.

Despite increasing survival of patients with NHL, there remains several sources of survival inequality such as deprivation levels and comorbidity score. This shows the inequity in survival between deprivation groups is still apparent. Patients with NHL in England are expected to have a poorer prognosis if they live in more deprived areas or have a previously diagnosed comorbid condition. The inequity in survival shows the need for the current framework of the National Health Service to embrace custom patient management systems for those with underlying health conditions.

References

1. World Health Organisation. International Agency for Research on Cancer. Non-Hodgkin Lymphoma Incidence (2018). Available at: <http://gco.iarc.fr/today/home>. (Accessed: 24th March 2020)
2. Buntinx, F., Campbell, C. & van den Akker, M. Cancer in the Elderly. *J Cancer Epidemiol* **2014**, (2014).
3. Smith, A. *et al.* Lymphoma incidence, survival and prevalence 2004-2014: sub-type analyses from the UK's Haematological Malignancy Research Network. *Br J Cancer* **112**, 1575–1584 (2015).
4. Department of Health. *The NHS cancer plan: a plan for investment: a plan for reform*. (Department of Health, 2000).
5. Department of Health. *Cancer reform strategy*. (Department of Health, 2007).
6. Cancer Research UK. *Our strategy to beat cancer sooner*. (2014).
7. National Cancer Equality Initiative. *Cancer Equalities*. (2015).
8. Rachet, B., Mitry, E., Shah, A., Cooper, N. & Coleman, M. P. Survival from non-Hodgkin lymphoma in England and Wales up to 2001. *Br. J. Cancer* **99**, S104–S106 (2008).
9. Quaresma, M., Coleman, M. P. & Rachet, B. 40-year trends in an index of survival for all cancers combined and survival adjusted for age and sex for each cancer in England and Wales, 1971–2011: a population-based study. *Lancet* **385**, 1206–1218 (2015).
10. van Spronsen, D. J., Janssen-Heijnen, M. L., Lemmens, V. E., Peters, W. G. & Coebergh, J. W. Independent prognostic effect of co-morbidity in lymphoma patients: results of the population-based Eindhoven Cancer Registry. *Eur J Cancer* **41**, 1051–1057 (2005).
11. Smith, A. *et al.* Impact of age and socioeconomic status on treatment and survival from aggressive lymphoma: a UK population-based study of diffuse large B-cell lymphoma. *Cancer Epidemiol* **39**, 1103–1112 (2015).
12. Charlson, M. E., Pompei, P., Ales, K. L. & MacKenzie, C. R. A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation. *J. Chronic Dis.* **40**, 373–383 (1987).
13. Sogaard, M., Thomsen, R. W., Bossen, K. S., Sorensen, H. T. & Norgaard, M. The impact of comorbidity on cancer survival: a review. *Clin Epidemiol* **5**, 3–29 (2013).

14. Rachet, B. *et al.* Socioeconomic inequalities in cancer survival in England after the NHS cancer plan. *Br J Cancer* **103**, 446–453 (2010).
15. gov.uk. National Cancer Registry and Analysis Service. (2017). Available at: <https://www.gov.uk/guidance/national-cancer-registration-and-analysis-service-ncras>. (Accessed: 4th October 2019)
16. NHS Digital. Hospital Episode Statistics. (2015). Available at: 2015. (Accessed: 4th October 2019)
17. International Agency for Research on Cancer. International Classification of Diseases for Oncology. (2013). Available at: <http://codes.iarc.fr/>. (Accessed: 4th October 2019)
18. Armitage, J. N. & van der Meulen, J. H. Identifying co-morbidity in surgical patients using administrative data with the Royal College of Surgeons Charlson Score. *Br J Surg* **97**, 772–781 (2010).
19. Maringe, C., Fowler, H., Rachet, B. & Luque-Fernandez, M. A. Reproducibility, reliability and validity of population-based administrative health data for the assessment of cancer non-related comorbidities. *PLoS One* **12**, e0172814 (2017).
20. gov.uk. Indices of deprivation. (2015). Available at: <https://www.gov.uk/government/statistics/english-indices-of-deprivation-2015>. (Accessed: 24th March 2020)
21. Lister, T. A. *et al.* Report of a committee convened to discuss the evaluation and staging of patients with Hodgkin's disease: Cotswolds meeting. *J Clin Oncol* **7**, 1630–1636 (1989).
22. Elliss-Brookes, L. *et al.* Routes to diagnosis for cancer - determining the patient journey using multiple routine data sets. *Br J Cancer* **107**, 1220–1226 (2012).
23. Perme, M. P., Stare, J. & Estève, J. On Estimation in Relative Survival. *Biometrics* **68**, 113–120 (2012).
24. Esteve, J., Benhamou, E. & Raymond, L. Statistical methods in cancer research. Volume IV. Descriptive epidemiology. *IARC Sci Publ* 1–302 (1994).
25. Brenner, H. & Rachet, B. Hybrid analysis for up-to-date long-term survival rates in cancer registries with delayed recording of incident cases. *Eur. J. Cancer* **40**, 2494–2501 (2004).
26. Cancer Survival Group. UK Life Tables. (2012). Available at: <https://csg.lshtm.ac.uk/tools-analysis/uk-life-tables/>. (Accessed: 4th October 2019)
27. Corazziari, I., Quinn, M. & Capocaccia, R. Standard cancer patient population for age standardising

- survival ratios. *Eur. J. Cancer* **40**, 2307–2316 (2004).
28. Pokhrel, A., Dyba, T. & Hakulinen, T. A Greenwood formula for standard error of the age-standardised relative survival ratio. *Eur J Cancer* **44**, 441–447 (2008).
 29. Clerc-Urmès, I., Grzebyk, M. & Hédelin, G. Net survival estimation with stns. *Stata J.* **14**, 87–102 (2014).
 30. Delarue, R. *et al.* Dose-dense rituximab-CHOP compared with standard rituximab-CHOP in elderly patients with diffuse large B-cell lymphoma (the LNH03-6B study): a randomised phase 3 trial. *Lancet Oncol* **14**, 525–533 (2013).
 31. McGowan, J. V *et al.* Anthracycline Chemotherapy and Cardiotoxicity. *Cardiovasc. Drugs Ther.* **31**, 63–75 (2017).
 32. National Institute for Health and Care Excellence. *Non-Hodgkin's lymphoma: diagnosis and management*. (National Institute for Health and Care Excellence, 2016).
 33. Janssen-Heijnen, M. L. *et al.* A population-based study of severity of comorbidity among patients with non-Hodgkin's lymphoma: prognostic impact independent of International Prognostic Index. *Br J Haematol* **129**, 597–606 (2005).
 34. NHS Digital. The processing cycle and HES data quality: automatic data cleaning and derivation rules. (2020). Available at: <https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/hospital-episode-statistics/the-processing-cycle-and-hes-data-quality#top>.

Appendix

Supplementary Table 1: Comorbidities and their diagnostic ICD-10 codes

Comorbidity	ICD-10
Myocardial infarction	I21.x, I22.x, I25.2
Congestive heart failure	I11.0, I13.0, I13.2, I25.5, I42.0, I42.5–I42.9, I43.x, I50.x, P29.0
Peripheral vascular disease	I70.x, I71.x, I73.1, I73.8, I73.9, I77.1, I79.0, I79.2, K55.1, K55.8, K55.9, Z95.8, Z95.9
Cerebrovascular disease	G45.x, G46.x, H34.0, I60.x–I69.x
Dementia	F00.x–F03.x, F05.1, G30.x, G31.1
Chronic obstructive pulmonary disease	I27.9, J40.x–J47.x, J60.x–J67.x, J68.4, J70.1, J70.3
Rheumatic disease	M05.x, M06.x, M31.5, M32.x–M34.x, M35.1, M35.3, M36.0
Liver disease	B18.x, K70.0–K70.3, K70.9, K71.3–K71.5, K71.7, K73.x, K74.x, K76.0, K76.2–K76.4, K76.8, K76.9, Z94.4, K71.1, K72.1, K72.9, K76.5, K76.6, K76.7, I85.0, I85.9, I86.4, I98.2, K70.4,
Diabetes without chronic complication	E10.0, E10.1, E10.6, E10.8, E10.9, E11.0, E11.1, E11.6, E11.8, E11.9, E12.0, E12.1, E12.6, E12.8, E12.9, E13.0, E13.1, E13.6, E13.8, E13.9, E14.0, E14.1, E14.6, E14.8, E14.9
Diabetes with chronic complication	E10.7, E11.2–E11.5, E11.7, E12.2–E12.5, E12.7, E13.2–E13.5, E13.7, E14.2–E14.5, E14.7
Hemiplegia or paraplegia	G04.1, G11.4, G80.1, G80.2, G81.x, G82.x, G83.0–G83.4, G83.9
Renal disease	I12.0, I13.1, N03.2–N03.7, N05.2–N05.7, N18.x, N19.x, N25.0, Z49.0–Z49.2, Z94.0, Z99.2
AIDS/HIV	B20.x–B22.x, B24.x

ICD-10: International Classification of Diseases, 10th Revision

Diabetes with/without chronic complication is combined in the RCS Charlson Comorbidity Score

Supplementary Table 2. Distribution of non-Hodgkin lymphoma subtypes for patients diagnosed in England between 2005-2013, with respective morphology and topography ICD-O-3 codes.

Index	Site group (subtype)	Progression	Topography	Morphology	n	%
1	CLL/SLL*	Indolent	C000-C809	9670, 9823	4,060	4.74
2	Waldenstrom macroglobulinemia	Indolent	C000-C809	9761	2,459	2.87
3	Mantle cell	Indolent	C000-C809	9673	3,559	4.16
4	Diffuse large B-cell	Aggressive	C000-C809	9680, 9688, 9737-9738	30,930	36.13
5	Burkitt	Aggressive	C000-C809	9687, 9826	1,083	1.27
6	Follicular	Indolent	C000-C809	9690-9691, 9695, 9698	15,645	18.27
7	Mature T cell	Aggressive	C000-C809	9702	6,096	7.12
8	Marginal zone B-cell	Indolent	C000-C809	9689, 9699, 9760, 9764, 9699	4,622	5.4
9	Not Otherwise Specified	n/a	C000-C809	9591, 9675, 9735	10,766	12.58
10	Other***	n/a	C000-C809	9591, 9675, 9735	6,378	7.45
Total					85,598	100.00**

n/a – not applicable; there was no morphological information
* Chronic lymphocytic leukaemia/Small-cell lymphocytic lymphoma
** Percentages may not equate to 100.00 due to rounding
*** The morphology code specifies these patients are diagnosed with NHL. However, the description states 'other'; these patients are classified similarly to 'Not Otherwise Specified'

Supplementary Table 3: Distribution of patient characteristics amongst patients diagnosed with ‘Not Otherwise Specified’ lymphoma subtypes in England between 2005 and 2013.

	Subtype (Morphology)		Odds Ratio (95% CI)	Wald p-value
	NOS n = 17,144 (20%)	Specified n = 68,454 (80%)		
Age (continuous)				
<i>Mean (SD)</i>	69.7 (15.2)	66.4 (14.9)	1.016 (1.01, 1.02)	<0.001
Age categories				
15-44	1,328 (7.8)	6,409 (9.4)	Ref	-
44-54	1,398 (8.2)	7,432 (10.9)	0.91 (0.84, 0.99)	0.021
55-64	2,822 (16.5)	14,118 (20.6)	0.96 (0.90, 1.04)	0.325
65-74	4,207 (24.5)	18,816 (27.5)	1.08 (1.01, 1.15)	0.028
75 or older	7,389 (43.1)	21,679 (31.7)	1.64 (1.54, 1.75)	<0.001
Comorbidity				
None	15,277 (89.1)	61,954 (90.5)	Ref	-
Mild	931 (5.4)	3,325 (4.9)	1.14 (1.05, 1.22)	0.001
Severe	936 (5.5)	3,175 (4.6)	1.20 (1.11, 1.29)	<0.001
Gender				
Male	9,234 (53.9)	37,153 (54.3)	Ref	-
Female	7,910 (46.1)	31,301 (45.7)	1.02 (0.98, 1.05)	0.332
Deprivation				
Affluent	3,583 (20.9)	14,985 (21.9)	Ref	-
2	3,620 (21.1)	15,172 (22.2)	1.00 (0.95, 1.05)	0.935
3	3,613 (21.1)	14,509 (21.2)	1.04 (0.99, 1.10)	0.122
4	3,528 (20.6)	13,205 (19.3)	1.12 (1.06, 1.18)	<0.001
Deprived	2,800 (16.3)	10,583 (15.5)	1.11 (1.04, 1.17)	<0.001
Ethnicity				
White	8,710 (50.8)	47,558 (69.5)	Ref	-
Black	197 (1.2)	798 (1.2)	1.35 (1.15, 1.58)	<0.001
Asian	286 (1.7)	1,549 (2.3)	1.01 (0.89, 1.15)	0.901
Other	123 (0.7)	535 (0.8)	1.25 (1.03, 1.53)	0.024
Missing	7,828 (45.7)	18,014 (26.3)	N/A	N/A
Stage				
I	450 (2.6)	3,742 (5.5)	Ref	-
II	214 (1.3)	2,201 (3.2)	0.81 (0.68, 0.96)	0.015
III	329 (1.9)	2,663 (3.9)	1.03 (0.88, 1.19)	0.726
IV	1,022 (6.0)	6,316 (9.2)	1.35 (1.20, 1.51)	<0.001
Missing	15,129 (88.3)	53,532 (78.2)	N/A	N/A
Route				
GP referral	5,490 (32.0)	24,087 (35.2)	Ref	-
Emergency	5,751 (33.6)	16,022 (23.4)	1.57 (1.51, 1.64)	<0.001
Inpatient elective	396 (2.3)	1,594 (2.3)	1.10 (0.97, 1.22)	0.138
Other outpatient	2,122 (12.4)	7,921 (11.6)	1.18 (1.11, 1.24)	<0.001
TWW	1,956 (11.4)	14,315 (20.9)	0.60 (0.57, 0.63)	<0.001
Unknown	818 (4.8)	2,949 (4.3)	1.22 (1.12, 1.32)	<0.001
Missing	611 (3.6)	1,566 (2.3)	N/A	N/A

Percentages may not sum to 100.0% due to rounding

Supplementary Table 4: Age-standardised 5-year net survival (ASNS) estimates by socio-demographic characteristics for patients diagnosed with ‘Not Otherwise Specified’ subtype in England during 2005-2010 and 2012-2013.

	‘Not otherwise specified’ subtype (5-year net survival %)		
	2005-2010	2012-2013	Difference
Comorbidity*			
<i>None</i>	55.8 (54.8-56.9)	78.2 (77.3-79.1)	+22.4
<i>Mild</i>	39.6 (34.8-45.1)	63.4 (58.9-68.3)	+23.8
<i>Severe</i>	37.8 (33.2-43.1)	60.8 (56.4-65.6)	+23.0
Deprivation			
<i>Affluent</i>	58.4 (56.3-60.5)	79.2 (77.5-80.9)	+20.8
2	56.1 (54.0-58.4)	77.0 (75.3-78.8)	+20.9
3	54.1 (51.9-56.4)	78.3 (76.5-80.1)	+24.2
4	52.8 (50.5-55.2)	74.3 (72.4-76.4)	+21.5
<i>Deprived</i>	49.5 (47.0-52.1)	71.2 (68.8-73.6)	+21.7
Gender			
<i>Male</i>	52.4 (51.0-53.9)	74.4 (73.2-75.7)	+22.0
<i>Female</i>	56.9 (55.5-58.4)	78.7 (77.6-79.9)	+21.8
ANS - 5-year age-standardised net survival (CI)			
NOS – not otherwise specified			
* As measured by the RCS Charlson comorbidity index score			

A.5.2 Association between patient and healthcare pathway characteristics on survival of patients with non-Hodgkin lymphoma

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	lsh1601639	Title	Mr
First Name(s)	Matthew		
Surname/Family Name	Smith		
Thesis Title	Survival of patients with non-Hodgkin lymphoma: investigating the socioeconomic inequalities		
Primary Supervisor	Edmund Njeru Njagi		

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?	Nature: Scientific Reports
Please list the paper's authors in the intended authorship order:	Matthew J. Smith, Aurélien Belot, Matteo Quartagno, Miguel Angel Luque Fernandez, Audery Bonaventure, Susan Gachau, Sara Benitez Majano, Bernard Rachet, Edmund Njeru Njagi
Stage of publication	Submitted

SECTION D – Multi-authored work

<p>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)</p>	<p>MJS, ENN and BR contributed to the conception of the study and designed the study. ENN, ABe, BR, SG, and MQ provided advice on statistical methods. MJS conducted the analyses of the data and prepared the draft of the manuscript, tables and figures. ENN, BR and MALF supervised the study and provided comments on the manuscript draft. ENN, BR, MQ, MALF, SG, SBM, ABe and ABo provided comments on the final draft of the manuscript. All authors read and approved the final manuscript.</p>
---	--

SECTION E

Student Signature	Matthew J. Smith
Date	7th June 2021

Supervisor Signature	
Date	

Title

Excess mortality by multimorbidity, socioeconomic, and healthcare factors, amongst patients diagnosed with diffuse large B–cell or follicular lymphoma in England

Authors

Matthew J. Smith^{1*}, Aurélien Belot¹, Matteo Quartagno², Miguel Angel Luque Fernandez^{1,3,4}, Audrey Bonaventure⁵, Susan Gachau^{6,7}, Sara Benitez Majano¹, Bernard Rchet¹, Edmund Njeru Njagi¹

Authors' affiliations

¹ Inequalities in Cancer Outcomes Network, Department of Non-Communicable Disease Epidemiology, London School of Hygiene and Tropical Medicine, Keppel Street, London, WC1E 7HT, UK

² MRC Clinical Trials Unit, Institute of Clinical Trials and Methodology, University College London, London WC1V 6LJ, UK

³ Noncommunicable Disease and Cancer Epidemiology Group, Instituto de Investigación Biosanitaria de Granada, Ibs.GRANADA, Andalusian School of Public Health, Granada, Spain

⁴ Centro de Investigación Biomédica en Red de Epidemiología y Salud Pública (CIBER of Epidemiology and Public Health, CIBERESP), Madrid, Spain

⁵ CRESS, Université de Paris, INSERM, UMR 1153, Epidemiology of Childhood and Adolescent Cancers Team, Villejuif, France

⁶ Health Services Unit, Kenya Medical Research Institute-Wellcome Trust Research Programme, Nairobi, Kenya

⁷ School of Mathematics, University of Nairobi, Nairobi, Kenya

Corresponding author*

Matthew J. Smith
LSHTM, Keppel Street, London, WC1E 7HT, UK
Email: matthew.smith1@lshtm.ac.uk

Word count: Abstract: 213; Text: 2638; Tables: 3; Figures: 6

Simple Summary

Diffuse large B-cell (DLBCL) and follicular lymphoma (FL) account for most non-Hodgkin lymphoma diagnoses: around 35% and 20% in England, respectively. Despite the vast contrast in survival between the subtypes, similar socioeconomic inequalities in survival have persisted over the past two decades, possibly due to the presence of comorbidity. The aim of our study was to assess the association between socioeconomic status and survival from DLBCL or FL accounting for the patients, and health system's, characteristics. We found that, for both DLBCL and FL, most deprived patients, and those with any comorbidity, had a higher excess mortality hazard compared to least deprived patients without any comorbidity. Comorbidities should be considered when planning public health interventions that target haematological malignancies in England, and further research is needed to identify specific comorbidities contributing to lower survival amongst patients with lymphomas.

Abstract

Background: Socioeconomic inequalities of survival in patients with lymphoma persists, which may be explained by patients' comorbidities. We aimed to assess the association between comorbidity and survival of patients diagnosed with diffuse large B-cell (DLBCL) or follicular lymphoma (FL) in England accounting for other socio-demographic characteristics.

Methods: Population-based cancer registry data was linked to Hospital Episode Statistics. We used a flexible multilevel excess hazard model to estimate excess mortality and net survival by patient's comorbidity status adjusted for sociodemographic, economic, healthcare factors, and accounting for the patient's area of residence. We used the latent normal joint modelling multiple imputation approach for missing data.

Results: Overall, 15,516 and 29,898 patients were diagnosed with FL and DLBCL in England between 2005-2013, respectively. Amongst DLBCL and FL, respectively, those with comorbidities had 1.23 (95% Confidence Interval -CI: 1.14–1.32) and 1.52 (95% CI 1.25–1.84) times higher excess mortality hazard compared to those without comorbidities. Patients in most deprived areas showed 1.22 (95% CI 1.18–1.27) and 1.45 (95% CI 1.30 – 1.62) times higher excess mortality hazard compared to those in least deprived areas.

Conclusion: Co/multimorbidities are consistently associated with poorer survival among patients diagnosed with DLBCL or FL. Comorbidities and multimorbidity need to be considered when planning public health interventions targeting haematological malignancies in England.

Key words: Cancer Epidemiology, Diffuse Large B-cell Lymphoma, Follicular Lymphoma, Survival Analysis, Comorbidity, Multimorbidity, Socioeconomic Status

Introduction

Non-Hodgkin lymphoma (NHL) is a heterogeneous group of malignancies, and is currently the 6th most commonly diagnosed cancer in England: in 2014, approximately 32 males and 23 females per 100,000 person-years were diagnosed.¹ The heterogeneity in morphology leads to variation in survival probability; for instance, 5-year survival of Follicular Lymphoma (FL) (86.3%) is higher than Diffuse Large B-cell Lymphoma (DLBCL) (54.8%).²

The healthcare system in England aims to offer equitable access to care for all patients. However, variability in health outcomes amongst patients with similar cancers and sociodemographic characteristics still occur;²⁻⁴ convincing reasons for the variability remains a topic of interest. In 2001, the National Health Service (NHS) Cancer Plan⁵ recognised, and aimed to reduce, the disparities in survival. Since implementation, there is no evidence the Plan had an impact on the inequalities.^{6,7} The deprivation-gap in survival is still apparent, despite the Plan and successive policies,^{5,8-10} illustrating the little understanding of the mechanisms underlying these inequalities and raising the concern that these policies have missed the relevant targets.

Patients' comorbidity status may impact timely diagnosis, possibly leading to treatment with more adverse effects;¹¹ comorbidities are, on average, more prevalent and severe amongst more deprived patients.¹² However, recent evidence indicates that comorbidity explains little of the differential cancer survival between socioeconomic groups.¹³⁻¹⁵ Variations in healthcare access, such as location of residence, could partly explain the inequalities.¹⁶⁻²⁰

Since population-based cancer registries rarely hold reliable information on the cause of death, cancer-specific mortality estimates can be estimated with relative survival methods. These methods compare

the mortality hazard (i.e., excess mortality hazard) observed in a population of cancer patients to the mortality hazard observed in the general population with identical demographic characteristics. In this context, the survival estimate derived from the excess mortality hazard is termed net survival (or cancer survival), which is interpreted as the survival where death is due directly, or indirectly, to the cancer studied, and death from other causes has been removed.²¹

Overall, the association between comorbidity and cancer survival in patients with DLBCL and FL, accounting for other socio-demographic characteristics and the area of residence, remains unclear. We aim to describe the association between comorbidities and cancer survival amongst DLBCL or FL patients, while accounting for sociodemographic and economic factors, hypothesizing that the presence of comorbidities is associated with poorer survival.

Methods

Study design, participants, and data sources

We developed a population-based multilevel cohort study of adult patients diagnosed with DLBCL or FL between 1st January 2005 and 31st December 2013 in England. Patients were followed up until death or the end of the study at the 31st of December 2015, whichever occurred first.

DLBCL and FL were defined according to the 10th revision of the International Statistical Classification of Diseases and Related Problems (ICD-10 codes C82.0-C85.9).²² Morphology (cell type) and topography (tumour site) were defined using renewed updates of the ICD for Oncology (ICD-O); ICD-O-3²³ was used for diagnoses up to 2010, and ICD-O-3.1²⁴ for diagnoses after 2011. Information on patients with DLBCL or FL was collected from the linkage of English cancer registry data, the Cancer Analysis System²⁵ (CAS) and Hospital Episode Statistics²⁶ (HES) data sets within the

national cancer registry and analysis service (NCRAS). These datasets contained detailed information on patient's and tumour's characteristics (see details below).

Outcome, exposure, and patients' sociodemographic characteristics

The outcome of the study was the time to death, or censoring, among DLBCL and FL patients 5 years after cancer diagnosis. Net survival was deduced after estimating the excess mortality hazard. Hence, we used England life tables stratified by deprivation, sex, age and calendar year (2005-2013) to account for the overall mortality rate from the background population.²⁷ As follow up of patients ended in 2015 and life tables were available until 2013, we assumed that the expected mortality rates plateau for 2014 and 2015.

Comorbidity status was the main exposure. We defined comorbidity as the existence of other chronic medical disorders, in addition to cancer, the primary disease of interest, which are causally unrelated to the primary disease.^{28,29} Records from HES were used to identify patients' comorbidity status based on a computational algorithm published elsewhere.³⁰ The algorithm seeks for the presence of comorbidities retrospectively and defines a time window of 6 to 24 months prior to cancer diagnosis where comorbidities are recorded to avoid bias due to the presence of comorbidities related to cancer (i.e., cardiological comorbidities due to DLBCL or FL cancer treatment). Patient's comorbidity status was adapted from the original Charlson comorbidity index³¹ (CCI). We used the Royal College of Surgeons (RCS) modified Charlson Score (**Appendix Table A1**).³² The score removes patients with a previous malignancy to avoid bias, does not assign different weights to comorbidities, and categorises comorbidities as: no comorbidities, one comorbidity and two or more comorbidities (multimorbidity).

Socio-demographic and economic characteristics were collected from the HES dataset. Age was specified at time of diagnosis. Sex is recorded as male or female. Ethnicity was recorded as white or

other. Area-level deprivation, classified into one of five quintiles, was determined by the Index of Multiple Deprivation³³ (IMD), which was based on the Lower Super Output Area³⁴ (LSOA) residence of the patient at the time of cancer diagnosis. LSOA is a geographical location with a median of 1500 inhabitants. We also include the information regarding patients' diagnosis path (route to diagnosis), a UK specific programme, classified as: accident and emergency room diagnosis, general practitioner referral (routine and urgent referrals where the patient was not referred under two-week-wait), two-week-wait (urgent GP referral with a suspicion of cancer), and secondary care diagnosis (other outpatient and inpatient elective routes).³⁵

Statistical Analysis

We tabulated the sociodemographic characteristics by DLBCL and FL. To estimate the excess mortality hazard, we used a multilevel excess hazard regression model (EHM) with a cubic B-spline with two knots placed at 1 year and at 3 years after diagnosis for the baseline hazard $\lambda_0(t)$. We accounted for the hierarchical structure of the data via the inclusion of a random effect.³⁶ The statistical contribution of the random effect to the overall goodness of fit of the model was tested using a likelihood ratio test statistic with a Chi-square mixture distribution.³⁷ From the estimated excess hazard, we could deduce the net survival via the classical relationship between hazard and survival.³⁸ Net survival is the survival associated with the cancer under study, after eliminating the other cause of death.

In the EHM we included the following variables: age, sex, comorbidities (categorical, 3 categories), deprivation (categorical, 5 categories), lymphoma subtype, ethnicity, route of cancer diagnosis. We included the non-linear effect of age using a regression spline (defined using a truncated power basis) with one knot located at 70 years of age. Furthermore, we assumed a time-dependent effect of age at diagnosis, represented by the interaction between B-spline function of time and age. The parameter

estimates for the variables were interpreted conditionally on the random effect, i.e., they have a cluster-specific interpretation, where a cluster refers to a given LSOA. From the model we derived the excess mortality hazard ratios (EMHR) and their respective 95% confidence intervals (CI) for all the categorical variables, and the variance of the random effect for the LSOA. Empirical Bayes estimates of the random effect were used to explore the between-LSOA variability in the excess mortality hazard from DLBCL or FL. The random effect was tested for using a likelihood ratio test, with the reference distribution being a mixture of chi-squared distributions with 0 and 1 degrees of freedom, to account for the well-known boundary problem for random effects variances.^{39,40}

Missing data analysis

We explored the missing data mechanism for each of the three variables with missing data (ethnicity [FL 24.9%, DLBCL: 22.7%], route [FL: 7.8%, DLBCL 5.0%]). Due to clustered data and partially observed categorical variables, we used the latent normal joint modelling multiple imputation approach, under a missing at random assumption (MAR).⁴¹ The imputation model included all fully- and partially-observed variables, vital status indicator, the Nelson-Aalen estimate of the cumulative overall hazard, and accounted for clustering of patients within lower-super output areas. We generated 10 imputed datasets. The multilevel EHM was fitted to each of these datasets, and results combined using Rubin's rules.^{42,43} Overall tests for the effects of age after multiple imputation were done using the F-based procedure for the test of multiple parameters after multiple imputation.⁴¹

We used R software for all data analyses; the *mexhaz*³⁶ package was used for excess hazard modelling and the *jomo*⁴⁴ package for multiple imputation.

Results

Overall, 15,516 (34.2%) patients were diagnosed with FL and 29,898 (65.8%) diagnosed with DLBCL in England between 2005 and 2013 (**Table 1**). The prevalence of at least one comorbidity was higher amongst DLBCL (10.7%) compared to FL (7.5%). The average age was lower amongst FL compared to DLBCL, 63.9 compared to 67.4 years, respectively. The prevalence of DLBCL was higher amongst deprived areas (16.0%) than FL (14.4%). ‘White’ was the most prevalent ethnicity for both FL (94.9%) and DLBCL (94.1%). GP referral was the most common route to diagnosis amongst FL (44.0%); whereas, amongst DLBCL, A&E was most common (33.8%).

Table 1: Distribution of cancer subtypes by patient and healthcare system characteristics for patients (n=45,414) diagnosed with non-Hodgkin lymphoma in England during the period 2005-2013.

	Subtype of NHL			
	FL N = 15,516		DLBCL N = 29,898	
Age (mean, SD)	63.9 (13.6)		67.4 (14.9)	
Sex, n(%)				
<i>Male</i>	7,318	(47.2%)	16,215	(54.2%)
<i>Female</i>	8,198	(52.8%)	13,683	(45.8%)
Deprivation quintiles (Q), n(%)				
<i>Least deprived (Q1)</i>	3,547	(22.9%)	6,340	(21.2%)
<i>Q2</i>	3,517	(22.7%)	6,663	(22.3%)
<i>Q3</i>	3,294	(21.2%)	6,246	(20.9%)
<i>Q4</i>	2,925	(18.9%)	5,863	(19.6%)
<i>Most deprived (Q5)</i>	2,233	(14.4%)	4,786	(16.0%)
Comorbidity status, n(%)				
<i>No comorbidity</i>	14,343	(92.4%)	26,718	(89.4%)
<i>One comorbidity</i>	641	(4.1%)	1,570	(5.3%)
<i>Multimorbidity</i>	532	(3.4%)	1,610	(5.4%)
Route of diagnosis, n(%)				
<i>GP referral</i>	6,297	(44.0%)	8,157	(28.7%)
<i>A & E</i>	1,869	(13.1%)	9,617	(33.8%)
<i>Secondary care</i>	2,222	(15.5%)	3,724	(13.1%)
<i>TWW</i>	3,912	(27.4%)	6,918	(24.4%)
<i>Missing*</i>	1,216	(7.8%)	1,482	(5.0%)
Ethnicity, n(%)				
<i>White</i>	11,052	(94.9%)	21,739	(94.1%)
<i>Others</i>	600	(5.2%)	1,369	(5.9%)
<i>Missing*</i>	3,864	(24.9%)	6,790	(22.7%)

GP: general practitioner referral, A&E: accident and emergency room, TWW: two-week-wait

Complete case analysis: missing ethnicity 23.5%; missing route to diagnosis 5.9%

* Proportions are of the total number of patients

In the multivariable analysis (**Table 2a**), amongst DLBCL, and after multiple imputation, patients with comorbidity and multimorbidity showed 23% and 40% increased excess mortality compared to patients without comorbidity (i.e., EMHR: 1.23; 95% CI: 1.14 – 1.32, and EMHR: 1.40; CI: 1.01 – 1.94, respectively). Patients living in the most deprived areas had 1.22 (95% CI: 1.18 – 1.27) times higher excess mortality than those living in the least deprived areas. Patients diagnosed through A&E had nearly three times a higher excess mortality compared to GP referral (i.e., EMHR: 2.75; 95% CI: 2.54 – 2.98). Females had a significantly lower excess mortality compared to males (i.e., EMHR 0.93; 95% CI: 0.90–0.96). There was, however, no evidence of a difference in excess mortality by ethnicity (**Table 2a**). Using a likelihood ratio test (a mixture of chi-square distributions), there was strong evidence ($p<0.001$) that including the random effect improved the fit of the model.

Table 2a: Adjusted excess mortality hazard ratios for age, sex, deprivation, comorbidity, cancer subtype, route of diagnosis, ethnicity, and LSOA as random intercept for (i) complete-case analysis, and (ii) after multiple imputation for patients (n=29,898) diagnosed with **diffuse large B-cell lymphoma** in England during the period 2005-2013.

		Model (i): Complete Case			Model (ii): After Imputation		
		HR	CI	p-value	HR	CI	p-value
Sex							
	<i>Male</i>	Ref	Ref		Ref	Ref	
	<i>Female</i>	0.93	0.89 – 0.98	0.003	0.93	0.90 – 0.96	<0.001
Ethnicity							
	<i>White</i>	Ref	Ref		Ref	Ref	
	<i>Other</i>	0.97	0.87 – 1.08	0.556	0.99	0.91 – 1.08	0.809
Deprivation quintiles (Q)							
	<i>Least deprived Q1</i>	Ref	Ref		Ref	Ref	
	<i>Q2</i>	1.03	0.96 – 1.11	0.372	1.00	0.93 – 1.08	0.922
	<i>Q3</i>	1.08	1.00 – 1.16	0.045	1.07	1.00 – 1.14	0.045
	<i>Q4</i>	1.17	1.08 – 1.26	<0.001	1.13	1.04 – 1.23	0.003
	<i>Most deprived Q5</i>	1.26	1.16 – 1.37	<0.001	1.22	1.18 – 1.27	<0.001
Comorbidity status							
	<i>No comorbidity</i>	Ref	Ref		Ref	Ref	
	<i>One comorbidity</i>	1.26	1.15 – 1.38	<0.001	1.23	1.14 – 1.32	<0.001
	<i>Multimorbidity</i>	1.50	1.38 – 1.64	<0.001	1.40	1.01 – 1.94	0.043
Route of diagnosis							
	<i>GP referral</i>	Ref	Ref		Ref	Ref	
	<i>A & E</i>	2.75	2.60 – 2.91	<0.001	2.75	2.54 – 2.98	<0.001
	<i>Secondary Care</i>	1.43	1.22 – 1.67	<0.001	1.23	1.11 – 1.36	<0.001
	<i>TWW</i>	1.33	1.23 – 1.45	<0.001	0.83	0.56 – 1.24	0.362
Random Effect							
	SD (SE)	0.48 (0.08)	-	-	0.39 (0.04)	-	-

GP: general practitioner referral. A&E: accident and emergency room. TWW: two-week-wait

In the multivariable analysis (**Table 2b**), amongst FL, patients with comorbidity and multimorbidity showed 1.52 and 2.19 times the excess mortality compared to patients without comorbidity (i.e., EMHR: 1.52; 95% CI: 1.25 – 1.84, and EMHR: 2.19; CI: 1.45 – 3.31, respectively). Patients living in the most deprived areas had 1.45 (95% CI: 1.30 – 1.62) times higher excess mortality than those living in the least deprived areas. Patients diagnosed through A&E had nearly three times a higher excess mortality compared to GP referral (i.e., EMHR: 3.32; 95% CI: 2.49 – 4.43). Females had a significantly lower excess mortality compared to males (i.e., EMHR 0.89; 95% CI: 0.81–0.97). There was, however, no evidence of a difference in excess mortality by ethnicity (**Table 2b**). Using likelihood ratio test (a mixture of chi-square distributions), there was strong evidence ($p < 0.001$) that including the random effect improved the fit of the model.

Table 2b: Adjusted excess mortality hazard ratios adjusted for age, sex, deprivation, comorbidity, cancer subtype, route of diagnosis, ethnicity, and LSOA as random intercept for (i) complete-case analysis, and (ii) after multiple imputation for patients (n=15,516) diagnosed with **follicular lymphoma** in England during the period 2005-2013.

		Model (i): Complete Case			Model (ii): After Imputation		
		HR	CI	p-value	HR	CI	p-value
Sex							
	<i>Male</i>	Ref	Ref		Ref	Ref	
	<i>Female</i>	0.86	0.76 – 0.96	0.010	0.89	0.81 – 0.97	0.009
Ethnicity							
	<i>White</i>	Ref	Ref		Ref	Ref	
	<i>Other</i>	0.59	0.41 – 0.83	0.003	0.76	0.60 – 0.96	0.019
Deprivation quintiles (Q)							
	<i>Least deprived Q1</i>	Ref	Ref		Ref	Ref	
	<i>Q2</i>	1.09	0.91 – 1.31	0.364	1.10	0.92 – 1.32	0.309
	<i>Q3</i>	1.23	1.02 – 1.48	0.030	1.11	0.96 – 1.29	0.166
	<i>Q4</i>	1.37	1.13 – 1.65	0.001	1.34	1.06 – 1.69	0.015
	<i>Most deprived Q5</i>	1.69	1.38 – 2.06	<0.001	1.45	1.30 – 1.62	<0.001
Comorbidity status							
	<i>No comorbidity</i>	Ref	Ref		Ref	Ref	
	<i>One comorbidity</i>	1.51	1.19 – 1.91	<0.001	1.52	1.25 – 1.84	<0.001
	<i>Multimorbidity</i>	2.38	1.90 – 3.00	<0.001	2.19	1.45 – 3.31	<0.001
Route of diagnosis							
	<i>GP referral</i>	Ref	Ref		Ref	Ref	
	<i>A & E</i>	3.18	2.69 – 3.76	<0.001	3.32	2.49 – 4.43	<0.001
	<i>Secondary Care</i>	1.27	0.86 – 1.90	0.233	1.22	0.96 – 1.55	0.107
	<i>TWW</i>	1.17	0.98 – 1.40	0.084	1.06	0.63 – 1.78	0.830
Random Effect							
	SD (SE)	0.87 (0.14)	-	-	0.69 (0.16)	-	-

GP: general practitioner referral. A&E: accident and emergency room. TWW: two-week-wait

Figures 1 and 2 show the EMHR for patients with DLBCL and FL, respectively., according to age at diagnosis at different time since diagnosis (figures 1A and 2A), and according to time since diagnosis for different age at diagnosis (figures 1B and 2B). The excess mortality hazard for DLBCL and FL patients for different values of age at diagnosis is shown in the appendix (figures A1 and A2, respectively). These plots were obtained from the 3-dimensional plots of EMHR, as shown in the appendix (figures A3 and A4, respectively). For DLBCL (**figure 1**), the EMHR was higher for older patients whatever the follow-up time (**figure 1A**). For those of older or younger ages, in comparison to 70-year-olds, the EMHR was markedly different immediately after, or at 5 years since, diagnosis, but was most similar around 18 months after diagnosis (**figure 1B**).

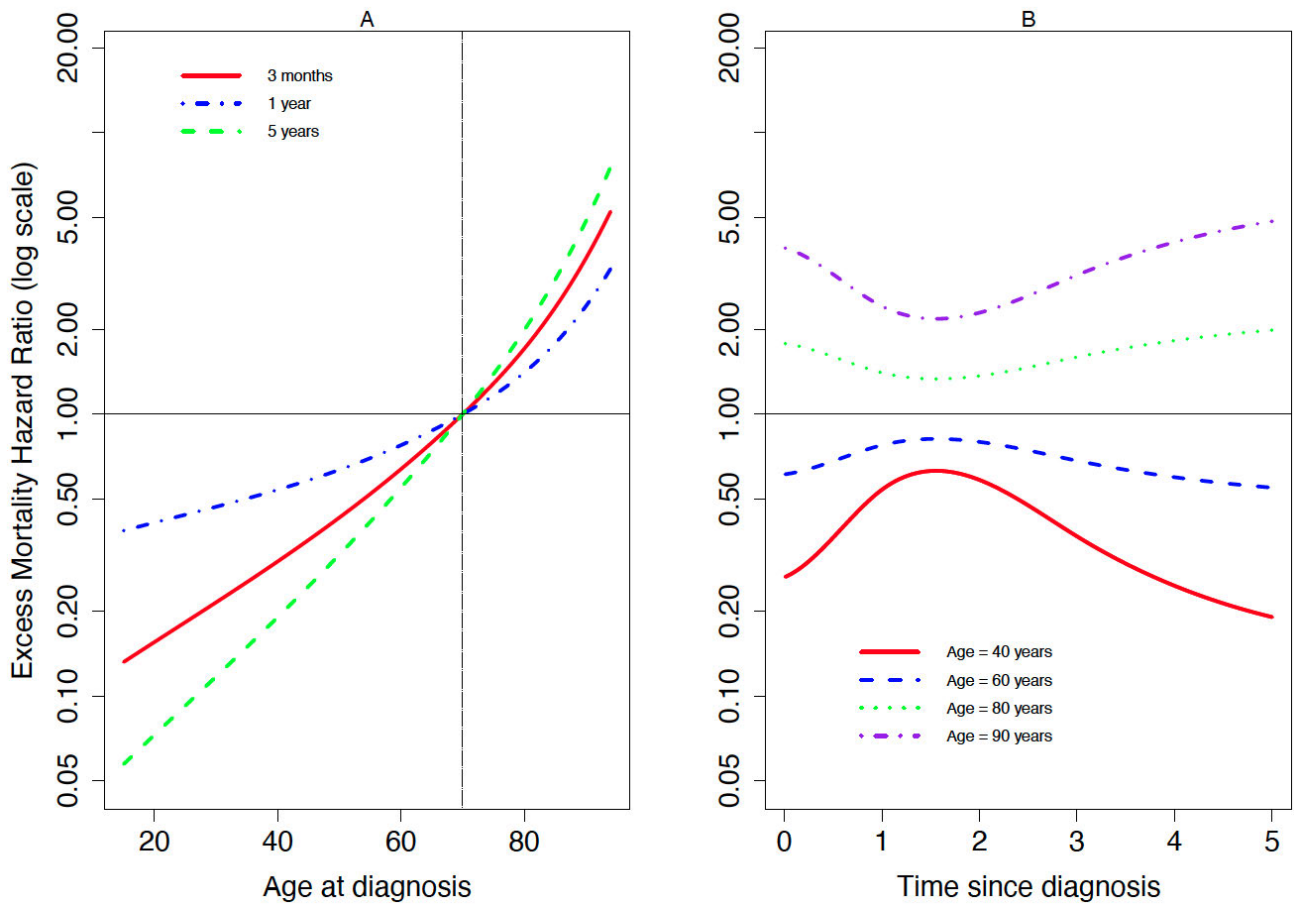


Figure 1: Excess mortality hazard ratios according to (A) age at diagnosis at different time since diagnosis (3 months, 1- and 5-years), and (B) time since diagnosis for different age groups, amongst patients diagnosed with **diffuse large B-cell lymphoma** (n=29,898) in England during 2005-2013.

For FL (**figure 2**), the non-linear effect of age was almost similar whatever the time since diagnosis; being older is associated with a higher excess mortality hazard (**figure 2A**). For those of older or younger ages, in comparison to 70-year-olds, the EMHR was markedly different immediately after or 5 years since diagnosis but was most similar around 18 months after diagnosis (**figure 2B**).

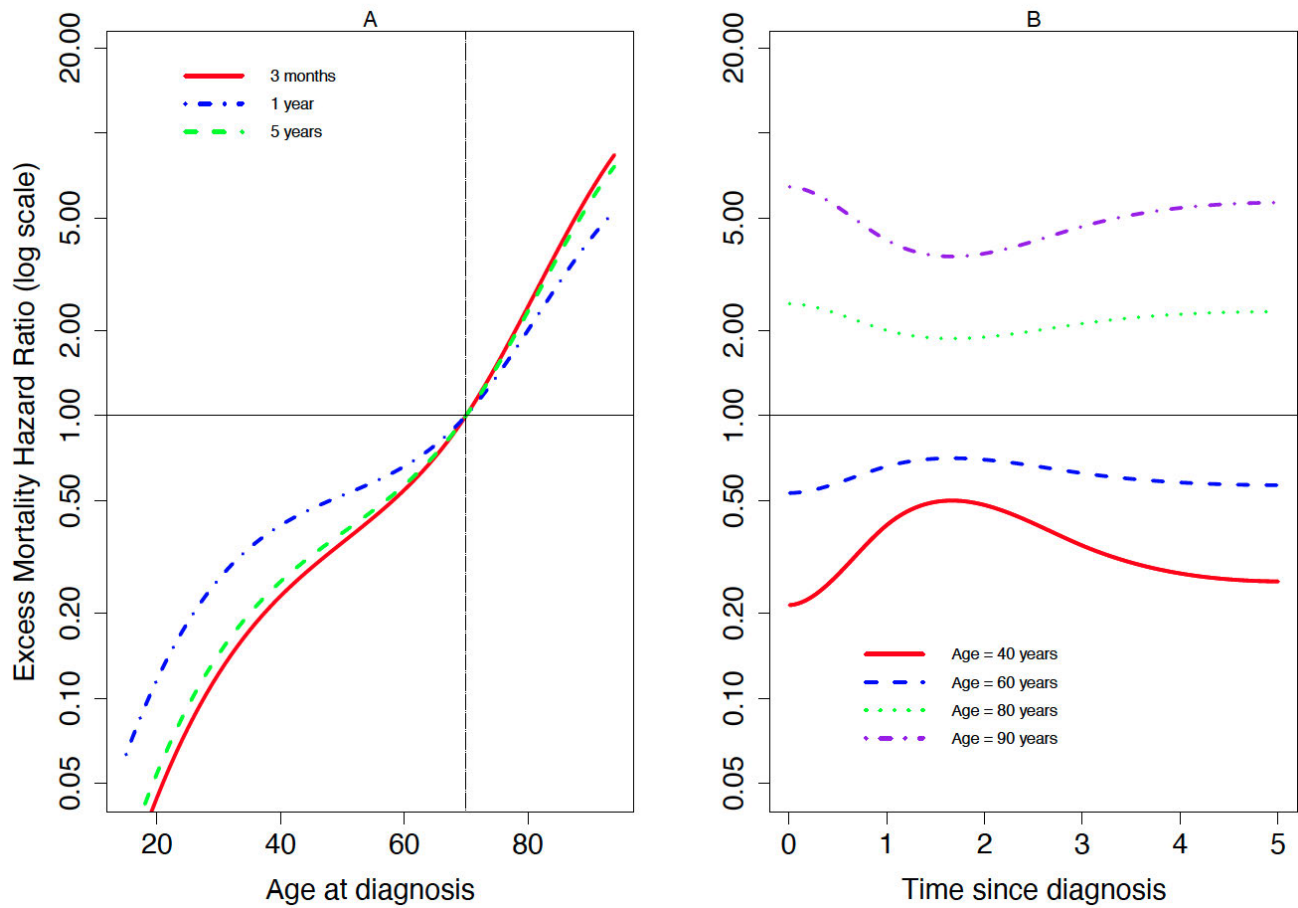


Figure 2: Excess mortality hazard ratios according to (A) age at diagnosis at different time since diagnosis (3 months, 1- and 5-years), and (B) time since diagnosis for different age groups, amongst patients diagnosed with **follicular lymphoma** (n=15,516) in England during 2005-2013.

Figures 3 and 4 show the net survival probability as predicted from the regression model amongst patients with DLBCL and FL, respectively. Amongst DLBCL patients (**figure 3**), those living in more deprived areas experienced approximately 7% lower 5-year survival compared to patients in least deprived areas (e.g., 5-year net survival, amongst those without comorbidities, was 56% for least deprived compared to 49% for most deprived). Amongst FL patients (**figure 4**), those living in more deprived areas experienced approximately 4% lower 5-year survival compared to those living in least deprived areas (e.g., 5-year net survival, amongst those without comorbidities, was 86% for least deprived compared to 82% for most deprived). For DLBCL only (**figure 3**), the deprivation gap in survival was apparent from approximately 6 months after diagnosis regardless the comorbidity status.

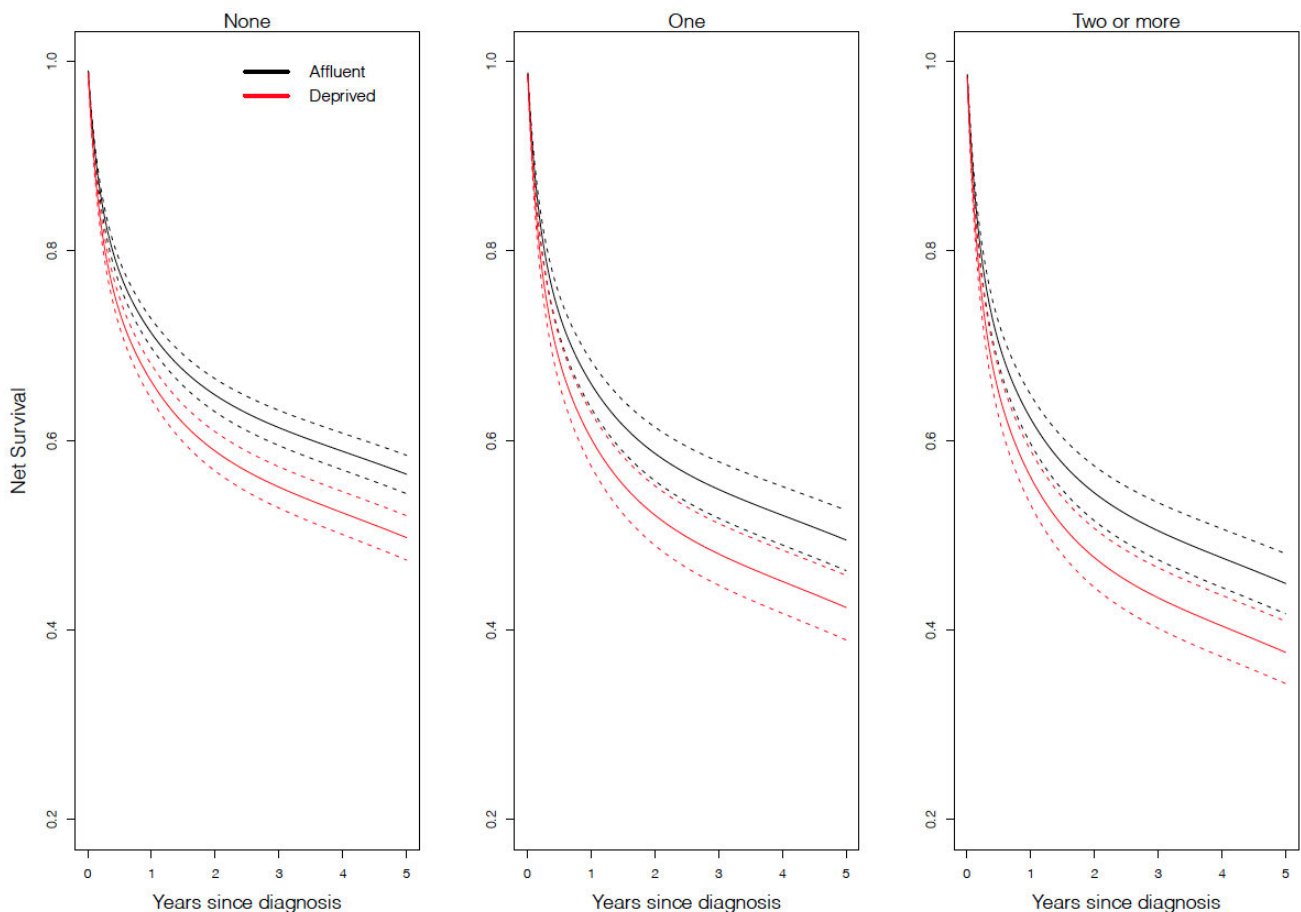


Figure 3: Net survival model-based prediction for **diffuse large B-cell lymphoma** for each comorbidity status by deprivation level (n=29,898) in England between 2005-2013.

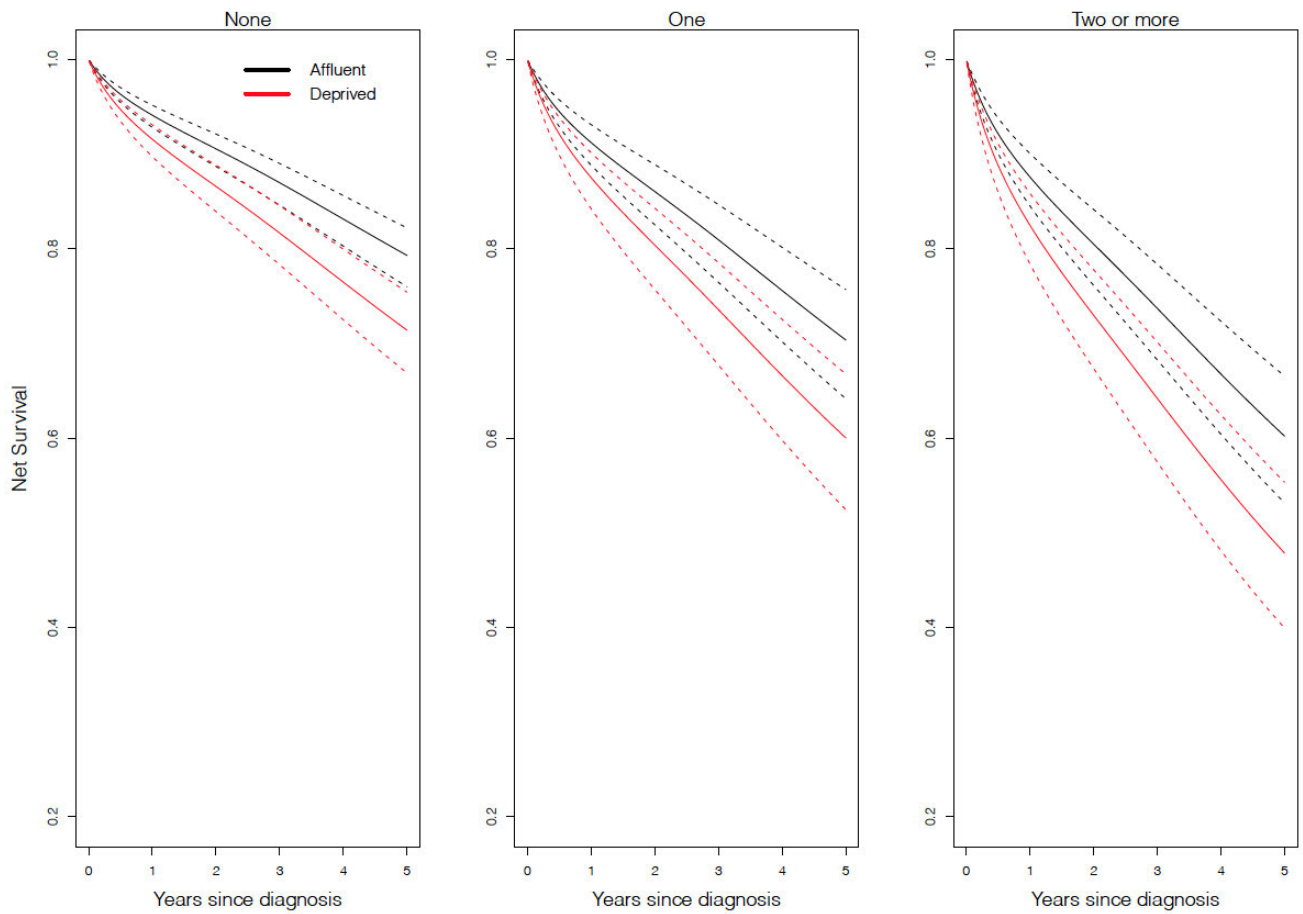


Figure 4: Net survival model-based prediction for **follicular lymphoma** for each comorbidity status by deprivation level (n=15,516) in England between 2005-2013.

We graphically illustrate the empirical Bayes (EB) estimates of the LSOA random effect for the excess mortality hazard from DLBCL and FL (figures 5 and 6, respectively). A positive EB estimate indicated a higher excess mortality hazard for a patient from that LSOA in comparison to a patient who has similar observed characteristics but from a LSOA with either a less positive, or negative EB estimate. The EB estimates were grouped by deprivation level to which the LSOA contributed. For both DLBCL and FL (figures 5 and 6), the results showed there were no outliers and approximately equal distribution of the EB estimates for each deprivation level.

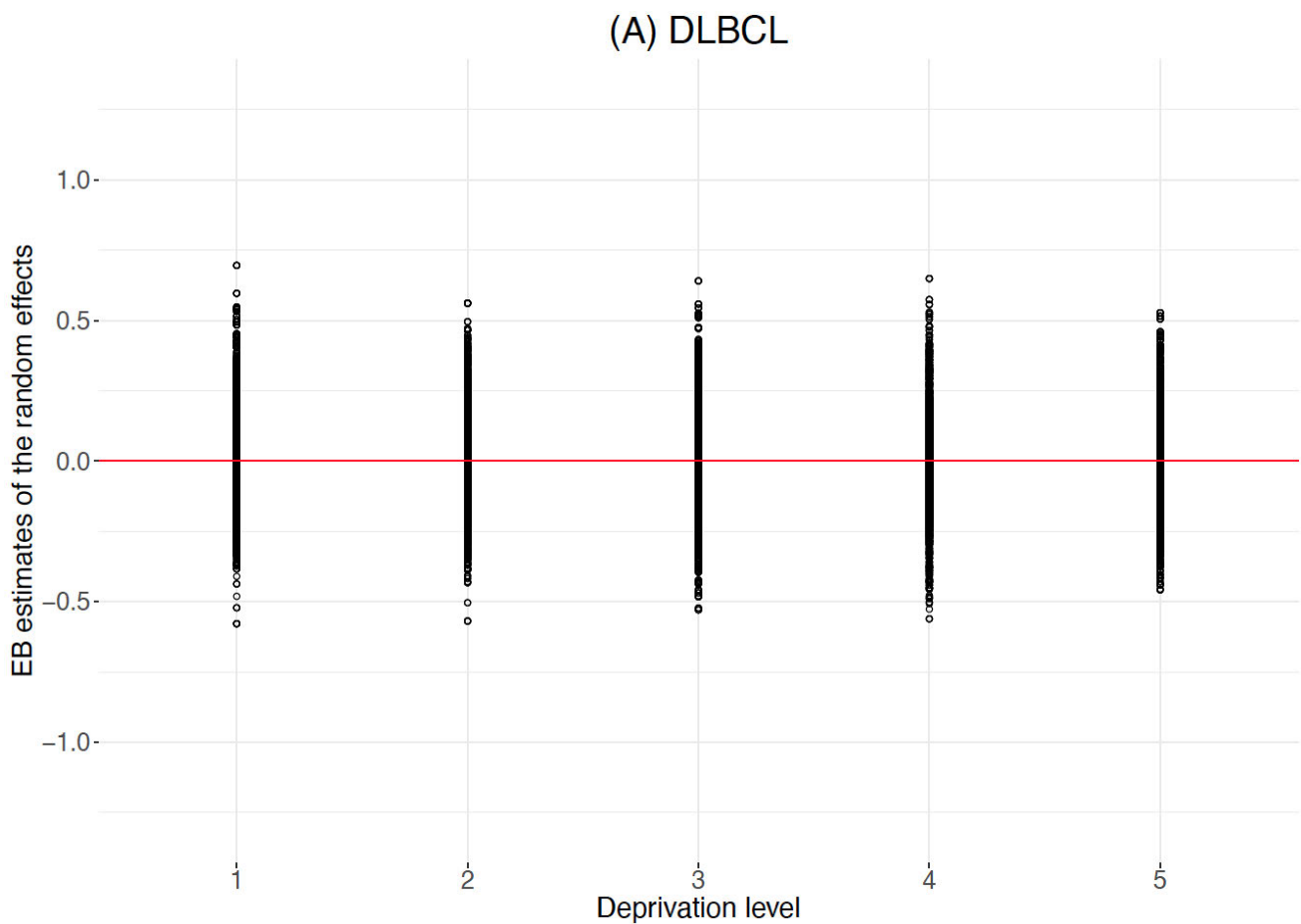


Figure 5: Empirical Bayes estimates of the random effect of LSOA from the excess mortality hazard model for patients diagnosed with **diffuse large B-cell lymphoma** (n=29,898) in England during 2005-2013.

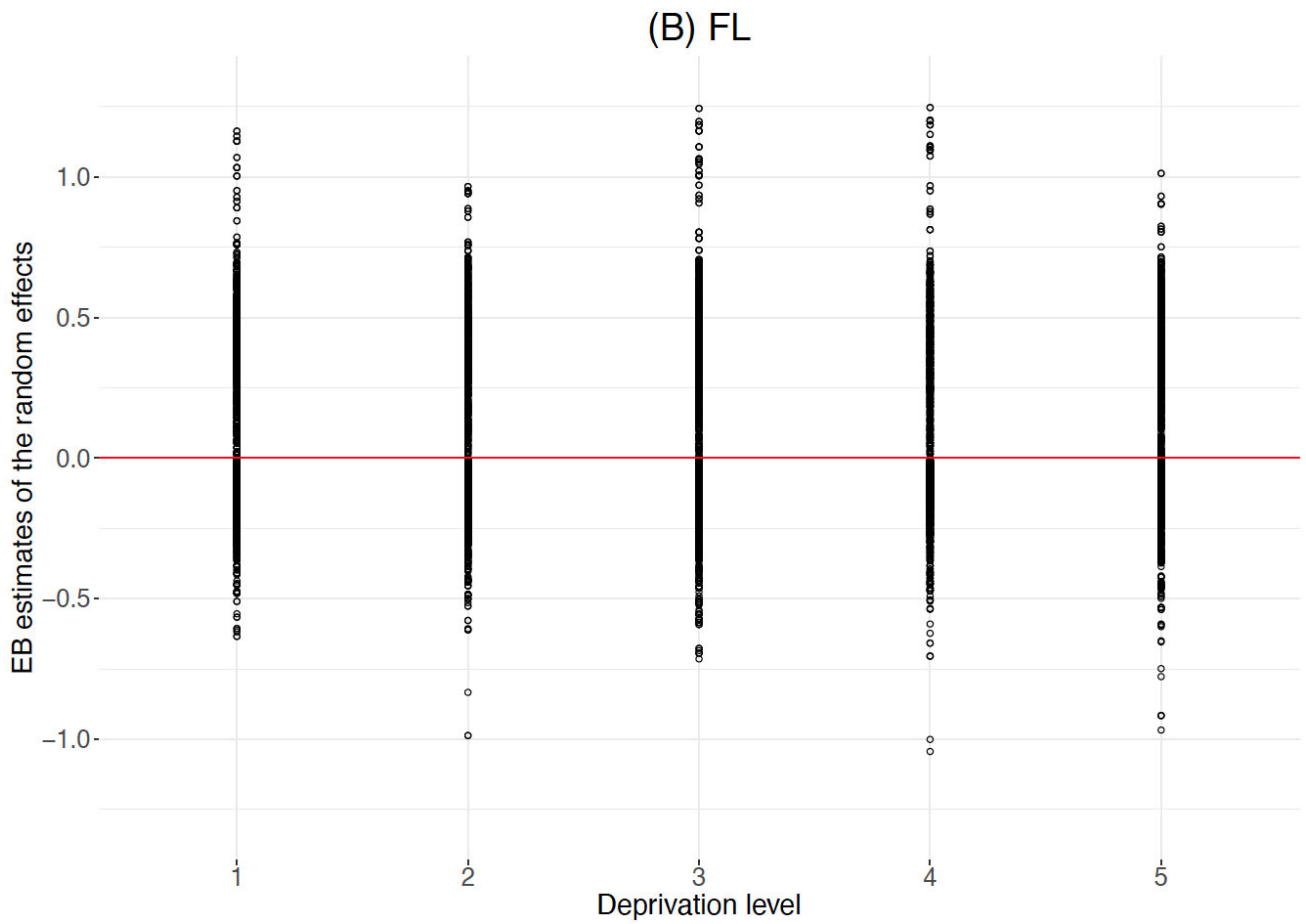


Figure 6: Empirical Bayes estimates of the random effect of LSOA from the excess mortality hazard model for patients diagnosed with **follicular lymphoma** (n=15,516) in England during 2005-2013.

Discussion

We found strong evidence of a higher excess mortality amongst DLBCL, and FL, patients diagnosed with comorbidities compared to patients without comorbidities after adjusting for age, deprivation level, ethnicity, route to diagnosis and accounting for the patient's area of residence; we also found a noticeable deprivation gap in cancer survival.

Differences in access to treatments, or risk of adverse effects, may explain some of the disparities in survival among DLBCL patients. Immunotherapies (rituximab) for the treatment of aggressive lymphomas (e.g. DLBCL) was shown to be effective for those of an advanced age.⁴⁵⁻⁴⁷ Rituximab is often used in combination with doxorubicin, an increase in dosage of which is associated with an increased incidence of adverse effects (cardiotoxicity), such as congestive heart failure.⁴⁸ Guidelines based on National Institute for Health and Care Excellence (NICE) recommend that patients at risk of cardiotoxicity, or low tolerance of intensive therapy, consider a less-intensive treatment regimen.⁴⁹⁻⁵¹ This less-intensive treatment allocation may partly explain the comorbidity inequalities in survival from DLBCL. For patients with FL, the standard management is 'watch-and-wait'; thus, in the absence of a treatment, the comorbidity inequalities in survival may be largely explained by the presence of a comorbidity itself rather than being explained by the effect of comorbidity on treatment.

For FL patients, we showed that the excess mortality hazard among older patients compared to younger patients is highest just after 4 years since diagnosis (**figure 2B**). Since we accounted for background population mortality, and adjusted for comorbidity, the higher excess hazard could be because of histological transformation from lower to higher grades of FL. Studies suggest the risk of histological transformation increases by 3% per year since cancer onset.⁵² Therefore, the increased excess hazard amongst older patients may be because histological transformation complicates the treatment and management of FL.

The importance of understanding the association of comorbid conditions on cancer patients' outcomes has been well documented.⁵³ To our knowledge, this is the first study of England cancer registry data that investigates survival by comorbidity status among DLBCL and FL patients. Our results are consistent with previous findings from a Danish study that showed the hazard of death increased with severity of comorbidity status;⁵⁴ however, the study did not account for missing data and the association with comorbidities was potentially overestimated. Indeed, the EHR associated with comorbidity decreased after accounting for missing data. The deprivation-gap in survival persists even after accounting for prognostic factors such as comorbidity.⁵⁴⁻⁵⁶ Smith *et al.*³ reported no deprivation-gap in survival; however, their study may have lacked power, and the study used the relative survival ratio, which can be biased over longer-term follow up.⁵⁷

Consistent with previous studies,⁴ survival after GP referral (non-emergency) diagnosis is significantly better compared to A&E. However, our study also finds that patients diagnosed through TWW, who would be expected to have worse symptoms and survival, showed no evidence of a difference in survival compared to GP referral. There are two possible reasons for the absence of a difference in the associations. Firstly, GPs could advocate for a prompt referral even though the patient is not on the TWW pathway: resulting in patients with similar access to healthcare facilities. Secondly, on the other hand, patients referred through the TWW pathway have more severe symptoms and expected to have a higher excess hazard. Our results show no difference in the excess mortality indicating that TWW pathway prevents patients with more severe symptoms from having a higher excess hazard: suggesting the performance of TWW pathway is at least as beneficial to a patient's survival as GP referral. Other studies have suggested ways to improve outcomes for patients diagnosed with comorbidities, which include: novel treatment strategies,⁵⁸ inclusion of elderly patients in clinical trials,^{59,60} and investigation of dose-allocation amongst those with higher comorbidity scores.⁶¹ However, further

factors associated with the interactions between comorbidities and health care systems leading to poorer survival among DLBCL and FL cancer patients need to be studied.

The strengths of this study are that, firstly, we used a large population-based sample size obtained from cancer registry databases linked to HES, which encompasses all patients in England with a diagnosis of DLBCL and FL between 2005 and 2013. HES data encapsulates a national coverage of comorbidities diagnosed during a hospital admission and may have missed comorbidities diagnosed during primary care (e.g., diabetes diagnosed during a GP consultation). However, the addition of information provided from comorbidity records captured during primary care does not improve the prediction of cancer patient survival beyond what is captured in HES data.⁶² For example, information on comorbidities, such as diabetes, diagnosed outside of hospital admission are likely to have a minimal impact on the prediction of the patient survival beyond information captured in HES. Secondly, we used the Royal College of Surgeons' adaptation³² of the Charlson comorbidity score, which provides a more valid measure of the patient's comorbidity status because it was developed within the England population healthcare data setting. Thirdly, we used a latent normal joint modelling multiple imputation to treat missing data in ethnicity and route of diagnosis. This approach allows imputation of a mix of variable types, while accounting for multilevel structures arising from clustering of patients within LSOAs.^{41,63,64} We assumed that missing data on partially observed variables were missing at random, given the observed variables; further analysis could explore the violation to this assumption and impute under a missing not at random assumption.

This study has its limitations. Firstly, individual-level socioeconomic measures are recommended in addition to area-level measures.⁶⁵ Information on individual-level socioeconomic status was unavailable, but using area-level measures captures the multidimensional composition of a patient's deprivation level in addition to the contextual level.^{33,66} Furthermore, using area-level measures, there

is greater consistency in the measurement of deprivation between time periods because deprivation scores have a high concordance amongst updates.³³ Secondly, due to data availability, we did not include tumour stage, which may have partly explained the socioeconomic inequalities in survival. However, even though reliable estimates can be obtained after multiple imputation of partially observed variables with high proportions of missing data,⁶⁷ the inclusion of tumour stage may not have provided further information for the prediction of survival beyond that of diagnostic route because late cancer stage is strongly associated with delayed diagnostic route.⁶⁸

Survival at 1- and 5-years since diagnosis of DLBCL and FL in England trails that of other European countries;⁶⁹ however, restricting estimates to those surviving at least 1 year after diagnosis (conditional survival) shows a comparable 5-year survival.⁷⁰ This indicates that long-term survival differences are largely explained by the increased short-term mortality. Understanding long-term survival from FL is more complex due to the histological transformation of indolent lymphomas, which would require an adaptation of the treatment, support, and management from healthcare facilities. This adaptation could be compounded by the patient's susceptibility to cardiotoxic treatments. Further studies could focus on the mechanisms and inequalities of short-term mortality, long-term survival of patients with transformed lymphomas, and survival of patients at risk of cardiotoxicity.

Conclusion

After accounting for sociodemographic factors, healthcare factors, socioeconomic deprivation, and the patient's area of residence, comorbidities were consistently associated with poorer survival and an increased excess mortality amongst patients with DLBCL or FL in England. Furthermore, survival inequalities between socioeconomic levels in patients with DLBCL or FL persist after accounting for the presence of comorbidities and multimorbidities. These results show the need for the current framework of the National Health Service to improve the survival of DLBCL and FL patients in the

most deprived areas of England, and further consideration is needed for patient-tailored management plans amongst patients with comorbidity or multimorbidity.

Acknowledgements

We would like to thank Adrian Turculet, Data Manager of the LSHTM Inequalities in Cancer Outcomes Network, for his support and assistance with the data linkage.

Author's contributions

Conceptualization, Matthew Smith, Aurelien Belot, Bernard Rachet and Edmund Njeru Njagi; Formal analysis, Matthew Smith; Methodology, Aurelien Belot, Matteo Quartagno, Audrey Bonaventure, Susan Gachau, Sara Benitez Majano, Bernard Rachet and Edmund Njeru Njagi; Supervision, Aurelien Belot, Miguel Angel Luque Fernandez, Bernard Rachet and Edmund Njeru Njagi; Writing – original draft, Matthew Smith; Writing – review & editing, Aurelien Belot, Matteo Quartagno, Miguel Angel Luque Fernandez, Audrey Bonaventure, Susan Gachau, Sara Benitez Majano, Bernard Rachet and Edmund Njeru Njagi.

Declarations

Funding: This research was funded by Cancer Research UK grant number C7923/A18525. The authors declare no support from any organisations for the submitted work. The design of the study, the analyses, and the writing of the manuscript were solely the responsibility of the authors. The findings and conclusions in this manuscript are those of the authors and do not necessarily represent the views of Cancer Research UK.

Availability of data and materials: The data that support the findings of this study are available via application to the Public Health England Office for Data Release, but restrictions apply to the availability of these data.

Ethics approval and consent to participate: We obtained the statutory approvals required for this research from the Confidentiality Advisory Group (CAG) of the Health Research Authority (HRA): PIAG 1–05(c) 2007. Ethical approval was obtained from the Research Ethics Committee (REC) of the Health Research Authority (HRA): 07/MRE01/52. Informed consent from participants was waived by the ethics committee. We used anonymised National Cancer Registry and Hospital Episode Statistics data. All methods were carried out in accordance with relevant guidelines and regulations.

Consent for publication: Not applicable

Competing interests: The authors declare no potential conflicts of interest.

Code availability: Not applicable

References

1. Smittenaar, C. R., Petersen, K. A., Stewart, K. & Moitt, N. Cancer incidence and mortality projections in the UK until 2035. *Br. J. Cancer* **115**, 1147–1155 (2016).
2. Smith, A. *et al.* Lymphoma incidence, survival and prevalence 2004-2014: sub-type analyses from the UK's Haematological Malignancy Research Network. *Br J Cancer* **112**, 1575–1584 (2015).
3. Smith, A. *et al.* Impact of age and socioeconomic status on treatment and survival from aggressive lymphoma: a UK population-based study of diffuse large B-cell lymphoma. *Cancer Epidemiol* **39**, 1103–1112 (2015).
4. Kane, E. *et al.* Emergency admission and survival from aggressive non-Hodgkin lymphoma: A report from the UK's population-based Haematological Malignancy Research Network. *Eur. J. Cancer* **78**, 53–60 (2017).
5. Department of Health and Social Care. *The NHS cancer plan: a plan for investment: a plan for reform.* (Department of Health, 2000).
6. Exarchakou, A., Rachet, B., Belot, A., Maringe, C. & Coleman, M. P. Impact of national cancer policies on cancer survival trends and socioeconomic inequalities in England, 1996-2013: population based study. *BMJ* **360**, k764–k764 (2018).
7. Maringe, C., Li, R., Mangtani, P., Coleman, M. P. & Rachet, B. Cancer survival differences between South Asians and non-South Asians of England in 1986–2004, accounting for age at diagnosis and deprivation. *Br. J. Cancer* **113**, 173 (2015).
8. Department of Health. *Improving Outcomes: a strategy for cancer.* (2011).
9. National Institute for Health and Care Excellence. *Improving outcomes in haematological cancers: the manual.* (2003).
10. National Institute for Health and Care Excellence. *Haematological cancers: improving outcomes.* (2016).

11. Renzi, C., Lyratzopoulos, G., Hamilton, W., Maringe, C. & Rachet, B. Contrasting effects of comorbidities on emergency colon cancer diagnosis: a longitudinal data-linkage study in England. *BMC Health Serv. Res.* **19**, 311 (2019).
12. Fowler, H. *et al.* Comorbidity prevalence among cancer patients: a population-based cohort study of four cancers. *BMC Cancer* **20**, 2 (2020).
13. Li, R., Daniel, R. & Rachet, B. How much do tumor stage and treatment explain socioeconomic inequalities in breast cancer survival? Applying causal mediation analysis to population-based data. *Eur. J. Epidemiol.* **31**, 603–611 (2016).
14. Fowler, H. *et al.* Persistent inequalities in 90-day colon cancer mortality: an English cohort study. *Br. J. Cancer* **117**, 1396 (2017).
15. Belot, A. *et al.* Association between age, deprivation and specific comorbid conditions and the receipt of major surgery in patients with non-small cell lung cancer in England: A population-based study. *Thorax* **74**, 51 LP – 59 (2019).
16. Rachet, B. *et al.* Socioeconomic inequalities in cancer survival in England after the NHS cancer plan. *Br J Cancer* **103**, 446–453 (2010).
17. Comber, H. *et al.* Affluence and private health insurance influence treatment and survival in non-Hodgkin's lymphoma. *PLoS One* **11**, (2016).
18. Woods, L. M., Rachet, B. & Coleman, M. P. Origins of socio-economic inequalities in cancer survival: a review. *Ann. Oncol.* **17**, 5–19 (2005).
19. Quaglia, A. *et al.* Socio-economic factors and health care system characteristics related to cancer survival in the elderly: A population-based analysis in 16 European countries (ELDCARE project). *Crit. Rev. Oncol. Hematol.* **54**, 117–128 (2005).
20. Afshar, N., English, D. R. & Milne, R. L. Rural–urban residence and cancer survival in high-income countries: A systematic review. *Cancer* **125**, 2172–2184 (2019).
21. Belot, A. & Pohar-Perme, M. Social Disparities in Cancer Survival: Methodological

- Considerations. in *Social Environment and Cancer in Europe: Towards an Evidence-Based Public Health Policy* (eds. Launoy, G., Zadnik, V. & Coleman, M. P.) 39–54 (Springer International Publishing, 2021). doi:10.1007/978-3-030-69329-9_5
22. International Agency for Research on Cancer. International Classification of Diseases for Oncology. (2013). Available at: <http://codes.iarc.fr/>. (Accessed: 4th October 2019)
 23. Fritz, A. *et al.* *International Classification of Diseases for Oncology*. (World Health Organisation, 2000).
 24. Campo, E. *et al.* The 2008 WHO classification of lymphoid neoplasms and beyond: evolving concepts and practical applications. *Blood* **117**, 5019–5032 (2011).
 25. gov.uk. National Cancer Registry and Analysis Service. (2017). Available at: <https://www.gov.uk/guidance/national-cancer-registration-and-analysis-service-ncras>. (Accessed: 4th October 2019)
 26. NHS Digital. Hospital Episode Statistics. (2015). Available at: <https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/hospital-episode-statistics>. (Accessed: 4th October 2019)
 27. Rachet, B. *et al.* Multivariable flexible modelling for estimating complete, smoothed life tables for sub-national populations. *BMC Public Health* **15**, 1240 (2015).
 28. Porta, M. *A Dictionary of Epidemiology*. (Oxford University Press, 2014).
 29. Feinstein, A. R. The pre-therapeutic classification of co-morbidity in chronic disease. *J Chronic Dis* **23**, 455–468 (1970).
 30. Maringe, C., Fowler, H., Rachet, B. & Luque-Fernandez, M. A. Reproducibility, reliability and validity of population-based administrative health data for the assessment of cancer non-related comorbidities. *PLoS One* **12**, e0172814 (2017).
 31. Charlson, M. E., Pompei, P., Ales, K. L. & MacKenzie, C. R. A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation. *J. Chronic Dis*.

- 40**, 373–383 (1987).
32. Armitage, J. N. & van der Meulen, J. H. Identifying co-morbidity in surgical patients using administrative data with the Royal College of Surgeons Charlson Score. *Br J Surg* **97**, 772–781 (2010).
 33. gov.uk. Indices of Multiple Deprivation. (2015). Available at: <https://www.gov.uk/government/statistics/english-indices-of-deprivation-2015>. (Accessed: 4th October 2019)
 34. National Health Service: data dictionary. Lower Super Output Area. (2018). Available at: https://www.datadictionary.nhs.uk/data_dictionary. (Accessed: 4th October 2019)
 35. Elliss-Brookes, L. *et al.* Routes to diagnosis for cancer - determining the patient journey using multiple routine data sets. *Br J Cancer* **107**, 1220–1226 (2012).
 36. Charvat, H. *et al.* A multilevel excess hazard model to estimate net survival on hierarchical data allowing for non-linear and non-proportional effects of covariates. *Stat Med* **35**, 3066–3084 (2016).
 37. Verbeke, G. & Molenberghs, G. *Linear Mixed Models for Longitudinal Data*. (Springer-Verlag New York, 2000).
 38. Belot, A. *et al.* Summarizing and communicating on survival data according to the audience: a tutorial on different measures illustrated with population-based cancer registry data. *Clin. Epidemiol.* **11**, 53–65 (2019).
 39. Molenberghs, G. & Verbeke, G. *Models for Discrete Longitudinal Data*. (Springer-Verlag New York, 2005).
 40. Agresti, A. *Categorical Data Analysis*. (John Wiley & Sons, Inc., 2002).
 41. Carpenter, J. R. & Kenward, M. G. *Multiple Imputation and Its Application*. (John Wiley & Sons, Ltd, 2013).
 42. Little, R. J. & Rubin, D. B. *Statistical Analysis with Missing Data*. (John Wiley & Sons, Inc.,

- 1987).
43. Rubin, D. B. *Multiple imputation for nonresponse in surveys*. (Wiley, 1987).
 44. Quartagno, M. & Carpenter, J. R. jomo: A package for multilevel joint modeling multiple imputation. (2016).
 45. Coiffier, B. *et al.* CHOP Chemotherapy plus Rituximab Compared with CHOP Alone in Elderly Patients with Diffuse Large-B-Cell Lymphoma. *N. Engl. J. Med.* **346**, 235–242 (2002).
 46. Coiffier, B. Rituximab in combination with CHOP improves survival in elderly patients with aggressive non-Hodgkin's lymphoma. *Semin Oncol* **29**, 18–22 (2002).
 47. Delarue, R. *et al.* Dose-dense rituximab-CHOP compared with standard rituximab-CHOP in elderly patients with diffuse large B-cell lymphoma (the LNH03-6B study): a randomised phase 3 trial. *Lancet Oncol* **14**, 525–533 (2013).
 48. McGowan, J. V *et al.* Anthracycline Chemotherapy and Cardiotoxicity. *Cardiovasc. Drugs Ther.* **31**, 63–75 (2017).
 49. National Institute for Health and Care Excellence. *Non-Hodgkin's lymphoma: diagnosis and management*. (National Institute for Health and Care Excellence, 2016).
 50. Tilly, H. *et al.* Diffuse large B-cell lymphoma (DLBCL): ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Ann. Oncol.* **26**, v116–v125 (2015).
 51. Bröckelmann, P. J. *et al.* Patient and physician preferences for first-line treatment of classical Hodgkin lymphoma in Germany, France and the United Kingdom. *Br. J. Haematol.* **184**, 202–214 (2019).
 52. Lossos, I. S. & Gascoyne, R. D. Transformation of follicular lymphoma. *Best Pract. Res. Clin. Haematol.* **24**, 147–163 (2011).
 53. Sogaard, M., Thomsen, R. W., Bossen, K. S., Sorensen, H. T. & Norgaard, M. The impact of comorbidity on cancer survival: a review. *Clin Epidemiol* **5**, 3–29 (2013).

54. Frederiksen, B. L., Dalton, S. O., Osler, M., Steding-Jessen, M. & de Nully Brown, P. Socioeconomic position, treatment, and survival of non-Hodgkin lymphoma in Denmark--a nationwide study. *Br J Cancer* **106**, 988–995 (2012).
55. Rachet, B., Mitry, E., Shah, A., Cooper, N. & Coleman, M. P. Survival from non-Hodgkin lymphoma in England and Wales up to 2001. *Br. J. Cancer* **99**, S104–S106 (2008).
56. Bray, C., Morrison, D. S. & McKay, P. Socio-economic deprivation and survival of non-Hodgkin lymphoma in Scotland. *Leuk. Lymphoma* **49**, 917–923 (2008).
57. Pohar Perme, M., Estève, J. & Rachet, B. Analysing population-based cancer survival - settling the controversies. *BMC Cancer* **16**, 933 (2016).
58. Kobayashi, Y. *et al.* Charlson Comorbidity Index is an independent prognostic factor among elderly patients with diffuse large B-cell lymphoma. *J Cancer Res Clin Oncol* **137**, 1079–1084 (2011).
59. Saygin, C. *et al.* Impact of comorbidities on outcomes of elderly patients with diffuse large B-cell lymphoma. *Am. J. Hematol.* **92**, 989–996 (2017).
60. Chihara, D. *et al.* Management strategies and outcomes for very elderly patients with diffuse large B-cell lymphoma. *Cancer* **122**, 3145–3151 (2016).
61. Janssen-Heijnen, M. L. *et al.* A population-based study of severity of comorbidity among patients with non-Hodgkin's lymphoma: prognostic impact independent of International Prognostic Index. *Br J Haematol* **129**, 597–606 (2005).
62. Crooks, C. J., West, J. & Card, T. R. A comparison of the recording of comorbidity in primary and secondary care by using the Charlson Index to predict short-term and long-term survival in a routine linked data cohort. *BMJ Open* **5**, e007974 (2015).
63. Carpenter, J., Goldstein, H. & Kenward, M. REALCOM-IMPUTE software for multilevel multiple imputation with mixed response types. *J. Stat. Softw.* **45**, (2011).
64. Quartagno, M. & Carpenter, J. R. Multiple imputation for discrete data: Evaluation of the joint

- latent normal model. *Biom. J.* **61**, 1003–1019 (2019).
65. Ingleby, F. C. *et al.* Assessment of the concordance between individual-level and area-level measures of socio-economic deprivation in a cancer patient cohort in England and Wales. *BMJ Open* **10**, e041714 (2020).
66. Belot, A. *et al.* Describing the association between socioeconomic inequalities and cancer survival: methodological guidelines and illustration with population-based data. *Clinical epidemiology* **10**, 561–573 (2018).
67. Madley-Dowd, P., Hughes, R., Tilling, K. & Heron, J. The proportion of missing data should not be used to guide decisions on multiple imputation. *J. Clin. Epidemiol.* **110**, 63–73 (2019).
68. Carnerio, I. *et al.* Variation in the Routes to Cancer Diagnosis and Stage for Ten Cancer Sites. *Conference: Cancer Data and Outcomes conference* (2016). Available at: https://www.researchgate.net/publication/311602679_Variation_in_the_Routes_to_Cancer_Diagnosis_and_Stage_for_Ten_Cancer_Sites. (Accessed: 15th August 2021)
69. Allemani, C. *et al.* Global surveillance of trends in cancer survival 2000–14 (CONCORD-3): analysis of individual records for 37,513,025 patients diagnosed with one of 18 cancers from 322 population-based registries in 71 countries. *Lancet* **391**, 1023–1075 (2018).
70. Thomson, C. S. & Forman, D. Cancer survival in England and the influence of early diagnosis: what can we learn from recent EURO CARE results? *Br. J. Cancer* **101**, S102–S109 (2009).

Appendix

Table A1: Comorbidities and their diagnostic ICD-10 codes

Comorbidity	ICD-10
Myocardial infarction	I21.x, I22.x, I25.2
Congestive heart failure	I11.0, I13.0, I13.2, I25.5, I42.0, I42.5–I42.9, I43.x, I50.x, P29.0
Peripheral vascular disease	I70.x, I71.x, I73.1, I73.8, I73.9, I77.1, I79.0, I79.2, K55.1, K55.8, K55.9, Z95.8, Z95.9
Cerebrovascular disease	G45.x, G46.x, H34.0, I60.x–I69.x
Dementia	F00.x–F03.x, F05.1, G30.x, G31.1
Chronic obstructive pulmonary disease	I27.9, J40.x–J47.x, J60.x–J67.x, J68.4, J70.1, J70.3
Rheumatic disease	M05.x, M06.x, M31.5, M32.x–M34.x, M35.1, M35.3, M36.0
Liver disease	B18.x, K70.0–K70.3, K70.9, K71.3–K71.5, K71.7, K73.x, K74.x, K76.0, K76.2–K76.4, K76.8, K76.9, Z94.4, K71.1, K72.1, K72.9, K76.5, K76.6, K76.7, I85.0, I85.9, I86.4, I98.2, K70.4,
Diabetes without chronic complication	E10.0, E10.1, E10.6, E10.8, E10.9, E11.0, E11.1, E11.6, E11.8, E11.9, E12.0, E12.1, E12.6, E12.8, E12.9, E13.0, E13.1, E13.6, E13.8, E13.9, E14.0, E14.1, E14.6, E14.8, E14.9
Diabetes with chronic complication	E10.7, E11.2–E11.5, E11.7, E12.2–E12.5, E12.7, E13.2–E13.5, E13.7, E14.2–E14.5, E14.7
Hemiplegia or paraplegia	G04.1, G11.4, G80.1, G80.2, G81.x, G82.x, G83.0–G83.4, G83.9
Renal disease	I12.0, I13.1, N03.2–N03.7, N05.2–N05.7, N18.x, N19.x, N25.0, Z49.0–Z49.2, Z94.0, Z99.2
AIDS/HIV	B20.x–B22.x, B24.x

ICD-10: International Classification of Diseases, 10th Revision

Diabetes with/without chronic complication is combined in the RCS Charlson Comorbidity Score

Figure A1: Excess mortality hazard (i.e., white males, least deprived, no comorbidities, diagnosed through general practitioner referral within an average LSOA [random effect of zero]) over time since diagnosis, for different ages, amongst those diagnosed with **diffuse large B-cell lymphoma** (n=29,898) in England during 2005-2013.

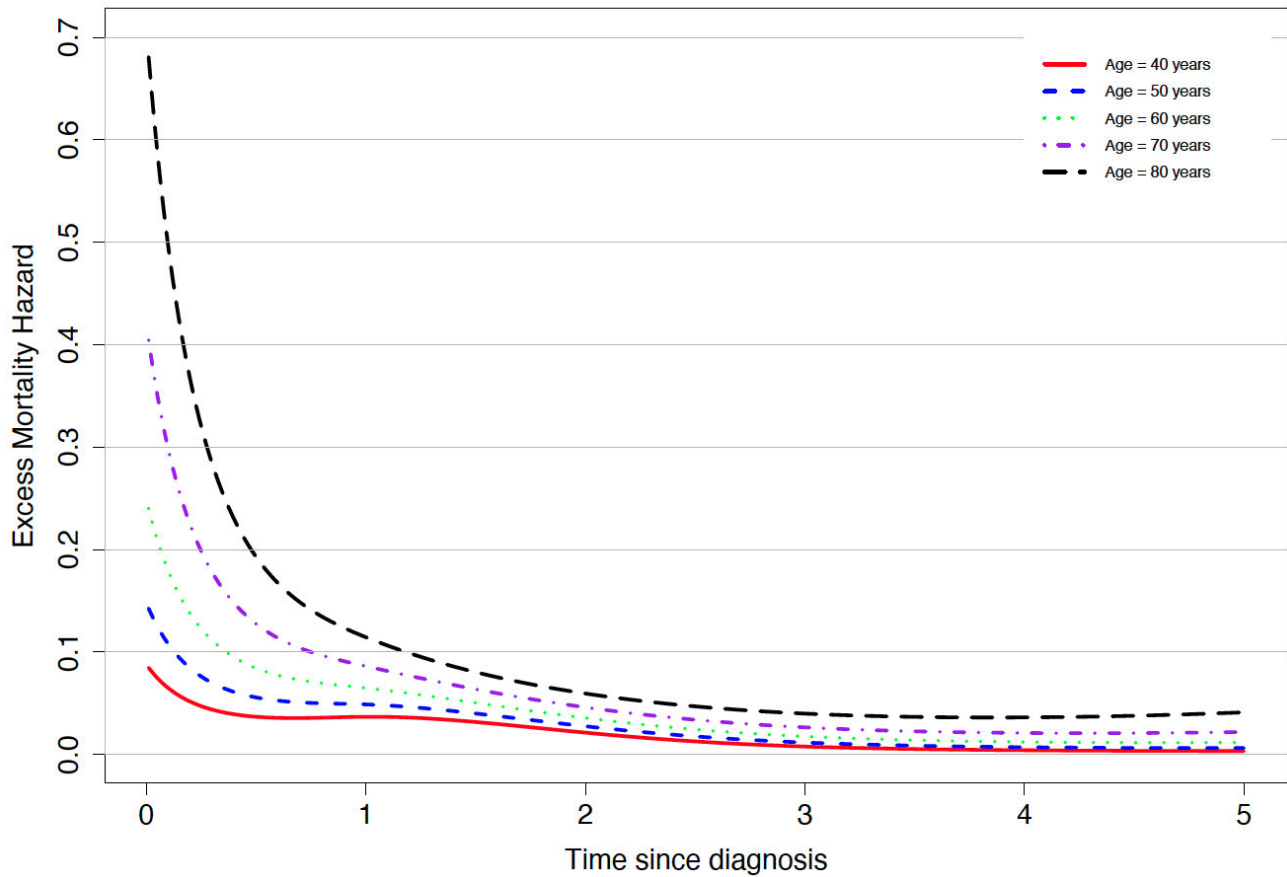


Figure A2: Excess mortality hazard (i.e., white males, least deprived, no comorbidities, diagnosed through general practitioner referral within an average LSOA [random effect of zero]) over time since diagnosis, for different ages, amongst those diagnosed with **follicular lymphoma** (n=15,516) in England during 2005-2013.

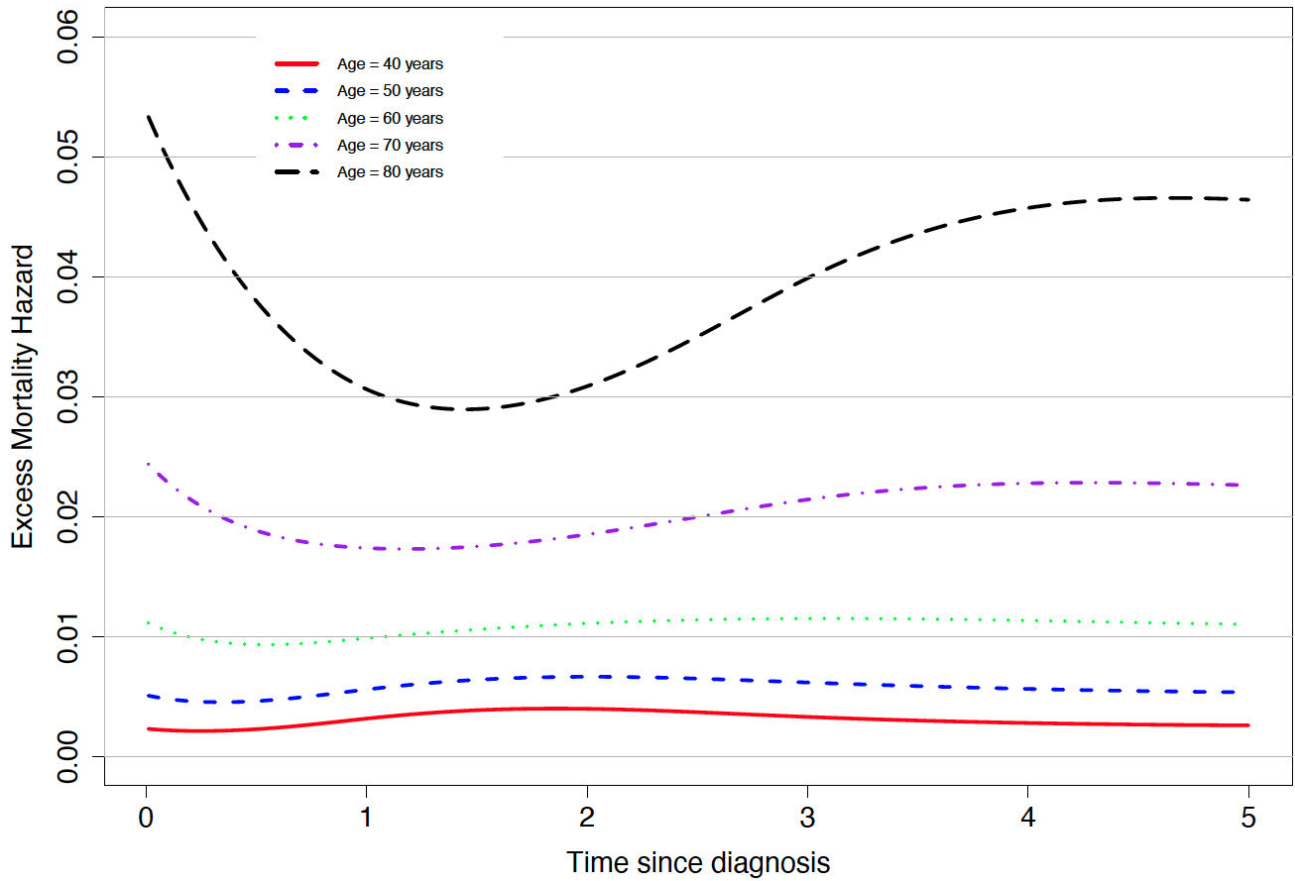


Figure A3: Excess mortality hazard ratio according to age at diagnosis and time since diagnosis for patients diagnosed with **diffuse large B-cell lymphoma** (n=29,898) in England between 2005 and 2013.

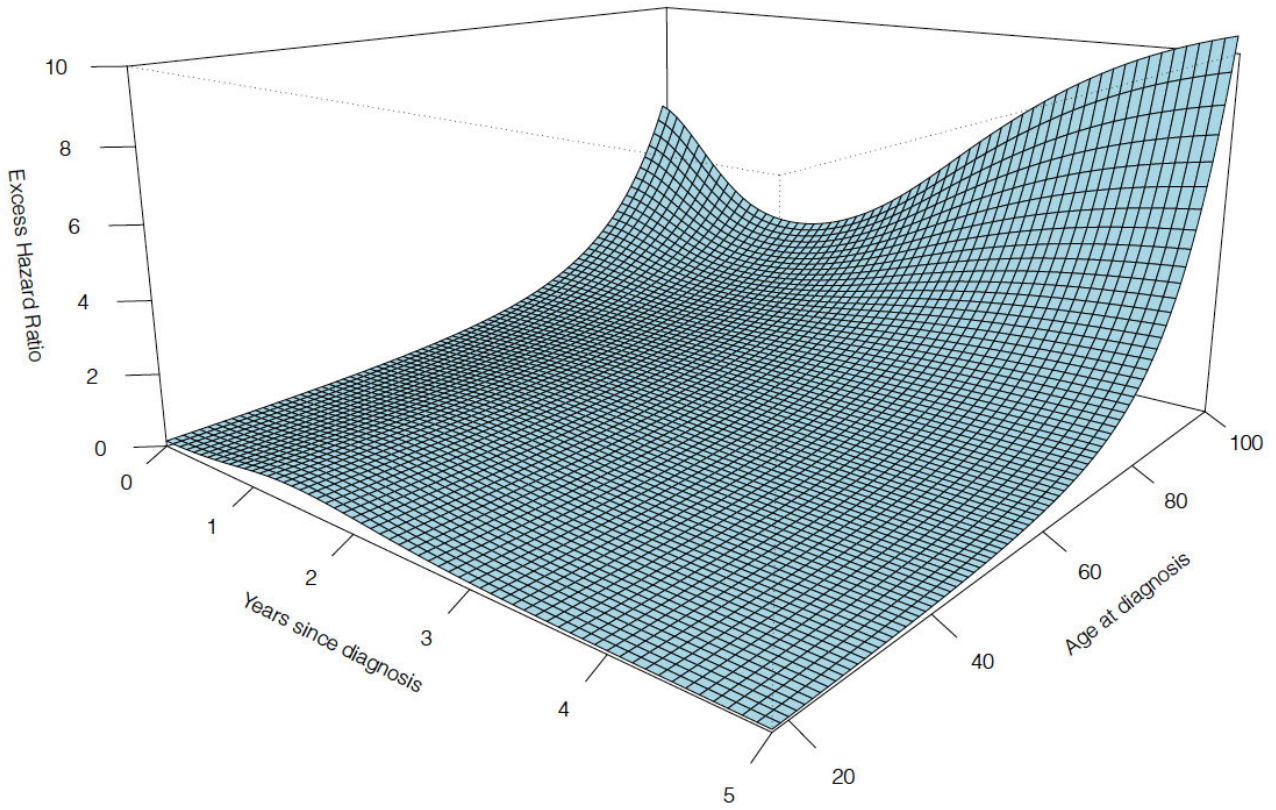
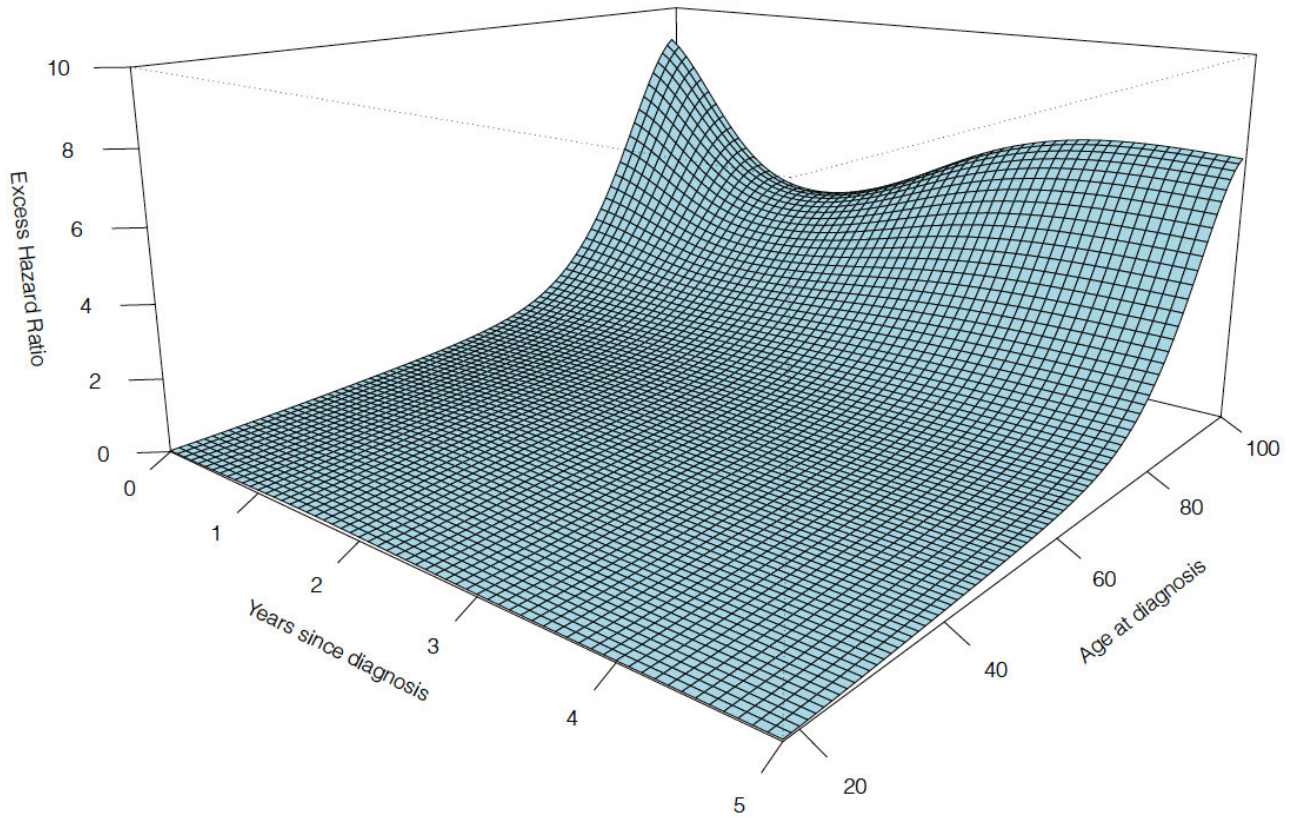


Figure A4: Excess mortality hazard ratio according to age at diagnosis and time since diagnosis for patients diagnosed with **follicular lymphoma** (n=15,516) in England between 2005 and 2013.



A.5.3 Association between comorbidity and short-term mortality

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	lsh1601639	Title	Mr
First Name(s)	Matthew		
Surname/Family Name	Smith		
Thesis Title	Survival of patients with non-Hodgkin lymphoma in England: investigating the socioeconomic inequalities		
Primary Supervisor	Edmund Njeru Njagi		

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?	British Medical Journal: Open
Please list the paper's authors in the intended authorship order:	Matthew J. Smith, Edmund Njeru Njagi, Aurélien Belot, Clémence Leyrat, Audrey Bonaventure, Sara Benitez Majano, Bernard Ratchet, Miguel Angel Luque Fernandez
Stage of publication	Submitted

SECTION D – Multi-authored work

<p>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)</p>	<p>MS, MALF, ENN and BR contributed to the conception of the study and designed the study. ENN, BR, MALF, ABe and CL provided advice on statistical methods. MS conducted the analyses of the data and prepared the draft of the manuscript, tables and figures. MALF, ENN and BR supervised the study and provided comments on the manuscript draft. ENN, BR, MALF, MQ, SBM, ABe and ABo provided comments on the final draft of the manuscript. All authors read and approved the final manuscript.</p>
---	---

SECTION E

Student Signature	Matthew J. Smith
Date	7th June 2021

Supervisor Signature	
Date	

Title

Association between multimorbidity and socioeconomic deprivation on short-term mortality amongst patients with Diffuse Large B-cells or Follicular lymphomas in England: a nationwide cohort study

Authors

Matthew J. Smith^{1*}, Edmund Njeru Njagi¹, Aurélien Belot¹, Clémence Leyrat^{1,2}, Audrey Bonaventure³, Sara Benitez Majano¹, Bernard Rachet¹, Miguel Angel Luque Fernandez^{1,4,5}

Authors' affiliations

¹ Inequalities in Cancer Outcomes Network, Department of Non-Communicable Disease Epidemiology, London School of Hygiene and Tropical Medicine, Keppel Street, London, WC1E 7HT, UK

² Department of Medical Statistics, London School of Hygiene and Tropical Medicine, Keppel Street, London, WC1E 7HT, UK

³ CRESS, Université de Paris, INSERM, UMR 1153, Epidemiology of Childhood and Adolescent Cancers Team, Villejuif, France

⁴ Noncommunicable Disease and Cancer Epidemiology Group, Instituto de Investigación Biosanitaria de Granada, Ibs.GRANADA, Andalusian School of Public Health, Granada, Spain

⁵ Centro de Investigación Biomédica en Red de Epidemiología y Salud Pública (CIBER of Epidemiology and Public Health, CIBERESP), Madrid, Spain

Corresponding author*

Matthew J. Smith
Keppel Street, London, WC1E 7HT, UK
ORCID: 0000-0002-8502-0056
Email: *matthew.smith1@lshtm.ac.uk*

Word count: Abstract: 193; Text: 2807; Tables: 4; Figures: 2

Abstract

Objectives: We aimed to assess the association between multimorbidity and deprivation on short-term mortality amongst DLBCL and FL patients in England.

Setting: The association of multimorbidity and socioeconomic deprivation on survival among patients diagnosed with Diffuse Large B-cell (DLBCL) and Follicular lymphoma (FL) in England between 2005 and 2013. We linked the English population-based cancer registry with electronic health records databases and estimated adjusted mortality rate ratios by multimorbidity and deprivation status. Using flexible hazard-based regression models, we computed DLBCL and FL standardised mortality risk by deprivation and multimorbidity at 1 year.

Results: Overall, 41,422 patients aged 45-99 years were diagnosed with DLBCL or FL in England during 2005-2015. Most deprived FL patients with multimorbidities had three times higher hazard of 1 year mortality (HR: 3.3, CI: 2.48 – 4.28, $p < 0.001$) than least deprived patients without comorbidity; amongst DLBCL there was approximately twice the hazard (HR: 1.9, CI: 1.70– 2.07, $p < 0.001$).

Conclusions: Multimorbidity, deprivation, and their combination, are strong and independent predictors of an increased short-term mortality risk amongst DLBCL and FL patients in England. Public health measures targeting the reduction of multimorbidity amongst most deprived DLBCL and FL patients are needed to reduce the short-term mortality gap.

Key words: Cancer epidemiology, Diffuse Large B-cell lymphoma, Follicular lymphoma, multimorbidity, deprivation, survival analysis

Strengths and limitations of this study

- Data contains a large sample size of high-quality population-based clinical records with a high national coverage of information on all patients diagnosed with Diffuse Large B-cell or Follicular lymphomas in England during 2005 to 2013.
- Population based administrative hospital discharge data was used for the assessment of comorbid conditions, and selection bias was reduced by restricting records of comorbidities to occurring between 6- and 24-months prior to the date of cancer diagnosis.
- 1-year cumulative mortality hazard was modelled using a flexible parametric modelling approach and included restricted cubic splines to account for non-linear effects of continuous variables.
- Modern methods (i.e., standardisation) were used to control for confounding of patient baseline characteristics; information on lifestyle characteristics was unavailable.
- As there was missing data we performed a sensitivity analysis with multiple imputation using chained equations, we found consistency of our conclusions under different missing data assumptions.

Introduction

In England, non-Hodgkin lymphoma (NHL) is the sixth most commonly diagnosed cancer in England with an incidence rate of 23.2 per 100,000 people.¹ Apart from lung cancer, survival estimates of NHL (79.4% survival probability at 1 year) are amongst the lowest of the six most common cancers.^{2,3} NHL encompasses a heterogeneous group of malignancies with diverse histological patterns; in addition, the commonest NHL subtypes are diffuse large B-cell lymphoma (DLBCL) and follicular lymphoma (FL),⁴ patients of which show a large variation in survival.⁵ Cancer survival in England is lower than other European countries,⁶ but is similar when restricting to those surviving after 1 year:⁷ identifying, and influencing, the factors of short-term mortality could reduce the gap in survival.

Over the past decades, patient survival of FL has steadily improved and stagnated for DLBCL;³ furthermore, disparities in survival between deprivation levels remains.⁸ Public health policies have increased awareness and set targets, such as minimising the length of time from referral to treatment, to reduce the inequalities;⁹ however, since their implementation, there has been no evidence that the National Health Service (NHS) Cancer Plan had an impact on the inequalities in cancer survival.¹⁰ The deprivation-gap in survival is still apparent, despite the NHS Cancer Plan and successive policies.⁹

Comorbidities, which refers to the presence of a long-term health condition additional, but unrelated, to the underlying cancer,¹¹ tends to mask cancer symptoms, delaying cancer diagnosis and decreases survival.¹² Older age groups and those living in more deprived areas experience more comorbidities. With a global aging population, the prevalence of comorbidities is expected to increase.¹³ The association between multimorbidity and survival is described for other cancers,¹⁴ but for DLBCL and FL this relationship remains unclear.

We aim to identify the association between multimorbidity and risk of short-term mortality amongst DLBCL and FL patients in England. We hypothesise that multimorbidity and deprivation, independently and combined, contribute to an increased risk of death after DLBCL or FL diagnosis.

Methods

Study Design, participants, data and setting

We used the data from a retrospective population-based cohort study with DLBCL and FL patients diagnosed between 1st January 2005 and 31st December 2013 and followed up to 31st December 2015. DLBCL and FL diagnoses were made according to the International Classification of Diseases for Oncology, 3rd edition, based on codes C82.0-C85.9¹⁵ (**Supplementary Table S1** shows the subtype categorisation). Patients entered the study on the date of their diagnosis and were followed up until death or censored at 1 year, whichever occurred first.

Data was obtained from population-based cancer registries within the English National Cancer Registry and Analysis Service (CAS)¹⁶ and linked to patient's electronic health records from Hospital Episode Statistics (HES). CAS contains patient and tumour variables including relevant dates (birth, diagnosis, and vital status), sex, age at diagnosis, deprivation, cancer site and morphology. We used population based administrative hospital discharge data for the assessment of comorbid conditions; we analysed HES data (containing comorbid conditions records) according to the International Classification of Diseases, 10th revision (**Supplementary Table S2**), for the period 2003 to 2015. HES contains clinical, administrative, and demographic information about individual patients. To avoid selection bias including cancer related comorbidities, we restricted retrospective records of comorbidities to occurring between 6- and 24-months prior to the date of DLBCL and FL diagnosis.¹⁷

Patient and public involvement

No patient or public involvement.

Outcome, exposure and other variables

The outcome of this study was the time since diagnosis up to death observed within the first year after diagnosis of DLBCL and FL patients (patients alive were censored at survival time defined with the date of last known vital status), the main exposures were multimorbidity status and deprivation. Due to data availability and clinical reasoning, we include as confounders age, sex, and ethnicity. Due to the positivity assumption, and the chances of having a multimorbidity, we included patients aged above 45 years at diagnosis (**Figure 1**). The positivity assumption states that there is a nonzero probability of receiving any level of comorbidity status for every combination of values of the independent variables among the patients in the population.¹⁸

Comorbidity status was classified according to the Royal College of Surgeons (RCS) Charlson score (an adaptation of the Charlson comorbidity index¹⁹) that includes 12 categories for comorbidities, excludes a category (peptic ulcer disease) and groups diseases together (e.g. diabetes mellitus codes with or without complications were grouped into a single category). The score was categorised into those with none, one comorbidity (whatever the type), or two or more comorbidities (defined as *multimorbidity*). The score does not weight the comorbidities assuming that any comorbidity has the same impact on short-term mortality.²⁰

Area-level deprivation was categorised into one of five quintiles (5th is most deprived). Deprivation was used as a proxy of individual level socioeconomic status. We used the Index of Multiple Deprivation²¹ (IMD), which is an area-level deprivation score based on the Lower Super Output Area²² (LSOA) residence of the patient at the time of cancer diagnosis. LSOA is a geographical location with a median of 1500 inhabitants.

Ethnicity due to data sparsity amongst ethnic minorities was recorded as either white or other. Route to diagnosis (CAS dataset), although not considered in our analysis because it was on the causal pathway, is included in the imputation models for missing data.

Statistical Analysis

We described the characteristics of DLBCL and FL patients using counts and proportions, and calculated odds ratios of having a lymphoma type along with Wald test p-values (**Supplementary Table S3**). We assessed the

unadjusted association between having multimorbidity and patients' characteristics, using chi-square tests. Then, we compute the number of deaths, person-time at risk, and unadjusted rates of deaths per 100 person-years and rate ratios with 95% confidence intervals (CI) by patients' characteristics by DLBCL and FL subtypes.

The follow-up time for those who died is from the date of diagnosis until death, for those alive it was until administrative censoring at 12 months (no lost to follow-up before 12 months was observed). To describe the multimorbidity and deprivation short-term mortality risk amongst DLBCL and FL patients, we computed the one year cumulative hazard obtained using the non-parametric Nelson-Aalen estimator.²³ Then, we computed adjusted short-term mortality risk by patient characteristics using a flexible parametric modelling approach to model the non-linear change in mortality risk over one year. We included restricted-cubic spline to model the baseline hazard,²⁴ with three knots located at the 25th, 50th, and 75th percentiles of the log event times. To define the model, let t be the time since DLBCL or FL diagnosis until death or censoring. We define the log cumulative mortality hazard as

$$\ln[H(t|\mathbf{x}_i, A_i, Dep_i)] = \gamma_0 + \gamma_1 z_{1i} + \gamma_2 z_{2i} + \gamma_3 A_{1i} + \gamma_4 A_{2i} + \sum_{k=2}^5 \beta_{0k} \cdot Dep_{ik} + \sum_{m=1}^M \beta_m x_{im}$$

where z_i are the indicators for the three knots of the baseline restricted cubic splines. The model specification included the restricted cubic splines for the continuous variable age are given by A_i where the knots are placed at 1 and 6 months, deprivation, and the vector x_{im} of categorical covariates (i.e., multimorbidity, ethnicity and sex). Restricted cubic splines were included to minimise residual confounding and to account for the non-linear association between age and the cumulative hazard. From the model, we derived the cumulative incidence of death at 1 year by comorbidity status (i.e. none, one, or multimorbidity) standardized to the empirical distribution of age, sex, ethnicity and deprivation.^{25,26}

We used the same modelling approach to evaluate the linearly combined effect of multimorbidity and deprivation on short-term mortality, assuming the effect of multimorbidity is constant across levels of deprivation.²⁷ We assume that both effects (multimorbidity and deprivation) are independently associated with the mortality rate and that their effect is constant and the rate is increasing linearly.

In sensitivity analysis, we evaluated the robustness of our results to the missing ethnicity records by utilising multiple imputation using chained equations (the fully conditional specification multiple imputation approach), under a missing at random assumption. We included route of cancer diagnosis and Ann Arbor²⁸ cancer stage as partially observed auxiliary variables because these variables were predictive of the probability of missing values (established via exploratory analysis of the missing data indicators) and are predictive of the underlying values themselves (established from clinical and epidemiological reasoning).²⁹ We generated 20 imputed datasets. The imputation model for the partially observed variable (ethnicity) was defined as a logistic regression model including all explanatory variables in the substantive model, the vital status and Nelson-Aalen estimate of the cumulative hazard, and the auxiliary variables (route to, and stage at, diagnosis). We then fit the substantive model to each of the 20 imputed datasets and from these estimates we predicted the standardised survival (using *stpm2* package), derived the cumulative hazard (-log[Survival probability]) for each imputed dataset and combined the survival probability estimates using Rubin's rules.³⁰

Analyses were performed in Stata version 16 (StataCorp, College Station, Texas, U.S.); the *stpm2* package was used to estimate flexible parametric survival models, and the *standsurv* command to compute standardized mortality risks. The *mi impute* command was used for multiple imputation and the *mi estimate* command to combine estimates.

Results

Overall, 41,422 patients in England, aged from 45 to 99 years, were diagnosed with either DLBCL or FL between 2005 and 2013 in England. Of 14,043 patients with FL, amongst those who died within 1-year, the proportion of patients with multimorbidity compared to no comorbidity was 3.2 times higher (22.0% vs. 6.9%, respectively) (**Table 1**); of 27,379 patients diagnosed with DLBCL, this comparison was 1.6 times higher (49.7% vs. 31.5%, respectively). For both DLBCL and FL subtypes, the proportion of multimorbidity was higher amongst those living in more deprived areas. In DLBCL only, there was a higher proportion of having any comorbidity amongst males compared to females ($p < 0.001$); for FL only, this proportion was slightly higher amongst females. As expected, multimorbidity was more prevalent amongst older age groups.

Table 1: Vital status, age, sex, deprivation level and ethnicity according to the comorbidity status amongst n = 41,422 patients with non-Hodgkin lymphoma in England between 2005 and 2013 (27,379 DLBCL cases and 14,043 FL cases).

	No comorbidity N (%)	Comorbidity N(%)	Multimorbidity N(%)	p-value *	
Diffuse Large B-cell lymphoma (n = 27,379)					
Vital status at 1 year					
	<i>Alive</i>	16,621 (68.5)	872 (56.7)	791 (50.3)	<0.001
	<i>Dead</i>	7,648 (31.5)	666 (43.3)	781 (49.7)	
Sex					
	<i>Male</i>	12,904 (53.2)	794 (51.6)	954 (60.7)	<0.001
	<i>Female</i>	11,365 (46.8)	744 (48.4)	618 (39.3)	
Age at diagnosis (y)					
	<i>45-54</i>	2,685 (11.1)	80 (5.2)	96 (6.1)	<0.001
	<i>55-64</i>	5,154 (21.2)	257 (16.7)	200 (12.7)	
	<i>65-74</i>	7,337 (30.2)	432 (28.1)	439 (27.9)	
	<i>75+</i>	9,093 (37.5)	769 (50.0)	837 (53.2)	
Deprivation					
	<i>Least deprived</i>	5,348 (22.0)	291 (18.9)	256 (16.3)	<0.001
	<i>2</i>	5,586 (23.0)	318 (20.7)	334 (21.3)	
	<i>3</i>	5,115 (21.1)	324 (21.1)	317 (20.2)	
	<i>4</i>	4,665 (19.2)	337 (21.9)	339 (21.6)	
	<i>Most deprived</i>	3,555 (14.7)	268 (17.4)	326 (20.7)	
Ethnicity					
	<i>White</i>	17,831 (95.5)	1,204 (96.6)	1,169 (92.6)	<0.001
	<i>Other</i>	848 (4.5)	43 (3.5)	93 (7.4)	
Follicular lymphoma (n = 14,043)					
Vital status at 1 year					
	<i>Alive</i>	12,003 (93.1)	546 (87.6)	407 (78.0)	<0.001
	<i>Dead</i>	895 (6.9)	77 (12.4)	115 (22.0)	
Sex					
	<i>Male</i>	5,980 (46.4)	275 (44.1)	257 (49.2)	0.227
	<i>Female</i>	6,918 (53.6)	348 (55.9)	265 (50.8)	
Age at diagnosis (y)					
	<i>45-54</i>	2,246 (17.4)	42 (6.7)	33 (6.3)	<0.001
	<i>55-64</i>	3,769 (29.2)	140 (22.5)	82 (15.7)	
	<i>65-74</i>	3,952 (30.6)	199 (31.9)	159 (30.5)	
	<i>75+</i>	2,931 (22.7)	242 (38.8)	248 (47.5)	
Deprivation					
	<i>Least deprived</i>	3,091 (24.0)	113 (18.1)	80 (15.3)	<0.001
	<i>2</i>	3,025 (23.5)	122 (19.6)	81 (15.5)	
	<i>3</i>	2,759 (21.4)	136 (21.8)	118 (22.6)	
	<i>4</i>	2,356 (18.3)	123 (19.7)	130 (24.9)	
	<i>Most deprived</i>	1,667 (12.9)	129 (20.7)	113 (21.7)	
Ethnicity					
	<i>White</i>	9,226 (95.7)	498 (96.3)	386 (92.8)	0.012
	<i>Other</i>	412 (4.3)	19 (3.7)	30 (7.2)	
Missing values: Ethnicity n(%) : DLBCL = 6,191 (22.6%), FL = 3,472 (24.7%)					
* Chi-squared test of association between the baseline characteristic and comorbidity status					

Tables 2a and 2b show the person-time at risk and unadjusted mortality rate of death for DLBCL and FL at one year after diagnosis. Amongst FL and DLBCL patients, 1087 (7.7%) and 9,095 (33.2%) died before 1 year, respectively (**Table 2a and 2b**). Amongst FL, those with one comorbidity or multimorbidity had 1.9 (95% CI: 1.46 – 2.33, $p < 0.001$) or 3.5 (95% CI: 2.90 – 4.27, $p < 0.001$) times the mortality rate, respectively, compared to those with no comorbidity (**Table 2a**). Amongst DLBCL, those with one comorbidity or multimorbidity had 1.5 (95% CI: 1.41 – 1.66, $p < 0.001$) or 1.9 (95% CI: 1.78 – 2.06, $p < 0.001$) times the mortality rate, compared to no comorbidity (**Table 2b**).

The unadjusted mortality rate of death increased with each increase in deprivation level: those living in the most deprived areas had 1.5 (95% CI: 1.22 – 1.82, $p < 0.001$) and 1.3 (95% CI: 1.22 – 1.40, $p < 0.001$) times the mortality rate compared to those living in the least deprived areas, for FL and DLBCL, respectively.

Table 2A: One-year unadjusted mortality rates and rate ratios by sex, age, deprivation, ethnicity, and comorbidity status amongst patients with **Follicular lymphoma** in England between 2005-2015 (n = 14,043; 1,087 deaths at one-year).

	Deaths / person years	Mortality rate* (95% CI)	Mortality RR	95% CI	p-value
One-year mortality (n = 1,087)					
Sex					
Male	526/6218.12	8.5 (7.77 – 9.21)	Ref		
Female	561/7225.93	7.8 (7.15 – 8.43)	0.92	(0.82 – 1.03)	0.157
Age at diagnosis (y) †					
10-year increase	-	-	2.20	(2.08 – 2.34)	<0.001
Age at diagnosis (y)					
45-54	46/2300.66	2.0 (1.50 – 2.67)	Ref		
55-64	129/3926.24	3.3 (2.77 – 3.90)	1.64	(1.17 – 2.30)	0.004
65-74	274/4157.52	6.6 (5.86 – 7.42)	3.30	(2.41 – 4.50)	<0.001
75+	638/3059.62	20.9 (19.30 – 22.54)	10.43	(7.73 – 14.07)	<0.001
Deprivation					
Least deprived	211/3175.55	6.7 (5.81 – 7.60)	Ref		
2	227/3102.18	7.3 (6.43 – 8.33)	1.10	(0.91 – 1.33)	0.313
3	230/2884.20	8.0 (7.01 – 9.08)	1.20	(1.00 – 1.45)	0.055
4	240/2471.30	9.7 (8.56 – 11.02)	1.46	(1.23 – 1.76)	<0.001
Most deprived	179/1810.81	9.9 (8.54 – 11.45)	1.49	(1.22 – 1.82)	<0.001
Ethnicity					
White	742/9712.99	7.6 (7.11 – 8.21)	Ref		
Other	21/450.96	4.7 (3.04 – 7.14)	0.61	(0.40 – 0.94)	0.024
Comorbidity status					
None	895/12412.44	7.5 (7.03 – 7.97)	Ref		
One	77/577.89	13.3 (10.66 – 16.66)	1.85	(1.46 – 2.33)	<0.001
Multimorbidity	115/453.71	25.4 (21.11 – 30.43)	3.52	(2.90 – 4.27)	<0.001

* per 100 person-years
† continuous form of age (for each 10-year increase in age)
CI: Confidence interval, **RR:** Rate Ratio, **Missing values:** ethnicity n(%) = 3,472 (24.7%)

Table 2B: One-year unadjusted mortality rates by sex, age, deprivation, ethnicity, and comorbidity status amongst patients with **DLBCL** in England between 2005-2015 (n = 27,379; 9,095 deaths at One-year).

	Deaths/person years	Mortality rate* (95% CI)	Mortality RR	95% CI	p-value
One-year mortality (n = 9,095)					
Sex					
Male	4867/11318.43	43.0 (41.81 – 44.23)	Ref	Ref	Ref
Female	4228/9784.59	43.2 (41.93 – 44.53)	1.01	(0.96 – 1.05)	0.817
Age at diagnosis (y) †					
10-year increase	-	-	1.50	(1.47 – 1.53)	<0.001
Age at diagnosis (y)					
45-54	430/2615.81	16.4 (14.96 – 18.07)	Ref	Ref	Ref
55-64	1074/4937.54	21.8 (20.49 – 23.09)	1.32	(1.18 – 1.48)	<0.001
65-74	2365/6604.47	35.8 (34.39 – 37.28)	2.18	(1.97 – 2.41)	<0.001
75+	5226/6945.19	75.2 (73.23 – 77.31)	4.58	(4.15 – 5.05)	<0.001
Deprivation					
Least deprived	1765/4684.43	37.7 (35.96 – 38.48)	Ref		
2	1955/4898.09	39.9 (38.18 – 41.72)	1.06	(0.99 – 1.13)	0.079
3	1948/4424.61	44.0 (42.11 – 46.03)	1.17	(1.10 – 1.25)	<0.001
4	1908/4017.48	47.5 (45.41 – 49.67)	1.26	(1.18 – 1.35)	<0.001
Most deprived	1519/3078.40	49.3 (46.92 – 51.89)	1.31	(1.22 – 1.40)	<0.001
Ethnicity					
White	6351/15900.73	39.9 (38.97 – 40.94)	Ref		
Other	270/808.61	33.4 (29.64 – 37.62)	0.84	(0.74 – 0.94)	0.004
Comorbidity status					
None	7648/19007.45	40.2 (39.35 – 41.15)	Ref		
One	666/1081.20	61.6 (57.09 – 66.46)	1.53	(1.41 – 1.66)	<0.001
Multimorbidity	781/1014.36	77.0 (71.78 – 82.59)	1.91	(1.78 – 2.06)	<0.001

* per 100 person-years
† continuous form of age (for each 10-year increase in age)
CI: Confidence interval, **RR:** Rate Ratio, **Missing values:** ethnicity n(%) = 6,191 (22.6%)

Table 3 shows, before and after multiple imputation, the mortality hazard amongst DLBCL and FL patients at 1 year adjusted for comorbidity status, sex, age, deprivation and ethnicity. After multiple imputation, patients with multimorbidity had 2.2 (CI 1.78 – 2.64, $p < 0.001$) and 1.4 (CI 1.34 – 1.55, $p < 0.001$) times the mortality hazard compared to those without a comorbidity, for FL and DLBCL, respectively. Patients in more deprived areas had 1.5 (CI 1.23 – 1.84, $p < 0.001$) and 1.3 (CI 1.21 – 1.40, $p < 0.001$) times the mortality hazard compared to those living in the least deprived areas, for FL and DLBCL, respectively. There was evidence of a linear trend in mortality hazard by deprivation level for FL ($p < 0.001$) and DLBCL ($p < 0.001$). The direction and magnitude of the hazard ratios after multiple imputation were similar to complete case analysis.

Table 3: Adjusted hazard ratios of death (before and after multiple imputation) for all patient characteristics amongst patients with (A) Follicular or (B) DLBCL in England between 2005-2015.

		Complete Case			After multiple imputation		
		HR*	95% CI	p-value	HR*	95% CI	p-value
(A) Follicular							
Sex							
	<i>Male</i>	Ref	Ref	-	Ref	Ref	-
	<i>Female</i>	1.08	0.85 – 1.38	0.540	0.84	0.74 – 0.94	<0.001
Comorbidity status							
	<i>None</i>	Ref	Ref	-	Ref	Ref	-
	<i>One</i>	1.52	1.15 – 2.03	<0.001	1.28	1.01 – 1.61	<0.041
	<i>Multimorbidity</i>	2.36	1.85 – 3.02	<0.001	2.17	1.78 – 2.64	<0.001
Deprivation[†]							
	<i>Least</i>	Ref	Ref	-	Ref	Ref	-
	<i>2</i>	1.03	0.70 – 1.53	0.873	1.09	0.90 – 1.31	0.378
	<i>3</i>	1.47	1.00 – 2.16	0.051	1.15	0.95 – 1.39	0.143
	<i>4</i>	1.30	0.87 – 1.94	0.200	1.44	1.19 – 1.73	<0.001
	<i>Most</i>	1.63	1.11 – 2.41	0.013	1.50	1.23 – 1.84	<0.001
Ethnicity							
	<i>White</i>	Ref	Ref	-	Ref	Ref	-
	<i>Other</i>	0.49	0.27 – 0.88	0.017	0.64	0.40 – 1.01	0.053
(B) Diffuse large B-cell							
Sex							
	<i>Male</i>	Ref	Ref	-	Ref	Ref	-
	<i>Female</i>	0.91	0.84 – 0.99	0.170	0.90	0.86 – 0.93	<0.001
Comorbidity status							
	<i>None</i>	Ref	Ref	-	Ref	Ref	-
	<i>One</i>	1.29	1.18 – 1.42		1.24	1.15 – 1.35	<0.001
	<i>Multimorbidity</i>	1.61	1.47 – 1.76	<0.001	1.44	1.34 – 1.55	<0.001
Deprivation[‡]							
	<i>Least</i>	Ref	Ref	-	Ref	Ref	-
	<i>2</i>	1.17	1.03 – 1.33	0.013	1.05	0.98 – 1.12	0.155
	<i>3</i>	1.15	1.01 – 1.30	0.029	1.13	1.06 – 1.20	<0.001
	<i>4</i>	1.25	1.11 – 1.42	<0.001	1.23	1.15 – 1.31	<0.001
	<i>Most</i>	1.32	1.16 – 1.51	<0.001	1.30	1.21 – 1.40	<0.001
Ethnicity							
	<i>White</i>	Ref	Ref	-	Ref	Ref	-
	<i>Other</i>	0.93	0.77 – 1.13	0.463	1.03	0.91 – 1.16	0.678

CI: Confidence interval, HR: Hazard Ratio
Missing values: (A) ethnicity n(%) = 3,472 (24.7%), (B) ethnicity n(%) = 6,191 (22.6%)
* Adjusted for sex, comorbidity status, deprivation, ethnicity and the restricted cubic splines of age
[†] [‡] Likelihood ratio test for the overall effect of deprivation (p<0.001)

Table 4 shows the linearly combined effect¹⁴ between comorbidity status and deprivation on short-term mortality by DLBCL and FL. Overall, at 1 year since diagnosis, amongst FL (**Table 4a**), patients who were most deprived with multimorbidity have 3.26 (CI 2.48 – 4.28) times higher short-term mortality hazard than patients without comorbidities and least deprived. Amongst DLBCL (**Table 4b**), and for the same comparison, the short-term mortality hazard was 1.88 (CI 1.70 – 2.07) times higher at 1 year.

Table 4: Linearly combined adjusted hazard ratio of comorbidity status with deprivation level on short-term mortality (after multiple imputation) amongst [A] FL (deaths at 1 year: n = 1,087) and [B] DLBCL (1 year: n = 9,095) for patients in England from 2005-2015.

	Comorbidity status		
	None HR* (95% CI)	One HR* (95% CI)	Multimorbidity HR* (95% CI)
[A] Follicular			
Deprivation			
<i>Least deprived</i>	Ref	1.28 (1.01 – 1.61)	2.17 (1.78 – 2.64)
2	1.09 (0.90 – 1.31)	1.39 (1.03 – 1.88)	2.36 (1.80 – 3.10)
3	1.15 (0.95 – 1.39)	1.47 (1.09 – 1.98)	2.50 (1.91 – 3.26)
4	1.44 (1.19 – 1.73)	1.84 (1.36 – 2.47)	3.12 (2.40 – 4.06)
<i>Most deprived</i>	1.50 (1.23 – 1.84)	1.92 (1.42 – 2.59)	3.26 (2.48 – 4.28)
[B] DLBCL			
Deprivation			
<i>Least deprived</i>	Ref	1.24 (1.15 – 1.35)	1.44 (1.34 – 1.55)
2	1.05 (0.98 – 1.12)	1.30 (1.18 – 1.44)	1.51 (1.37 – 1.67)
3	1.13 (1.06 – 1.20)	1.40 (1.27 – 1.55)	1.63 (1.48 – 1.80)
4	1.23 (1.15 – 1.31)	1.52 (1.38 – 1.69)	1.78 (1.61 – 1.95)
<i>Most deprived</i>	1.30 (1.21 – 1.40)	1.62 (1.46 – 1.80)	1.88 (1.70 – 2.07)

CI: Confidence interval. HR: Hazard Ratio.

*Adjusted for sex, ethnicity and the restricted cubic splines of age.

Figure 2 shows the unadjusted (Nelson-Aalen non-parametric estimate) and standardised risks of death up to 1 year since diagnosis for FL and DLBCL by comorbidity status and deprivation. **Supplementary Tables S4a** and **S4b** show results of complete case and after multiple imputation. Standardised to age, sex, deprivation and ethnicity, the risk of death over the first year was consistently higher amongst those with multimorbidity compared to those with one comorbidity or none. For both FL and DLBCL, the unadjusted analysis showed that patients with multimorbidity had consistently higher cumulative incidence of death compared to those with one comorbidity or none (log rank test $p < 0.001$).

Discussion

We aimed to explore the association between comorbidity status, and deprivation and their combination, on short-term mortality for patients with FL or DLBCL. We found that multimorbidity and deprivation and their combined effect are strong independent predictors of short-term mortality amongst patients with DLBCL and FL in England during 2005-2015.

To our knowledge, this is the first study that investigates the association between multimorbidity, and deprivation, on short-term mortality amongst patients with DLBCL and FL in England. Despite the scarcity of research within England, our findings are consistent with previous evidence from other countries. A Swedish study found that higher comorbidity status was independently associated with a higher risk of mortality amongst patients with diffuse large B-cell lymphoma.³¹ Additionally, more deprived, compared to least deprived, patients had a higher risk of DLBCL-related mortality and there was evidence of a significant linear trend across the quintiles of deprivation. A Danish study found that higher comorbidity status was independently associated with shorter survival lengths amongst patients with any type of NHL (Hazard Ratio 1.60, CI 1.45-1.75).³² However, these studies used a non-cancer specific comorbidity score, which underperforms (in comparison to cancer-specific scores) when using predictive models for short-term outcomes.²⁰ These studies suggest that the effect of comorbidity mainly occurs prior to, and shortly after, cancer diagnosis. Further studies could assess the effect of these prognostic factors on longer term survival from DLBCL or FL, using deprivation-specific life tables to minimise the inaccuracy of expected mortality when life tables are not also stratified by comorbidity status.³³

As the association between deprivation and NHL survival is not studied as widely as solid tumours, it was unclear whether there was an association between deprivation and short-term mortality for haematological malignancies. Previous studies have described a deprivation-gap in survival comparing the least- to most-deprived,^{8,10,34} but have not assessed the association. Although explored for Hodgkin's lymphoma,³⁵ to our knowledge, this is the first study to explore the association between deprivation and short-term mortality for NHL in England: our study provides evidence of a strong and independent association.

There are several dynamics that may explain the association observed in this study. Firstly, the presence of a comorbidity is known to affect the timely diagnosis of DLBCL and FL,³⁶ such that comorbidities presenting with similar symptoms to DLBCL or FL may delay the diagnosis and dissimilar symptoms may hasten the diagnosis. Moreover, the prevalence of comorbidities increases with age, and amongst older patients with DLBCL and FL this prevalence is consistently over 60%,^{37,38} which may partly explain the delay in diagnosis amongst older ages. Further research is needed to identify comorbidities that alter the timely diagnosis

Secondly, guidelines of lymphoma management focus on a single-disease standard regimen, but there is little guidance on multi-disease management.³⁹ A systematic review found the majority of patients with a comorbidity did not receive the standard regimen and were allocated alternative, less-intense treatments.⁴⁰ Cancer care could be improved by defining clear guidelines that recommend a comorbidity-specific treatment regimen and provide an accurate definition, and a measure, of the dose-intensity.

Thirdly, differences in access to treatments, or risk of adverse effects, may partly explain the multimorbidity gap in survival from DLBCL and FL; clinicians may abstain from allocating a treatment associated with a higher risk of adverse events because it can exacerbate the complex management of cancer care. Patients without comorbidities, after receiving standard treatment regimens, still experience an increased risk of cardiovascular events.⁴¹ A first-line standard treatment for DLBCL and FL is a combination of chemotherapy and immunotherapies, such as rituximab, and is known to be effective for those of an advanced age. Rituximab is often used in combination with anthracyclines (e.g. doxorubicin), which is associated with an increase in the incidence of adverse events (e.g. cardiotoxicity) commonly in the form of congestive heart failure.⁴²

Lastly, the association between deprivation and short-term mortality, that is not explained by patient characteristics, might be explained by the association between deprivation and use of emergency services or population density,^{43,44} or between population density and the use of emergency services.⁴⁵ For example, population dense areas may accumulate high demands that current facilities of healthcare services are unable to accommodate. Therefore, emergency services (e.g., emergency diagnostic route), which is associated with a

late stage of cancer, may explain the higher mortality hazard observed amongst more deprived patients. Further research could investigate the demand and availability of healthcare services in densely populated areas.

The strengths of this study include the large sample size within a database of high-quality population-based clinical records with a high national coverage. We linked clinical records with the HES database, which encompasses all patients in England with a diagnosis of DLBCL and FL between 2005 and 2013. The objective data sources provide information on patients that is gathered prospectively. Furthermore, the standardised risk provides an interpretation of the risk of death that is averaged over the entire population.

Due to data availability, our study has some limitations. Firstly, we did not include tumour stage, route to diagnosis (e.g., general practitioner referral), or treatment plan; consequently, further research is needed to dissect the effects of comorbidity, stage and treatment on survival. Since tumour stage, route and treatment allocation are considered to be on the causal pathway between comorbidity status and short-term mortality, causal inference mediation analysis is required to estimate the proportion of the effect of comorbidity status on survival that is explained by said mediators.

Secondly, recent research highlights the interest in using individual-level socioeconomic measures for assessing patient health outcomes in addition to area-level measures of deprivation.⁴⁶ However, information on individual-level socioeconomic measures were unavailable, so we used only an area-level measure of socioeconomic status, which encapsulates the multidimensional composition of a patient's deprivation level in addition to the contextual level.²¹ Furthermore, there is better concordance between area- and individual-level measures of education when assessing patient health outcomes.⁴⁶ The observed deprivation level of a patient in our study is likely to be consistent had they been diagnosed at a different time; this is because deprivation scores have a high concordance between updates (i.e., IMD of 2007, 2010, and 2015).²¹ Our results are comparable to studies using this area-level measure of deprivation.

Thirdly, Hospital Episode Statistics (HES) data contains information on all patients admitted to a hospital (secondary care) in England. It is possible that some comorbidities were not observed because they were

diagnosed, and treated, during primary care (e.g., general practitioner consultations). However, the Royal College of Surgeons' comorbidity index, amongst other indices, are constructed based on the impact of the comorbidity on the risk of mortality; in other words, severe comorbidities that require hospitalisation. Comorbidities of the RCS comorbidity index are those that often require hospitalisation, leading to a record within HES data. Previous research has shown that combining primary care records to secondary care data identifies a greater proportion of comorbidity within the population; however, the inclusion of comorbidities identified from primary care records does not have a large effect on predicted cancer survival beyond results obtained using secondary care data.⁴⁷

Lastly, as complete case analysis may lead to selection bias, we performed multiple imputation under a missing at random assumption. We obtained the same conclusions under a complete case analysis and after multiple imputation. Since the missing at random assumption is untestable,⁴⁸ further work could conduct a sensitivity analysis to departures from the missing at random assumption, through techniques for imputing under a missing not at random assumption.⁴⁹

In conclusion, multimorbidity and deprivation, combined and independently, are strong predictors of an increased risk of short-term mortality at 1 year since diagnosis amongst patients with DLBCL or FL in England. Therefore, public health prevention strategies are needed to reduce the short-term mortality gap due to socioeconomic inequalities and comorbidities amongst NHL patients.

Figure Legends

Figure 1: Overlap plots for the density of predicted probabilities of comorbidity status amongst patients (n=41,422), aged 45-99, in England diagnosed with non-Hodgkin lymphoma during 2005-2013. Propensity score: relates to the predicted probability of having any comorbidity level as measured by a multinomial logistic regression model conditioning on the independent variables (i.e., age at diagnosis, sex, deprivation level, and ethnicity).

Figure 2. Risk of short-term mortality for Follicular lymphoma (n=14,043) and DLBCL (n=27,379) by comorbidity status and deprivation level in England between 2005-2015. (Solid: Aalen-Nelson approach, Dash: standardised to the empirical distribution of age, sex, and ethnicity).

Additional information

Acknowledgements: We thank Professor Paul Lambert (Leicester University, UK) for his advice on using the *stpm2* Stata software package in combination with multiple imputation. We also thank Adrian Turculet (Data Manager, LSHTM Inequalities in Cancer Outcomes Network), for his support with the data linkage.

Author's contributions: MS, ENN and BR contributed to the conception of the study and designed the study. ENN, BR, CL, and MALF provided advice on statistical methods. MS conducted the analyses of the data and prepared the draft of the manuscript, tables and figures. ENN and BR supervised the study and provided comments on the manuscript draft. ENN, BR, MALF, CL, SBM, ABe and ABo provided comments on the final draft of the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate: We obtained the statutory approvals required for this research from the Confidentiality Advisory Group (CAG) of the Health Research Authority (HRA): PIAG 1–05(c) 2007. Ethical approval was obtained from the Research Ethics Committee (REC) of the Health Research Authority (HRA): 07/MRE01/52. This work uses data provided by patients and collected by the National Health Service as part of their care and support. We used anonymised National Cancer Registry and Hospital Episode Statistics data. No consent to participate was sought from patients.

Availability of data and materials: The data that support the findings of this study are available via application to the Public Health England Office for Data Release, but restrictions apply to the availability of these data.

Funding: This research was funded by Cancer Research UK grant number C7923/A18525. The authors declare no support from any organisations for the submitted work. The design of the study, the analyses and the writing of the manuscript were solely the responsibility of the authors. The findings and conclusions in this manuscript are those of the authors and do not necessarily represent the views of Cancer Research UK.

Competing interests: The authors declare there are no competing interests.

References

1. Office for National Statistics. *Cancer registration statistics, England: 2017*. (2019).
2. Cancer Research UK. Non-Hodgkin lymphoma statistics. (2017). Available at: <https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/non-hodgkin-lymphoma>. (Accessed: 24th April 2020)
3. Quaresma, M., Coleman, M. P. & Rachet, B. 40-year trends in an index of survival for all cancers combined and survival adjusted for age and sex for each cancer in England and Wales, 1971–2011: a population-based study. *Lancet* **385**, 1206–1218 (2015).
4. Shankland, K. R., Armitage, J. O. & Hancock, B. W. Non-Hodgkin lymphoma. *Lancet* **380**, 848–857 (2012).
5. Epidemiology & Cancer Statistics Group University of York. Haematological Malignancy Research Network. (2016).
6. Allemani, C. *et al.* Global surveillance of trends in cancer survival 2000–14 (CONCORD-3): analysis of individual records for 37,513,025 patients diagnosed with one of 18 cancers from 322 population-based registries in 71 countries. *Lancet* **391**, 1023–1075 (2018).
7. Thomson, C. S. & Forman, D. Cancer survival in England and the influence of early diagnosis: what can we learn from recent EURO CARE results? *Br. J. Cancer* **101**, S102–S109 (2009).
8. Rachet, B., Mitry, E., Shah, A., Cooper, N. & Coleman, M. P. Survival from non-Hodgkin lymphoma in England and Wales up to 2001. *Br. J. Cancer* **99**, S104–S106 (2008).
9. National Institute for Health and Care Excellence. *Haematological cancers: improving outcomes*. (2016).
10. Exarchakou, A., Rachet, B., Belot, A., Maringe, C. & Coleman, M. P. Impact of national cancer policies on cancer survival trends and socioeconomic inequalities in England, 1996–2013: population based study. *BMJ* **360**, k764–k764 (2018).
11. Porta, M. *A Dictionary of Epidemiology*. (Oxford University Press, 2014).
12. Salika, T., Lyratzopoulos, G., Whitaker, K. L., Waller, J. & Renzi, C. Do comorbidities influence help-seeking for cancer alarm symptoms? A population-based survey in England. *J. Public Health (Bangkok)*. **40**, 340–349 (2017).

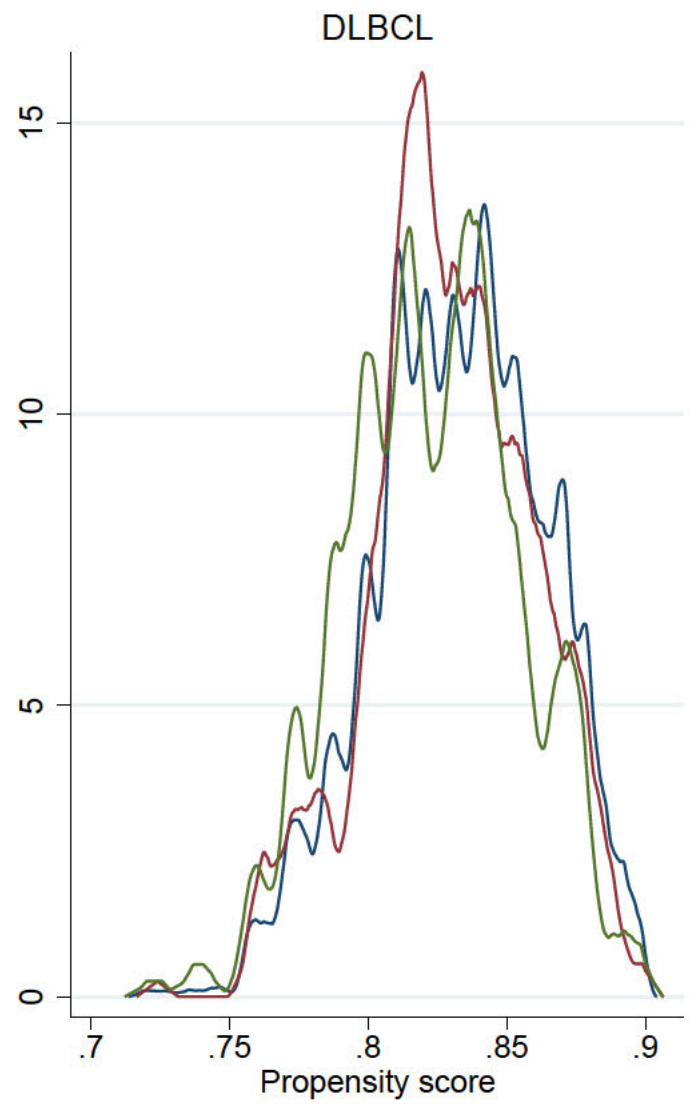
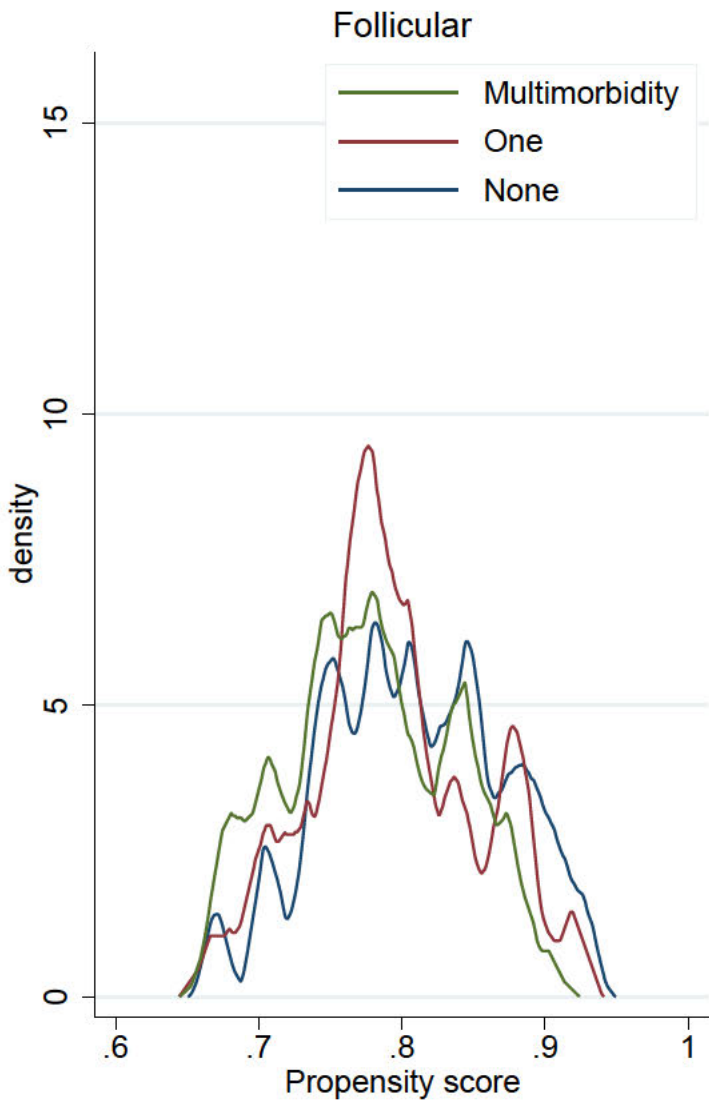
13. Fowler, H. *et al.* Comorbidity prevalence among cancer patients: a population-based cohort study of four cancers. *BMC Cancer* **20**, 2 (2020).
14. Luque-Fernandez, M. A. *et al.* Multimorbidity and short-term overall mortality among colorectal cancer patients in Spain: A population-based cohort study. *Eur. J. Cancer* **129**, 4–14 (2020).
15. International Agency for Research on Cancer. International Classification of Diseases for Oncology. (2013). Available at: <http://codes.iarc.fr/>. (Accessed: 4th October 2019)
16. gov.uk. National Cancer Registry and Analysis Service. (2017). Available at: <https://www.gov.uk/guidance/national-cancer-registration-and-analysis-service-ncras>. (Accessed: 4th October 2019)
17. Maringe, C., Fowler, H., Rachet, B. & Luque-Fernandez, M. A. Reproducibility, reliability and validity of population-based administrative health data for the assessment of cancer non-related comorbidities. *PLoS One* **12**, e0172814 (2017).
18. Hernán, M. Á. & Robins, J. M. Estimating causal effects from epidemiological data. *J. Epidemiol. Community Health* **60**, 578–586 (2006).
19. Charlson, M. E., Pompei, P., Ales, K. L. & MacKenzie, C. R. A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation. *J. Chronic Dis.* **40**, 373–383 (1987).
20. Armitage, J. N. & van der Meulen, J. H. Identifying co-morbidity in surgical patients using administrative data with the Royal College of Surgeons Charlson Score. *Br J Surg* **97**, 772–781 (2010).
21. gov.uk. Indices of Multiple Deprivation. (2015). Available at: <https://www.gov.uk/government/statistics/english-indices-of-deprivation-2015>. (Accessed: 4th October 2019)
22. National Health Service: data dictionary. Lower Super Output Area. (2018). Available at: https://www.datadictionary.nhs.uk/data_dictionary. (Accessed: 4th October 2019)
23. Armitage, P., Berry, G. & Matthews, J. *Statistical methods in medical research*. (Blackwell Science, 2002).
24. Royston, P. & Lambert, P. *Flexible Parametric Survival Analysis Using Stata: Beyond the Cox Model*. (2011).
25. Robins, J. M., Hernán, M. Á. & Brumback, B. Marginal Structural Models and Causal Inference in

- Epidemiology. *Epidemiology* **11**, (2000).
26. Cole, S. & Hernan, M. A. Adjusted survival curves with inverse probability weights. *Comput. Methods Programs Biomed.* **75**, 45–49 (2004).
 27. Dupont, W. D. *Statistical Modeling for Biomedical Researchers: A Simple Introduction to the Analysis of Complex Data*. (Cambridge University Press, 2009).
 28. Armitage, J. O. Staging Non-Hodgkin Lymphoma. *CA. Cancer J. Clin.* **55**, 368–376 (2005).
 29. Halpern, M. T. *et al.* Association of insurance status and ethnicity with cancer stage at diagnosis for 12 cancer sites: a retrospective analysis. *Lancet Oncol.* **9**, 222–231 (2008).
 30. Little, R. J. & Rubin, D. B. *Statistical Analysis with Missing Data*. (John Wiley & Sons, Inc., 1987).
 31. Wästerlid, T. *et al.* Impact of comorbidity on disease characteristics, treatment intent and outcome in diffuse large B-cell lymphoma: a Swedish lymphoma register study. *J. Intern. Med.* **285**, 455–468 (2019).
 32. Frederiksen, B. L., Dalton, S. O., Osler, M., Steding-Jessen, M. & de Nully Brown, P. Socioeconomic position, treatment, and survival of non-Hodgkin lymphoma in Denmark—a nationwide study. *Br J Cancer* **106**, 988–995 (2012).
 33. Rubio, F. J., Rachet, B., Giorgi, R., Maringe, C. & Belot, A. On models for the estimation of the excess mortality hazard in case of insufficiently stratified life tables. *Biostatistics* **22**, 51–67 (2021).
 34. Rachet, B. *et al.* Socioeconomic inequalities in cancer survival in England after the NHS cancer plan. *Br J Cancer* **103**, 446–453 (2010).
 35. Rafiq, M. *et al.* Socioeconomic deprivation and regional variation in Hodgkin’s lymphoma incidence in the UK: A population-based cohort study of 10 million individuals. *BMJ Open* **9**, (2019).
 36. Howell, D. A. *et al.* Time-to-diagnosis and symptoms of myeloma, lymphomas and leukaemias: a report from the Haematological Malignancy Research Network. *BMC Blood Disord.* **13**, 9 (2013).
 37. van Spronsen, D. J., Janssen-Heijnen, M. L., Breed, W. P. & Coebergh, J. W. Prevalence of co-morbidity and its relationship to treatment among unselected patients with Hodgkin’s disease and non-Hodgkin’s lymphoma, 1993-1996. *Ann Hematol* **78**, 315–319 (1999).
 38. Janssen-Heijnen, M. L. *et al.* A population-based study of severity of comorbidity among patients with non-Hodgkin’s lymphoma: prognostic impact independent of International Prognostic Index. *Br J*

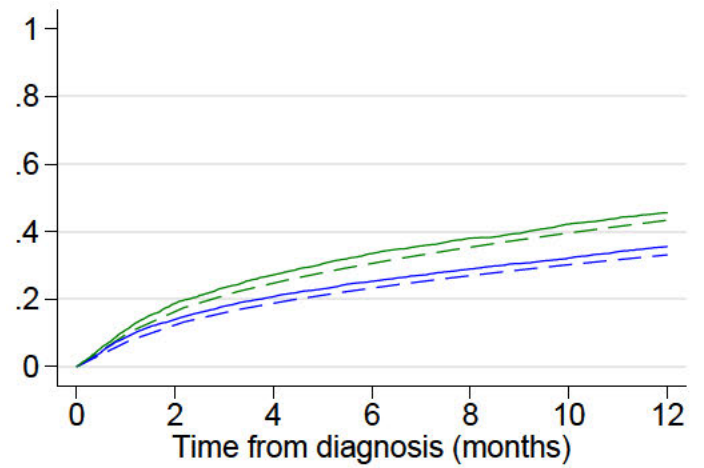
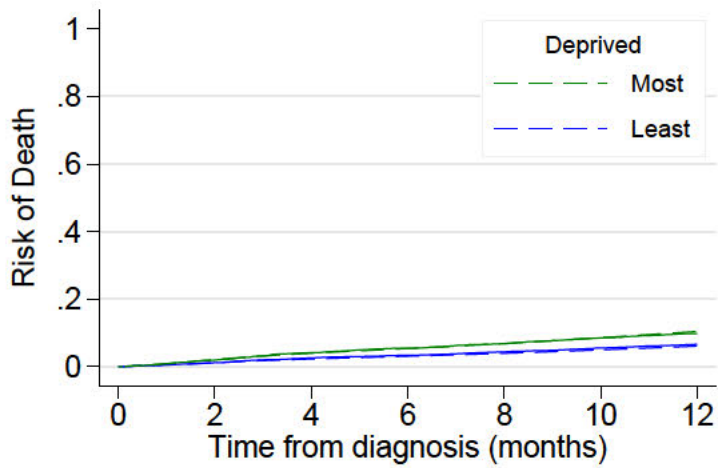
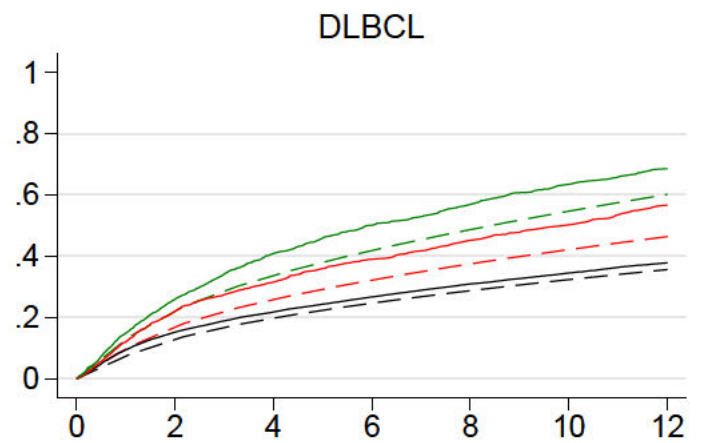
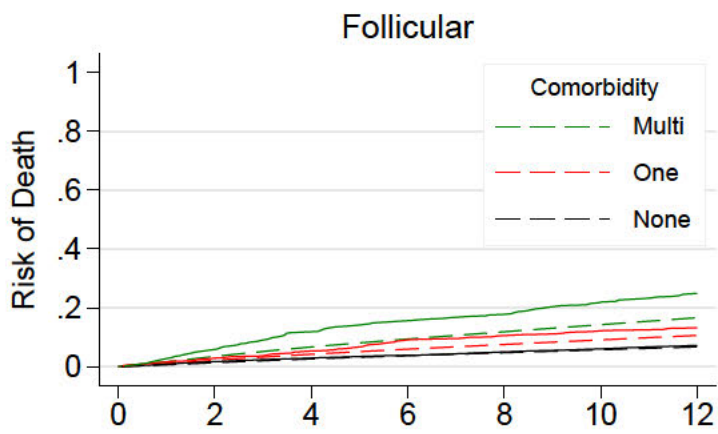
Haematol **129**, 597–606 (2005).

39. National Institute for Health and Care Excellence. *Non-Hodgkin's lymphoma: diagnosis and management*. (National Institute for Health and Care Excellence, 2016).
40. Terret, C., Albrand, G., Rainfray, M. & Soubeyran, P. Impact of comorbidities on the treatment of non-Hodgkin's lymphoma: a systematic review. *Expert Rev. Hematol.* **8**, 329–341 (2015).
41. Linschoten, M. *et al.* Cardiovascular adverse events in patients with non-Hodgkin lymphoma treated with first-line cyclophosphamide, doxorubicin, vincristine, and prednisone (CHOP) or CHOP with rituximab (R-CHOP): a systematic review and meta-analysis. *Lancet Haematol.* **7**, e295–e308 (2020).
42. McGowan, J. V *et al.* Anthracycline Chemotherapy and Cardiotoxicity. *Cardiovasc. Drugs Ther.* **31**, 63–75 (2017).
43. Carlisle, R., Groom, L. M., Avery, A. J., Boot, D. & Earwicker, S. Relation of Out of Hours Activity by General Practice and Accident and Emergency Services with Deprivation in Nottingham: Longitudinal Survey on JSTOR. *BMJ Br. Med. J.* **316**, 520–523 (1998).
44. Venerandi, A., Quattrone, G. & Capra, L. A scalable method to quantify the relationship between urban form and socio-economic indexes. *EPJ Data Sci.* **7**, 4 (2018).
45. Peacock, P. J. & Peacock, J. L. Emergency call work-load, deprivation and population density: An investigation into ambulance services across England. *J. Public Health (Bangkok).* **28**, 111–115 (2006).
46. Ingleby, F. C. *et al.* Assessment of the concordance between individual-level and area-level measures of socio-economic deprivation in a cancer patient cohort in England and Wales. *BMJ Open* **10**, e041714 (2020).
47. Crooks, C. J., West, J. & Card, T. R. A comparison of the recording of comorbidity in primary and secondary care by using the Charlson Index to predict short-term and long-term survival in a routine linked data cohort. *BMJ Open* **5**, e007974 (2015).
48. Carpenter, J. R. & Kenward, M. G. *Multiple Imputation and Its Application*. (John Wiley & Sons, Ltd, 2013).
49. Gachau, S. *et al.* Handling missing data in modelling quality of clinician-prescribed routine care: Sensitivity analysis of departure from Missing at Random (MAR) assumption. *Stat. Methods Med. Res.*

29, 3076 (2020).



Risk of short-term mortality for non-Hodgkin lymphoma subtypes



Supplementary Tables

This page is intentionally blank. Please move to next page.

Supplementary Table S1. Distribution of non-Hodgkin lymphoma subtypes for patients in England diagnosed from 2005-2013, with respective morphology and topography ICD-O-3 codes.

Index	Site group (subtype)	Progression	Topography	Morphology	n	%
1	CLL/SLL*	Indolent	C82.0-C85.9	9670, 9823	3,875	5.08
2	Waldenstrom macroglobulinemia	Indolent	C82.0-C85.9	9761	2,398	3.14
3	Mantle cell	Indolent	C82.0-C85.9	9673	3,458	4.53
4	Diffuse large B-cell	Aggressive	C82.0-C85.9	9680, 9688, 9737-9738	27,379	35.89
5	Burkitt	Aggressive	C82.0-C85.9	9687, 9826	695	0.91
6	Follicular	Indolent	C82.0-C85.9	9690-9691, 9695, 9698	14,043	18.41
7	Mature T-cell	Aggressive	C82.0-C85.9	9702	5,127	6.72
8	Marginal zone B-cell	Indolent	C82.0-C85.9	9689, 9699, 9760, 9764, 9699	4,277	5.61
Subtotal					61,252	80.30
9	Not Otherwise Specified	n/a	C82.0-C85.9	9591, 9675, 9735	9,581	12.56
10	Other***	n/a	C82.0-C85.9	9591, 9675, 9735	5,449	7.14
Total					76,282	100.00**

n/a – not applicable; there was no subtype information
* Chronic lymphocytic leukaemia/Small-cell lymphocytic lymphoma
** Percentages may not equate to 100 0% due to rounding
*** The morphology code specifies these patients are diagnosed with NHL. However, the description states ‘other’ these patients are classified similarly to ‘Not Otherwise Specified’

Supplementary Table S2: Comorbidities and their diagnostic ICD-10 codes

Comorbidity	ICD-10
Myocardial infarction	I21.x, I22.x, I25.2
Congestive heart failure	I11.0, I13.0, I13.2, I25.5, I42.0, I42.5–I42.9, I43.x, I50.x, P29.0
Peripheral vascular disease	I70.x, I71.x, I73.1, I73.8, I73.9, I77.1, I79.0, I79.2, K55.1, K55.8, K55.9, Z95.8, Z95.9
Cerebrovascular disease	G45.x, G46.x, H34.0, I60.x–I69.x
Dementia	F00.x–F03.x, F05.1, G30.x, G31.1
Chronic obstructive pulmonary disease	I27.9, J40.x–J47.x, J60.x–J67.x, J68.4, J70.1, J70.3
Rheumatic disease	M05.x, M06.x, M31.5, M32.x–M34.x, M35.1, M35.3, M36.0
Liver disease	B18.x, K70.0–K70.3, K70.9, K71.3–K71.5, K71.7, K73.x, K74.x, K76.0, K76.2–K76.4, K76.8, K76.9, Z94.4, K71.1, K72.1, K72.9, K76.5, K76.6, K76.7, I85.0, I85.9, I86.4, I98.2, K70.4,
Diabetes without chronic complication	E10.0, E10.1, E10.6, E10.8, E10.9, E11.0, E11.1, E11.6, E11.8, E11.9, E12.0, E12.1, E12.6, E12.8, E12.9, E13.0, E13.1, E13.6, E13.8, E13.9, E14.0, E14.1, E14.6, E14.8, E14.9
Diabetes with chronic complication	E10.7, E11.2–E11.5, E11.7, E12.2–E12.5, E12.7, E13.2–E13.5, E13.7, E14.2–E14.5, E14.7
Hemiplegia or paraplegia	G04.1, G11.4, G80.1, G80.2, G81.x, G82.x, G83.0–G83.4, G83.9
Renal disease	I12.0, I13.1, N03.2–N03.7, N05.2–N05.7, N18.x, N19.x, N25.0, Z49.0–Z49.2, Z94.0, Z99.2
AIDS/HIV	B20.x–B22.x, B24.x

ICD-10: International Classification of Diseases, 10th Revision

Diabetes with/without chronic complication is combined in the RCS Charlson Comorbidity Score

Supplementary Table S3. Sociodemographic characteristics, route of diagnosis, comorbidity status and Ann Arbor cancer stage distribution, of Follicular (n=14,043) and DLBCL (n=27,379) lymphomas in England, 2005-2013.

	Follicular N = 14,043	DLBCL N = 27,379	OR* (95% CI)	p-value
Age (mean; SD)**	66.7 (11.0)	70.8 (11.3)	1.37 (1.35 – 1.40)	<0.001
Gender				
<i>Male</i>	6,512 (46.4)	14,652 (53.5)	Ref	Ref
<i>Female</i>	7,531 (53.6)	12,727 (46.5)	0.75 (0.72 - 0.78)	<0.001
Deprivation				
<i>Least deprived</i>	3,284 (23.4)	5,895 (21.5)	Ref	Ref
2	3,228 (23.0)	6,238 (22.8)	1.08 (1.01 – 1.14)	0.016
3	3,013 (21.5)	5,756 (21.0)	1.06 (1.00 – 1.13)	0.047
4	2,609 (18.6)	5,341 (19.5)	1.14 (1.07 – 1.22)	<0.001
<i>Most deprived</i>	1,909 (13.6)	4,149 (15.2)	1.21 (1.13 – 1.30)	<0.001
Comorbidity status				
<i>Non</i>	12,898 (91.9)	24,269 (88.6)	Ref	Ref
<i>One</i>	623 (4.4)	1,538 (5.6)	1.60 (1.45 – 1.77)	<0.001
<i>Multimorbidity</i>	522 (3.7)	1,572 (5.7)	1.58 (1.43 – 1.75)	<0.001
Route				
<i>Elective</i>	11,211 (86.7)	17,280 (66.3)	Ref	Ref
<i>Emergency</i>	1,727 (13.4)	8,799 (33.7)	3.31 (3.12 – 3.50)	<0.001
<i>Missing</i>	1,105 (7.9)	1,300 (4.8)	-	-
Stage				
<i>I</i>	848 (26.6)	1,680 (27.8)	Ref	Ref
<i>II</i>	473 (14.8)	1,169 (19.3)	1.25 (1.09 – 1.43)	0.001
<i>III</i>	822 (25.7)	1,046 (17.3)	0.64 (0.57 – 0.73)	<0.001
<i>IV</i>	1050 (32.9)	2,152 (35.6)	1.03 (0.93 – 1.16)	0.548
<i>Missing</i>	10,850 (77.3)	21,332 (77.9)	-	-
Ethnicity				
<i>White</i>	10,110 (95.6)	20,204 (95.4)	Ref	Ref
<i>Other</i>	461 (4.4)	984 (4.6)	1.07 (0.95 – 1.20)	0.254
<i>Missing</i>	3,472 (24.7)	6,191 (22.6)	-	-

* Odds ratio from a complete case analysis comparing the odds of DLBCL to Follicular lymphoma

** 10-year increase in age

P-values calculated from Wald tests

Supplementary Table S4a: Risk of short-term mortality amongst patients diagnosed with Follicular lymphomas (n=14,043) by comorbidity status in England between 2005 and 2013

Month	Complete Case Analysis						After Multiple Imputation					
	Comorbidity status						Comorbidity status					
	None		One		Multimorbidity		None		One		Multimorbidity	
	CH	95% CI	CH	95% CI	CH	95% CI	CH	95% CI	CH	95% CI	CH	95% CI
1	0.6	0.4–0.9	1.0	0.7–1.4	1.6	1.0–2.3	0.8	0.7–1.0	1.3	1.1–1.5	2.4	1.6–3.2
2	1.4	1.0–1.7	2.2	1.6–2.7	3.5	2.3–4.6	1.7	1.5–1.9	2.6	2.2–3.0	4.8	3.4–6.3
3	2.0	1.6–2.4	3.2	2.6–3.9	5.2	3.6–6.7	2.4	2.2–2.7	3.7	3.2–4.2	6.8	4.9–8.7
4	2.6	2.2–3.1	4.2	3.5–4.9	6.7	4.8–8.6	3.0	2.7–3.3	4.6	4.0–5.2	8.4	6.1–10.8
5	3.2	2.6–3.7	5.1	4.3–5.9	8.1	5.9–10.3	3.6	3.3–3.9	5.4	4.8–6.1	9.9	7.2–12.6
6	3.7	3.1–4.3	5.9	5.0–6.9	9.4	6.9–11.9	4.1	3.8–4.4	6.2	5.5–7.0	11.3	8.3–14.3
7	4.2	3.6–4.9	6.7	5.7–7.7	10.6	7.8–13.4	4.6	4.2–5.0	7.0	6.2–7.9	12.7	9.3–16.0
8	4.7	4.0–5.5	7.5	6.4–8.6	11.8	8.8–14.9	5.1	4.8–5.5	7.8	6.9–8.8	14.1	10.4–17.8
9	5.2	4.5–6.0	8.3	7.2–9.4	13.1	9.7–16.4	5.7	5.3–6.1	8.7	7.6–9.7	15.5	11.5–19.5
10	5.7	5.0–6.5	9.1	7.9–10.3	14.3	10.6–17.9	6.2	5.8–6.7	9.5	8.4–10.6	17.0	12.6–21.3
11	6.2	5.4–7.0	9.9	8.6–11.1	15.4	11.5–19.4	6.8	6.4–7.3	10.3	9.2–11.5	18.4	13.7–23.1
12	6.7	5.9–7.5	10.6	9.2–12.0	16.6	12.3–20.9	7.4	6.9–7.8	11.2	9.9–12.4	19.8	14.8–24.9

CH – cumulative hazard of death (e x 10²)

95% CI – confidence interval (e x 10²)

Supplementary Table S4b: Risk of short-term mortality amongst patients diagnosed with DLBCL (n=27,379) by comorbidity status in England between 2005 and 2013

Month	Complete Case Analysis						After Multiple Imputation					
	Comorbidity status						Comorbidity status					
	None		One		Multimorbidity		None		One		Multimorbidity	
	CH	95% CI	CH	95% CI	CH	95% CI	CH	95% CI	CH	95% CI	CH	95% CI
1	7.4	6.9–7.9	9.7	9.0–10.4	13.2	11.5–14.9	9.2	8.9–9.6	12.0	11.4–12.5	16.1	14.5–17.7
2	12.9	12.2–13.7	17.0	16.0–18.0	23.0	20.4–25.7	15.3	14.8–15.7	19.7	18.8–20.5	26.3	23.9–28.8
3	16.8	16.0–17.6	22.0	20.8–23.2	30.0	26.6–33.3	19.1	18.6–19.7	24.6	23.6–25.6	32.8	29.8–35.8
4	19.8	18.9–20.7	25.9	24.5–27.2	35.4	31.5–39.3	22.1	21.6–22.7	28.4	27.2–29.5	37.7	34.3–41.1
5	22.4	21.4–23.3	29.2	27.7–30.7	39.9	35.6–44.3	24.8	24.1–25.4	31.7	30.4–32.9	42.0	38.2–45.8
6	24.7	23.7–25.7	32.2	30.6–33.8	43.9	39.2–48.7	27.1	26.5–27.8	34.7	33.3–36.0	45.9	41.8–50.0
7	26.8	25.8–27.8	34.9	33.3–36.6	47.5	42.4–52.6	29.3	28.6–30.0	37.5	36.0–38.9	49.6	45.2–53.9
8	28.8	27.7–29.8	37.5	35.8–39.2	50.8	45.4–56.2	31.4	30.7–32.1	40.1	38.5–41.6	53.0	48.3–57.6
9	30.6	29.5–31.7	39.9	38.1–41.7	53.8	48.2–59.5	33.3	32.6–34.0	42.5	40.9–44.1	56.2	51.2–61.1
10	32.4	31.2–33.5	42.2	40.3–44.0	56.7	50.8–62.6	35.1	34.4–35.9	44.8	43.2–46.5	59.2	54.0–64.4
11	34.0	32.8–35.2	44.3	42.4–46.3	59.4	53.3–65.5	36.9	36.1–37.7	47.1	45.3–48.8	62.1	56.7–67.6
12	35.6	34.4–36.9	46.4	44.3–48.5	62.0	55.6–68.4	38.6	37.7–39.4	49.2	47.4–51.0	64.9	59.2–70.6

CH – cumulative hazard (e x 10²)

95% CI – confidence interval (e x 10²)

A.5.4 Introduction to computational causal inference

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	lsh1601639	Title	Mr
First Name(s)	Matthew		
Surname/Family Name	Smith		
Thesis Title	Survival of patients with non-Hodgkin lymphoma in England: investigating the socioeconomic inequalities		
Primary Supervisor	Edmund Njeru Njagi		

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?	Statistics In Medicine
Please list the paper's authors in the intended authorship order:	Matthew J. Smith, Mohammad Ali Mansournia, Camille Maringe, Paul N. Zivich, Stephen R. Cole, Clemence Leyrat, Aurelien Belot, Bernard Rchet, Miguel Angel Luque Fernandez
Stage of publication	Submitted

SECTION D – Multi-authored work

<p>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)</p>	<p>The article arises from the motivation to disseminate the principles of modern epidemiology amongst clinicians and applied researchers. MJS and MALF developed the concept, designed the first draft of the article and developed the computing code. All authors interpreted and reviewed the code, and drafted and revised the manuscript. All authors read and approved the final version of the manuscript. MALF is the guarantor of the article.</p>
---	--

SECTION E

Student Signature	Matthew J. Smith
Date	7th June 2021

Supervisor Signature	[REDACTED]
Date	

RESEARCH ARTICLE

Educational notes: Introduction to computational causal inference using reproducible Stata, R and Python code

Matthew J. Smith¹ | Mohammad Ali Mansournia² | Camille Maringe¹ | Paul N. Zivich^{3,4} | Stephen R. Cole³ | Clémence Leyrat¹ | Aurélien Belot¹ | Bernard Rachet¹ | Miguel Angel Luque-Fernandez^{*1,5,6}

¹Inequalities in Cancer Outcomes Network, Department of Non-communicable Disease Epidemiology, London School of Hygiene and Tropical Medicine, London, U.K.

²Department of Epidemiology and Biostatistics, Tehran University of Medical Sciences, Tehran, Iran.

³Department of Epidemiology, University of North Carolina at Chapel Hill, North Carolina, U.S.

⁴Carolina Population Center, University of North Carolina at Chapel Hill, North Carolina, U.S.

⁵Non-communicable Disease and Cancer Epidemiology Group, Instituto de investigacion Biosanitaria de Granada (ibs.GRANADA), Andalusian School of Public Health, University of Granada, Granada, Spain

⁶Biomedical Network Research Centers of Epidemiology and Public Health (CIBERESP), Madrid, Spain.

Correspondence

* Miguel Angel Luque-Fernandez, Email: miguel-angel.luque@lshtm.ac.uk

Present Address

Keppel St, Bloomsbury, London WC1E 7HT, United Kingdom

Abstract The main purpose of many medical studies is to estimate the effects of a treatment or exposure on an outcome. However, it is not always possible to randomise the study participants to a particular treatment, therefore observational study designs may be used. There are major challenges with observational studies; one of which is confounding. Controlling for confounding is commonly performed by direct adjustment of measured confounders; although, sometimes this approach is suboptimal due to modelling assumptions and misspecification. Recent advances in the field of causal inference have dealt with confounding by building on classical standardisation methods. However, these recent advances have progressed quickly with a relative paucity of computational-oriented applied tutorials contributing to some confusion in the use of these methods among applied researchers. In this tutorial, we show the computational implementation of different causal inference estimators from a historical perspective where new estimators were developed to overcome the limitations of the previous estimators (i.e., nonparametric and parametric g-formula, inverse probability weighting, double-robust, and data-adaptive estimators). We illustrate the implementation of different methods using an empirical example from the Connors study based on intensive care medicine, and most importantly, we provide reproducible and commented code in Stata, R and Python for researchers to adapt in their own observational study. The code can be accessed at https://github.com/migariane/Tutorial_Computational_Causal_Inference_Estimators

KEYWORDS:

Causal Inference; Regression adjustment; G-methods; g-formula; Propensity score; Inverse probability weighting; Double-robust methods; Machine learning; Targeted maximum likelihood estimation

1 | INTRODUCTION

Often, questions that motivate studies in the health, social and behavioral sciences are causal. However, these research questions are usually answered using classical statistical methods, including multivariable outcome regression, to assess the relationship between an exposure and an outcome. For example, in a given population, what is the mortality risk difference amongst those patients who received surgery for colorectal cancer versus those who did not?¹ Often, the associations between a treatment and

an outcome assessed using classical methods cannot be interpreted as causal. Randomised clinical trials (RCT) are considered the gold standard for causal inference because randomisation ensures the outcome is independent of the treatment assignment. RCT are not always feasible (i.e., for ethical reasons or when the interest lies in the estimation of real-world effects) or may fail when randomisation does not work. Therefore, when causality cannot be guaranteed by design (i.e., in observational studies) or when the randomisation procedure fails, causal inference methods must be used. Based on the randomised experiment setting, Rubin introduced the potential outcomes framework: extending causal inference from randomised experiments to observational data.² Then, these methods were extended to observational settings with time-varying confounders.³

One of the aims when designing an observational study is to answer a scientific question that characterises the effect of a treatment on an outcome. This question is translated to an *estimand* (a target), which is the as yet unknown quantity we are interested in. Then, we use the *estimator* (a method), which is an algorithm that uses the values of the observations in the sample (in other words, a function of the random variables) to generate the *estimate* (the quantitative value generated for the estimator). The estimators are represented by algebraic equations that explicitly describe a function of the realised observations. Over the years, rapid ongoing advances in the field of causal inference have resulted in several algorithms that improve upon classical methods (i.e., outcome regression adjustment) to estimate the causal effect of a treatment on an outcome. These methods incorporate estimators using propensity scores, g-computation, or a combination of both (i.e., double-robust estimators). G-computation methods model the outcome mechanism, whereas propensity-score based methods model the treatment allocation, thus balancing the treatment groups in terms of the confounders. Often, double-robust estimators are preferred over classical single-robust regression approaches when the research question is causal.^{4,5}

In this tutorial we introduce the estimators mentioned above and show their computational implementation in regards to their chronological development (i.e., the methods were developed to address the limitations of the previous approaches). However, these methods are also introduced from a practical computational perspective, allowing readers to learn by using the replicable code. We use the Stata statistical software (*StataCorp. 2020. College Station, TX: StataCorp LLC, USA*), R statistical software (*R Development Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria*) and Python software (*Python Software Foundation (2020)*). All materials are available at a GitHub repository for reuse and replication of our examples at <https://github.com/migariane/TutorialComputationalCausalInferenceEstimators>. All examples in this paper use Stata code, but examples using R and Python are provided on the GitHub repository.

In the following sections, we will illustrate the computational implementation of different estimators that are computing the same estimand (i.e., the average treatment effect (ATE), a.k.a risk difference for a binary treatment and outcome). We will not focus on the assessment of heterogeneous treatment effects. In section 2, we briefly introduce the setting to estimate the ATE using Connors' study. In section 3, we introduce the g-computation based on the g-formula; and in section 4 we introduce the methods based on the Inverse Probability of Treatment Weights (IPTW). Afterwards, in section 5, we describe the computation of double-robust methods including the Augmented Inverse Probability of Treatment Weighting (AIPTW), and in section 6 we present Targeted Maximum Likelihood Estimation (TMLE). Finally, in section 7, we compare the performance of the various estimators using a single simulated data set.

2 | SETTING TO ESTIMATE THE ATE

To illustrate the implementation of the most common causal inference estimators we use an empirical data set from the prospective cohort study of Connors *et al* (1996).⁶ We use the data within the aforementioned GitHub repository; the original data are available at: <https://hbiostat.org/data/>, however some variables will need to be recoded (see Box 1). The study was set within intensive care units of five United States teaching hospitals between 1989 and 1994, and evaluated the effectiveness of right heart catheterisation (RHC) on short-term mortality (30 days) of 5,735 critically ill adult patients (2,184 treated and 3,551 untreated) receiving care for 1 of 9 prespecified disease categories.

A common estimand in causal inference is the ATE. The ATE is defined by an average of the difference of two random variables (i.e., the potential outcomes $Y(1)$ and $Y(0)$).^{3,7,8} For a binary treatment, each patient in the study has two potential outcomes

(i.e., $Y(a)$), where $Y(1)$ denotes the potential outcome if they received RHC, and $Y(0)$ denotes the potential outcome if they did not receive RHC (Appendix 1).^{2,3,7,8} More detailed introductions of the causal language used for the potential outcomes, and the assumptions needed to estimate causal effects using observational data, we refer readers to a recently published tutorial.⁹ In our illustration the outcome is short-term mortality (a binary variable) defined as mortality within 30 days after intensive care unit (ICU) admission; the main intervention was RHC. We define the vector (\mathbf{W}) to include the set of predefined confounders. To estimate the ATE (i.e., the standardised short-term risk difference of death for those patients who received RHC versus those who did not), we compute different estimators using the prospective cohort study of Connors *et al* (1996).⁶

Figure 1 is a directed acyclic graph representing the causal relationship between the vector of predefined confounders (i.e., \mathbf{W} : sex, age, education, race, and cancer), the intervention (A : receipt of RHC during their stay at the ICU), and the outcome (Y : vital status of the patient in an ICU at 30 days after admission). Note that throughout the article, we refer to A as the 'treatment', but it can be used interchangeably with the terms 'exposure' or 'intervention' depending on the context.

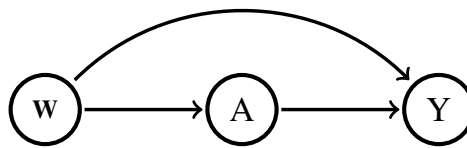


FIGURE 1 Y : outcome; A : treatment; \mathbf{W} : sufficient set of variables to control for confounding, as outlined in Connors *et al.* (1996)

To estimate the ATE of the intervention (RHC) on short-term mortality, we assume counterfactual consistency, conditional exchangeability, non-interference, and positivity (see Appendix 1). Furthermore, all the variables included in \mathbf{W} are confounders of the effect of A on Y ; there are no intermediate variables (i.e., mediators or colliders); and there is no residual confounding. Therefore, we assume, for illustrative purposes, that the set of covariates included in \mathbf{W} suffices, implying that the assumption of conditional mean independence holds (i.e., sufficient control for confounding).

3 | G-COMPUTATION METHODS BASED ON THE G-FORMULA

3.1 | Nonparametric g-formula

Regression adjustment is used to estimate the main effect of a risk factor on an outcome. It is one of the classical methods used in epidemiology to control for confounding. When a regression model does not include interactions the use of regression adjustment to control for confounding makes the assumption that the effect is constant across levels of confounders (\mathbf{W}) included in the model.⁵ Note that we focus on a binary outcome and treatment, thus "classical methods" will involve logistic regression adjustment to estimate the conditional odds ratio (OR) for the association between the treatment and the outcome. However, the OR is a non-collapsible measure of association, which means that the conditional OR cannot be used to estimate the marginal ATE.⁵ Furthermore, in observational and randomised studies, the estimate of the effect measure can be confounded given the different distribution of individual characteristics by treatment levels; thus, causal inference methods are needed to correct for the imbalance. For example, in the instance of differential age distributions between two treatment groups, classical methods will approach the problem using multivariable regression adjustment. However, causal inference methods use the g-formula, a generalisation of the classical standardisation procedure, which allows obtaining an unconfounded marginal estimation of the ATE. For a binary treatment, the g-formula is given by:³

$$\text{ATE} = \sum_{\mathbf{w}} [P(Y = 1 | A = 1, \mathbf{W} = \mathbf{w}) - P(Y = 1 | A = 0, \mathbf{W} = \mathbf{w})] P(\mathbf{W} = \mathbf{w}), \quad (1)$$

where

$$P(Y = 1 | A = a, \mathbf{W} = \mathbf{w}) = \frac{P(\mathbf{W} = \mathbf{w}, A = a, Y = 1)}{\sum_y P(\mathbf{W} = \mathbf{w}, A = a, Y = y)}$$

is the conditional probability of the outcome $Y = 1$, given the treatment $A = a$, and the set of confounders $\mathbf{W} = w$. Note, the implementation of the g-formula requires the use of the total law of probability.² In probability theory, the law of total probability is a fundamental rule relating marginal probabilities to conditional probabilities.

In the following set of boxes we show how to estimate the marginal causal effect (i.e., effect of RHC on short-term mortality) using the nonparametric and parametric g-formula in Stata. The Stata code, and the implementation of the same computational approach using R and Python, is provided in a GitHub repository: https://github.com/migariane/Tutorial_Computational_Causal_Inference_Estimators. For now, in the first 9 boxes, we use sex as the sole confounder, namely "c" (sex: 0 female, 1 male). It is an oversimplification for pedagogical purposes, which allows readers to readily appreciate the implementation of the computation of the parametric and nonparametric g-formula using G-computation methods. In boxes 10-13 we extend the methods by including multiple confounders. In contrast to classical methods (regression-based methods), the way we adjust for confounding based on the generalisation of standardisation (g-formula) is more coherent as we assume that the effect of RHC on short-term mortality can differ by sex. Classical methods, by including an interaction term in the model, can allow the effect to differ by sex but this hampers the interpretation of the main effect of RHC.⁵ It is a subtle difference but provides a richer adjustment for confounding.

In Box 1, we declare the global variables Y, A, C, and W to match the presented algebraic nomenclature (i.e., Y: outcome, A: treatment, C: one unique confounder, and W: a set of confounders). We use these global variables throughout the implementation of the different methods.

Box 1: Setting the data

```

1  clear
2  set more off
3  use "rhc.dta" // (From GitHub repo: https://github.com/migariane/
   TutorialComputationalCausalInferenceEstimators)
4  * Define the outcome (Y), exposure (A), confounder (C), and confounders (W)
5  lab def rhc 0 "No RHC" 1 "RHC", modify // Define labels for the rhc variable
6  lab val rhc rhc // Assign the label to the rhc variable
7  global Y death_d30 // Outcome: 30-day mortality
8  global A rhc // Treatment: Right Heart Catheterisation
9  global C sex // One unique confounder of the set of W
10 global W i.sex c.age c.edu i.race i.carcinoma // A set of five confounders

```

We first introduce, in Box 2, a naïve approach to estimate the ATE: we regress the outcome over the treatment (using a linear model) and adjust for the confounder (i.e., sex). In the naïve regression adjustment, the interpretation of the value for the regression coefficient of the treatment in the model is assumed to be constant for a fixed level of the confounder (i.e., sex). In Box 2 the result for the naïve regression adjustment shows strong evidence ($p < 0.001$) that the risk difference of death within 30 days is 7.35% higher amongst those with RHC (95% confidence interval [CI]: 4.84 – 9.86), conditional on sex. Results for this method, and for all of the methods in this tutorial, are shown in Table 1).

Box 2: Adjusted Regression

```

1  regress $Y $A $C // Risk difference = 7.35%; 95% CI: (4.84 - 9.86); p<0.001
2  // Bootstrap 95% CI
3  qui bootstrap, reps(1000) seed(1): regr $Y $A $C

```

For the first causal inference method we use, in Box 3, we compute the marginal probability of the confounder, save it, and generate two new variables named *sexf* for females and *sexm* for males (i.e., the marginal proportion of females was 44%, thus 56% are males, which shows unequal probability of being assigned the treatment by sex). We then compute, and save in a matrix, the expected conditional probabilities of the outcome by levels of the treatment and the confounder. We substitute the results of the matrix into the g-formula, given in equation 1, and compute the ATE.

Box 3: Nonparametric g-formula for the ATE

```

1  proportion $C // Marginal probability of C (sex)
2  matrix m=e(b)

```

```

3  gen sexf = m[1,1]
4  sum sexf
5  gen sexm = m[1,2]
6  sum sexm
7  ssc install sumup
8  sumup $Y, by($A $C) // Expected conditional probabilities of the outcome by levels of A and C
9  // from sumup command extract the conditional means by the given levels of A and C (i.e. zero and one)
10 matrix y00 = r(Stat1) // [6,1] matrix for E(Y|A=0,C=0)
11 matrix y01 = r(Stat2) // [6,1] matrix for E(Y|A=0,C=1)
12 matrix y10 = r(Stat3) // [6,1] matrix for E(Y|A=1,C=0)
13 matrix y11 = r(Stat4) // [6,1] matrix for E(Y|A=1,C=1)
14 // see "matrix list y00": position subscript [3,1] is the one of interest
15 // Applying the g-formula
16 gen EY1 = ((y11[3,1]-y01[3,1])*sexm // E(Y|A=1)
17 gen EY0 = ((y10[3,1]-y00[3,1])*sexf // E(Y|A=0)
18 qui: mean EY1 EY0
19 matrix ATE = r(table)
20 display "The ATE is: " ATE[1,1] + ATE[1,2] // Applying the g-formula
21 drop EY1 EY0
22 // Also one can try
23 gen ATE = ((y11[3,1]-y01[3,1])*sexm + ((y10[3,1]-y00[3,1])*sexf
24 qui sum ATE
25 drop ATE
26
27 // Check that Stata "teffects" command obtains the same estimate
28 teffects ra ($Y $C) ($A)
29 // The ATE from "teffects" implementation is: 7.37 (95% CI 4.83 - 9.91)

```

TABLE 1 Estimates of ATE from the different computational methods

Method	ATE	95% CI	ATE	Bootstrap 95% CI
One confounder				
Regression	7.35	4.84 – 9.86	7.35	4.78 – 9.92
NPG - 1C	n/a	n/a	7.37	4.72 – 9.89
NPG - FS	7.37	4.83 – 9.91	7.37	4.68 – 9.84
PG - 1C	7.37	4.83 – 9.91	7.37	4.68 – 9.84
Multiple confounders				
Regression	8.26	5.77 – 10.75	8.26	5.69 – 10.83
PG - FS	8.36	5.83 – 10.88	8.36	5.68 – 10.84
IPW - PS	8.33	5.81 – 10.85	8.33	5.65 – 10.81
MSM	8.33	5.77 – 10.89	8.33	5.74 – 10.62
IPW - RA	8.35	5.82 – 10.87	8.35	5.74 – 10.63
AIPW	8.35	5.82 – 10.87	8.39	5.78 – 10.66
TMLE	8.45	5.92 – 10.97	n/a	n/a
ELTMLE	8.35	5.82 – 10.87	n/a	n/a

1C = One confounder, **NPG** = Non-Parametric g-formula, **FS** = Fully saturated, **PG** = Parametric g-formula, **IPW** = Inverse Probability Weighting, **PS** = Propensity Score, **RA** = Regression Adjustment, **MSM** = Marginal Structural Model, **AIPW** = Augmented Inverse Probability Weighting, **TMLE** = Targeted Maximum Likelihood Estimation by hand, **ELTMLE** = Ensemble Learning Targeted Maximum Likelihood Estimation using Stata *eltmlle* package.

(n/a): 95% CI were not computed for the NPG-1C because the normal approximation was not appropriate. Bootstrap 95% CIs for the TMLE and ELTMLE estimators are not theoretically supported.

For the case of only one confounder, the results from the naïve regression adjustment and g-formula approaches are the same to one decimal place and nearly the same for the multivariable setting (i.e., multiple confounders (Table 1)). However, this is due to the use of a teaching data set with limited residual confounding (i.e., good balance of treatment across the levels of the confounders). Note that in real settings it will not be the case and more importantly the results and interpretation will differ

(i.e., conditional vs. marginal estimate). The naïve approach (Box 2) is a conditional estimation interpreted as the individual risk for treated vs. non-treated, holding the levels of the confounder constant. Whereas, the g-formula is a marginal contrast (Box 3) and therefore it must be interpreted at a population level.

The interpretation of the estimate from the naïve approach (Box 2) is difficult to conceptualise because we are holding the value of the confounder constant, and it requires the assumption that the effect of the treatment is the same for males and females (i.e., constant across the levels of the confounders).^{5,10} However, in observational studies, the ATE within strata of confounders may differ. Therefore, the g-formula has become a powerful alternative to the multivariable regression adjustment when controlling for confounding and evaluating the effects of treatments.³

3.1.1 | Statistical inference: The bootstrap

When constructing confidence intervals from an estimate obtained from a causal inference estimator, model-based standard errors (SE) are incorrect. This is because the model-based SE do not account for the different steps we need to take when we balance the confounders between treatment groups. We use the bootstrap procedure for inference implemented in Stata with the command `bootstrap`.¹¹ The bootstrap is a resampling method used to approximate the variance of the estimate (e.g., G-computation for the ATE).^{11,12} When estimating the variance using the bootstrap method, the observed data is thought of as representing the entire target population, and each draw (with replacement) from the data mimics the sampling variability. Under certain assumptions, this set of draws will return estimates of the sampling distribution that are equivalent to having actually repeated the sampling from the original target population.¹¹ Typically, for procedures that use parametric models, the bootstrap is a reliable estimator of the variance (i.e., the bootstrap uses the standard deviation of the bootstrap estimates of the ATE as a plug-in for the SE and the computation of the confidence intervals). However, note, it does not account for the bias engendered by model misspecification, so it only provides sampling variability for whatever the estimator is estimating.¹¹ The accuracy with which the bootstrap distribution estimates the sampling distribution depends on the number of observations in the original sample and the number of replications in the bootstrap.

To implement the bootstrap procedure in Stata we need to define a program that estimates the nonparametric g-formula and then samples (with replacement) the ATE to derive the confidence intervals for the ATE. In Box 4 we provide the code to compute the SE for the ATE using Stata.

Box 4: Bootstrap 95% Confidence Intervals (CI) for the ATE estimated using the Nonparametric g-formula

```

1  capture drop program ATE
2  program define ATE, rclass           // As before but now define a program to estimate the ATE
3      capture drop ATE
4      sumup $Y, by($A $C)
5      matrix y00 = r(Stat1)
6      matrix y01 = r(Stat2)
7      matrix y10 = r(Stat3)
8      matrix y11 = r(Stat4)
9      gen ATE = ((y11[3,1]-y01[3,1])*sexm + ((y10[3,1]-y00[3,1])*sexf
10     qui sum ATE
11     return scalar ate = `r(mean)'
12 end
13 qui bootstrap r(ate), reps(1000) seed(1): ATE // Bootstrap 1000 estimates of the ATE
14 estat boot, all
15 drop ATE

```

Based on the nonparametric g-formula, the estimate of the ATE was 7.37%. Using the command "estat boot, all", Stata gives three sets of CIs for the ATE; by default the bootstrap procedure will only provide the Normal-based CI. The first (N) is an approximation based on the Normal distribution (95% CI: 4.79 – 9.94). The naïve approach also uses the Normal approximation based on the central limit theorem giving asymptotic CIs. It is observed that the performances of the bootstrap CIs are better than the asymptotic confidence intervals in terms of the nominal coverage. Furthermore, the average length of bootstrap CIs is slightly larger than those of asymptotic CIs.^{13,14} The second (P) is based on the percentile of the bootstrap distribution (95% CI: 4.59 – 9.82), and the third (BC) is based on the bias-corrected (95% CI: 4.72 – 9.89) (Table 1). Note that the percentile interval is a simple "first-order" interval that is formed from quantiles of the bootstrap distribution. However, it is based only on bootstrap samples and does not adjust for skewness in the bootstrap distribution, unlike the bias-corrected. Thus, we will

report the BC 95% CI.¹⁴

For an alternative implementation of the nonparametric g-formula, we turn our attention to computing the ATE using a fully saturated regression model (still with only one confounder) using the full information of the sample (including the interactions between the treatment and the confounders). In Stata there are two different approaches that we illustrate in boxes 5 and 6 using the commands *predictnl* and *margins*.

To estimate the ATE using a fully saturated regression model we need to include all the possible interactions between the treatment and the different levels of the confounders (if categorical, otherwise with a continuous confounder) (Box 5). We do this by using the hashtag "#" symbol in Stata to include the interaction between A and C. The Stata prefix "ibn." specifies estimation of a categorical variable without the use of a base level (use with the *noconstant* option). The prefix "c.()" indicates that the confounder (i.e., sex) is to be used as a continuous variable (it does not matter for continuous or binary variables, but will matter for categorical variables). The *coeflegend* option asks Stata to provide the list of the labels of the variables in the analysis. The labels are then used for the *predictnl* command, which allows the computation of the nonparametric predictions based on the combination of the conditional probabilities from the regression coefficients. Finally, we average over the predictions to get the nonparametric estimate for the ATE. Note that the approach introduced in Box 5, in contrast to the approach presented in Box 3, is less computationally intensive in terms of time and code.

Box 5: Nonparametric g-formula using a fully saturated regression model in Stata (A)

```

1 regress $Y ibn.$A ibn.$A#$C , noconstant vce(robust) coeflegend
2 predictnl ATE = (_b[1.rhc] + _b[1.rhc#1.sex]*sex) - (_b[0bn.rhc] + _b[0bn.rhc#1.sex]*sex)
3 qui sum ATE
4 display "The ATE is: " `r(mean)'
5 drop ATE
6
7 // Bootstrap 95% CI
8 capture program drop ATE
9 program define ATE, rclass
10 capture drop ATE
11 regress $Y ibn.$A ibn.$A#$C , noconstant vce(robust) coeflegend
12 predictnl ATE = (_b[1.rhc] + _b[1.rhc#1.sex]*sex) - (_b[0bn.rhc] + _b[0bn.rhc#1.sex]*sex)
13 qui sum ATE
14 return scalar ate = `r(mean)'
15 end
16 qui bootstrap r(ate), reps(1000) seed(1): ATE
17 estat boot, all
18 drop ATE

```

A simpler option for the nonparametric g-formula would be to use the *margins* command to estimate the marginal probabilities using the option *vce(unconditional)* (Box 6). Then, the difference in marginal probabilities between the treated versus non-treated is implemented using the contrast option from the *margins* command. Note that here we obtain the same estimate of the ATE as 7.37% (95%: CI 4.83 – 9.91) but the appropriate 95% CI has been calculated using the Delta method (Table 1). The Delta method is a statistical approach to derive the SE of an asymptotically normally distributed estimator. It uses a first-order Taylor approximation, which is how we approximate the distribution of a function using a tangent line (i.e., the first derivative).¹⁵ Therefore, using the Delta method here we assume that the ATE estimate from the G-computation is normally distributed.¹⁶

Box 6: Nonparametric g-formula using a fully saturated regression model in Stata (B)

```

1 regress $Y ibn.$A ibn.$A#$C, noconstant vce(robust) // Fully saturated model specification
2 margins $A , vce(unconditional) // Marginal probability for A
3 margins r.$A , contrast(nowald) // Difference in marginal probability between treatment groups

```

3.2 | Parametric g-formula

In contrast to the nonparametric methods (i.e., probability distribution free or infinite dimensions), parametric methods are not affected by the curse of dimensionality.¹⁷ However, to compute the ATE parametrically we have to assume there is a particular probability distribution that fits the distribution of our data. To compute the ATE, we first regress (using a simple linear regression model) the outcome over the confounder(s) separately for each treatment group. We then predict the probability of treatment

and contrast the difference in the expected probabilities between the two treatment groups (note that every individual has two predicted probabilities corresponding to the two estimated potential outcomes). The algebraic form of the ATE under the G-computation is given by

$$\text{ATE} = \frac{1}{n} \sum_{i=1}^n (E(Y_i | A_i = 1, \mathbf{W}_i) - E(Y_i | A_i = 0, \mathbf{W}_i)) \quad (2)$$

In box 7, we provide the code to compute, by hand, the ATE based on the parametric g-formula for one confounder using parametric regression adjustment (based on formula 2).

Box 7: Parametric regression adjustment implementation of the g-formula

```

1 regress $Y $C if $A==1 // Expected probability amongst those with RHC
2 predict double y1hat
3 regress $Y $C if $A==0 // Expected probability amongst those without RHC
4 predict double y0hat
5 mean y1hat y0hat // Difference between the expected probabilities
6 lincom _b[y1hat] - _b[y0hat] // ATE and biased confidence interval

```

The risk of mortality amongst those with RHC is 7.37%, higher compared to those without RHC. Note that using a simple linear combination (i.e., `lincom` command in Stata) to compute a 95% CI for the linear contrast between the marginal potential outcomes results in a biased CI that does not account for the two-step procedure to get the marginal probabilities.

Box 8: Parametric regression adjustment using Stata's *teffects*

In Box 8, we confirm the result we obtained (by hand in Box 7) using the STATA's *teffects* command and including the *'ra'* option to perform the regression adjustment. Note the difference between the naïve and the *teffects* 95% CIs. The *teffects* uses the Delta method to correct for the uncertainty for each of the two models (i.e., $E(Y|A = 1, C)$ and $E(Y|A = 0, C)$), and provide appropriate statistical inference.

```

1 teffects ra ($Y $C) ($A) // Parametric g-formula implementation in Stata

```

With the *teffects* command in Stata the ATE is 7.37%, which is the same as we obtained by hand (Box 7). However, note that the 95% CI for the ATE using the command from Stata (*teffects*) is more conservative than using the naïve approach without accounting for the uncertainty of the two regression models to predict the marginal probabilities (i.e., 95% CI: 4.83 – 9.91 and 95% CI: 7.35 – 7.39, respectively for the *teffects* and the naïve approaches) (Table 1).

Box 9: Bootstrap for the parametric regression adjustment

Again, if we want to compute the 95% CI by hand using Stata we could use the bootstrap procedure (refer to Box 4 for an explanation).

```

1 capture program drop ATE
2 program define ATE, rclass // Define the program that will run the bootstrap
3 capture drop y1 // Drop any previously defined variable 'y1'
4 capture drop y0 // Drop any previously defined variable 'y0'
5 reg $Y $C if $A==1 // Regress the outcome amongst those with A=1
6 predict double y1, xb // Generate a variable (y1) to hold the predicted values
7 quiet sum y1 // Summarise the predicted value of the regression model
8 reg $Y $C if $A==0 // Regress the outcome amongst those with A=0
9 predict double y0, xb // Generate a variable (y0) to hold the predicted values
10 quiet sum y0 // Summarise the predicted value of the regression model
11 mean y1 y0 // Check the mean for y1 and for y0
12 lincom _b[y1]-_b[y0] // Calculate the difference in mean between y1 and y0
13 return scalar ate = `r(estimate)´ // Save the value of the difference in mean in a scalar called 'ate'
14 end
15 qui bootstrap r(ate), reps(1000) seed(1): ATE // Bootstrap 1000 times to generate the standard errors
16 estat bootstrap, all // Reports a table summarising the results of the bootstrap
17 drop ATE

```

After bootstrapping, the estimate of the ATE is 7.37% and the bias-corrected 95% CI: (4.68 – 9.84) (Table 1).

Box 10: Parametric multivariable regression adjustment implementation of the g-formula

As is often the case, there is almost always more than one confounder. The parametric computation of the g-formula can easily be extended to include more than one confounder: remember that **W** includes a set of five confounders.

```

1 regress $Y $W if $A==1 // Regression model with all confounders for those with RHC
2 predict double y1hat
3 regress $Y $W if $A==0 // Regression model with all confounders for those without RHC
4 predict double y0hat
5 mean y1hat y0hat // ATE is the difference in expectations
6 lincom _b[y1hat] - _b[y0hat] // We use lincom for the contrast
7 // but it gives a biased confidence interval for the ATE

```

The ATE of those with RHC (i.e. risk of mortality amongst those with RHC) is 8.36%, (95% CI: 8.25 – 8.47) higher compared to those without RHC in contrast to the naïve regression multivariable adjustment of 8.26% (Table 1). Note, the 95% CI provided by the *lincom* Stata command is biased as it is not accounting for the two-step estimation procedure to derive the ATE.

Box 11: Parametric multivariable regression adjustment using Stata's *teffects* command

In box 11, we use Stata's *teffects* command to confirm our results. We now include **W** instead of the single confounder **C**.

```

1 teffects ra ($Y $W) ($A)

```

We obtain the same results with Stata's *teffects* command as with our calculations by hand (ATE 8.36%; 95% CI: 5.83 – 10.88). However, note again the difference between the 95% CI estimated naïvely and using the *teffects* command (Table 1).

Box 12: Parametric multivariable regression adjustment using Stata's *margins* command

In box 12, we show another way of obtaining the ATE under the parametric g-formula approach using the Stata *margins* command after fitting a fully saturated regression model. First, we regress the dependent variable (Y) over the treatment (A), including the interaction of A with all of the other confounders (**W**). We do this using the same approach as in Box 5 (i.e., `ibn.$A#c.($W)`) to include the interaction between all levels of A and a vector of all of the other confounders included in the model. Then, the *margins* command calculates the predicted value of the expectation of the outcome given the treatment and the confounders, and reports the mean value of those predictions for each level of the treatment (A) (i.e., $E(Y|A=1, \mathbf{W})$ and $E(Y|A=0, \mathbf{W})$). Finally, to compute the ATE and provide corrected 95% CI based on the Delta method, we use the *contrast* option to compute the ATE. The ATE is the difference in the average 30-day mortality between those treated with RHC and those who were not (i.e., $E(Y|A=1, \mathbf{W}) - E(Y|A=0, \mathbf{W})$). Note the results are the same as before using the *teffects* command (i.e., ATE 8.36%; 95% CI: 5.83 – 10.88) (Table 1).

```

1 regress $Y ibn.$A ibn.$A#($W), noconstant vce(robust) // Fully saturated model
2 margins $A, vce(unconditional) // E(Y|A=1,W), E(Y|A=0,W) and Delta method for the standard errors (i.e., vce
3 // unconditional) and 95%CI
4 margins r.$A, contrast(nowald) // ATE and Delta method for the standard error and 95%CI

```

Box 13: Bootstrap for the multivariable parametric regression adjustment

Finally, in box 13 we show how to compute the bootstrap 95% CIs for the G-computation implementation of the g-formula by hand using regression adjustment in Stata.

```

1 capture program drop ATE
2 program define ATE, rclass
3   capture drop y1
4   capture drop y0
5   reg $Y $W if $A==1
6   predict double y1, xb
7   quiet sum y1
8   reg $Y $W if $A==0
9   predict double y0, xb
10  quiet sum y0
11  mean y1 y0
12  lincom _b[y1] - _b[y0]

```

```

13     return scalar ate = `r(estimate)`
14 end
15 qui bootstrap r(ate), reps(1000) seed(1): ATE dots
16 estat boot, all
17 drop ATE

```

After bootstrapping, the estimate of the ATE is 8.36%. The bootstrapped bias-corrected 95% CI is (5.68 – 10.84) (Table 1).

4 | INVERSE PROBABILITY OF TREATMENT WEIGHTING

4.1 | Inverse probability weighting based on the propensity score

In observational studies, some individuals will be more likely than others to be treated ($A=1$) due to their characteristics. Suppose some individuals who were treated were unlikely to be treated based on a specific set of features encapsulated in a particular vector of confounders (\mathbf{W}). To balance the differences in characteristics between treatment groups, we re-weight the outcome variable of these individuals by the inverse of their probability of the treatment (A) actually received (i.e., propensity score). Originally, the weights were motivated from the classical Horvitz and Thompson survey estimator used to re-weight the outcome variable by the inverse probability that it is observed, thus accounting for the sampling process.¹⁸ The result of this weighting procedure is that, among the treated we up-weight those who had a low probability of being treated, and among the untreated we up-weight those who were unlikely to be untreated; that is, the individuals underrepresented in their treatment group. As a consequence, the weighted set of data is unchanged apart from A and \mathbf{W} are now conditionally independent. Therefore, a comparison of $Y_w(1)$ to $Y_w(0)$ gives a marginal causal effect under the three identification assumptions (Appendix 1) whilst also assuming the propensity score model is correctly specified. The inverse probability of treatment weighting (IPTW), and the g-formula when targeting the same estimand (i.e., the ATE), are equivalent in the nonparametric setting.^{3,19} In appendix 2 we provide a proof of the equivalence between IPTW and G-computation procedures using the law of total expectation.

Departing from the identification assumptions of the ATE for the regression adjustment G-computation estimand ($ATE = E_w(E(Y|A=1, \mathbf{W}) - E_w(Y|A=0, \mathbf{W}))$), we can rewrite the same estimand as a function of the distribution of A given \mathbf{W} (i.e., $P(A = 1|\mathbf{W})$ a.k.a propensity score or treatment mechanism).

Therefore, the estimator is given by

$$ATE = \frac{1}{n} \sum_{i=1}^n \left(\frac{A_i}{P(A_i = 1 | \mathbf{W}_i)} - \frac{1 - A_i}{(1 - P(A_i = 1 | \mathbf{W}_i))} \right) Y_i. \quad (3)$$

There is a modified version (i.e., Hájek type)²⁰ of the IPTW estimator (equation 3) consisting of stabilised weights, which is more commonly used in practice when treatment and exposure vary over time (i.e., time dependent confounding). However, stabilised weights should have a mean of 1, but some values could be higher (i.e., large weights). The stabilised version of the IPTW estimator is given by

$$ATE = \frac{\sum \left(\frac{AY}{P(A=1|\mathbf{W})} \right)}{\sum \left(\frac{A}{P(A=1|\mathbf{W})} \right)} - \frac{\sum \left(\frac{(1-A)Y}{1-P(A=1|\mathbf{W})} \right)}{\sum \left(\frac{(1-A)}{1-P(A=1|\mathbf{W})} \right)}. \quad (4)$$

In box 14, we show how to compute the IPTW by hand in two steps:

- First, the propensity score model is fitted in rows 1-4 (i.e., a logistic regression model for a binary treatment)
- Then the sampling weights are generated based on the inverse probability of treatment actually received. Note, the weights are just the implementation of the classical Horvitz-Thompson survey estimator,¹⁸ (see rows 3 and 4) also known as unstabilised weights (rows 5-9).

When there are near violations of the positivity assumption, the unstabilised weights can have large values, forcing the variance to increase and exacerbate the uncertainty of the ATE estimation. Therefore, it is advisable to explore the distribution of the

weights to evaluate the extent to which they balance the distribution of confounders across the levels of the treatment (i.e., equally distributed). It is common to provide a table with the unweighted and weighted differences of the standardised means of the confounders by the levels of the treatment. Also, it is common to visualise an overlap of the propensity scores by the level of the treatment to identify and visualise positivity or near positivity violations and to explore the descriptive distribution of the weights (i.e., mean, minimum and maximum values). Lastly, while we are showing the use of logistic regression, the propensity score model may alternatively be estimated using nonparametric approaches (e.g., the *twang*²¹ R package uses generalised boosted regression modelling).

Box 14: Computation of the IPTW estimator for the ATE

```

1  logit $A $W, vce(robust) nolog           // Propensity score model for the treatment
2  predict double ps                       // Propensity score prediction
3  generate double ipw1 = ($A==1)/ps       // Sampling weights for the treated group
4  generate double ipw0 = ($A==0)/(1-ps)   // Sampling weights for the non-treated group
5  mean $Y [pw=ipw1], coeflegend          // Weighted outcome probability among treated
6  scalar Y1 = _b[death_d30]
7  mean $Y [pw=ipw0], coeflegend          // Weighted outcome probability among non treated
8  scalar Y0 = _b[death_d30]
9  display "ATE =" Y1 - Y0

```

The risk difference between those with RHC and those without is 8.33%. Re-weighting the individuals generates a pseudo-population (weighted population) from which the data generation does not follow a theoretical distribution and individuals are no longer independent. Therefore the 95% CI is estimated using the bootstrap procedure in Box 15.

Box 15: Bootstrap computation for the IPTW estimator

As before, we can obtain confidence intervals using the bootstrap procedure.

```

1  capture program drop ATE
2  program define ATE, rclass
3      capture drop y1
4      capture drop y0
5      capture drop ipw0
6      capture drop ipw1
7      capture drop ps
8      logit $A $W, vce(robust) nolog // propensity score model for the exposure
9      predict double ps // propensity score predictions
10     generate double ipw1 = ($A==1)/ps // Sampling weights for the treated group
11     generate double ipw0 = ($A==0)/(1-ps) // Sampling weights for the non-treated group
12     regress $Y [pw=ipw1] // Weighted outcome probability among treated
13     matrix y1 = e(b)
14     gen double y1 = y1[1,1]
15     regress $Y [pw=ipw0] // Weighted outcome probability among non-treated
16     matrix y0 = e(b)
17     gen double y0 = y0[1,1]
18     mean y1 y0
19     lincom _b[y1]-_b[y0]
20     return scalar ate = 'r(estimate)'
21 end
22 qui bootstrap r(ate), reps(1000) seed(1): ATE
23 estat boot, all
24 drop ATE

```

After bootstrapping, the estimate of the ATE is 8.33%. The bootstrapped bias-corrected confidence interval is: (5.65 – 10.81) (Table 1).

Box 16: Computation of the IPTW estimator for the ATE using Stata's *teffects* command

We now confirm this result in box 16 using Stata's *teffects* command. Note that the Horvitz-Thompson estimator is implemented using the *ipw* option. We obtain the same point estimate for the ATE and slightly different, but consistent, 95% CI based on the robust SE derived from the functional Delta method (i.e., ATE 8.33%; 95% CI: 5.81 – 10.85) (Table 1).

```
teffects ipw ($Y) ($A $W, logit), nolog vsquish
```

Box 17: Assessing IPTW balance

In box 17, we show how to explore the balance of the confounders after weighting the contributions of individuals using IPTW

(i.e., that the distribution of the confounders are balanced between those with RHC and those without). When applying weights, we must be careful as we are assuming that the treatment has been balanced across the levels of the confounders. In Stata, we use the *tebalance* option after using the *teffects* command but the balance can be assessed by hand as well.

```

1  qui teffects ipw ($Y) ($A $W)
2  tebalance summarize // Stata's tebalance
3
4  // tebalance by hand (sex)
5  egen sexst = std(sex) // Standardisation
6  logistic $A $W // Propensity score
7  predict double ps
8  gen ipw = .
9  replace ipw=($A==1)/ps if $A==1
10 replace ipw=($A==0)/(1-ps) if $A==0
11 regress sexst $A // Raw difference
12 regress sexst $A [pw=ipw] // standardised difference

```

After weighting, the two treatment groups appear to be well-balanced. Prior to weighting, there was some imbalance (absolute values of the standardised differences close to, or beyond, 0.10) on sex, education level and presence/extent of cancer between treatment groups.²² A variance ratio (i.e., the ratio of the standardised distribution of the confounders by the levels of the treatment) equal to 1 before and after weighting informs us that the distribution of the confounders across the levels of the treatments is the same (i.e., perfectly balanced). Note, the weighted variance ratio for the continuous variable age is 0.79, which is slightly further from 1 than the variance ratio for the original (unweighted) sample (i.e., 0.82); this slight change is possibly because the weighted mean for age might have greater sampling variance than the unweighted mean (Table 2).²³

TABLE 2 Distribution of the treatment before and after applying weights

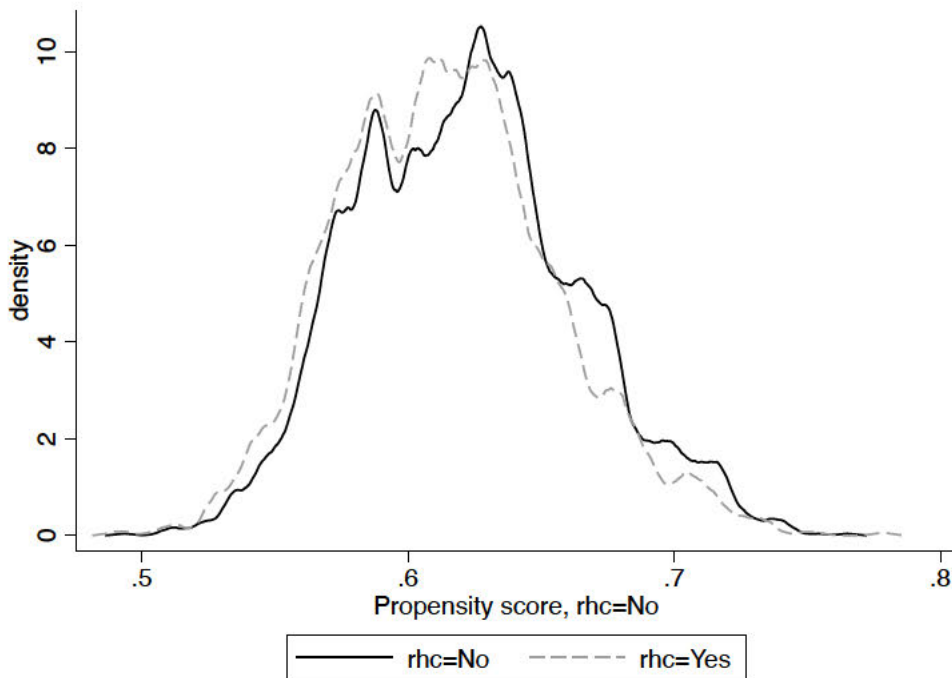
Confounder	Standardised differences		Variance ratio	
	Raw	Weighted	Raw	Weighted
Sex	0.093	0.000	0.977	1.000
Age	-0.061	-0.004	0.817	0.791
Education	0.091	-0.002	1.015	1.027
Race - Black	-0.031	0.002	0.944	1.003
Race - Other	0.020	0.001	1.078	1.004
Cancer - Metastatic	-0.069	-0.000	0.780	1.000
Cancer - Localised	-0.072	0.000	0.879	0.999

Reference groups: race - white, cancer - none

There is no definitive value at which the treatment is considered unbalanced; however, as a guideline, a variance ratio less than 0.5 indicates that the data is not balanced and the potential for the positivity violation must be explored (i.e., when $P(A = a | C = c)$ is near to zero or one). An additional strategy is to check the distribution of the weights: if there are very large weights this indicates the violation of the positivity assumption but, also, it can be due to parametric modelling misspecification. Again there is no clear consensus but, when there are very large weights, researchers often set the weights to a less extreme value. This is done by trimming or removing the data at the extremes of the distribution of the weights (e.g., the 5th and 95th percentiles).²⁴ Trimming the weights reduces variance (i.e., omitting the largest weights and making the positivity assumption more plausible), but at the expense of introducing bias.²⁵ However, another alternative without dropping observations is truncation, whereby all the values of the weights, larger than a user-specified maximum value or percentile (e.g., 1st and 99th or 5th and 95th), are replaced by that threshold value.^{25,26} In extreme cases, when the weights are extremely large, changing the estimand could be another solution (e.g., estimating the ATE in a subset of the sample, just like among only those

treated, representing the average treatment effect among the treated -ATT-).

FIGURE 2 Propensity score overlap by treatment status



Box 18: Assessing IPTW overlap by hand

It is also important to check the overlap of the propensity scores of the two treatment groups. The "overlap" gives a visual identification regarding the strength of confounding and whether it is acceptable. In box 18 we show how to visualise the "overlap" using a kernel density estimate of the treatment assignment by the levels of the treatment. Figure 2 shows there is a suitable amount of overlap.

```

1  sort $A
2  by $A: summarize ps
3  kdensity ps if $A==1, generate(x1pointsa d1A) nograph n(10000) // Nonparametric kernel density estimate of
   the distribution of the propensity score among treated individuals
4  kdensity ps if $A==0, generate(x0pointsa d0A) nograph n(10000) // Nonparametric kernel density estimate of
   the distribution of the propensity score among non-treated individuals
5  label variable d1A "density for RHC=1"
6  label variable d0A "density for RHC=0"
7  twoway (line d0A x0pointsa , yaxis(1))(line d1A x1pointsa , yaxis(2))

```

Box 19: Assessing overlap using teffects overlap

The overlap plots can be obtained with Stata's *overlap* command after calling the *teffects* command. We are using data where there is a good balance and overlap but in real-world observational data the balance and overlap, before using any weighting procedure, are not likely to be well balanced.

```

1  qui: teffects ipw ($Y) ($A $W, logit), nolog vsquish
2  teffects overlap

```

4.2 | Marginal structural model with stabilised weights

We now introduce the marginal structural model (MSM) as a transition to the double-robust methods.²⁷ A MSM is a marginal mean model. A popular method for estimating the parameters of the MSM is weighted regression modelling that estimates the marginal distributions of the counterfactuals.^{27,28} In the MSM, the coefficient for the treatment is the estimate of treatment effect, usually the ATE. The MSM uses an updated version of the Horvitz-Thompson weights, commonly used in sampling theory.¹⁸ The weights represent the inverse of the probability of treatment (a.k.a propensity score). In box 20 we show how to compute a MSM:

- First, in rows 1 to 18, we compute the propensity score and the weights.
- In row 20 we fit the MSM using the unstabilised weight, and in row 21 using the stabilised version. The approach to compute the weights is equivalent to the one presented in Box 14 where re-weighting the individuals generated a pseudo-population and classical statistical inference does not hold.²⁹ Thus, for statistical inference we use the *vce(robust)* option, which implements the Delta method, to estimate the appropriate SE for the ATE.¹⁷ However, using the bootstrap procedure is also a valid option.
- Finally, in rows 21 to 45 we show how to implement the bootstrap procedure to compute the 95% CI.

The ATE derived from the MSM is 8.33%, and the 95% CI using the Delta method:(5.77 – 10.89) and (5.84 – 10.85) using the bootstrap procedure (Table 1).

Box 20: Computation of the IPTW estimator for the ATE using a MSM

```

1 // baseline treatment probabilities
2 logit $A, vce(robust) nolog
3 predict double nps, pr
4 // propensity score model
5 logit $A $W, vce(robust) nolog
6 predict double dps, pr
7 // Unstabilised weight
8 gen ipw = .
9 replace ipw=($A==1)/dps if $A==1
10 replace ipw=($A==0)/(1-dps) if $A==0
11 sum ipw
12
13 // Stabilised weight
14 gen sws = .
15 replace sws = nps/dps if $A==1
16 replace sws = (1-nps)/(1-dps) if $A==0
17 sum sws
18
19 // MSM
20 reg $Y $A [pw=ipw], vce(robust) // MSM unstabilised weight
21 reg $Y $A [pw=sws], vce(robust) // MSM stabilised weight
22
23 // Bootstrap the 95% confidence intervals
24 capture program drop ATE
25 program define ATE, rclass
26 capture drop nps
27 capture drop dps
28 capture drop sws
29 // Baseline treatment probabilities
30 logit $A, vce(robust) nolog
31 predict double nps, pr
32 // propensity score model
33 logit $A $W, vce(robust) nolog
34 predict double dps, pr
35 // Stabilized weight
36 gen sws = .
37 replace sws = nps/dps if $A==1
38 replace sws = (1-nps)/(1-dps) if $A==0
39 sum sws
40 // MSM
41 reg $Y $A [pw=sws], vce(robust)
42 return scalar ate = e(b)[1,1]
43 end
44 qui bootstrap r(ate), reps(1000) seed(1): ATE

```

```
45 estat boot, all
```

5 | DOUBLE-ROBUST METHODS

5.1 | Inverse probability weighting plus regression adjustment

The IPTW-RA is an estimator using a G-computation regression adjustment (RA) that incorporates the estimated stabilised IPTW. It has been shown that the IPTW-RA helps to correct the estimator when the regression function is misspecified, provided that the propensity score model for the treatment is correctly specified. When the regression function is correctly specified, the weights do not affect the consistency of the estimator even if the model from which they are derived is misspecified.³⁰ Note that combining both, the IPTW and the RA approaches, the IPTW-RA estimator has the special property that it is consistent as long as at least one of the two models (i.e. IPTW and RA) is correctly specified, it is why estimators that combine both modelling approaches are named double-robust.³¹ When one uses G-computation methods only, they rely on extrapolation of the treatment effects when there are identifiability issues due to data sparsity and near-positivity violations. Adding the IPTW to the regression adjustment allows evaluation of the balance of the treatment and of possible positivity violations, increasing the researcher's awareness of the limitations of causal inference modelling. It is encouraged, when possible, to explore the implementation of the nonparametric g-formula (using the important confounders) and identify potential problems with the data relating to the curse of dimensionality from finite samples (i.e., zero empty cells for a given combination of conditional probabilities from the different variables included in analysis needed to implement the g-formula).

Although IPTW with regression adjustment (IPTW-RA) is usually more efficient than IPTW, it also relies on different parametric modelling assumptions: (i) a parametric G-computation regression adjustment model, and (ii) a model for the propensity score of binary treatments. The G-computation weighted model uses the weights calculated from the predictions of the propensity score logistic model. An estimated propensity score that is close to 0 or 1 is problematic, since it implies that some individuals will receive a very large weight leading to imprecise and unstable estimates (i.e., near positivity assumption violation). Therefore, the use of stabilised weights is suggested (see code from Box 20), and the bootstrap for statistical inference.

Box 21: Computation of the IPTW-RA estimator for the ATE and bootstrap for statistical inference

```
1 capture program drop ATE
2 program define ATE, rclass
3     capture drop y1
4     capture drop y0
5     // Weighted (stabilised weights) regression adjustment among the treated
6     reg $Y $W if $A==1 [pw=sws]
7     predict double y1, xb
8     quiet sum y1
9     return scalar y1=`r(mean)'  
10    // Weighted (stabilised weights) regression adjustment among the non-treated
11    reg $Y $W if $A==0 [pw=sws]
12    predict double y0, xb
13    quiet sum y0
14    return scalar y0=`r(mean)'  
15    mean y1 y0
16    // ATE
17    lincom _b[y1]-_b[y0]
18    return scalar ate =`r(estimate)'  
19 end
20 qui bootstrap r(ate), reps(1000) seed(1): ATE // Bootstrapping for statistical inference
21 estat boot, all
```

After bootstrapping, the estimate of the IPTW-RA ATE is 8.35%, bias-corrected 95% CI (5.83 - 10.87). The results are very similar to those obtained using the Stata's *teffects* command with the option *ipwra* presented in box 22 (i.e., ATE: 8.35% and 95% CI: 5.82 – 10.87) (Table 1).

Box 22: Computation of the IPTW-RA estimator for the ATE using Stata's *teffects*

Note that using *ipwra* we specify two models (i.e., the model for the outcome and the model for the treatment).

```
1 teffects ipwra ($Y $W) ($A $W), nolog vsquish
```


5.2 | Augmented inverse probability of treatment weighting

The AIPTW estimator is an improved IPTW estimator that includes an augmentation term, which corrects the estimator when the treatment model is misspecified. When the treatment model is correctly specified, the augmentation term vanishes as the sample size becomes large. Thus, the AIPTW estimator is more efficient than the IPTW. However, like the IPTW, the AIPTW does not perform well when the predicted treatment probabilities are too close to zero or one (i.e., near positivity violations). Under correct modelling specification, the augmentation term has expectation zero and includes the expectation of the propensity score and the regression adjustment outcome. Thus, the AIPTW combines two parametric models (i.e., a model for the outcome and a model for the treatment).^{32,33} The AIPTW estimator produces a consistent estimate of the ATE if either of the two models has been correctly specified.^{30,33}

Focusing on the IPTW estimator for the ATE in equation 3, let $\hat{\mu}_a$ be the expectation of the ATE using IPTW, more formally this is

$$\hat{\mu}_a = E \left(\frac{I(A = a)}{g(A | \mathbf{W})} Y \right),$$

where I is the indicator function and $g(\cdot)$ refers to the treatment mechanism.

We can rewrite the equation in the form of an estimating equation (see glossary) as,

$$\frac{1}{n} \sum_{i=1}^n \left(\frac{I(A_i = a) Y_i}{g(A_i | \mathbf{W}_i)} - \mu_a \right) = 0,$$

As long as the estimating function has mean zero then $\hat{\mu}$ is a consistent estimator of μ , where $\mu_a = E(Y | A = a, \mathbf{W})$. If we augment the estimating function using a mean-zero term,

$$\frac{I(A = a) - g(A = a | \mathbf{W})}{g(A = a | \mathbf{W})},$$

including the propensity score expectation ($g(A = a | \mathbf{W})$), we have integrated both the estimation of the treatment mechanism and the mean outcome ($E(Y | A = a, \mathbf{W})$), then

$$E \left(\frac{I(A = a) Y}{g(A = a | \mathbf{W})} - \left(\frac{I(A = a) - g(A = a | \mathbf{W})}{g(A = a | \mathbf{W})} \right) E(Y | A = a, \mathbf{W}) \right) - \mu_a = 0.$$

Rearranging the equation we can see that the AIPTW estimator is a combination of inverse weighting and outcome regression defined for a binary treatment as

$$\underbrace{\frac{1}{n} \sum_{i=1}^n (E(Y_i | A_i = 1, \mathbf{W}_i) - E(Y_i | A_i = 0, \mathbf{W}_i))}_{\text{G-computation-Regression-Adjustment}} + \underbrace{\frac{1}{n} \sum_{i=1}^n \left(\frac{A_i [Y_i - E(Y_i | A_i = 1, \mathbf{W}_i)]}{g(A_i = 1 | \mathbf{W}_i)} - \frac{(1 - A_i) [Y_i - E(Y_i | A_i = 1, \mathbf{W}_i)]}{g(A_i = 0 | \mathbf{W}_i)} \right)}_{\text{Zero-expectation}}, \quad (5)$$

where the ATE from the AIPTW estimator is defined as

$$\text{AIPTW-ATE} = \mu_1 - \mu_0$$

$$\mu_1 = \frac{1}{n} \sum_{i=1}^n \left(E(Y_i | A_i = 1, \mathbf{W}_i) + \frac{A_i[Y_i - E(Y_i | A_i = 1, \mathbf{W}_i)]}{g(A_i = 1 | \mathbf{W}_i)} \right), \quad (6)$$

$$\mu_0 = \frac{1}{n} \sum_{i=1}^n \left(E(Y_i | A_i = 0, \mathbf{W}_i) + \frac{(1 - A_i)[Y_i - E(Y_i | A_i = 1, \mathbf{W}_i)]}{g(A_i = 0 | \mathbf{W}_i)} \right).$$

The second term in equation 5 can be interpreted as playing the role of two nuisance parameters of the AIPTW estimating function. The nuisance parameters are represented as a weighted sum of the residuals for the conditional mean of the outcome.¹⁶ Equation 5 shows that the AIPTW estimator equals the g-formula estimator if the outcome model is correctly specified irrespective of the treatment model. Likewise, the point estimate will be equal to the IPTW estimator if the treatment model is correctly specified, irrespective of the outcome model.^{33,31}

In Box 23 we show how to compute the AIPTW estimator for the ATE using Stata:

- Step 1: First we predict the mean outcome by treatment status using G-computation regression adjustment (rows 1-7).
- Step 2: Then we compute the inverse of treatment weights (rows 9-16).
- Step 3: Using equation 5 we compute the ATE (rows 18-26).
- Step 4: Finally, we compute 95% CI using the bootstrap procedure in Stata (rows 28-52).

Box 23: Computation of the AIPTW estimator for the ATE and bootstrap for statistical inference

```

1 // Step (i) prediction model for the outcome using G-computation regression adjustment
2 qui glm $Y $A $W, fam(bin)
3 predict double QAW, mu
4 qui glm $Y $W if $A==1, fam(bin)
5 predict double Q1W, mu
6 qui glm $Y $W if $A==0, fam(bin)
7 predict double QOW, mu
8
9 // Step (ii): prediction model for the treatment
10 qui logit $A $W
11 predict double dps, pr
12 qui logit $A
13 predict double nps, pr
14 gen sws = .
15 replace sws = nps/dps if $A==1
16 replace sws = (1-nps)/(1-dps) if $A==0
17
18 // Step (iii): Estimation equation based on analytical formula 5
19 gen double y1 = (sws*($Y-QAW) + (Q1W))
20 quiet sum y1
21 gen double y0 = (sws*($Y-QAW) + (QOW))
22 quiet sum y0
23 mean y1 y0
24 lincom _b[y1] - _b[y0]
25
26 // step (iv) Bootstrap confidence intervals
27 capture program drop ATE
28 program define ATE, rclass
29 capture drop y1
30 capture drop y0
31 capture drop Q*
32 qui glm $Y $A $W, fam(bin)
33 predict double QAW, mu
34 qui glm $Y $W if $A==1, fam(bin)
35 predict double Q1W, mu
36 qui glm $Y $W if $A==0, fam(bin)
37 predict double QOW, mu
38 gen double y1 = (sws*($Y-QAW) + (Q1W))
39 quiet sum y1

```

```

40 return scalar y1=`r(mean)'
41 gen double y0 = (sws*($Y-QAW) + (QOW))
42 quiet sum y0
43 return scalar y0=`r(mean)'
44 mean y1 y0
45 lincom _b[y1] - _b[y0]
46 return scalar ate =`r(estimate)'
47 end
48 qui bootstrap r(ate), reps(1000) seed(1): ATE
49 estat boot, all
50 drop ATE

```

After bootstrapping the ATE is 8.39% and the bias-corrected 95% CI confidence intervals are: (5.87 – 10.89) (Table 1). In Box 24, we show the same results using Stata’s *teffects* command with the *aipw* option. Note that we have to specify the model for the treatment and the model for the outcome.

Box 24: Computation of the AIPTW estimator for the ATE using Stata’s *teffects*

```

1 teffects aipw ($Y $W) ($A $W, logit), nolog vsquish

```

The ATE is 8.35%, 95% CI: 5.82 – 10.87 (Table 1).

6 | DATA-ADAPTIVE ESTIMATION: ENSEMBLE LEARNING TARGETED MAXIMUM LIKELIHOOD ESTIMATION

Targeted Maximum Likelihood Estimation (TMLE) is a plug-in, semi-parametric, double-robust method that reduces the bias of an initial estimate by allowing for flexible estimation using nonparametric data-adaptive machine-learning methods to target an estimate closer to the true model specification.⁴ There are several TMLE tutorials published elsewhere,^{34,35,36,37,38} but here we provide a brief introduction. To learn more about the algorithm, readers can refer to van der Laan and Rose’s TMLE book⁴, and from a practical perspective to a step-by-step tutorial illustrated in a realistic cancer epidemiology scenario published by Statistics in Medicine in 2018.³⁸ The advantages of TMLE have been demonstrated in both simulation studies and applied analyses.^{4,39} Evidence shows that TMLE can provide the least biased ATE estimate compared with other double-robust estimators such as the IPTW-RA and AIPTW. In particular, while TMLE and AIPW estimators are asymptotically equal, TMLE enjoys better finite sample properties. Separately, TMLE is often implemented with ensemble machine learning, which can relax model specification constraints.^{4,39}

In Box 25, we provide the computational implementation of TMLE by hand (without data-adaptive estimation) to guide and interpret the different steps involved in the TMLE. A description of the theory behind these steps can be found elsewhere.³⁸

- Step 1: We estimate the expected outcome given treatment and confounders ($E(Y|A, \mathbf{W})$): this is called the plug-in initial estimate of the estimator obtained via G-computation, namely Q_0 (Box 25: rows 1-15).
- Step 2: We define the expected treatment given the confounders as we did previously for the estimation of the propensity score in box 14, namely g_0 . Steps 1 and 2 are similar to the double-robust methods of AIPTW; however, we now come to the advantage of TMLE (Box 25: rows 17-21).
- Step 3: We regress the predicted treatment values and predicted outcome introduced in the model as an offset on the observed outcome. The parameter estimates (epsilon) for that regression are used to correct the initial estimations of Q_0 (Box 25: rows 24-27). In other words, we reduce the residual bias and optimised the bias-variance trade-off for the estimate of the ATE so that we can obtain valid statistical inference. Note that the TMLE framework adds the possibility to estimate the Q_0 and g_0 models using data adaptive machine learning algorithms and selecting the best model or an ensemble of the models.⁴ It has been shown that using machine learning algorithms reduces misspecification bias.⁴⁰ Note, in box 25, the residual bias is reduced by solving an equation that calculates how much to update, or fluctuate, our initial outcome estimates

$$E_*[Y|A, \mathbf{W}] = \text{logit}(E[Y|A, \mathbf{W}]) + \epsilon H(A, \mathbf{W}),$$

where $E_*(Y|A, \mathbf{W})$ represents the updated initial expectation of the outcome (Y) given the treatment status (A) and the set of confounders (\mathbf{W}). To solve this equation, we fit an intercept-free logistic regression (using H as the only predictor of the observed outcome) and the initially predicted outcome (under the observed treatment) as an offset (step 3 rows 24-27) as a targeting step aimed to reduce bias. Fitting the logistic regression, using maximum likelihood procedures, TMLE yields many useful statistical properties, such as: (1) the final estimate is consistent as long as either the outcome or treatment model are estimated correctly (consistently); (2) if both of these models are estimated consistently, the final estimate achieves "semi-parametric efficiency" i.e., variance reduction as the sample size approaches infinity. Also the AIPTW is semi-parametric efficient.

- Step 4: We added the coefficient ϵ of the clever covariate H in the previous step to the expected outcome for all observations from the model fitted in Step 1 using (step 4: rows 29-31), updating the Q_0 model predictions to Q_1 .

$$Q_1(A = 1, \mathbf{W}) = E_*[Y|A = 1, \mathbf{W}] = \text{expit}(\text{logit}(E[Y|A = 1, \mathbf{W}]) + \epsilon H(1, \mathbf{W})), \text{ and}$$

$$Q_1(A = 0, \mathbf{W}) = E_*[Y|A = 0, \mathbf{W}] = \text{expit}(\text{logit}(E[Y|A = 0, \mathbf{W}]) + \epsilon H(0, \mathbf{W})).$$

- Step 5: We compute the ATE as the difference between expectations of the updated Q_1 predictions in the previous step (i.e., $E[Y|A = 1, \mathbf{W}] - E[Y|A = 0, \mathbf{W}]$) (Box 25: rows 33-36). It is worth noting that Steps 3 and 4, which are improvements to AIPTW and IPTW-RA estimators, are the very concepts that make TMLE more robust against near positivity violations and force the estimator to respect the boundaries of the limits of the parameter space (i.e., the probabilities stay between 0 and 1). For example, to estimate the ATE using the AIPTW estimator the researcher sets the estimation equation equal to zero. However, solving the estimating equation when there are near violations of the positivity assumption can cause the estimator to fall outside the boundaries of the parameter space (i.e., 0 and 1). Using TMLE, the ATE estimate is 8.34%, 95% CI: 5.82 – 10.98 (Table 1), which is consistent with all the previous estimates using different estimators.
- Finally in step 6, we provide statistical inference using the functional Delta method and the Influence Function (IF).^{4,41,16,42} In the next section we briefly introduce these concepts.

Box 25: Computational implementation of TMLE by hand

```

1  * Step 1: prediction model for the outcome Q0 (G-computation)
2  glm $Y $A $W, fam(binomial)
3  predict double QAW_0, mu
4  gen aa=$A
5  replace $A = 0
6  predict double QOW_0, mu
7  replace $A = 1
8  predict double Q1W_0, mu
9  replace $A = aa
10 drop aa
11
12 // Q to logit scale
13 gen logQAW = log(QAW / (1 - QAW))
14 gen logQ1W = log(Q1W / (1 - Q1W))
15 gen logQOW = log(QOW / (1 - QOW))
16
17 * Step 2: prediction model for the treatment g0 (IPTW)
18 glm $A $W, fam(binomial)
19 predict gw, mu
20 gen double H1W = $A / gw
21 gen double H0W = (1 - $A) / (1 - gw)
22
23 * Step 3: Computing the clever covariate H(A,W) and estimating the parameter (epsilon) (MLE)
24 glm $Y H1W H0W, fam(binomial) offset(logQAW) noconstant
25 mat a = e(b)
26 gen eps1 = a[1,1]
27 gen eps2 = a[1,2]
28
29 * Step 4: update from QOW and Q1W to QOW_1 and Q1W_1
30 gen double Q1W_1 = exp(eps1 / gw + logQ1W) / (1 + exp(eps1 / gw + logQ1W))
31 gen double QOW_1 = exp(eps2 / (1 - gw) + logQOW) / (1 + exp(eps2 / (1 - gw) + logQOW))
32
33 * Step 5: Targeted estimate of the ATE

```

```

34 gen ATE = (Q1W_1 - Q0W_1)
35 summ ATE
36 global ATE = `r(mean)'
37
38 * Step 6: Statistical inference (functional Delta method): Influence function
39 qui sum(Q1W_1)
40 gen EY1tmle = `r(mean)'
41 qui sum(Q0W_1)
42 gen EY0tmle = `r(mean)'
43
44 gen d1 = (($A * ($Y - Q1W_1)/gw)) + Q1W_1 - EY1tmle
45 gen d0 = ((1 - $A) * ($Y - Q0W_1)/(1 - gw)) + Q0W_1 - EY0tmle
46
47 gen IF = d1 - d0
48 qui sum IF
49 gen varIF = r(Var) / r(N)
50
51 global LCI = $ATE - 1.96*sqrt(varIF)
52 global UCI = $ATE + 1.96*sqrt(varIF)
53 display "ATE: " %05.4f $ATE _col(15) "95%CI: " %05.4f $LCI ", " %05.4f $UCI

```

6.1 | Statistical inference for data-adaptive estimators: Functional Delta Method

We used the bootstrap procedure and Delta method for statistical inference presetting the previous estimators. Although both approaches are commonly used in practice, and show good statistical properties in a wide range of settings, they have some limitations. The bootstrap procedure is computationally intensive for large data sets and the use of the Delta method will not always be appropriate (i.e., nonparametric settings). Furthermore, when data-adaptive estimation is used, the bootstrap procedure is not supported theoretically, and the functional Delta method based on the IF is required. The IF is a fundamental object of semi-parametric theory that allows us to characterise a wide range of estimators and their efficiency.^{4,16,42} The IF of a regular asymptotic and linear estimator $\hat{\psi}$ of $\psi(\theta)$, where θ is a random variable based on independent and identically distributed samples O_i which capture the first order asymptotic behaviour of $\hat{\psi}$, such that

$$n^{\frac{1}{2}}\hat{\psi} - \psi(\theta) = n^{-\frac{1}{2}} \sum_{i=1}^n IF(O_i; \theta) + o_p(1)$$

where $o_p(1)$ represents the remainder term from the first order approximation that converges to zero (in terms of the probability) when the sample size converges to infinity. Mathematically, we can identify the IF as being the second term of a first degree Taylor approximation.^{41,43} From the variance of the IF we derive the SE of the ATE from the TMLE estimator. Therefore, the functional Delta method based on the IF readily allows the application of the Central Limit Theorem and, therefore, to compute Wald-type confidence intervals.⁴ However, using the IF for statistical inference may require larger sample sizes to avoid finite-sample issues. Recent research and theoretical developments support the use of double-robust cross-fit estimators to retain valid statistical inference when using machine learning algorithms that are non-Donsker.⁴⁴ The computation of the IF is provided in Box 25 (step 6: rows 38-53).

In Box 26, we outline how to compute the ATE using data-adaptive procedures implemented in the *eltmle* user-written Stata command.⁴⁵ This command implements the TMLE framework for the ATE of the marginal risk ratio and odds ratio for a binary or continuous outcome and a binary treatment. It also includes the use of data-adaptive estimation of the propensity score g_0 and regression outcome Q_0 models via ensemble learning,⁴⁶ which is implemented by calling the *SuperLearner* package v.2.0-21 from R.^{46,47} The super-learner uses 5-fold cross-validation by default to assess the performance of prediction regarding the potential outcomes and the propensity score as weighted averages of a set of machine learning algorithms. The *SuperLearner* has default algorithms implemented in the base installation of the *tmle-R* package v.1.2.0-5.³⁵ The default algorithms include the following: (i) stepwise selection, (ii) generalised linear modeling (GLM), (iii) a GLM variant that includes second order polynomials and two-by-two interactions of the main terms included in the model. Additionally, *eltmle* has an option to include Bayes Generalised Linear Models and Generalised Additive Models as additional algorithms.

Box 26: TMLE and data-adaptive estimation with Stata's user written *eltmle*

```
ssc install eltmle //install via ssc or "github install migariane/eltmle" via GitHub
```

```

2  help eltmle // Description of the command
3  clear
4  set more off
5  use "rhc.dta", clear
6  global W sex age edu race carcinoma
7  eltmle $Y $A $W, tml bal // check balance

```

The ATE is 8.35%, 95% CI: 5.82 – 10.87 (Table 1).

7 | SIMULATION

The motivation of this section is to compare all of the different methods provided in the tutorial under a simple Monte Carlo simulated experiment. For simplicity and pedagogical purposes, we only simulate one sample. However, we provide the results and code in R of a Monte Carlo experiment with 1,000 samples based on the same template as the one presented here and available at <https://github.com/migariane/TutorialComputationalCausalInferenceEstimators>. In Box 27, we outline the data generation process to create random variables including the confounders, the treatment, and the outcome. Afterward, we estimate the simulated value for the ATE, and compute the ATE using all the aforementioned different estimators under a scenario of forced near-positivity violation and model misspecification. Lastly, we compare their performance based on the relative bias with respect to the value of the simulated ATE (note that this approximates bias, as we only simulate 1 data set). Note that other metrics to assess performance can also be used, including the variance of the estimate. The simulation setting includes a binary outcome (Y), potential outcomes (i.e., $Y(1)$ and $Y(0)$), and a binary treatment (A). The vector of confounders \mathbf{W} reflect the commonly analysed cancer patient characteristics: deprivation level (w1, five categories), age at diagnosis (w2, binary), cancer stage (w3, four categories) and comorbidity (w4, four categories).

Box 27: Data generation for the Monte Carlo experiment

```

1 // Data generation
2 clear
3 set obs 1000
4 set seed 777
5 gen w1 = round(runiform(1, 5)) //Quintiles of Socioeconomic Deprivation
6 gen w2 = rbinomial(1, 0.45) //Binary: probability age >65 = 0.45
7 gen w3 = round(runiform(0, 1) + 0.75*(w2) + 0.8*(w1)) //Stage
8 recode w3 (5/6=1) //Stage (TNM): categorical 4 levels
9 gen w4 = round(runiform(0, 1) + 1.2*(w2) + 0.2*(w1)) //Comorbidities: categorical four levels
10 gen A = (rbinomial(1, invlogit(-3 - 0.5*(w4) + 1.5*(w2) + 0.75*(w3) + 0.25*(w1) + 0.8*(w2)*(w4)))) //
    Binary treatment
11 gen Y1 = (invlogit(-3 + 1 + 0.25*(w4) + 0.75*(w3) + 0.8*(w2)*(w4) + 0.05*(w1))) // Potential outcome 1
12 gen Y0 = (invlogit(-3 + 0 + 0.25*(w4) + 0.75*(w3) + 0.8*(w2)*(w4) + 0.05*(w1))) // Potential outcome 2
13 gen psi = Y1-Y0 // Simulated ATE
14 gen Y = A*(Y1) + (1 - A)*Y0 // Binary outcome (consistency)
15
16 // Estimate the true simulated ATE
17 mean psi
18
19 // ATE estimation
20 * Regression adjustment
21 teffects ra (Y i.w1 i.w2 i.w3 i.w4) (A)
22 estimates store ra
23
24 * IPTW
25 teffects ipw (Y) (A i.w1 i.w2 i.w3 i.w4)
26 estimates store ipw
27
28 * IPTW-RA
29 teffects ipwra (Y i.w1 i.w2 i.w3 i.w4) (A i.w1 i.w2 i.w3 i.w4)
30 estimates store ipwra
31
32 * AIPTW
33 teffects aipw (Y i.w1 i.w2 i.w3 i.w4) (A i.w1 i.w2 i.w3 i.w4)
34 estimates store aipw
35
36 * Results
37 qui reg psi

```

```

38 estimates store psi
39 estout psi ra ipw ipwra aipw
40
41 // Ensemble learning maximum likelihood estimation
42 eltmle Y A w1 w2 w3 w4, tmle bal
43
44 // Relative bias for the ATE
45
46 * Regression adjustment
47 display abs(0.1652804 - 0.1726079)/0.1652804
48 0.04433375 // 4.4% bias
49 * IPTW
50 display abs(0.1652804 - 0.1597895)/0.1652804
51 0.03322173 // 3.3% bias
52 * IPTW-RA
53 display abs(0.1652804 - 0.1673554)/0.1652804
54 0.01255442 // 1.2% bias
55 * AIPTW
56 display abs(0.1652804 - 0.1682798)/0.1652804
57 0.01814734 // 1.8% bias
58 * ELTMLE
59 display abs(0.1652804 - 0.1652167)/0.1652804
60 0.00038541 // 0% bias to 3 decimal places

```

For a single-instance simulated data set, compared to the true ATE of 0.165, all of the methods produced a biased estimate under near positivity violations and model misspecification (i.e., RA: 4.4% bias, IPTW: 3.3% bias, IPTW-RA: 1.2% bias, and AIPTW: 1.8% bias), but ELTMLE produces an estimate that is unbiased (i.e., ELTMLE: 0% bias to 3 decimal places) relative to the true ATE. The relative bias from only one simulated sample for the regression adjustment and IPTW estimator is large because they rely on the positivity assumption, which, in this simulation, is violated because there was a low number of individuals with a higher comorbidity value. Without correcting for this imbalance in the data, the methods that rely on this assumption will be vulnerable to bias.

8 | CONCLUSION

Overall, methods introduced here rely on the estimation of the g-formula (nonparametrically or parametrically), which is a generalisation of standardisation, the inverse probability of treatment weighting (IPTW), or their combination (i.e., double-robust methods).³ However, there are other estimators based on matching strategies that we did not cover here.¹⁹ Readers can find a more detailed overview of the propensity score and matching methods in a recently published article.⁴⁸

Table 2 shows the results of the ATE for all of the different causal inference estimators we introduced in the tutorial. Overall, all of the methods showed a consistent result for the ATE (Table 2). The RHC data (demonstrated in this paper) is used to teach causal inference methods because of its extremely well-balanced distribution of confounders across levels of the treatment (RHC). However, in most observational studies, data are not usually well-balanced and there are potentially near violations of the positivity assumption that must always be checked.

We introduced different estimators in regards to their chronological development: the methods were developed to answer the limitations of the previous approach. For example, parametric estimators were developed to address the curse of dimensionality. Then, issues related to extrapolation for the G-computation, and the instability of the estimations due to large weights for the IPTW estimators, encouraged the development of double-robust methods. AIPTW was a strong candidate to answer this issue by incorporating semi-parametric theory and methods to causal inference. However, it was known that it did not solve the estimation equation (i.e., equal to zero) due to the fact that it is not a substitution estimator or plug-in estimator (see glossary). Thus, to overcome this limitation of the AIPTW estimator, data-adaptive estimation using machine learning algorithms and ensemble learning to estimate the nuisance parameters from the regression and propensity score models, were combined to solve the estimation equation.⁴ Evidence shows that the double-robust estimators (particularly TMLE) obtain less biased estimates of the true causal effect in comparison to naive estimators such as multivariate regression.⁴

Evidence shows that when comparing the underlying properties of each method based on Monte-Carlo experiments, only TMLE provides the numerous properties of estimating the probability distribution that enable it to out-perform the others.

The properties of the estimator are: loss-based, well-defined, unbiased, efficient and can be used as a substitution estimator. Maximum likelihood estimation (MLE) based methods (stratification, propensity score and parametric regression) and other estimating equations (IPTW and AIPTW) do not have all of the properties of TMLE and evidence shows that they underperform in comparison to TMLE in selected samples. For more detailed comparisons between the different methods, the interested reader is referred to Chapter 6 of van der Laan and Rose's TMLE textbook.⁴ It is important to highlight that in contrast to the AIPTW estimator, TMLE respects the global constraints of the statistical model (i.e. $P_0(0 < Y < 1) = 1$) and solves the estimation equations being equal to zero.⁴

However, even if TMLE is less prone to errors due to misspecification than alternative methods (e.g., inverse probability weighting) there is some question regarding the validity of the robustness of inference produced by TMLE in nonparametric settings.⁴⁹ This is an area of ongoing work (i.e., double/debiased machine learning, cross-validated TMLE and cross-fit estimators).^{44,50,51} Furthermore, TMLE and the *SuperLearner* were originally developed in R.^{46,35} Outside R, there is a Python library implementing TMLE and the *SuperLearner* named *zEpid*,⁵² and a SAS library implementing the *SuperLearner*.⁵³ Also, there is a user written program for Stata (*elmlle*).⁴⁵ However, *elmlle* is not completely native to Stata but rather calls the *SuperLearner* R package to calculate the predictions of the treatment and outcome models. More work is required to continue implementing and improving the TMLE framework in other statistical software.³⁵

Causal inference is a growing field in rapid developments. Modern causal inference methods allow machine learning to be used when strong assumptions for parametric models are not reasonable. Overall, due to the difficulty of correctly specifying parametric models in high-dimensional data, we advocate for the use of double-robust estimators with ensemble learning. Using these approaches may require larger sample sizes to avoid finite-sample bias.^{16,54} However, recent developments support the use of cross-fit double-robust estimators for data adaptive estimation.^{44,50} Tutorials introducing the use and derivation of the functional Delta method and Influence Curve for applied researchers are needed. The tutorial presented here may help applied researchers to gain a better understanding of the computational implementation for different causal inference estimators.

FUNDING

MJS, CM, AB, BR are supported by a programme grant from Cancer Research UK (Reference C7923/A18525). MALF is supported by a Miguel Servet I Investigator Award (grant CP17/00206 EU-FEDER) from the National Institute of Health, Carlos III (ISCIII), Madrid, Spain. PNZ received training support from the National Institutes of Health (T32-HD091058 PI: Aiello, Hummer). CL is supported by the UK Medical Research Council (Skills Development Fellowship MR/T032448/1). Funders had no role in the study design, data collection, data analysis, data interpretation, or writing of the report.

AUTHORS CONTRIBUTIONS

The article arises from the motivation to disseminate the principles of modern epidemiology amongst clinicians and applied researchers. MALF developed the concept, designed the first draft of the article and the computing code. All authors interpreted and reviewed the code and the data, drafted and revised the manuscript. All authors read and approved the final version of the manuscript. MALF is the guarantor of the article.

ACKNOWLEDGEMENTS

We thank David Benkeser for his help updating the proof of the equivalence between the IPTW and G-computation estimators. We thank Fiona Ingleby and Cristina Renzi for their comments and testing the code included in the boxes of the article.

GLOSSARY

The glossary is adapted from the book "*Targeted learning: causal inference for observational and experimental data*" and a recent publication introducing the TMLE framework.^{55,4}

- **Data-generating process (DGP)**
The mechanism that generated the observed data, with the corresponding data-generating probability distribution which produces the observed samples that were collected.
- **Estimand**
A quantity we are interested in estimating from our data.
- **Estimator**
A function of the sample of observations (that is, a function of the random variables) that generates estimates. The estimators are represented by algebraic equations that explicitly describe a function of the realised observations.
- **Estimate**
The realised value of an estimator, or a function of the realised observations. It is the value of the quantity defined by the estimand.
- **Counterfactual**
A contrary-to-fact value said to arise from hypothetically imposing an intervention on a system represented by a structural causal model. For example, the potential outcome $Y(a)$ is a counterfactual that arises from a hypothetical intervention that sets the treatment A to level a .
- **Statistical model**
A set (family) of probability distributions that could describe the data-generating process. Note that, outside simulation exercises, the true data-generating process is unknown.
- **Saturated model**
A saturated model fits the data perfectly and it includes the main terms plus the higher order interactions between the factors included in the model. Usually, the number of parameters is equal to the number of the possible combinations between the levels of the distinct covariates included in the model.
- **Model misspecification**
A scenario in which the statistical model, which is postulated to contain the distribution describing the data-generating process, fails to actually contain the corresponding true data-generating distribution.
- **Parametric statistical model**
A family of probability distributions indexed by a finite set of model parameters. For example, a linear model traditionally assumes the outcome is a linear function of covariates plus a normally distributed error term with constant variance. Its parameters are the coefficients on the covariates and the variance of the error term.
- **Nonparametric statistical model**
A family of probability distributions that cannot be indexed by a finite set of parameters. That is, the set of parameters indexing this family of distributions is infinite-dimensional. Most often, when making minimal assumptions, the data-generating process cannot be defined by a finite set of parameters, making the set of parameters infinite-dimensional. For example, if all we know about the data-generating process is that we have access to n independent and identically distributed (i.i.d.) samples, then the statistical model for the data-generating process is a nonparametric statistical model.
- **Target estimand or target parameter**
A function of the true (unknown) data-generating process that one is interested in estimating, and represents the mathematical formulation of the motivating question of interest.
- **Maximum likelihood estimation**
The most common method for estimating parameters in a finite-dimensional model (i.e., parametric statistical model).

As the name implies, such estimates are generated by finding a set of parameter estimates that maximise the likelihood function of the observed data.

- Score equation

The gradient (i.e., multi-variable generalisation of the derivative) of the log-likelihood function of the data with respect to the parameter(s). This equation provides information on the degree of change resulting from very small perturbations of the parameter values.

- Regular estimator

A class of estimators that converge in distribution to some limit distribution even if one samples from a slightly perturbed data distribution. Such estimators, if also asymptotically linear, accommodate inference by way of their asymptotic convergence to a Normal distribution.

- Plug-in (substitution) estimator

An estimator that generates an estimate of the true parameter value by “plugging in” estimates of relevant parts of the data-generating distribution into the parameter mapping. This method is commonly referred to as the plug-in principle. For example, “plugging in” targeted Super Learner fit of the conditional mean under $A = 1$ and $A = 0$ generates an estimate of the average treatment effect.

References

1. Pearl Judea. Causal Inference in Statistics: An Overview. *Statistics Surveys*. 2009;3:96–146.
2. Rubin Donald B. Estimating causal effects of treatments in randomized and nonrandomized studies.. *Journal of educational Psychology*. 1974;66(5):688.
3. Robins James. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling*. 1986;7(9):1393–1512.
4. Laan M J, Rose Sherri. *Targeted learning: causal inference for observational and experimental data*. New York: Springer Series in Statistics; 2011.
5. Luque-Fernandez Miguel Angel, Redondo-Sanchez Daniel, Schomaker Michael. Effect Modification and Collapsibility in Evaluations of Public Health Interventions. *American Journal of Public Health*. 2019;109(3):e12-e13.
6. Connors Alfred F., Speroff Theodore, Dawson Neal V., et al. The effectiveness of right heart catheterization in the initial care of critically ill patients. *Journal of the American Medical Association*. 1996;276(11):889–897.
7. Rubin Donald B. The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Stat Med*. 2007;26(1):20–36.
8. Gutman Roe, Rubin Donald B. Estimation of causal effects of binary treatments in unconfounded studies. *Stat Med*. 2015;34(26):3381–3398.
9. Goetghebeur Els, Cessie Saskia, De Stavola Bianca, Moodie Erica E.M., Waernbaum Ingeborg. Formulating causal questions and principled statistical answers. *Statistics in Medicine*. 2020;39(30):4922–4948.
10. Keil Alexander P., Edwards Jessie K., Richardson David B., Naimi Ashley I., Cole Stephen R.. The parametric g-formula for time-to-event data: Intuition and a worked example. *Epidemiology*. 2014;.
11. Efron Bradley, Tibshirani Robert. *An introduction to the bootstrap*. New York: Chapman & Hall; 1993.
12. Efron Bradley. *The jackknife, the bootstrap and other resampling plans*. SIAM; 1982.
13. Wasserman Larry. *All of Nonparametric Statistics*. Springer New York; 2006.

14. Jung Kwanghee, Lee Jaehoon, Gupta Vibhuti, Cho Gyeongcheol. Comparison of Bootstrap Confidence Interval Methods for GSCA Using a Monte Carlo Simulation. *Frontiers in Psychology*. 2019;10:2215.
15. Oehlert Gary W.. A note on the delta method. *American Statistician*. 1992;46(1):27–29.
16. Kennedy Edward H. *Semiparametric theory and empirical processes in causal inference*. 2016.
17. Boos Dennis D, Stefanski L A. *Essential statistical inference: theory and methods*. Springer; 2013.
18. Horvitz D. G., Thompson D. J.. A Generalization of Sampling Without Replacement From a Finite Universe. *Journal of the American Statistical Association*. 1952;47(260):663.
19. Rosenbaum Paul R, Rubin Donald B. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70(1):41–55.
20. Hajek J.. *Comment on An essay on the logical foundations of survey sampling by Basu, D. in Foundations of Statistical Inference (Godambe, V.P. and Sprott, D.A. eds.)*. Holt, Rinehart and Winston; 1971.
21. Ridgeway G, McCaffrey D, Morral A, Griffin B A, Burgette L. *twang: Toolkit for Weighting and Analysis of Nonequivalent Groups. R package version 1.3-20*. 2013.
22. Austin Peter C.. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in Medicine*. 2009;28(25):3083–3107.
23. Cattaneo Matias D.. Efficient semiparametric estimation of multi-valued treatment effects under ignorability. *Journal of Econometrics*. 2010;155(2):138–154.
24. Stürmer Til, Rothman Kenneth J., Avorn Jerry, Glynn Robert J.. Treatment effects in the presence of unmeasured confounding: Dealing with observations in the tails of the propensity score distribution—A simulation study. *American Journal of Epidemiology*. 2010;172(7):843–854.
25. Cole Stephen R., Hernán Miguel A.. *Constructing inverse probability weights for marginal structural models*. 2008.
26. Xiao Yongling, Moodie Erica E.M., Abrahamowicz Michal. Comparison of approaches to weight truncation for marginal structural Cox models. *Epidemiologic Methods*. 2013;2(1):1–20.
27. Hernán Miguel Ángel, Brumback Babette, Robins James M.. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology*. 2000;11(5):561–570.
28. Robins James M., Hernán Miguel Ángel, Brumback Babette. Marginal structural models and causal inference in epidemiology. *Epidemiology*. 2000;11(5):550–560.
29. Tsiatis Anastasios A. *Semiparametric Theory and Missing Data*. Springer Science and Business Media LLC; 2006.
30. Tsiatis Anastasios A, Davidian Marie, Kang Joseph D.Y. Y, Schafer Joseph L.. Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data. *Statistical Science*. 2007;22(4):523–539.
31. Daniel Rhian M.. Double Robustness. In: Wiley 2018 (pp. 1–14).
32. Robins James M, Rotnitzky Andrea, Zhao Lue Ping. Estimation of Regression Coefficients When Some Regressors Are Not Always Observed. *Journal of the American Statistical Association*. 1994;89(427):846–866.
33. Bang Heejung, Robins James M. Doubly robust estimation in missing data and causal inference models. *Biometrics*. 2005;61(4):962–973.
34. Gruber Susan, Laan Mark van der. Targeted Maximum Likelihood Estimation: A Gentle Introduction. *U.C. Berkeley Division of Biostatistics Working Paper Series*. 2009;.

35. Gruber Susan, Laan Mark. tmle: An R Package for Targeted Maximum Likelihood Estimation. *U.C. Berkeley Division of Biostatistics Working Paper Series*. 2011;.
36. Gruber Susan, Laan Mark J.. tmle: An R package for targeted maximum likelihood estimation. *Journal of Statistical Software*. 2012;51(13):1–35.
37. Schuler Megan S., Rose Sherri. Targeted maximum likelihood estimation for causal inference in observational studies. *American Journal of Epidemiology*. 2017;185(1):65–73.
38. Luque-Fernandez Miguel Angel, Schomaker Michael, Rachet Bernard, Schnitzer Mireille E.. Targeted maximum likelihood estimation for a binary treatment: A tutorial. *Statistics in Medicine*. 2018;37(16):2530–2546.
39. Luque-Fernandez Miguel Angel, Belot Aurélien, Valeri Linda, Cerulli Giovanni, Maringe Camille, Rachet Bernard. Data-Adaptive Estimation for Double-Robust Methods in Population-Based Cancer Epidemiology: Risk Differences for Lung Cancer Mortality by Emergency Presentation. *American Journal of Epidemiology*. 2018;187(4):871–878.
40. Naimi Ashley I, Mishler Alan E, Kennedy Edward H. Challenges in Obtaining Valid Causal Effect Estimates with Machine Learning Algorithms. *arXiv*. 2017;.
41. Luque-Fernandez M A. Delta Method in Epidemiology: An Applied and Reproducible Tutorial. *GitHub repository*. 2020;.
42. Vaart A W van der. *Asymptotic statistics*. Cambridge, UK: Cambridge University Press; 1998.
43. MA Mansournia, M Nazemipour, AI Naimi, GS Collins, MJ Campbell. Reflection on modern methods: demystifying robust standard errors for epidemiologists. *International journal of epidemiology*. 2021;50(1):346–351.
44. Zivich Paul N, Breskin Alexander. Machine learning for causal inference: on the use of cross-fit estimators. *arXiv*. 2020;.
45. Luque-Fernandez Miguel Angel. *migariane/meltmle: Ensemble Learning Targeted Maximum Likelihood Estimation for Stata users* | Zenodo. 2019.
46. Van Der Laan Mark J., Polley Eric C., Hubbard Alan E., Laan Mark J, Polley Eric C., Hubbard Alan E.. Super learner. *Statistical applications in genetics and molecular biology*. 2007;6(1).
47. Polley Eric, Laan Mark van der. Super Learner In Prediction. *U.C. Berkeley Division of Biostatistics Working Paper Series*. 2010;.
48. Webster-Clark Michael, Stürmer Til, Wang Tiansheng, et al. Using propensity scores to estimate effects of treatment initiation decisions: State of the science. *Statistics in Medicine*. 2021;40(7):1718–1735.
49. Díaz Iván. Machine learning in the estimation of causal effects: targeted minimum loss-based estimation and double/debiased machine learning. *Biostatistics (Oxford, England)*. 2020;21(2):353–358.
50. Chernozhukov Victor, Chetverikov Denis, Demirer Mert, et al. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*. 2018;21(1):C1–C68.
51. Levy Jonathan. An Easy Implementation of CV-TMLE. *arXiv*. 2018;.
52. Zivich Paul N. Python for Epidemiologists (v0.9.0). *Zenodo*. 2020;.
53. Keil Alexander P., Westreich Daniel, Edwards Jessie K., Cole Stephen R.. Super learning in the SAS system. *arXiv*. 2019;.
54. Gill Richard D. Non- and Semi-Parametric Maximum Likelihood Estimators and the Von Mises Method (Part 1). *Scandinavian Journal of Statistics*. 1989;16(2):97–128.
55. Coyle Jeremy R., Hejazi Nima S., Malenica Ivana, et al. Targeting Learning: Robust Statistics for Reproducible Research. *arXiv*. 2020;.



1 | SUPPLEMENTARY WEB MATERIALS

1.1 | Appendix 1: Potential outcomes framework, causal assumptions, and g-formula

To illustrate the framework we use an empirical example based on intensive care medicine.¹ The study, set in intensive care units of five United States teaching hospitals between 1989 and 1994, evaluated the effectiveness of right heart catheterisation (RHC) on short-term mortality (30 days) of 5,735 critically ill adult patients (2,184 received a RHC and 3,551 did not received it) receiving care for 1 of 9 prespecified disease categories. In our illustration, the outcome is short-term mortality defined as 30 days after ICU admission, and RHC was the main intervention henceforth the treatment, and we define (\mathbf{W}) to include the set of confounders. Let Y denote the vital status of the patient in an intensive care unit (ICU) at 30 days after admission. Let A denote the treatment of whether or not the patient received RHC during their stay at the ICU, and let C denote a binary confounder.

For a binary treatment, each patient in the study has two potential outcomes (i.e., $Y(a)$), where $Y(1)$ denotes the potential outcome if they received RHC, and $Y(0)$ denotes the potential outcome if they did not receive RHC.² However, only one of the potential outcomes can be observed since a patient can only ever receive one of the treatments and only one of the outcomes (they cannot both live and die after 30 days). As an example from Table 1, imagine that Matthew has two potential outcomes: firstly, $Y(0) = 1$ says that if Matthew were not to receive RHC then he would die within 30 days, and secondly, $Y(1) = 0$ says that if Matthew were to receive RHC then he would not die within 30 days. Likewise, for the rest of the individuals their potential outcomes are presented and the ATE can be estimated as the contrast between the *potential outcomes* under different treatment levels (i.e., the difference between $E[Y(1)] - E[Y(0)]$).³

Patient	Y	A	C	Y(0)	Y(1)
Matthew	1	0	0	1	0
Camille	1	1	1	1	1
Aurelien	1	1	1	0	1
Paul	0	1	0	0	0
Mohammad	1	0	1	1	1
Steve	0	1	1	0	0
Miguel	1	0	0	1	1
Bernard	0	1	1	0	0
Clemence	1	1	0	1	1

TABLE 1 Potential outcomes framework: C = Binary confounder, A = Binary treatment, Y = Binary outcome, $Y(0)$ = Potential outcome when untreated, $Y(1)$ = Potential outcome when treated

However, we must make certain assumptions to identify potential outcomes from the observed data and then estimate the ATE.⁴ Given that the potential outcomes are not necessarily directly observed from the data, to identify the ATE from observable data (i.e., from Table 1) the following three assumptions are made:

1. Counterfactual consistency

Counterfactual consistency holds if the observed outcome for all treated individuals equals their outcome if they had been treated, and likewise for untreated individuals. For example, in Table 1, Matthew's observed outcome equals his potential outcome if he had not been treated ($Y = Y(0) = 1$). This means that the definition of the treatment, and outcome, is consistent for Matthew (the same applies for all the other patients). Analytically, consistency is represented by:

$$Y = AY(1) + (1 - A)Y(0)$$

We further assume observations are independent (e.g., no interference) and there is no measurement error.

2. Conditional exchangeability

In randomised studies, conditional and marginal exchangeability holds because the treated individuals, had they not been treated, would have had the same average potential outcomes as the untreated, and vice versa. This cannot be guaranteed in observational studies but it can be assumed to hold if the unmeasured risk factors of the outcome are equally distributed between the treated and the untreated groups conditional on the measured confounders. Thus, using the language of the potential outcomes, the conditional exchangeability assumption (a.k.a conditional independence, unconfoundedness or ignorability) is given:

$$Y(a) \perp\!\!\!\perp A \mid C \forall a \in \{0, 1\}$$

Hence, the conditional mean independence is given

$$E[Y_a \mid A = 1, C = c] = E[Y_a \mid A = 0, C = c] = E[Y_a \mid C = c] \forall a \in \{0, 1\}$$

3. Positivity

Positivity holds if the conditional probability of being treated (and similarly for being untreated) is greater than zero. Therefore, if $P(C = c) > 0$ then $P(A = a \mid C = c) > 0 \forall C \in \mathbf{c}, a \in \{0, 1\}$. When this assumption is violated, it is typically because the target population is poorly defined (trying to estimate the effect of a treatment on people who would never receive it anyway).

With these assumptions, the observed data can then be used to estimate the average treatment effect as follows:

By the law of total probability

$$P[Y(a) = 1] = \sum_c P[Y(a) = 1 \mid C = c] P(C = c)$$

By conditional exchangeability the right hand side is

$$\sum_c P[Y(a) = 1 \mid A = a, C = c] P(C = c)$$

This is possible since we are assuming that, within levels of C , the predictors of the outcome are equally distributed between treated (e.g., RHC) and non-treated (e.g., non-RHC) groups: that is we have achieved what would happen if patients were randomised to each treatment within stratum of C . If we assume consistency the right hand side is

$$\sum_c P[Y = 1 \mid A = a, C = c] P(C = c)$$

The ATE is defined as

$$P(Y(1) = 1) - P(Y(0) = 1)$$

and, under the preceding assumptions, can be estimated by

$$\sum_c P[Y = 1 \mid A = 1, C = c] \Pr(C = c) - \sum_c P[Y = 1 \mid A = 0, C = c] P(C = c). \quad (1)$$

We have transitioned from the (unobserved) potential outcomes to a setting where we can estimate our causal estimand, from the distribution of the observed data, using equation 1, namely the g-formula.⁵

1.2 | Appendix 2: Equivalence between IPTW and G-computation

By repeated use of the law of total expectation, the IPTW and the G-computation regression adjustment estimators for the ATE are equivalent as given by

$$\underbrace{E\left(\frac{I(a=1)}{P(A=1|W)}Y\right)}_{\text{IPTW}} =$$

By definition of expectations...

$$= \sum_{w,a,y} \frac{I(a=1)}{P(A=1|W=w)} y P(Y=y, A=a, W=w)$$

By the law of total probability...

$$= \sum_{w,a,y} \frac{I(a=1)}{P(A=1|W=w)} y P(Y=y|A=a, W=w) P(A=a|W=w) P(W=w)$$

Cancellation by evaluating at A=1...

$$= \sum_{w,y} y P(Y=y|A=1, W=w) P(W=w)$$

By definition of expectations...

$$= \sum_w E(Y|A=1, W=w) P(W=w)$$

Finally, again by definition of expectations...

$$= \underbrace{E[E(Y|A=1, W)]}_{\text{G-computation}}$$

References

1. Connors Alfred F., Speroff Theodore, Dawson Neal V., et al. The effectiveness of right heart catheterization in the initial care of critically ill patients. *Journal of the American Medical Association*. 1996;276(11):889–897.
2. Rubin Donald B. Estimating causal effects of treatments in randomized and nonrandomized studies.. *Journal of educational Psychology*. 1974;66(5):688.
3. Rubin Donald B. Causal inference using potential outcomes. *Journal of the American Statistical Association*. 2005;100(469):322–331.
4. Robins James. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling*. 1986;7(9):1393–1512.
5. Robins James M.. Association, Causation, and Marginal Structural Models. *Synthese*. 1999;121(1/2):151–179.



A.5.5 Association between patient characteristics and delayed diagnosis of patients with non-Hodgkin lymphoma

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	lsh1601639	Title	Mr
First Name(s)	Matthew		
Surname/Family Name	Smith		
Thesis Title	Survival of patients with non-Hodgkin lymphoma in England: investigating the socioeconomic inequalities		
Primary Supervisor	Edmund Njeru Njagi		

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?	British Journal of Cancer
Please list the paper's authors in the intended authorship order:	Matthew J. Smith, Miguel Angel Luque Fernandez, Aurélien Belot, Matteo Quartagno, Audrey Bonaventure, Sara Benitez Majano, Bernard Rachet, Edmund Njeru Njagi
Stage of publication	Submitted

SECTION D – Multi-authored work

<p>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)</p>	<p>MJS, ENN and BR contributed to the conception of the study and designed the study. ENN, BR, ABe, MALF and MQ provided advice on statistical methods. MJS conducted the analyses of the data and prepared the draft of the manuscript, tables and figures. ENN and BR supervised the study and provided comments on the manuscript draft. ENN, BR, MALF, MQ, SBM, ABe and ABo provided comments on the final draft of the manuscript. All authors read and approved the final manuscript.</p>
---	---

SECTION E

Student Signature	Matthew J. Smith
Date	7th June 2021

Supervisor Signature	
Date	

Title

Investigating the inequalities in route to diagnosis amongst patients with Diffuse Large B-cell or Follicular lymphoma in England

Authors

Matthew J. Smith^{1*}, Miguel Angel Luque Fernandez^{1,2}, Aurélien Belot¹, Matteo Quartagno³, Audrey Bonaventure⁴, Sara Benitez Majano¹, Bernard Rachet¹, Edmund Njeru Njagi¹

Authors' affiliations

¹ Inequalities in Cancer Outcomes Network, Department of Non-Communicable Disease Epidemiology, London School of Hygiene and Tropical Medicine, Keppel Street, London, WC1E 7HT, UK

² Noncommunicable Disease and Cancer Epidemiology Group, Instituto de Investigación Biosanitaria de Granada, Ibs.GRANADA, Andalusian School of Public Health, Granada, Spain

³ MRC Clinical Trials Unit, Institute of Clinical Trials and Methodology, University College London, London WC1V 6LJ, UK

⁴ CRESS, Université de Paris, INSERM, UMR 1153, Epidemiology of Childhood and Adolescent Cancers Team, Villejuif, France

Corresponding author*

Matthew J. Smith
LSHTM, Keppel Street, London, WC1E 7HT, UK
Email: matthew.smith1@lshtm.ac.uk

Word count: Abstract: 201; Text: 3547; Tables: 4; Figures: 2

Additional information

Acknowledgements

We would like to thank Adrian Turculet, Data Manager of the LSHTM Inequalities in Cancer Outcomes Network, for his support and assistance with the data linkage.

Author's contributions

MS, ENN and BR contributed to the conception of the study and designed the study. ENN, BR, ABe, MALF and MQ provided advice on statistical methods. MS conducted the analyses of the data and prepared the draft of the manuscript, tables and figures. ENN and BR supervised the study and provided comments on the manuscript draft. ENN, BR, MALF, MQ, SBM, ABe and ABo provided comments on the final draft of the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate: We obtained the statutory approvals required for this research from the Confidentiality Advisory Group (CAG) of the Health Research Authority (HRA): PIAG 1–05(c) 2007. Ethical approval was obtained from the Research Ethics Committee (REC) of the Health Research Authority (HRA): 07/MRE01/52. This work uses data provided by patients and collected by the National Health Service as part of their care and support. We used anonymised National Cancer Registry and Hospital Episode Statistics data. No consent to participate was sought from patients.

Consent for publication: Not applicable

Availability of data and materials: The data that support the findings of this study are available via application to the Public Health England Office for Data Release, but restrictions apply to the availability of these data.

Competing interests: The authors declare that they have no competing interests.

Funding: This research was funded by Cancer Research UK grant number C7923/A18525. The authors declare no support from any organisations for the submitted work. The design of the study, the analyses and the writing of the manuscript were solely the responsibility of the authors. The findings and conclusions in this manuscript are those of the authors and do not necessarily represent the views of Cancer Research UK.

Abstract

Introduction

Diagnostic delay is associated with lower chances of cancer survival. Underlying comorbidities are known to affect the timely diagnosis of cancer. Diffuse Large B-cell (DLBCL) and Follicular lymphomas (FL) are primarily diagnosed amongst older patients, who are more likely to have comorbidities. Characteristics of clinical commissioning groups (CCG) are also known to impact diagnostic delay. We assess the association between comorbidities and diagnostic delay amongst patients with DLBCL or FL in England during 2005-2013.

Methods

Multivariable generalised linear mixed-effect models were used to assess the main association. Empirical Bayes estimates of the random effects were used to explore between-cluster variation. Latent normal joint modelling multiple imputation approach was used to account for partially-observed variables.

Results

We included 30,078 and 15,551 patients diagnosed with DLBCL or FL, respectively. Amongst patients from the same CCG, having multimorbidity was strongly associated with emergency route to diagnosis (DLBCL: Odds Ratio 1.56, CI 1.40 – 1.73; FL: Odds ratio 1.80, CI 1.45 – 2.23). Amongst DLBCL patients, diagnostic delay was possibly correlated with CCGs that had higher population densities .

Conclusions

Underlying comorbidity is associated with diagnostic delay amongst patients with DLBCL or FL. Results suggest a possible correlation between CCGs with higher population densities and diagnostic delay of aggressive lymphomas.

Key words: Cancer epidemiology, diffuse large B-cell lymphoma, follicular lymphoma, comorbidity, deprivation, clinical commissioning groups

Introduction

Non-Hodgkin lymphoma is a heterogeneous disease comprising over sixty morphological entities with diverse histological patterns.¹ The most common of which are diffuse large B-cell (DLBCL) and follicular lymphomas (FL), exhibiting an annual rate of 8.2 and 3.3 cases (respectively) per 100,000 people in the UK. These subtypes are relatively common in adults, with incidence increasing amongst older ages.² Each of these subtypes have markedly differing treatments and health outcomes.¹

Survival of DLBCL or FL patients in England has steadily increased over the past decades,^{3,4} yet the proportion of patients surviving trails that of other European countries.⁵ Evidence has highlighted that diagnostic delay (compared to an earlier diagnosis) is associated with a less intensive treatment plan, which then impacts on the chances of survival.⁶ Public health policies have aimed to increase awareness, encourage more patient and healthcare system interactions, and set targets for earlier cancer diagnosis.⁷⁻¹⁰

In the United Kingdom, the cancer diagnostic route is defined as the first of eight possible points of contact between the patient and the healthcare system.¹¹ Emergency diagnosis is defined as a diagnosis of cancer following presentation to an Accident and Emergency Unit, or following an emergency pathway for in/out-patients: it is used as an indicator of diagnostic delay for cancer patients.¹² Underlying comorbidities are known to affect the timely diagnosis of other cancers.¹³⁻¹⁵ A comorbidity expressing symptoms similar to cancer may delay the diagnosis: dissimilar symptoms may hasten the cancer diagnosis. For example, some symptoms are present in both lymphomas and other chronic diseases, such as swollen abdomen and fatigue in diabetes,¹⁶ chest pain in congestive heart failure,¹⁷ and shortness of breath in chronic obstructive pulmonary disease.¹⁸ Furthermore, all three of these diseases are prevalent amongst patients with lymphoma, which could explain misdiagnosis and diagnostic delay.^{19,20}

A universal healthcare system (UHS), such as the National Health Service (NHS) in England, aims to provide all residents with access to healthcare.²¹ However, variability in health outcomes amongst patients with the same lymphoma still occur.^{22,23} Clinical Commissioning Groups (CCGs) commission the hospital and community

NHS services, and decide on local priorities (informed by general practices), for their respective geographical areas; however, CCGs have shown variability in health outcomes since their inception,^{24,25} which may partly explain differences in diagnostic delay.

We aim to assess the association between pre-diagnosed comorbidities and diagnostic delay (i.e. route to diagnosis) amongst patients with DLBCL or FL, accounting for patient sociodemographic characteristics.

Methods

Study design, participants, data and setting

We developed a population-based cross-sectional study comprising all patients, aged 18 to 99 years, diagnosed with non-Hodgkin lymphoma (NHL) between 1st January 2005 and 31st December 2013. NHL was coded (C82.0-C85.9) according to the 10th revision of the International Statistical Classification of Diseases and Related Problems (ICD).²⁶ Morphology (cell type) and topography (tumour site) were defined using renewed updates of the ICD for Oncology (ICD-O); ICD-O-3²⁷ was used for diagnoses up to 2010, and ICD-O-3.1²⁸ for diagnoses after 2011. Patients diagnosed with either DLBCL or FL were included in the study and are hereby referred to as *subtype* (**Supplementary table S2**).²⁶

Information on patients' cancer diagnosis was collected by the national cancer registry and analysis service (NCRAS).²⁹ The NCRAS contains England national cancer registry data and Hospital Episode Statistics³⁰ (HES) datasets that are accessed via the Cancer Analysis System³¹ (CAS). Cancer registry (CAS dataset) contained information on subtype (morphology), age at diagnosis, ethnicity, gender and date of diagnosis. This was linked to HES, which contained information on patient's previous hospital admissions, accident and emergency presentations, outpatient appointments.

Variables

Route to diagnosis, obtained from NCRAS, was originally recorded as one of eight routes to diagnosis.¹¹ Patients with a 'death certificate only' route to diagnosis were excluded to remove bias. There is no nationally recognised screening programme for NHL and no patients were diagnosed via a 'screen-detected' route. An 'unknown' route to diagnosis was recoded as a missing record. The remaining routes were dichotomised into a binary variable indicating whether the patient was diagnosed following an emergency or elective presentation: elective

presentation consisted of patients diagnosed through two-week-wait, general practitioner referral, inpatient elective and other outpatient.

Comorbidity status, based on the Charlson comorbidity index³² (CCI), was defined as “the existence of disorders, in addition to a primary disease of interest, which are causally unrelated to the primary disease”.^{33,34} Comorbidities were coded within HES according to the International Classification of Diseases, 10th revision (**Supplementary Table S1**). Previous records of comorbidity were obtained from HES data. Patients with any previous malignancy were removed. For each patient, we defined a time window of 6 to 24 months prior to cancer diagnosis for a comorbidity to be recorded. A patient’s CCI was determined using an algorithm developed by Maringe *et al.*³⁵ CCI was classified according to the Royal College of Surgeons (RCS) Charlson Score,³⁶ which was categorised into three groups: 0 for no previous comorbidity, 1 for a single comorbidity, and 2 or more for multimorbidity. We tabulated the prevalence of comorbidity for DLBCL and FL (**Supplementary Table S3**).

Stage at diagnosis is based on the Ann Arbor classification system (CAS dataset).³⁷ A lower tumour stage is predictive of a higher survival outcome compared to a higher tumour burden. For NHL subtypes, stages I/II is a criterion for treatment of low tumour stage; stages III/IV is a criterion for treatment of high tumour stage.³⁸ Therefore, early stage was dichotomised as I/II, and late stage as III/IV.

Deprivation level (HES dataset) is based on the Lower Super Output Area³⁹ (LSOA) of residence of the patient at the date of cancer diagnosis. An LSOA is a geographical location with a median of 1500 inhabitants. From the Index of Multiple Deprivation⁴⁰ (IMD), income domain was classified into one of five quintiles based on the national distribution of ranked deprivation scores in the 32,844 LSOAs. Each patient was linked with one of the 209 Clinical Commissioning Groups (CCG) where their LSOA resides.⁴¹ Lastly, *ethnicity* (HES dataset) was recorded as either white or other.

Statistical analysis

We described the study population, tabulated the patient characteristics with diagnostic delay markers (route to diagnosis), and calculated unadjusted odds ratios (and 95% confidence intervals [CI]) with Wald test p-values.

We conducted analysis for DLBCL and FL separately. Univariable independent logistic regression models were used to explore the crude association between route to diagnosis and each of the patient characteristics. Then, multivariable generalised linear mixed-effect models (GLMM) were used to account for the dependency between patients $j = 1, \dots, n_i$ from CCG $i = 1, \dots, 209$. The GLMM model for route to diagnosis was defined as

$$\text{logit}(\pi_{ij}) = \beta_0 + b_i + \beta_1 A_{ij} + \beta_2 G_{ij} + \beta_3 E_{ij} + \sum_{k=2}^5 \beta_{4k} \cdot D_{ijk} + \sum_{k=2}^3 \beta_{5k} \cdot C_{ijk}$$

where $b_i \sim N(0, \sigma_b^2)$. The patient, and tumour, characteristics were age (A), gender (G), ethnicity (E), deprivation (D), and comorbidity score (C).

The model was estimated using maximum likelihood. Likelihood ratio tests were used to compare between models with and without each covariate and for linear trend. Note that these and subsequent estimates are for any given CCG as results from logistic mixed effects models have cluster-specific interpretation.^{42–45} Empirical Bayes estimates of the random effect \hat{b}_i were used to explore the between-CCG variability in the odds of emergency route to diagnosis. The random effect variance parameter was tested for using a mixture of chi-squares with 0 and 1 degrees of freedom.^{42,43} The mixture of chi-square test is a likelihood-ratio type test, where an appropriate reference distribution is used to account for the fact that the null hypothesis in this case is at the boundary of the parameter space.^{42,46} Combining likelihood ratio tests after multiple imputation requires derivation of a particularly modified likelihood ratio test statistic, which is compared with a particularly derived reference distribution. For tests of fixed effects parameters, the relevant methodology exists.⁴⁷ We are not aware of existing corresponding methodology for combining after multiple imputation likelihood-ratio type tests for random effect variance parameters.

Missing data analysis

Variables with missing data were the outcome (route to diagnosis [DLBCL: 1.9%, FL: 2.1%]), and ethnicity [DLBCL: 22.8%, FL: 24.9%]. Using logistic regression models, we explored the missing data mechanism for each partially observed variable. The imputation model included all fully- and partially-observed covariates and the cluster variable indicator. To reduce potential bias,⁴⁷ the auxiliary variables (patient's vital status, Nelson-Aalen estimate of the cumulative mortality hazard, and stage at diagnosis) were included as, per the missing data indicator model, they were predictive of the chance of missing values and, as per subject matter knowledge, associated with the underlying values themselves.⁴⁸ We used the latent normal joint modelling multiple imputation approach, under a missing at random assumption, and generated 10 imputed datasets. The multilevel logistic regression models for each outcome were fitted to each of these datasets and results combined using Rubin's rules.^{49,50}

We used *R* software for all analysis; the *glmer* function of the *lme4* package was used for generalised linear mixed effects models, and the *jomo*⁵¹ package for multiple imputation, which allows imputation of clustered data.

Results

Summary statistics

In this study, we included 45,629 patients diagnosed with DLBCL (30,078; 65.9%) or FL (15,551; 34.1%) between 1st January 2005 and 31st December 2013 (**Table 1a** and **1b**). The prevalence of emergency diagnostic route amongst those diagnosed with DLBCL or FL was 9,683 (34.1%) and 1,879 (12.3%), respectively, there was no evidence of a yearly trend. Amongst these patients, the average age at diagnosis was 68.2 and 66.3 years, respectively.

The prevalence of emergency diagnostic route (compared to elective) was higher amongst FL males, ethnic minorities in DLBCL, and those living in most deprived areas (both DLBCL and FL). Emergency route, compared to elective, was more common amongst those with multimorbidity: DLBCL (7.2% vs 4.6%, respectively) and FL (6.2% vs 3.1%, respectively). Similarly, for both DLBCL and FL, an increase in the crude odds of emergency route to diagnosis was strongly associated with an increase in age and living in most deprived areas, while for ethnic minority it was observed in DLBCL only. There was an increase in the odds of emergency route to diagnosis with each increase in deprivation level.

Table 1a: Summary statistics of **emergency route to diagnosis** amongst patients diagnosed with Diffuse Large B-cell lymphoma (n=30,078) in England during 2005-2013

	Route to diagnosis ²		cOR [†]	95% CI	p-value
	Elective N = 19,833 (65.9%)	Emergency N = 9,683 (34.1%)			
Age (mean, sd)	67.2 (14.8)	68.2 (15.5)	1.04 ¹	1.03 – 1.06	<0.001
Gender					
<i>Male</i>	10,658 (53.7)	5,292 (54.7)	<i>Ref</i>	<i>Ref</i>	<i>Ref</i>
<i>Female</i>	9,175 (46.3)	4,391 (45.4)	0.96	0.92 – 1.01	0.139
Ethnicity					
<i>White</i>	14,583 (94.8)	6,898 (92.6)	<i>Ref</i>	<i>Ref</i>	<i>Ref</i>
<i>Minorities</i>	802 (5.2)	549 (7.4)	1.44	1.29 – 1.62	<0.001
<i>Missing</i> ³	4,448 (22.4)	2,236 (23.1)	-	-	-
Deprivation					
<i>Least deprived</i>	4,410 (22.2)	1,823 (18.8)	<i>Ref</i>	<i>Ref</i>	<i>Ref</i>
2	4,455 (22.5)	2,105 (21.7)	1.14	1.06 – 1.23	<0.001
3	4,145 (20.9)	2,031 (21.0)	1.19	1.10 – 1.28	<0.001
4	3,806 (19.2)	1,993 (20.6)	1.27	1.17 – 1.37	<0.001
<i>Most deprived</i>	3,017 (15.2)	1,731 (17.9)	1.39	1.28 – 1.50	<0.001
Comorbidity					
<i>None</i>	17,957 (90.5)	8,396 (86.7)	<i>Ref</i>	<i>Ref</i>	<i>Ref</i>
<i>One</i>	970 (4.9)	590 (6.1)	1.30	1.17 – 1.45	<0.001
<i>Multimorbidity</i>	906 (4.6)	697 (7.2)	1.65	1.49 – 1.82	<0.001

¹ Increase in *odds* of emergency route for each 10-year increase in age.

² 562 (1.9%) missing route to diagnosis records

³ Proportions of missing records amongst all ethnicity records (including observed records)

† Crude odds ratios for emergency vs elective

Percentages may not sum to 100% due to rounding. **cOR** – crude odds ratio. **CI** – Confidence interval

Table 1b: Summary statistics of **emergency route to diagnosis** amongst patients diagnosed with Follicular Lymphoma (n=15,551) in England during 2005-2013

	Route to diagnosis ²		cOR [†]	95% CI	p-value
	Elective N = 13,353 (87.7%)	Emergency N = 1,879 (12.3%)			
Age (mean, sd)	63.5 (13.5)	66.3 (14.2)	1.17 ¹	1.13 – 1.21	<0.001
Gender					
<i>Male</i>	6,209 (46.5)	962 (51.2)	<i>Ref</i>	<i>Ref</i>	<i>Ref</i>
<i>Female</i>	7,144 (53.5)	917 (48.8)	0.83	0.75 – 0.91	<0.001
Ethnicity					
<i>White</i>	9,459 (94.9)	1,399 (94.8)	<i>Ref</i>	<i>Ref</i>	<i>Ref</i>
<i>Minorities</i>	510 (5.1)	77 (5.2)	1.02	0.80 – 1.31	0.870
<i>Missing</i> ³	3,384 (25.3)	403 (21.5)			
Deprivation					
<i>Least deprived</i>	3,100 (23.2)	375 (20.0)	<i>Ref</i>	<i>Ref</i>	<i>Ref</i>
2	3,040 (22.8)	405 (21.6)	1.10	0.95 – 1.28	0.205
3	2,857 (21.4)	375 (20.0)	1.09	0.93 – 1.26	0.292
4	2,462 (18.4)	412 (21.9)	1.38	1.19 – 1.61	<0.001
<i>Most deprived</i>	1,894 (14.2)	312 (16.6)	1.36	1.16 – 1.60	<0.001
Comorbidity					
<i>None</i>	12,410 (92.9)	1,667 (88.7)	<i>Ref</i>	<i>Ref</i>	<i>Ref</i>
<i>One</i>	536 (4.0)	95 (5.1)	1.32	1.05 – 1.65	0.015
<i>Multimorbidity</i>	407 (3.1)	117 (6.2)	2.14	1.73 – 2.65	<0.001

¹ Increase in *odds* of emergency route for each 10-year increase in age.

² 319 (2.1%) missing route to diagnosis records

³ Proportions of missing records amongst all ethnicity records (including observed records)

[†] Crude odds ratios for emergency vs elective

Percentages may not sum to 100% due to rounding. **cOR** – crude odds ratio. **CI** – Confidence interval

Multivariable Mixed Effect Logistic Regression Models

Tables 2a and 2b shows the results from the multivariable GLMM for odds of emergency route to diagnosis of DLBCL and FL, respectively. For both DLBCL and FL, under complete case analysis, we found that for any given CCG, the presence of a comorbidity was associated with emergency route to diagnosis: the association was largest amongst those with a comorbidity status of two or more (**table 2a and 2b**). Living in more deprived areas was strongly associated with emergency route to diagnosis.

After multiple imputation (**tables 2a and 2b**), there were similar conclusions to the complete case analysis. Amongst patients from the same CCG, having a comorbidity score of 2 or more, compared to no comorbidity, was strongly associated with an emergency route to diagnosis (DLBCL: OR 1.56, CI 1.40 – 1.73; FL: OR 1.80, CI 1.45 – 2.23). There was weak evidence of a trend for deprivation and comorbidity index amongst DLBCL ($p = 0.054$ and $p = 0.060$, respectively); however, there was no evidence of a trend amongst FL ($p = 0.206$ and $p = 0.113$, respectively).

Using a mixture Chi-square tests with 0 and 1 degree of freedom (i.e. half the p-value from a chi-square with 1 degree of freedom), we found strong evidence of between-CCG variability in the odds of emergency route to diagnosis (DLBCL: $p < 0.005$; FL: $p < 0.001$). The variance of the CCG random effects of the models for DLBCL and FL indicated some heterogeneity between CCGs in routes to diagnosis.

Table 2a: Multivariable GLMM for the odds of emergency route to diagnosis in (a) complete case analysis, (b) multiple imputation amongst patients (n=30,078) diagnosed with **Diffuse Large B-cell lymphoma** in England during 2005-2013

	(a) Complete case analysis (n=22,832)			(b) After multiple imputation (n=30,078)		
	OR	95% CI	P value	OR	95% CI	P value
Age*	1.03	1.02 – 1.04	0.002	1.05	1.04 – 1.06	<0.001
Gender						
<i>Male</i>	Ref	Ref		Ref	Ref	
<i>Female</i>	0.95	0.90 – 1.01	0.082	0.95	0.91 – 1.00	0.061
Ethnicity						
<i>White</i>	Ref	Ref		Ref	Ref	
<i>Minority</i>	1.44	1.28 – 1.62	<0.001	1.42	1.26 – 1.60	<0.001
Deprivation						
<i>Least deprived</i>	Ref	Ref		Ref	Ref	
2	1.14	1.04 – 1.24	0.003	1.13	1.05 – 1.22	0.001
3	1.18	1.08 – 1.29	<0.001	1.17	1.08 – 1.27	<0.001
4	1.23	1.12 – 1.34	<0.001	1.23	1.14 – 1.34	<0.001
<i>Most deprived</i>	1.24	1.13 – 1.36	<0.001	1.32	1.21 – 1.43	<0.001
Comorbidity						
<i>None</i>	Ref	Ref		Ref	Ref	
<i>One</i>	1.26	1.12 – 1.41	<0.001	1.27	1.14 – 1.41	<0.001
<i>Multimorbidity</i>	1.58	1.41 – 1.78	<0.001	1.56	1.40 – 1.73	<0.001
Variance of RE	0.007			0.008		
(Standard error)	(0.09)	-	-	(0.09)	-	-

*Increase in odds of emergency route to diagnosis for each 10-year increase in age at diagnosis
OR – odds ratio. **CI** – confidence interval

Table 2b: Multivariable GLMM for the odds of emergency route to diagnosis in (a) complete case analysis, (b) multiple imputation amongst patients (n=15,551) diagnosed with **Follicular lymphoma** in England during 2005-2013

	(a) Complete case analysis (n=11,445)			(b) After multiple imputation (n=15,551)		
	OR	95% CI	P value	OR	95% CI	P value
Age*	1.15	1.12 – 1.17	<0.001	1.17	1.15 – 1.19	<0.001
Gender						
<i>Male</i>	Ref	Ref		Ref	Ref	
<i>Female</i>	0.76	0.68 – 0.85	<0.001	0.80	0.73 – 0.89	<0.001
Ethnicity						
<i>White</i>	Ref	Ref		Ref	Ref	
<i>Minority</i>	1.03	0.80 – 1.32	0.835	1.03	0.81 – 1.29	0.833
Deprivation						
<i>Least deprived</i>	Ref	Ref		Ref	Ref	
2	1.16	0.98 – 1.38	0.084	1.11	0.95 – 1.29	0.190
3	1.09	0.92 – 1.30	0.312	1.07	0.92 – 1.24	0.396
4	1.42	1.20 – 1.69	<0.001	1.38	1.18 – 1.61	<0.001
<i>Most deprived</i>	1.38	1.14 – 1.66	<0.001	1.39	1.18 – 1.64	<0.001
Comorbidity						
<i>None</i>	Ref	Ref		Ref	Ref	
<i>One</i>	1.18	0.92 – 1.51	0.190	1.19	0.94 – 1.49	0.143
<i>Multimorbidity</i>	1.78	1.40 – 2.26	<0.001	1.80	1.45 – 2.23	<0.001
Variance of RE	0.016	-	-	0.017	-	-
(Standard error)	(0.128)			(0.130)		

*Increase in odds of emergency route to diagnosis for each 10-year increase in age at diagnosis

OR – odds ratio. CI – confidence interval

We graphically illustrate, from our analysis accounting for both clustering and missing data, the Empirical Bayes (EB) estimates of the CCG random effects for odds of emergency route to diagnosis (**figures 1 and 2**). These are used to explore the between-CCG variability. A positive EB estimate indicated higher probability of emergency route to diagnosis for a patient from that CCG in comparison to a patient who has similar observed characteristics but from a CCG with either a less positive, or a negative EB estimate. For DLBCL, there are possibly a few outlying CCGs with the lowest probabilities, and possibly an outlying one with the highest probability. For FL, there are possibly a few outlying ones with the highest probabilities. To explore possible patterns, the size of the markers were weighted by the population density for the respective CCG and have a lighter shade for a higher proportion of missing records of route to diagnosis.

For DLBCL (**figure 1**), the results show a slight pattern such that there were more CCGs with a larger population density (larger-sized markers) that had a higher probability for their patients being diagnosed through an emergency route to diagnosis (markers with EB estimates above 0). There was no apparent pattern for patients with FL (**figure 2**).

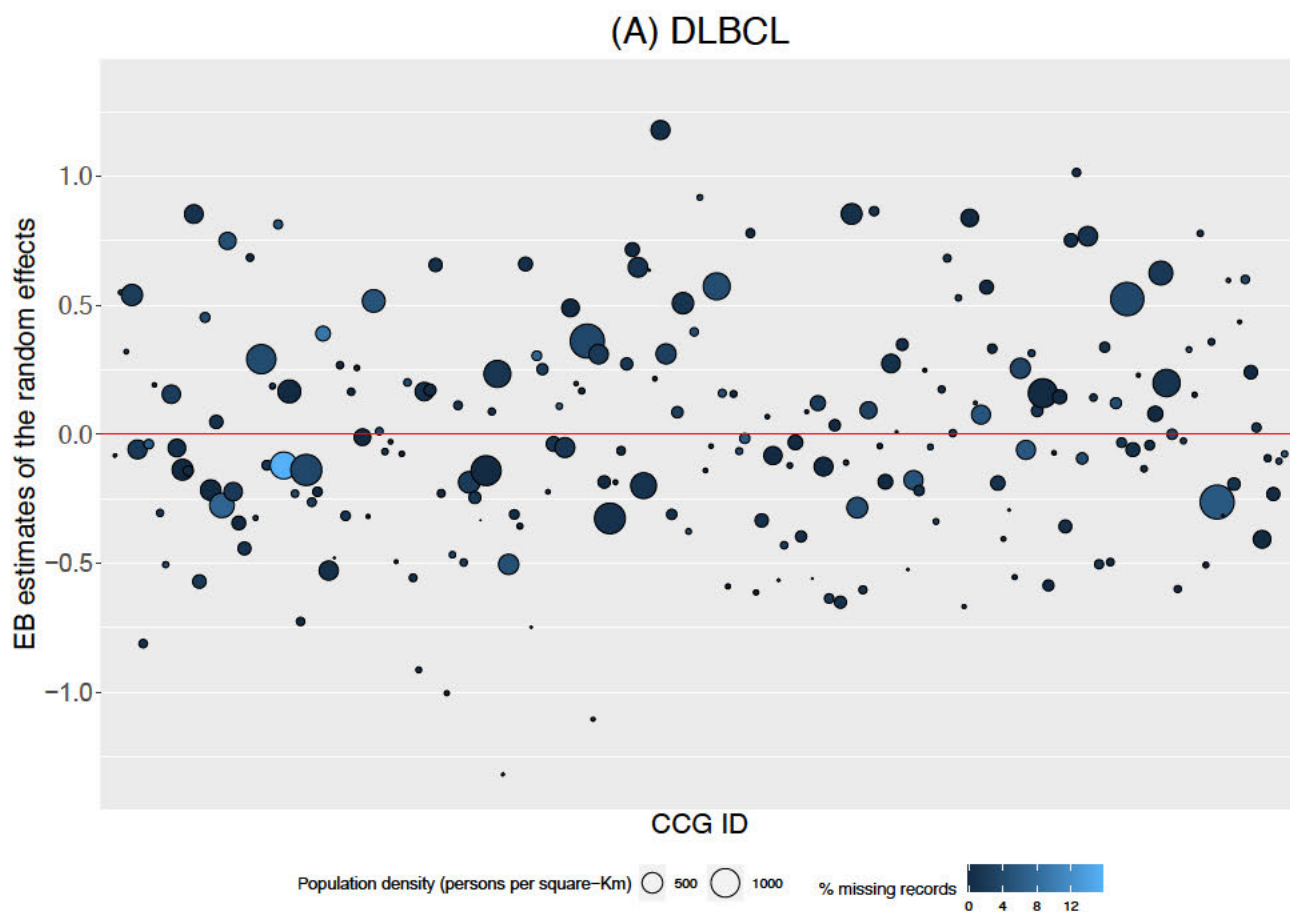


Figure 1 Empirical Bayes estimates of the random effects from the model for route to diagnosis, by each Clinical Commissioning Group amongst patients (n=30,078) diagnosed with **Diffuse Large B-cell lymphoma** in England during 2005-2013

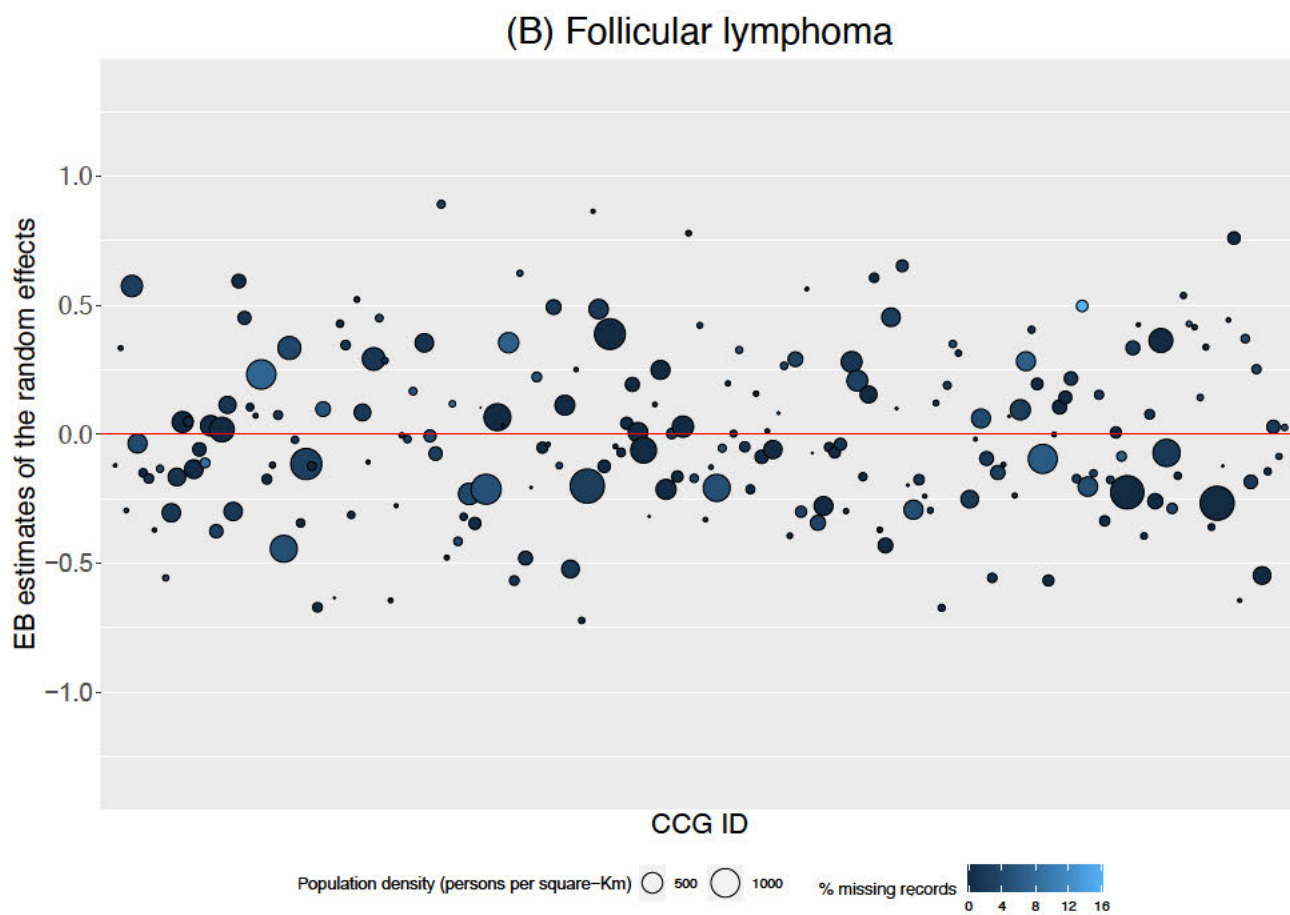


Figure 2 Empirical Bayes estimates of the random effects from the model for route to diagnosis, by each Clinical Commissioning Group amongst patients (n=15,551) diagnosed with **Follicular lymphoma** in England during 2005-2013

Discussion

We aimed to assess the association between comorbidity status and a marker of diagnostic delay (route to diagnosis), amongst patients diagnosed with non-Hodgkin lymphoma, adjusting for patient and healthcare pathway characteristics.

We found that comorbidity status was significantly associated with emergency route to diagnosis, after adjusting for age, gender, ethnicity, and deprivation and accounting for clustering due to CCG did not explain the relative difference. The more severe the comorbidity score, and those living in more deprived areas, increased the odds of emergency route to diagnosis. Our results are consistent with previous findings of an increase in the probability of emergency route to diagnosis,^{6,52} and, in other countries and for other cancers, comorbidities were associated with diagnostic delay.⁵³ Similar results were found amongst studies investigating colon cancer.^{54,55} Since the proportion of patients with emergency route remains stable over calendar time, this phenomenon is not thought to be time-dependent.

Deprivation level was a strong independent predictor of route to diagnosis after adjusting for comorbidity and other factors (table 2a and 2b); however, accounting for clustering increased the strength of the association for patients living in more deprived areas. This suggests that the difference in diagnostic delays between deprivation groups is partly explained by unobserved, and possibly unmeasured, characteristics of CCGs. A characteristic of CCGs, not explored in this study but for other cancers, could be accessibility to the healthcare system (e.g. accessibility to a GP appointment).⁵⁶ Previous studies⁵⁷ have found delays in diagnosis since first symptoms and suggested introducing rapid access to lymph node diagnostic clinics⁵⁸ and providing: less variability in the number of GP appointments attended before a diagnosis,^{59,60} clearer definitions of symptoms,⁶¹ and appropriate patient-oriented information when previous investigations rule out cancer.¹⁵ These unmeasured characteristics of CCGs could explain the large between-CCG variation in outcomes. In the United States, and for other malignancies, physician supply is associated with early detection of breast cancer,⁶² and higher primary care physician density is associated with a lower incidence of late-stage colorectal cancer.⁶³

Contrary to the assurances of a universal healthcare system, such as the NHS, our results suggest inequitable access to healthcare services between CCGs (i.e., more densely-populated CCGs appear to have patients with greater chance of diagnostic delay compared to less densely-populated CCGs). Patients diagnosed through emergency route are patients that either could not access a GP appointment or the GP appointment was inconclusive: during this waiting time the cancer can progress and the patient admitted themselves to emergency department. Inequalities may be due to a combination of competing demand and lack of clinical guidance regarding symptoms. However, lack of clinical guidance would be a nondifferential misclassification and this would not explain the inequalities in emergency route amongst patient characteristics.

Our results challenge previous research that did not find evidence of a difference in diagnostic delay between deprivation levels using unadjusted analyses; although, previous studies were based on a smaller sample size that were potentially underpowered in comparison to our study.⁶ We highlight that deprivation is predictive of diagnostic route if analyses do not account for CCGs that widely differ, among other dimensions, in healthcare provision.⁶⁴ Furthermore, late lymphoma stage at diagnosis seems associated with poorer survival. Evidence is limited due to the extended use of the FL and DLBCL International Prognostic Indices (FLIPI and IPI, respectively) and for lymphoma prognosis and survival outcomes. The indices, in addition to the lymphoma stage, integrates other prognostic factors such as serum lactate dehydrogenase, the number of nodal site involvement, patient ages, and haemoglobin. Evidence shows that a higher index score, and thus higher stage, is associated with poorer health outcomes and survival: highlighting the necessity of prompt management among patients at advanced stage.⁶⁵

We graphically illustrated that patients living in CCGs with more dense populations have a higher probability of emergency route to diagnosis. To our knowledge, there is yet no research into the relationship between population density and diagnostic delay of cancer in England. This study shows that NHL patients living in CCGs with higher population densities have a higher probability of emergency route to diagnosis. On one hand, deprivation tends to be correlated with high population density in England,⁶⁶ and is also associated with higher use of emergency services.⁶⁷ On the other hand, population density is independently associated with high emergency calls.⁶⁸ This could be because highly dense areas accumulate high demands that are not completely

covered by available healthcare resources; accordingly, this demand could be exacerbated by the association between deprivation and the prevalence of comorbidities. This association has not been well explored, but it is likely that cancers other than NHL are affected by the association between prevalence of emergency route to diagnosis and population density. Further research should be conducted to determine the need for greater availability of healthcare services in more populated areas.

Furthermore, there will be differences in availability and specialisation of cancer-specific resources between CCGs. For example, a CCG may have a specialised centre for breast cancer but not for another cancer. Additional analyses are needed to provide a full interpretation of these results. Densely populated areas may be associated with populations from less favourable backgrounds and potentially higher pressure on the healthcare system. CCGs were established from the Health and Social Care Act 2012 and replaced Primary Care Trusts (PCTs). However, CCGs and PCTs were constructed based on administrative boundaries, and the population size of CCGs are similar to the PCTs they replaced. Since 2013, the number of CCGs have reduced due to mergers,⁶⁹ and the proportion of late-staged lymphomas has increased,⁷⁰ possibly indicating competition for healthcare services.

Our study is strengthened by the large population-based sample capturing all patients with a diagnosis of DLBCL and FL between 2005 and 2013. To date, this is the largest study of diagnostic delay amongst patients with NHL. Patients were diagnosed according to the latest (ICD-O-3) well-defined WHO cancer classifications, and through a linkage of databases we obtained reliable information on comorbidity diagnosis prior to, and likely independent of, the cancer. The objective data sources provide information on patients that is gathered prospectively, preventing differential misclassification.

Despite the lack of well-defined guidance on which comorbidity index is the gold-standard depending on the setting of study, Charlson comorbidity index (CCI) is one of the most commonly used comorbidity indices in population-based cancer epidemiology.⁷¹ We used the Royal College of Surgeons' adaptation of the CCI, which provides a cancer-specific comorbidity indicator, and is advantageous in comparison to other indices that

measure underlying comorbidities as independent from each other.^{32,71,72} Computed algorithms were used to define comorbidity status, which strengthens the reliability of this study.³⁵

In this study, we had missing data in two dimensions: route to diagnosis (the outcome) and explanatory variables. Missing data in outcomes present less complexity when using a likelihood-based analysis such as a generalized linear mixed model, as the ignorability property assures validity of results from analysis of the complete cases, under a missing at random mechanism.^{42,47} With missing data additionally in explanatory variables, analyses are more complex, as multiple imputation is in general needed to achieve validity of results under a missing at random mechanism, if the outcome is included in the missingness mechanism for these variables. Research in missing data has shown that multiple imputation has potential to mitigate bias and loss of efficiency; whether multiple imputation provides gains over a complete case analysis cannot be simply determined from the proportion of incomplete cases in a single variable. Indeed, potential benefits from multiple imputation depend on factors such as whether missing data occur in the explanatory variable of interest or covariates, and interrelationships between the variables.⁷³ Lee and Carlin (2012)⁷³ and White and Carlin (2010)⁷⁴ have highlighted the importance of conducting both a complete case analysis and an analysis after multiple imputation, and to carefully compare results. We used the latent normal joint modelling multiple imputation approach under a missing at random assumption to account for the missing ethnicity and route to diagnosis. This approach allows imputation of a mix of variable types, while accounting for multilevel structures arising from clustering of patients.^{47,75,76} As with all missing data problems, it is impossible to distinguish between a missing at random and a missing not at random mechanism based on the observed data.^{47,77-79} Follow-up work will therefore involve assessing sensitivity of our results to departures from the missing at random mechanism, by imputing under a missing not at random assumption.

A limitation of this study is that route to diagnosis does not entirely encapsulate the patient's multifaceted experiences along the healthcare pathway prior to a cancer diagnosis. Information on performance status and education were not available but may have contributed to differences in diagnostic delay. Firstly, distinct from having a comorbidity, performance status measures the patient's ability to carry out everyday tasks, such as reaching the healthcare system, which may contribute to diagnostic delay.⁶ Secondly, the low average time

allocated for each GP appointment requires the patient to use the English language efficiently and describe important symptoms in a concise manner, which may hasten the cancer diagnosis.⁸⁰

Conclusion

Patients with DLBCL or FL are more likely to experience an emergency route to diagnosis if they have an underlying comorbidity. Differences in diagnostic delay indicators between deprivation levels are minimally explained by comorbidity status, and are further explained by differences in the healthcare provisions between clinical commissioning groups (CCG). DLBCL patients living in CCGs with higher population densities have a higher probability of emergency route to diagnosis.

References

1. Shankland, K. R., Armitage, J. O. & Hancock, B. W. Non-Hodgkin lymphoma. *Lancet* **380**, 848–857 (2012).
2. Haematological Malignancy Research Network. Incidence of non-Hodgkin lymphoma. (2016).
3. Rachet, B., Mitry, E., Shah, A., Cooper, N. & Coleman, M. P. Survival from non-Hodgkin lymphoma in England and Wales up to 2001. *Br. J. Cancer* **99**, S104–S106 (2008).
4. Exarchakou, A., Rachet, B., Belot, A., Maringe, C. & Coleman, M. P. Impact of national cancer policies on cancer survival trends and socioeconomic inequalities in England, 1996-2013: population based study. *BMJ* **360**, k764–k764 (2018).
5. Allemani, C. *et al.* Global surveillance of trends in cancer survival 2000–14 (CONCORD-3): analysis of individual records for 37,513,025 patients diagnosed with one of 18 cancers from 322 population-based registries in 71 countries. *Lancet* **391**, 1023–1075 (2018).
6. Kane, E. *et al.* Emergency admission and survival from aggressive non-Hodgkin lymphoma: A report from the UK's population-based Haematological Malignancy Research Network. *Eur. J. Cancer* **78**, 53–60 (2017).
7. Department of Health. *The NHS cancer plan: a plan for investment: a plan for reform.* (Department of Health, 2000).
8. Department of Health. *Improving Outcomes: a strategy for cancer.* (2011).
9. National Institute for Health and Care Excellence. *Improving outcomes in haematological cancers: the manual.* (2003).
10. National Institute for Health and Care Excellence. *Haematological cancers: improving outcomes.* (2016).
11. Elliss-Brookes, L. *et al.* Routes to diagnosis for cancer - determining the patient journey using multiple routine data sets. *Br J Cancer* **107**, 1220–1226 (2012).
12. Hamilton, W. Emergency admissions of cancer as a marker of diagnostic delay. *Br J Cancer* **107**, 1205–1206 (2012).
13. Gurney, J., Sarfati, D. & Stanley, J. The impact of patient comorbidity on cancer stage at diagnosis. *Br.*

- J. Cancer* **113**, 1375–1380 (2015).
14. Sarfati, D., Koczwara, B. & Jackson, C. The impact of comorbidity on cancer and its treatment. *CA. Cancer J. Clin.* **66**, 337–350 (2016).
 15. Salika, T., Lyratzopoulos, G., Whitaker, K. L., Waller, J. & Renzi, C. Do comorbidities influence help-seeking for cancer alarm symptoms? A population-based survey in England. *J. Public Health (Bangkok)*. **40**, 340–349 (2017).
 16. Mitri, J., Castillo, J. & Pittas, A. G. Diabetes and risk of Non-Hodgkin’s lymphoma: a meta-analysis of observational studies. *Diabetes Care* **31**, 2391–2397 (2008).
 17. Gowda, R. M. & Khan, I. A. Clinical Perspectives of Primary Cardiac Lymphoma. *Angiology* **54**, 599–604 (2003).
 18. Kim, J. H. *et al.* Primary Pulmonary Non-Hodgkin’s Lymphoma. *Jpn. J. Clin. Oncol.* **34**, 510–514 (2004).
 19. Fowler, H. *et al.* Comorbidity prevalence among cancer patients: a population-based cohort study of four cancers. *BMC Cancer* **20**, 2 (2020).
 20. Hester, L., Park, S. I. & Lund, J. L. Patterns of comorbidity among older U.S. patients with non-Hodgkin lymphoma. *J. Clin. Oncol.* **34**, 304 (2016).
 21. NHS Choices. *The principles and values of the NHS in England*.
 22. Smith, A. *et al.* Lymphoma incidence, survival and prevalence 2004-2014: sub-type analyses from the UK’s Haematological Malignancy Research Network. *Br J Cancer* **112**, 1575–1584 (2015).
 23. Smith, A. *et al.* Impact of age and socioeconomic status on treatment and survival from aggressive lymphoma: a UK population-based study of diffuse large B-cell lymphoma. *Cancer Epidemiol* **39**, 1103–1112 (2015).
 24. Office for National Statistics. *Index of cancer survival for Clinical Commissioning Groups in England: adults diagnosed 2001 to 2016 and followed up to 2017*. (2019).
 25. London School of Hygiene and Tropical Medicine. Expert comment on ONS cancer survival bulletins. Available at: https://www.lshtm.ac.uk/newsevents/news/2014/comment_cancer_survival.html. (Accessed: 30th March 2020)
 26. International Agency for Research on Cancer. International Classification of Diseases for Oncology.

- (2013). Available at: <http://codes.iarc.fr/>. (Accessed: 4th October 2019)
27. Fritz, A. *et al.* *International Classification of Diseases for Oncology*. (World Health Organisation, 2000).
 28. Campo, E. *et al.* The 2008 WHO classification of lymphoid neoplasms and beyond: evolving concepts and practical applications. *Blood* **117**, 5019–5032 (2011).
 29. Public Health England. National Cancer Registration and Analysis Service. (2019). Available at: <https://www.gov.uk/guidance/national-cancer-registration-and-analysis-service-ncras#cancer-registration>. (Accessed: 30th March 2020)
 30. NHS Digital. Hospital Episode Statistics. (2015). Available at: 2015. (Accessed: 4th October 2019)
 31. gov.uk. National Cancer Registry and Analysis Service. (2017). Available at: <https://www.gov.uk/guidance/national-cancer-registration-and-analysis-service-ncras>. (Accessed: 4th October 2019)
 32. Charlson, M. E., Pompei, P., Ales, K. L. & MacKenzie, C. R. A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation. *J. Chronic Dis.* **40**, 373–383 (1987).
 33. Porta, M. *A Dictionary of Epidemiology*. (Oxford University Press, 2014).
doi:10.1093/acref/9780195314496.001.0001
 34. Feinstein, A. R. The pre-therapeutic classification of co-morbidity in chronic disease. *J Chronic Dis* **23**, 455–468 (1970).
 35. Maringe, C., Fowler, H., Rachet, B. & Luque-Fernandez, M. A. Reproducibility, reliability and validity of population-based administrative health data for the assessment of cancer non-related comorbidities. *PLoS One* **12**, e0172814 (2017).
 36. Armitage, J. N. & van der Meulen, J. H. Identifying co-morbidity in surgical patients using administrative data with the Royal College of Surgeons Charlson Score. *Br J Surg* **97**, 772–781 (2010).
 37. Lister, T. A. *et al.* Report of a committee convened to discuss the evaluation and staging of patients with Hodgkin's disease: Cotswolds meeting. *J Clin Oncol* **7**, 1630–1636 (1989).
 38. European Society for Medical Oncology. European Clinical Practice Guidelines: Haematological Malignancies. (2019). Available at: <https://www.esmo.org/Guidelines/Haematological-Malignancies>. (Accessed: 6th October 2019)

39. National Health Service: data dictionary. Lower Super Output Area. (2018). Available at: https://www.datadictionary.nhs.uk/data_dictionary/nhs_business_definitions/l/lower_layer_super_output_area_de.asp?shownav=1. (Accessed: 4th October 2019)
40. gov.uk. Indices of Multiple Deprivation. (2015). Available at: <https://www.gov.uk/government/statistics/english-indices-of-deprivation-2015>. (Accessed: 4th October 2019)
41. Office for National Statistics. Clinical Commissioning Group population estimates. (2020). Available at: <https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/datasets/clinicalcommissioninggroupmidyearpopulationestimates>. (Accessed: 30th March 2020)
42. Molenberghs, G. & Verbeke, G. *Models for Discrete Longitudinal Data*. (Springer-Verlag New York, 2005).
43. Agresti, A. *Categorical Data Analysis*. (John Wiley & Sons, Inc., 2002).
44. Fitzmaurice, G. M., Laird, N. M. & Ware, J. H. *Applied Longitudinal Analysis*. (John Wiley & Sons, Inc., 2011).
45. Rabe-Hesketh, S. & Skrondal, A. *Multilevel and Longitudinal Modelling Using Stata, Volume II: Categorical Responses, Counts, and Survival*. (Stata Press, 2012).
46. Verbeke, G. & Molenberghs, G. *Linear Mixed Models for Longitudinal Data*. (Springer-Verlag New York, 2000).
47. Carpenter, J. R. & Kenward, M. G. *Multiple Imputation and Its Application*. (John Wiley & Sons, Ltd, 2013).
48. Haematological Malignancy Research Network. Survival of non-Hodgkin lymphoma. (2016). Available at: <https://www.hmrn.org/statistics/survival>. (Accessed: 7th May 2020)
49. Little, R. J. & Rubin, D. B. *Statistical Analysis with Missing Data*. (John Wiley & Sons, Inc., 1987).
50. Rubin, D. B. *Multiple imputation for nonresponse in surveys*. (Wiley, 1987).
51. Quartagno, M. & Carpenter, J. R. jomo: A package for multilevel joint modeling multiple imputation. (2016).
52. National Cancer Intelligence Network. Routes to diagnosis. 2006-2015 (2016). Available at:

http://www.ncin.org.uk/publications/routes_to_diagnosis. (Accessed: 1st April 2020)

53. Nikonova, A., Guirguis, H. R., Buckstein, R. & Cheung, M. C. Predictors of delay in diagnosis and treatment in diffuse large B-cell lymphoma and impact on survival. *Br. J. Haematol.* **168**, 492–500 (2015).
54. Renzi, C., Lyratzopoulos, G., Hamilton, W., Maringe, C. & Rachet, B. Contrasting effects of comorbidities on emergency colon cancer diagnosis: a longitudinal data-linkage study in England. *BMC Health Serv. Res.* **19**, 311 (2019).
55. Renzi, C. *et al.* Do colorectal cancer patients diagnosed as an emergency differ from non-emergency patients in their consultation patterns and symptoms? A longitudinal data-linkage study in England. *Br. J. Cancer* **115**, 866 (2016).
56. Jones, A. P. *et al.* Travel times to health care and survival from cancers in Northern England. *Eur. J. Cancer* **44**, 269–274 (2008).
57. Howell, D. A. *et al.* Time-to-diagnosis and symptoms of myeloma, lymphomas and leukaemias: a report from the Haematological Malignancy Research Network. *BMC Blood Disord.* **13**, 9 (2013).
58. Chau, I. *et al.* Rapid access multidisciplinary lymph node diagnostic clinic: analysis of 550 patients. *Br. J. Cancer* **88**, 354–361 (2003).
59. Howell, D. A., Smith, A. G. & Roman, E. Lymphoma: variations in time to diagnosis and treatment. *Eur. J. Cancer Care (Engl.)* **15**, 272–278 (2006).
60. Lyratzopoulos, G., Abel, G. A., McPhail, S., Neal, R. D. & Rubin, G. P. Measures of promptness of cancer diagnosis in primary care: secondary analysis of national audit data on patients with 18 common and rarer cancers. *Br J Cancer* **108**, 686–690 (2013).
61. Howell, D. A., Smith, A. G. & Roman, E. Help-seeking behaviour in patients with lymphoma. *Eur. J. Cancer Care (Engl.)* **17**, 394–403 (2008).
62. Ferrante, J. M., Gonzalez, E. C., Pal, N. & Roetzheim, R. G. Effects of Physician Supply on Early Detection of Breast Cancer. *J. Am. Board Fam. Pract.* **13**, 408 LP – 414 (2000).
63. Ananthakrishnan, A. N., Hoffmann, R. G. & Saeian, K. Higher Physician Density is Associated with Lower Incidence of Late-stage Colorectal Cancer. *J. Gen. Intern. Med.* **25**, 1164–1171 (2010).
64. Cookson, R., Propper, C., Asaria, M. & Raine, R. Socio-Economic Inequalities in Health Care in

- England. *Fisc. Stud.* **37**, 371–403 (2016).
65. Lee, S. F. & Luque-Fernandez, M. A. Prognostic value of lymphocyte-To-monocyte ratio and neutrophil-To-lymphocyte ratio in follicular lymphoma: A retrospective cohort study. *BMJ Open* **7**, (2017).
 66. Venerandi, A., Quattrone, G. & Capra, L. A scalable method to quantify the relationship between urban form and socio-economic indexes. *EPJ Data Sci.* **7**, 4 (2018).
 67. Carlisle, R., Groom, L. M., Avery, A. J., Boot, D. & Earwicker, S. Relation of Out of Hours Activity by General Practice and Accident and Emergency Services with Deprivation in Nottingham: Longitudinal Survey on JSTOR. *BMJ Br. Med. J.* **316**, 520–523 (1998).
 68. Peacock, P. J. & Peacock, J. L. Emergency call work-load, deprivation and population density: An investigation into ambulance services across England. *J. Public Health (Bangkok)*. **28**, 111–115 (2006).
 69. NHS Clinical Commissioners. NHS Clinical Commissioners: About CCGs.
 70. Public Health England. National Disease Registration Service: Staging data in England. (2018). Available at: https://www.cancerdata.nhs.uk/stage_at_diagnosis. (Accessed: 13th July 2020)
 71. Piccirillo, J. F., Tierney, R. M., Costas, I., Grove, L. & Spitznagel Jr., E. L. Prognostic importance of comorbidity in a hospital-based cancer registry. *Jama* **291**, 2441–2447 (2004).
 72. Elixhauser, A., Steiner, C., Harris, D. R. & Coffey, R. M. Comorbidity Measures for Use with Administrative Data. *Med. Care* **36**, 8–27 (1998).
 73. Lee, K. J. & Carlin, J. B. Recovery of information from multiple imputation: a simulation study. *Emerg. Themes Epidemiol.* **9**, 3 (2012).
 74. White, I. R. & Carlin, J. B. Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Stat. Med.* **29**, 2920–2931 (2010).
 75. Carpenter, J., Goldstein, H. & Kenward, M. REALCOM-IMPUTE software for multilevel multiple imputation with mixed response types. *J. Stat. Softw.* **45**, (2011).
 76. Quartagno, M. & Carpenter, J. R. Multiple imputation for discrete data: Evaluation of the joint latent normal model. *Biom. J.* **61**, 1003–1019 (2019).
 77. Molenberghs, G., Beunckens, C., Sotto, C. & Kenward, M. G. Every Missingness Not at Random

Model Has a Missingness at Random Counterpart with Equal Fit. *J. R. Stat. Soc. Ser. B (Statistical Methodol.* **70**, 371–388 (2008).

78. Verbeke, G. & Molenberghs, G. Arbitrariness of models for augmented and coarse data, with emphasis on incomplete data and random effects models. *Stat. Modelling* **10**, (2010).
79. Molenberghs, G., Njagi, E., Kenward, M. & Verbeke, G. Enriched-Data Problems and Essential Non-Identifiability. *Int. J. Stat. Med. Res.* 16–44 (2012). doi:10.6000/1929-6029.2012.01.01.02
80. Swann, R., Lyratzopoulos, G., Rubin, G., Pickworth, E. & McPhail, S. The frequency, nature and impact of GP-assessed avoidable delays in a population-based cohort of cancer patients. *Cancer Epidemiol.* **64**, 101617 (2020).

Appendix

Supplementary Table S1: Comorbidities and their diagnostic ICD-10 codes

Comorbidity	ICD-10
Myocardial infarction	I21.x, I22.x, I25.2
Congestive heart failure	I11.0, I13.0, I13.2, I25.5, I42.0, I42.5–I42.9, I43.x, I50.x, P29.0
Peripheral vascular disease	I70.x, I71.x, I73.1, I73.8, I73.9, I77.1, I79.0, I79.2, K55.1, K55.8, K55.9, Z95.8, Z95.9
Cerebrovascular disease	G45.x, G46.x, H34.0, I60.x–I69.x
Dementia	F00.x–F03.x, F05.1, G30.x, G31.1
Chronic obstructive pulmonary disease	I27.9, J40.x–J47.x, J60.x–J67.x, J68.4, J70.1, J70.3
Rheumatic disease	M05.x, M06.x, M31.5, M32.x–M34.x, M35.1, M35.3, M36.0
Liver disease	B18.x, K70.0–K70.3, K70.9, K71.3–K71.5, K71.7, K73.x, K74.x, K76.0, K76.2–K76.4, K76.8, K76.9, Z94.4, K71.1, K72.1, K72.9, K76.5, K76.6, K76.7, I85.0, I85.9, I86.4, I98.2, K70.4, E10.0, E10.1, E10.6, E10.8, E10.9, E11.0, E11.1, E11.6, E11.8, E11.9, E12.0, E12.1, E12.6, E12.8, E12.9, E13.0, E13.1, E13.6, E13.8, E13.9, E14.0, E14.1, E14.6, E14.8, E14.9
Diabetes without chronic complication	E10.7, E11.2–E11.5, E11.7, E12.2–E12.5, E12.7, E13.2–E13.5, E13.7, E14.2–E14.5, E14.7
Diabetes with chronic complication	G04.1, G11.4, G80.1, G80.2, G81.x, G82.x, G83.0–G83.4, G83.9
Hemiplegia or paraplegia	I12.0, I13.1, N03.2–N03.7, N05.2–N05.7, N18.x, N19.x, N25.0, Z49.0–Z49.2, Z94.0, Z99.2
Renal disease	B20.x–B22.x, B24.x
AIDS/HIV	

ICD-10: International Classification of Diseases, 10th Revision

Diabetes with/without chronic complication is combined in the RCS Charlson Comorbidity Score

Supplementary Table S2: Distribution of non-Hodgkin lymphoma subtypes for patients diagnosed from 2005-2013, with respective morphology and topography ICD-O-3 codes. DLBCL (index 4) and Follicular (index 6) lymphomas were included in this study.

Index	Site group (subtype)	Grade	Topography	Morphology	n	%
1	CLL/SLL*	Indolent	C82.0-C85.9	9670, 9823	4,043	4.78
2	Waldenstrom macroglobulinemia	Indolent	C82.0-C85.9	9761	2,453	2.90
3	Mantle cell	Indolent	C82.0-C85.9	9673	3,549	4.20
4	Diffuse large B-cell	Aggressive	C82.0-C85.9	9680, 9688, 9737-9738	30,750	36.39
5	Burkitt	Aggressive	C82.0-C85.9	9687, 9826	1,077	1.27
6	Follicular	Indolent	C82.0-C85.9	9690-9691, 9695, 9698	15,624	18.49
7	Mature T-cell	Aggressive	C82.0-C85.9	9702	6,066	7.18
8	Marginal zone B-cell	Indolent	C82.0-C85.9	9689, 9699, 9760, 9764, 9699	4,615	5.46
9	Not Otherwise Specified	n/a	C82.0-C85.9	9591, 9675, 9735	10,308	12.20
10	Other***	n/a	C82.0-C85.9	9591, 9675, 9735	6,019	7.12
Total					84,504	100.00**

n/a – not applicable; there was no subtype information
* Chronic lymphocytic leukaemia/Small-cell lymphocytic lymphoma
** Percentages may not equate to 100.0% due to rounding
*** The morphology code specifies these patients are diagnosed with NHL. However, the description states other ; these patients are classified similarly to Not Otherwise Specified .

Supplementary Table S3: Summary statistics of comorbidity amongst patients diagnosed with Diffuse Large B-cell lymphoma (n=30,078) or Follicular lymphoma (n=15,551) in England during 2005-2013.

	Diffuse Large B-cell Lymphoma			Follicular Lymphoma		
	None	Comorbidity	Multi-morbidity	None	Comorbidity	Multi-morbidity
Age (y, SD)	66.9 (15.2)	72.5 (12.6)	73.0 (12.8)	63.4 (13.6)	70.2 (11.7)	72.3 (11.0)
Gender						
<i>Male</i>	14,470 (53.9)	815 (51.4)	986 (60.7)	6,792 (47.3)	281 (43.8)	261 (48.8)
<i>Female</i>	12,398 (46.1)	770 (48.6)	639 (39.3)	7,582 (52.8)	361 (56.2)	274 (51.2)
Ethnicity*						
<i>White</i>	19,418 (94.1)	1,236 (96.2)	1,204 (92.1)	10,165 (94.9)	514 (96.4)	397 (92.8)
<i>Other</i>	1,218 (5.9)	49 (3.8)	103 (7.9)	550 (5.1)	19 (3.6)	31 (7.2)
<i>Missing</i>	6,232 (23.2)	300 (18.9)	318 (19.6)	3,659 (25.5)	109 (17.0)	107 (20.0)
Deprivation						
<i>Least</i>	5,808 (21.6)	300 (18.9)	262 (16.1)	3,358 (23.4)	113 (17.6)	81 (15.1)
2	6,035 (22.5)	323 (20.4)	344 (21.2)	3,314 (23.1)	127 (19.8)	84 (15.7)
3	5,616 (20.9)	334 (21.1)	333 (20.5)	3,040 (21.2)	139 (21.7)	123 (23.0)
4	5,223 (19.4)	343 (21.6)	344 (21.2)	2,671 (18.6)	129 (20.1)	132 (24.7)
<i>Most</i>	4,186 (15.6)	285 (18.0)	342 (21.1)	1,991 (13.9)	134 (20.9)	115 (21.5)

Percentages may not sum to 100.0% due to rounding

* Percentages are calculated based on observed data

A.6 R code for the Approximate F-test of Inference for Vector β

The following code was written to test for non-linear and time-dependent effects of parameters in an excess hazard model after performing multiple imputation and combining estimates using Rubin's rules.

```
#####  
# Inference for vector Beta  
#####  
  
# Test for the inclusion of all age parameters  
  
# Obtain the Beta parameters from the MI results  
library(data.table)  
install.packages("janitor")  
library(janitor)  
  
betaMIage <- fit[[1]]$coefficients[22:30]  
betaMIage <- data.frame(betaMIage)  
betaMIage <- transpose(betaMIage)  
colnames(betaMIage) <- names(fit[[1]]$coefficients[22:30])  
betaMIage  
  
for(m in 2:10) {  
  betaMIage <- rbind(betaMIage, fit[[m]]$coefficients[22:30])  
}  
  
Beta <- data.frame(colMeans(betaMIage))  
Beta <- transpose(Beta)  
colnames(Beta) <- names(fit[[1]]$coefficients[22:30])  
Beta  
  
# Calculate the variance  
# Within imputation variance  
WvarMIage <- fit[[1]]$std.errors[22:30]  
WvarMIage <- data.frame(WvarMIage)  
WvarMIage <- transpose(WvarMIage)  
colnames(WvarMIage) <- names(fit[[1]]$std.errors[22:30])  
WvarMIage  
  
for(m in 2:10) {  
  WvarMIage <- rbind(WvarMIage, fit[[m]]$std.errors[22:30])  
}  
  
WvarMIage <- WvarMIage^2  
  
W <- colMeans(WvarMIage)
```

```

W <- data.frame(W)
W <- transpose(W)
colnames(W) <- names(fit[[1]]$coefficients[22:30])
W

# Between imputation variance
# Difference between the parameter and the average of the parameter
BvarMIage <- NULL
for (m in 1:10) {
  BvarMIage <- rbind(BvarMIage, betaMIage[m,]-Beta)
}
BvarMIage

# Square the difference
BvarMIage <- BvarMIage^2
BvarMIage

# Sum and divide by K-1
B <- NULL
B
for (m in 1:9) {
  B <- rbind(B, sum(BvarMIage[,m])/9)
}
B

B <- data.frame(B)
B <- transpose(B)
colnames(B) <- names(fit[[1]]$coefficients[22:30])
B

# Overall imputation variance
VarMI <- W + (1+(1/10))*B
VarMI

## Calculate F
# Vector of Betas
Beta
mat1 <- c(0.5322972,0.07298481,0.00758753,-0.02203884,
          -0.0588019,-0.2666055,0.2941086,0.1843415,
          0.3396441)
mat1 <- matrix(mat1, ncol=1)
# Vector of variances
VarMI
mat2 <- c(0.0002700874,0.000129292,5.173005e-06,0.0001001533,
          0.000671041,0.000573426,0.004347233,0.009042637,
          0.01121909)
mat2 <- matrix(mat2, nrow=1)

```

```

# Value of F
Fvalue <- (t(mat1)/mat2)%*%mat1
Fvalue

# Calculate the degrees of freedom
# Calculate r
B
mat3 <- c(8.549323e-06,3.021064e-06,8.067058e-08,2.77921e-06,
          8.497425e-07,6.053896e-06,0.0001105716,6.75916e-05,
          0.0001063615)
mat3 <- matrix(mat3, ncol=1)
mat3
W
mat4 <- c(0.0002606831,0.0001259688,5.084267e-06,9.709613e-05,
          0.0006701063,0.0005667667,0.004225604,0.008968287,
          0.01110209)
mat4 <- matrix(mat4, nrow=1)
mat4
r <- ((1/9)*(1+(1/10)))*(sum(t(mat3)/mat4))
r

# Calculate t
t <- 9*(10-1)
t

# Calculate v-prime
v <- 4 + (t-4)*((1+((1-(2/t))/r))^2)
v

# Calculate F statistic
Fstat <- Fvalue/(9*(1+r))
Fstat

## Calculate p-value
pf(Fstat,9,v,lower.tail = F)          # p < 0.001

# Test for the inclusion of all stage parameters

## Obtain the Beta parameters from the MI results
betaMIage <- fit[[1]]$coefficients[c(15:17,31:45)]
betaMIage <- data.frame(betaMIage)
betaMIage <- transpose(betaMIage)
colnames(betaMIage) <- names(fit[[1]]$coefficients[c(15:17,
                                                    31:45)])

```

```

for(m in 2:10) {
  betaMIage <- rbind(betaMIage, fit[[m]]$coefficients[c(15:17,
                                                    31:45)])
}

Beta <- data.frame(colMeans(betaMIage))
Beta <- transpose(Beta)
colnames(Beta) <- names(fit[[1]]$coefficients[c(15:17,31:45)])
Beta

## Calculate the variance
### Within imputation variance
WvarMIage <- fit[[1]]$std.errors[c(15:17,31:45)]
WvarMIage <- data.frame(WvarMIage)
WvarMIage <- transpose(WvarMIage)
colnames(WvarMIage) <- names(fit[[1]]$std.errors[c(15:17,
                                                    31:45)])

for(m in 2:10) {
  WvarMIage <- rbind(WvarMIage, fit[[m]]$std.errors[c(15:17,
                                                    31:45)])
}

WvarMIage <- WvarMIage^2
WvarMIage

W <- colMeans(WvarMIage)
W <- data.frame(W)
W <- transpose(W)
colnames(W) <- names(fit[[1]]$coefficients[c(15:17,31:45)])
W

# Between imputation variance
# Difference between the parameter and the average of the parameter
BvarMIage <- NULL
for (m in 1:10) {
  BvarMIage <- rbind(BvarMIage, betaMIage[m,]-Beta)
}

#### Square the difference
BvarMIage <- BvarMIage^2
BvarMIage

#### Sum and divide by K-1
B <- NULL
for (m in 1:18) {
  B <- rbind(B, sum(BvarMIage[,m])/18)
}

```



```

}

B <- data.frame(B)
B <- transpose(B)
colnames(B) <- names(fit[[1]]$coefficients[c(15:17,31:45)])
B

### Overall imputation variance
VarMI <- W + (1+(1/10))*B

## Calculate F
# Vector of Betas
Beta
mat1 <- c(0.3018712,0.5874981,0.4601315,0.03797913,0.2495526,
          -0.2918091,0.2456404,-0.3605825,-0.08377665,
          0.3497999,-0.1396066,0.2993637,-0.2763175,0.2434308,
          0.6031144,0.2483417,0.3968832,-0.1698742)
mat1 <- matrix(mat1, ncol=1)
# Vector of variances
VarMI
mat2 <- c(0.007567455,0.005088523,0.004325247,0.02019479,
          0.02129246,0.1138764,0.199709,0.220644,0.01764837,
          0.01797975,0.09882063,0.1906772,0.1926898,0.0144931,
          0.01603803,0.08352774,0.1202591,0.1482508)
mat2 <- matrix(mat2, nrow=1)
# Value of F
Fvalue <- (t(mat1)/mat2)%*%mat1
Fvalue

### Calculate the degrees of freedom
# Calculate r
B
mat3 <- c(0.004300584,0.002533225,0.002290794,0.004857226,
          0.007017613,0.02144934,0.04660956,0.06649526,
          0.004907925,0.006178286,0.02376889,0.06483389,
          0.06791431,0.004708364,0.006920125,0.02638452,
          0.02543636,0.04608641)
mat3 <- matrix(mat3, ncol=1)
mat3
W
mat4 <- c(0.002836813,0.002301975,0.001805374,0.01485184,
          0.01357309,0.09028209,0.1484384,0.1474993,0.01224965,
          0.01118363,0.07267485,0.1193599,0.117984,0.009313896,
          0.008425894,0.05450477,0.09227913,0.09755572)
mat4 <- matrix(mat4, nrow=1)
mat4
r <- ((1/18)*(1+(1/10)))*(sum(t(mat3)/mat4))

```

```

r

# Calculate t
t <- 18*(10-1)
t

# Calculate v-prime
v <- 4 + (t-4)*((1+((1-(2/t))/r))^2)
v

### Calculate F statistic
Fstat <- Fvalue/(18*(1+r))
Fstat

### Calculate p-value
pf(Fstat,18,v,lower.tail = F)      # p <0.001

```