Variation in Cancer Outcomes Amongst Children and Young Adults in Yorkshire

Marlous van Laar

Submitted in accordance with the requirements for the degree of Doctor of Philosophy

The University of Leeds
School of Medicine

January 2015

Intellectual Property Statement

The candidate confirms that the work submitted is her own, except where work which has formed part of jointly authored publications has been included. The contribution of the candidate and the other authors to this work has been explicitly indicated below. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others.

Publications

Chapters 4 and 6 contain work based on the following publications:

1 van Laar, M., P.A. McKinney, D.P. Stark, A. Glaser, S.E. Kinsey, I.J. Lewis, S.V. Picton, M. Richards, P.D. Norman, and R.G. Feltbower. (2012). Survival trends of cancer amongst the south Asian and non-south Asian population under 30 years of age in Yorkshire, UK. *Cancer Epidemiology* 36(1): e13-e18.
Attributable content to Marlous van Laar: Data extraction, cleaning and analysis, interpretation of the results and writing of the manuscript.
Contribution of other authors: DPS, AG, SEK, IJL, SVP, MR and RGF contributed to the concept and provided input into the clinical aspects of the work for the background and discussion. PAM, PDN and RGF additionally provided input into the drafting of the manuscript.

2 van Laar, M., D. Greenwood, D. Stark, R.G. Feltbower. (2014). Missing data and survival analysis of central nervous system tumours amongst children and adolescents in Yorkshire, UK, 1990-2009. *Pediatric Blood & Cancer*. 61: S293-S293. Conference Abstract.
Attributable content to Marlous van Laar: Data extraction, cleaning and analysis, interpretation of the results and writing of the abstract.
Contribution of other authors: DG, DPS, and RGF contributed to the concept and proof read the abstract.

3 van Laar, M., D. Greenwood, D. Stark, R.G. Feltbower. (2013). Missing data and survival analysis of central nervous system tumours amongst children and young people in Yorkshire, 1990-2009. *European Journal of Cancer*. 60:3-3. Conference Abstract.
Attributable content to Marlous van Laar: Data extraction, cleaning and analysis, interpretation of the results and writing of the abstract.
Contribution of other authors: DG, DPS, and RGF contributed to the concept and proof read the abstract.

4 van Laar, M., D. Greenwood, D. Stark, R.G. Feltbower. (2011). Multiple imputation and survival analysis: an example using cancer registry data. *Journal of Epidemiology and Community Health* 65:A395-A395. Conference Abstract.
Attributable content to Marlous van Laar: Data extraction, cleaning and analysis, interpretation of the results and writing of the abstract.
Contribution of other authors: DG, DPS, and RGF contributed to the concept and proof read the abstract.

Chapters 4 and 8 contain work based on the following publications:

5 van Laar, M., R.G. Feltbower, C.P. Gale, D.T. Bowen, S.E. Oliver, and A. Glaser. (2014). Cardiovascular sequelae in long-term survivors of young peoples' cancer: a linked cohort study. *British Journal of Cancer* 110(5) 2014:1338-1341.
Attributable content to Marlous van Laar: Data extraction, cleaning and analysis, interpretation of the results and writing of the manuscript.
Contribution of other authors: RGF, CPG, DTB, SEO and AG contributed to the concept and helped draft the manuscript. In addition, CPG helped define which cardiovascular hospital admissions were to be included in the study.

6 Simms A.D., M. van Laar, R.J. Birch, R.G. Feltbower, C.P. Gale, D.T. Bowen, S.E. Oliver, A. Glaser. (2011). Cardiovascular sequelae in long term survivors of childhood and young adult cancer. *European Heart Journal* 32:544-544. Conference Abstract.
Attributable content to Marlous van Laar: Data extraction, cleaning and analysis, interpretation of the results and writing of the abstract.
Contribution of other authors: RGF, RJB, CPG, DTB, SEO and AG contributed to the concept and proof read the abstract. ADS presented the abstract at the conference.

7 van Laar, M., R.J. Birch, C.P. Gale, A. Glaser, D.T. Bowen, S.E. Oliver, R.G. Feltbower. (2010). Cardiovascular sequelae in long term survivors of childhood and young adult cancer. *Pediatric Blood & Cancer* 55:825-825. Conference Abstract.
Attributable content to Marlous van Laar: Data extraction, cleaning and analysis, interpretation of the results and writing of the abstract.
Contribution of other authors: RGF, RJB, CPG, DTB, SEO and AG contributed to the concept and proof read the abstract. ADS presented the abstract at the conference.

Abstract

This study set out to improve upon the quality of research addressing variation in cancer outcomes amongst children and young adults (CYAs) through the novel application of multiple imputation (MI) to the population based Yorkshire cancer register. The study also sought to determine whether there were inequalities in disease severity according to age, ethnicity or deprivation for CYAs with cancer for the first time in the UK and to quantify cardiovascular late effects amongst survivors of CYA cancer based on a unique data linkage approach to hospital admission data.

Key survival inequalities for CYAs with central nervous system (CNS) tumours (n=795), leukaemia (n=912) and germ cell tumours (GCTs) (n=846) between 1990 and 2009 were identified. Teenagers and young adults (TYA) compared to children across all three groups and those of non-white and non-Asian ethnicity diagnosed with CNS tumours had significantly poorer survival. Importantly, these survival inequalities could not be explained by differences in the stage at diagnosis according to age, ethnicity or deprivation. Survival rates from CNS tumours and leukaemia continued to improve over time. These improvements only became evident after using MI to adjust for missingness, which is often ignored by researchers. Despite no observed improvement for GCTs over time, the number of advanced stage tumours at diagnosis decreased significantly for this diagnostic group. For all cancers combined, the long term cardiovascular effects of cancer exist not only for children, but also for TYAs surviving their cancer.

Continued efforts should be made to ensure equal access to clinical trials and improved treatment protocols for TYAs. In addition, children as well as TYAs should be monitored for early signs of cardiovascular disease to maximise cardiovascular health. Finally, ignoring missing data can result in reduced study power and biased estimates, thus researchers should strive to use advanced techniques such as MI to account for missing data.

Acknowledgements

I would like to acknowledge first of all my supervisors Dr Richard Feltbower, Dr Darren Greenwood and Dr Daniel Stark for their support and guidance throughout my PhD. In particular their patience in reading my ‘almost finished’ and ‘not quite there’ chapters over and over again.

My PhD would not have been possible without funding from the Candlelighters Trust for my role as epidemiologist on the Yorkshire register, the support of the Division of Epidemiology and Biostatistics and the hard work of Paula Feltbower in the collection and cross checking of data.

I would also like to thank my colleagues, friends and family, especially Lorna Fraser, Jodie Singh, Claire Keeble and my mum, dad and sister for listening to my worries and stresses, supporting me in keeping me motivated, giving me advice and continuously asking me when I would finish!

Finally, I would like to thank Ben for his patience and support, the cups of tea and the background noise of warfare on the playstation while waiting, waiting and waiting until we could head out for weekend walks in the dales again.

Contents

List of Figures

List of Tables

List of Abbreviations

A&E . . . . . Accident and Emergency
ACCIS . . . . Automated Childhood Cancer Information System
AIC . . . . . Akaike's information criterion
AJCC . . . . . American Joint Committee on Cancer
ALL . . . . . Acute lymphoid leukaemia
AML . . . . . Acute myeloid leukaemia
AUROC . . . . Area under the receiver operator curve
BCCSS . . . . British Childhood Cancer Survivor Study
BIC . . . . . Bayesian information criterion
CCA . . . . . Complete case analysis
CCSS . . . . . North American Childhood Cancer Survivor Study
CI . . . . . Confidence interval
CIPS . . . . . Continuous inpatient spell
CML . . . . . Chronic myeloid leukaemia
CNS . . . . . Central nervous system
COSD . . . . . Cancer Outcomes Services Dataset
CRS . . . . . Cancer Reform Strategy
CYA . . . . . Children and young adults
DAG . . . . . Directed acyclic graph
df . . . . . Degrees of freedom
EHR . . . . . Excess hazard ratio
EM . . . . . Expectation-Maximization
EPV . . . . . Events per variable
FCE . . . . . Finished consultant episode
FCS . . . . . Fully conditional specification
FIGO . . . . . International Federation of Obstetricians and Gynaecologists
GCT . . . . . Germ cell tumour
GOF . . . . . Goodness of fit

GP . . . . . General Practioner
HES . . . . . Hospital episode statistics
HL . . . . . Hodgkin lymphoma
HL test . . . . Hosmer-Lemeshow test
HR . . . . . Hazard ratio
HRR . . . . . Hospitalisation rate ratio
HSCIC . . . . Health and Social Care Information Centre
ICCC . . . . . International Classification of Childhood Cancer
ICD-O . . . . . International Classification of Diseases for Oncology
IGCCCG . . . International Germ Cell Cancer Collaborative Group
IMD . . . . . Index of multiple deprivation
IPW . . . . . Inverse probability weighting
IQR . . . . . Interquartile range
IR . . . . . Incidence rate
IRR . . . . . Incidence rate ratio
KIT . . . . . Knowledge and Information Team
K-M . . . . . Kaplan-Meier
LE . . . . . Late effects
LOCF . . . . . Last observation carried forward
MAR . . . . . Missing at random
MC errors . . Monte Carlo errors
MCAR . . . . Missing completely at random
MI . . . . . Multiple imputation
MICE . . . . . Multiple imputation by chained equations
MLE . . . . . Maximum likelihood estimation
MNAR . . . . Missing not at random
NCIN . . . . . National Cancer Intelligence Network
NHL . . . . . non-Hodgkin lymphoma
NHS . . . . . National Health Service
NIGB . . . . . National Information Governance Board
ONS . . . . . Office for National Statistics
OR . . . . . Odds ratio
PH . . . . . Proportional hazards
PO . . . . . Proportional odds
PROM . . . . Patient reported outcome measures
pys . . . . . Person years

SEER . . . . . Surveillance Epidemiology and End Results
SMCFCS . . . Substantive model compatible fully conditional specification
STS . . . . . Soft tissue sarcoma
TNM . . . . . Tumour Node Metastesis
TYA . . . . . Teenagers and young adults
UK . . . . . United Kingdom
UKACR . . . . United Kingdom Association of Cancer Registries
US . . . . . United States
WBC . . . . . White blood cell
WHO . . . . . World Health Organization

Chapter 1

Introduction

1.1 Motivation

Cancer is predominantly a disease seen in older people, with over a third of all cancer diagnoses occurring in those aged over 75. In the UK, 40% of males and 37% of females are expected to develop cancer at some stage during their life, and as such, cancer has substantial public health implications [8]. Children and young adults (CYAs) with cancer make up a small proportion of the overall cancer burden, with 0.5% and $<$1% of all cancers being diagnosed amongst children and teenagers and young adults (TYAs) respectively in the UK [5, 9, 10]. Despite being rare, CYA cancer remains an important focus of research as it is the leading cause of death amongst children, contributing to 21% of all deaths amongst 0-14 year olds, and TYAs, contributing to 9% and 15% of death from disease for 15-24 year old males and females respectively [4, 5]. Furthermore, the potential years of life lost for CYAs is much greater than for adults.

CYAs with cancer are a unique population for a number of reasons, and therefore warrant research specifically focused on this population. Firstly, the types of cancer diagnosed amongst CYAs differ substantially from adult cancers. Secondly, the clinical behaviour is such that CYA cancers tend to grow more rapidly, but are also found to respond better to chemotherapy than adult cancers [9]. Due to its rarity, the pathway to diagnosis for CYA cancer tends to differ substantially than for adults. Presenting symptoms can be vague and attributable to a range of different health issues, and in addition, a GP is likely to encounter just one new case of CYA cancer once every 20 years [11, 12, 13].

Although cancer is the leading cause of death for CYAs, a high proportion of those diagnosed with cancer survive more than 5-years (current figures range from 78% to 85% for children and TYAs [4, 5, 14]). This is in contrast to adults; 54% of whom are said to survive beyond 5-years of diagnosis [15]. As the population of long term

survivors increases, however, there is a growing concern over the risk of serious co-morbidities experienced later in life as a result of intensive treatments received for cancer [16]. Furthermore, results from a European wide study between 1995 and 2002 showed that UK cancer survival for children was poorer than the European average (78% and 81% 5-year survival respectively) [17]. The same study also showed a difference for TYA survival (85% and 87.5% for the UK and Europe respectively). Continued disparities in survival across Europe are said to be due to late stage at presentation [18, 19, 20] which could indicate delayed diagnosis and inequalities in access to healthcare.

The NHS outcomes framework, originally developed in 2010 and updated on a yearly basis, highlights the importance of timely epidemiological research and reporting of outcomes to improve the quality of care across all areas of the NHS [21]. Improving survival and patient outcomes for cancer in the UK is a key priority across all age groups, as evidenced by the Improving Outcomes: A Strategy for Cancer report [22] which sets out the aim of saving an additional 5000 lives every year by 2015. The NHS outcomes framework emphasizes the need for preventing premature mortality and reducing the potential years of life lost. Following recommendations from the Children and Young People's Health Outcomes Forum [23], the NHS outcomes framework specifically highlighted the need to focus on survival from childhood cancer for the first time in 2013, by the introduction of a health indicator for 5-year survival from all cancers in children under the age of 15 years. However, importantly, the report does not specify the need for a similar indicator for teenagers and young adults (TYAs) with cancer, despite the TYA population facing similar challenges as the childhood population.

The purpose of this research project was to improve upon the quality of information available on cancer outcomes amongst CYAs within Yorkshire. The accuracy of the conclusions drawn from such an analysis depend on the quality of data and appropriateness of statistical techniques used. The problem of missing data is common within health services research. Within cancer registration, the stage of the disease at diagnosis is often poorly recorded, and as a result it is often not taken into account when analysing survival patterns of cancer, despite it being a key prognostic factor. Furthermore, historically in the UK, cancer registry data on ethnic group has not been routinely, or accurately, collected, despite evidence from the US that childhood cancer survival varies according to ethnic group based on detailed ethnic information held by the Surveillance Epidemiology and End Results (SEER) cancer registries (for example, in Kadan-Lottick et al. [24] and Linabery and Ross [25]). There are several available methods for handling missing data within epidemiological research, including advanced techniques such as imputation via the expectation maximization algorithm, multiple imputation and inverse probability weighting. Despite this availability, these techniques are not routinely used within medical research.

1.2 Summary of Aims and Objectives

This project has 4 key aims:

1. **To improve upon the quality of information available on cancer outcomes amongst CYAs within Yorkshire.**
   Existing methods for handling missing data were researched, and the method(s) which minimised bias and made most efficient use of the available data were applied to data on CYAs with cancer. An assessment of the extent that missing data affects the results and conclusions within the analysis was also determined.

2. **To describe variation in cancer survival amongst CYAs within Yorkshire.**
   Variation in survival according to diagnostic group, disease severity, gender, year of diagnosis, ethnicity and socioeconomic status was described for CYAs with cancer in Yorkshire.

3. **To describe variation in disease severity at presentation for CYAs with cancer in Yorkshire.**
   The variation in disease severity at diagnosis was quantified for CYAs with cancer in Yorkshire. In particular, potential inequalities in disease severity according to deprivation, ethnicity and age were explored whilst adjusting for gender and year of diagnosis.

4. **To describe the long term health effects of treatment for long term survivors of cancer amongst CYAs in Yorkshire.**
   The incidence and risk of long term health effects of treatment for long term survivors of cancer were described using population-based cancer registry data linked with national administrative data on hospital admissions.

1.3 Thesis Outline

To provide background information to the work presented in this thesis, Chapter 2 contains a critical review of the current literature on CYA cancer outcomes and provides the evidence base for this thesis highlighting key gaps in the current literature. Chapter 3 contains a critical review of missing data techniques and the use of these methods within the current literature inform the choice of missing data methods used throughout this thesis to reflect aim 1 as listed above. The specific methods used throughout the thesis are defined in Chapter 4 followed by detailed descriptive analysis of the cohort in Chapter 5. The main results of the project are presented in three chapters focusing on

variation in cancer survival (Chapter 6), inequalities in disease severity (Chapter 7) and long term effects amongst survivors of cancer (Chapter 8) to achieve aims 2, 3 and 4 of this study respectively. These results are then discussed in Chapter 9, highlighting the main contributions of this work in the context of previous research and policy as well as a discussion of the strengths and limitations and recommendations for future work.

Chapter 2

Background and Epidemiology

2.1 Introduction

The aim of this chapter is to provide background information on cancer registration processes and structures in the UK and describe the importance of such registration systems for epidemiological research. This section will also include information on the classification and staging of cancers in general as well as specific classification and staging of childhood and young adult (CYA) cancers in addition to a discussion of missing data problems within cancer registration.

Following this background information, the current literature and knowledge of CYA cancer epidemiology is critically reviewed, focusing on variation in survival of CYA cancer according to diagnostic group, age at diagnosis, gender, temporal changes, stage at diagnosis, ethnicity, deprivation and geographical trends. The current evidence base looking at variation in disease severity at diagnosis for CYA cancers is explored according to ethnicity, deprivation and age at diagnosis, followed by a literature review on longer term outcomes and co-morbidities for CYA cancer survivors.

Finally, the literature is summarised and key gaps in the knowledge are identified, forming the basis of the analysis within this thesis.

Chapter 3 will go on to describe how missing data can have an impact upon statistical analysis and describe in detail the methods which currently exist to deal with missing data, and which of these methods is most suitable for use in population based research.

2.2 Cancer Registration

Prior to critically reviewing the literature surrounding CYA cancer epidemiology (§2.3), cancer registration processes and classifications in the UK are described to provide relevant background information to the data sources from which the literature arises.

Cancer registration is "the process of maintaining a systematic collection of data on the occurrence and characteristics of malignant neoplasms and certain non-malignant tumours" [26]. Cancer registration is important to enable epidemiological research to describe the degree of access to treatment services as well as for health planning within a population; furthermore it can be used to identify specific targets for cancer screening [27]. Cancer registration in England was previously covered by eight regional registries which were all members of the UK Association of Cancer Registries (UKACR). These regional registries collected data on all people diagnosed with cancer within their geographical areas, adhering to a defined minimum dataset which was subsequently collated into the national cancer registry. As of January 2013, the national registry was replaced by the Cancer Outcomes Services Dataset (COSD) and the 8 regional registries were replaced by lead cancer knowledge and information teams (KITs), each of which led on one or more specific tumour site. The North West KIT is the lead KIT for all cancers amongst children and TYAs.

2.2.1 Specialist Cancer Registries

In addition to the COSD, there are several specialist registers across the country. These include the Manchester Childrens Cancer Registry which is the oldest childhood cancer registry in the UK dating back to 1953, the Northern Region Young Persons' Malignant Disease Registry based at the University of Newcastle, dating back to 1968 and the Yorkshire Specialist Register of Cancer in Children and Young People based at the University of Leeds, dating back to 1974. The historical data in combination with continued data collection makes these registers a great resource for large epidemiological studies in differing parts of the country. These registers differ from the national cancer register not only due to the level of detailed information they hold in addition to the core data items such as treatment and relapse data, but also because cases are followed up for a number of years making survival analysis a possibility.

The primary data source for this PhD is the Yorkshire Specialist Register of Cancer in Children and Young People (the Yorkshire register from here on in). The Yorkshire register is a regional population based register of cancer in CYAs. The register contains detailed demographic and clinical information on children (0-14 years) diagnosed with

cancer in Yorkshire since 1974. In 1998 the register expanded to include young adults aged between 15 and 29 years, and data were retrospectively collected for cases diagnosed since 1990 and prospectively from 1998. In 2006, there was a further expansion to include all cases diagnosed in the current Yorkshire and Humber Strategic Health Authority which previously did not include South Yorkshire. All cancer diagnoses on the Yorkshire register are classified according to the International Classification of Childhood Cancer (ICCC) [7] and cases are followed up every two years to determine whether or not they have survived, have received any subsequent treatment, or if they have experienced any relapses or secondary tumours. The core aim of the register is to conduct timely epidemiological research examining patterns and causes of cancer amongst this cohort of under 30 year olds in Yorkshire.

2.2.2 Classification of Cancer

Cancer is a term used to describe a collection of diseases which arise from mutated cells exhibiting abnormal cell growth and division. A mass of tumour can form as a result of abnormal cell behaviour, for example, if the cells do not die when they should or if new cells grow unnecessarily. Tumours that do not spread around the body are known as benign tumours, and although they can cause complications such as putting pressure on other body organs, these are not cancerous. Malignant tumours, or neoplasms, are those made up of cancerous cells that can spread to, and destroy, nearby tissues as well as invade other body parts [28, 29]. Cancers that invade other organs or body parts are said to have metastasised. There are over 200 types of cells in the human body; and as a result there are over 200 different types of cancer [15].

Typically, cancers can be classified according to the primary tumour site, for example, if bladder cancer metastasises to the lung, it will still be referred to as bladder cancer. However, more accurately, the International Classification of Diseases for Oncology, Third Edition (ICD-O-3), classifies cancer according to the primary site (topography) and the type of tissue in which the cancer originated (morphology or histology) [30]. ICD-O-3 is an internationally recognised classification scheme used for coding tumours in cancer registries and is widely used for cancers occurring in adults. However, the types of cancer observed amongst CYAs differ substantially from those amongst adults. These differences are described further in §2.3. Two classification schemes for children and TYAs respectively were designed to group tumours according to their histological similarities and reflect the common tumour groups observed amongst both age groups. The ICCC was developed in 1996 to classify tumours amongst children according to their histological similarities, as opposed to the site based scheme used for adult malignancies. It has since been updated to its 3rd revision (ICCC-3) [7]. These histological groupings

are a better reflection of the range of tumours more commonly seen amongst children. The ICCC-3 contains 12 main diagnostic groups as shown below. The full classification system is provided in Appendix A.

I. Leukaemia

II. Lymphoma and reticuloendothelial neoplasms

III. CNS and miscellaneous intracranial and intraspinal neoplasms

IV. Neuroblastoma

V. Retinoblastoma

VI. Renal tumours

VII. Hepatic tumours

VIII. Malignant bone tumours

IX. Soft tissue sarcomas

X. Germ-cell, trophoblastic and other gonadal neoplasms

XI. Carcinomas and other malignant epithelial neoplasms

XII. Other and unspecified malignant neoplasms

The Birch TYA Classification Scheme [31] was specifically designed for grouping cancers which occur amongst TYAs. These malignancies tend to differ from tumours amongst adults, and despite some similarities, also differ from childhood tumours (§2.3). The Birch TYA Classification Scheme defines 10 main diagnostic groups as outlined below. The full classification system is provided in Appendix B.

Group 1: Leukaemias

Group 2: Lymphomas

Group 3: CNS and other intracranial and intraspinal neoplasms

Group 4: Osseous and chondromatous neoplasms, Ewing sarcoma and other neoplasms of bone.

Group 5: Soft tissue sarcomas

Group 6: Germ cell and trophoblastic neoplasms

Group 7: Melanoma and skin carcinoma

Group 8: Carcinomas

Group 9: Miscellaneous neoplasms not elsewhere classified

Group 10: Unspecified malignant neoplasms not elsewhere classified

The Birch TYA classification scheme is similar to the ICCC with a focus on tumour histology rather than tumour site. The slight differences in categories reflect those tumours that are more common amongst TYAs compared to children, for example, melanomas and carcinomas form separate groups within the Birch TYA Classification as they are the second and third most common cancers amongst TYAs, however, amongst children, melanomas and carcinomas are much less common. Due to TYAs bridging the gap between childhood and adulthood, a mixture of childhood and adult cancers are seen within this group.

2.2.3 Staging Mechanisms

Staging of cancer is an important concept in the classification of tumours. The stage of a tumour defines the extent of the cancer and in general includes the location of the primary tumour, the size of the tumour, whether or not cancer has spread to lymph nodes and the presence or absence of metastases (the spread of the tumour to other body parts or organs). The stage determines the management of the tumour, for example whether surgery is appropriate or whether the patient is suitable for clinical trial enrollment. It also determines prognosis and allows researchers to study relatively homogeneous tumour groups. For the majority of tumours, staging mechanisms including the above elements of size and spread, are easily defined. The tumour node metastasis (TNM) staging system is used worldwide to classify stage of disease in terms of the primary tumour (T), whether regional lymph nodes are affected (N) and whether or not the tumour has metastasized (M). Although other staging mechanisms exist, TNM staging is said to be the most clinically useful [32]. This staging mechanism was first created in the 1940s, and is now maintained by the American Joint Committee on Cancer (AJCC) and the International Union for Cancer Control. The TNM staging system constitutes multiple algorithms specific to tumours of a particular anatomic site and histology. Once the TNM categories have been determined, this information is then used to classify the tumour into a grouped stage which generally ranges from stage I to stage IV. Table 2.1 gives an example of the TNM staging mechanism for soft tissue sarcomas (STS), which additionally includes a G code for the grade of the tumour. The grade of a tumour in this context refers to the degree of differentiation of the tumour. The degree of differentiation refers to the similarity of the cancer cells to normal cells, therefore, a well differentiated tumour is made up of cells which are similar to normal cells (low grade) and a poorly differentiated tumour means

the cells are less like normal cells (high grade). Table 2.2 shows how these individual parts of the staging mechanism are grouped into a combined stage for STS.

Table 2.1: TNM Staging Definitions: Soft Tissue Sarcomas

Primary Tumour (T)	
TX	Primary tumour cannot be assessed
T0	No evidence of primary tumour
T1	Tumour 5cm or less in greatest dimension
- T1a	- Superficial tumour
- T1b	- Deep tumour
T2	Tumour more than 5cm in greatest dimension
- T2a	- Superficial tumour
- T2b	- Deep tumour

Regional Lymph Nodes (N)	
NX	Regional lymph nodes cannot be assessed
N0	No regional lymph nodes
N1	Regional lymph node metastasis

Histologic Grade (G)	
GX	Grade cannot be assessed
G1	Grade 1
G2	Grade 2
G3	Grade 3

Table 2.2: Prognostic Groups based on TNM Staging: Soft Tissue Sarcomas

Grouped Stage	**Primary Tumour (T)**	**Regional Lymph Nodes (N)**	**Distant Metastasis (M)**	**Histologic Grade (G)**
Stage IA	T1a	N0	M0	G1, GX
	T1b	N0	M0	G1, GX
Stage IB	T2a	N0	M0	G1, GX
	T2b	N0	M0	G1, GX
Stage IIA	T1a	N0	M0	G2, G3
	T1b	N0	M0	G2, G3
Stage IIB	T2a	N0	M0	G2
	T2b	N0	M0	G2
Stage III	T2a, T2b	N0	M0	G3
	Any T	N1	M0	Any G
Stage IV	Any T	Any N	M1	Any G

Due to the nature of some cancers, any of the T, N and M values may not be relevant and sometimes alternative or additional measurements are required to stage a tumour. A TNM staging algorithm exists for cancers of almost all sites and histology, however, the AJCC staging manual does not include staging of CYA cancers specifically. As explained further in §2.3.1, CYA cancers are vastly different from adult cancers. Furthermore, for

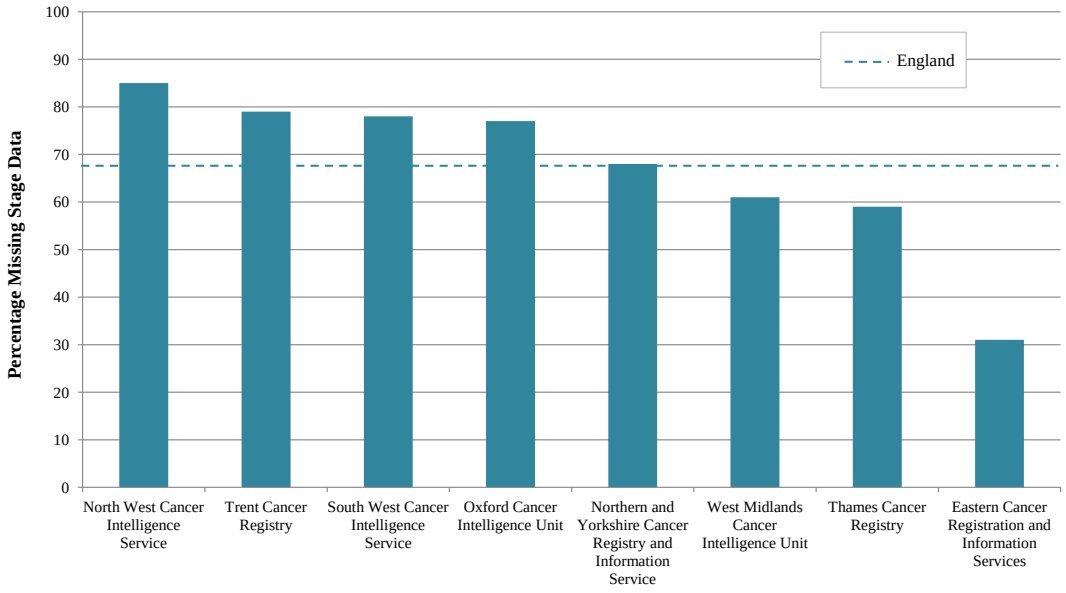
some common CYA cancers, such as leukaemia and brain tumours, the size and spread are not logical concepts and other measures of prognosis are required. For example, although brain tumours can spread to other parts of the central nervous system (CNS) they do not tend to spread to other organs. However, the growth of a brain tumour can have important implications as it can put pressure on parts of the brain and potentially damage brain function. Therefore, the WHO grade of a tumour is often used as a prognostic factor instead of its stage [6]. A detailed exploration of staging mechanisms for all CYA cancers is given in Chapter 5 alongside the descriptive data analysis.

2.2.4 Missing Data in Cancer Registration

This section focuses specifically upon the issue of missing data within cancer registration. The statistical implications of missing data are discussed in Chapter 3. The Department of Health's Cancer Reform Strategy (CRS) highlighted the importance of the collection and use of high quality data on cancer outcomes and survival to improve patient choice and service quality. The CRS emphasizes that although there has been some improvement in overall data collection, information on staging and co-morbidities, radiotherapy activity and chemotherapy delivery is not consistently recorded across the UK [33]. Both the 2010 National Audit Office report on delivering the CRS and the Department of Health's 'Improving Outcomes: A Strategy for Cancer' report of 2011 highlighted that incomplete and inconsistent data on stage of disease at the time of diagnosis remains a key gap in cancer intelligence [1, 22]. These data are not only vital for accurate analysis of cancer outcomes across differing populations but also enables improved understanding of the number of patients who are diagnosed late and whether this is an issue amongst particular subgroups of patients [32].

As described in §2.2.3, the staging of a tumour is not a simple algorithm that can be applied to all cancers, and the complicated nature of staging mechanisms has played a large role in the missing data problem. Individual tumours require different staging mechanisms, and some tumours, in particular amongst CYAs, do not have clearly defined staging mechanisms. Poor understanding of these mechanisms can lead to a poor translation of what is contained in the medical notes to cancer registry databases. For example, in some cases, medical notes may contain data about the metastases of a tumour which may imply directly that the tumour is stage IV, without explicitly stating the stage. Therefore, missing data on stage can vary depending on the level of knowledge and experience of the person responsible for collecting such data. The national cancer intelligence network (NCIN) has identified a need to educate non-medical staff involved in collecting and using cancer registry data within their report in 2009 [34]. Despite the recognition of these issues by large national organisations, missing data of stage within

cancer registries is still a large problem. Data from the NCIN shows that missing stage data in the former regional English cancer registries ranged from around 30% to 85% in 2007 (Figure 2.1).



Data Source: National Cancer Intelligence Network

Figure 2.1: Missing Stage Data in English Cancer Registries, 2007. Figure adapted from Morse [1]

Another area of concern within cancer registration is missing data on ethnicity. Many health outcomes, including cancer survival, are subject to inequalities based on ethnicity. Complete and accurate information on ethnicity therefore plays an important role in health services research. Historically, ethnicity data within the NHS has been both inaccurate and incomplete as highlighted by the NCIN's 2009 report on cancer incidence and survival by major ethnic group [35]. NHS data sets such as hospital episode statistics (HES) data contain multiple records per patient which can result in one person being assigned several different ethnicities. Further issues in data quality can arise from differences between ethnicity being determined by a health care professional to those that are self-reported. A study in the U.S. showed that the level of agreement between administrative data sources and self-reported ethnic groups could be as low as 15% and varied considerably by ethnic group (60% for Hispanics, African Americans and Whites, but less than 40% for Asians and Pacific Islanders) [36]. Several methods for improving the quality of data held on ethnicity include cross referencing the data with other sources using data linkage methods

and the use of name analysis programs [37].

2.3 Cancer Epidemiology

Cancer accounted for 7.6 million deaths in 2008, making it the leading cause of death worldwide. Worldwide deaths from cancer are predicted to rise to over 11 million by 2030, with the most common types being lung, stomach, liver, colorectal and breast cancer [38]. In the UK, incidence of cancer is higher amongst adult males than females, with the European age standardised rate of all cancers combined being 417 and 366 per 100,000 for males and females respectively in 2008. The most common types of cancers in the UK are breast, lung, colorectal and prostate cancer which account for over half of the UKs cancer burden [15]. The prognosis of cancer in adults varies hugely according to the type of cancer. For example, 5-year survival for people diagnosed with cancer in England between 2003 and 2007 was 83.3% for women with breast cancer, 7.3% and 8.7% for men and women respectively with lung cancer, 50.9% and 52.6% for men and women respectively with colorectal cancer and 79.7% for men with prostate cancer [39]. The variation in outcomes observed between different cancers is a reflection of the fact that cancer is not one single disease but many.

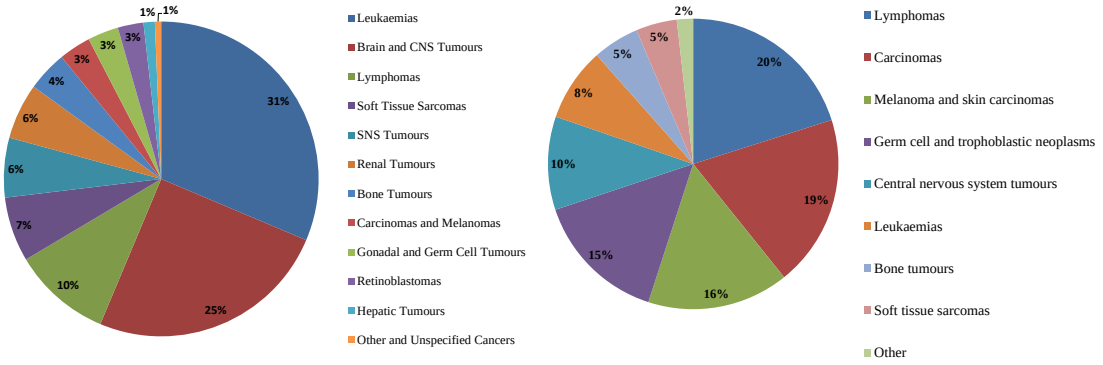
2.3.1 Childhood and Young Adult Cancer

Children and TYAs with cancer form a distinct population from adults with cancer due to a number of factors as outlined below. However, firstly, it is important to define the childhood and TYA age range, which varies across the UK and worldwide. The epidemiological definition of childhood cancer has been relatively consistent and generally includes children diagnosed under the age of 15 (0-14 years inclusive). This age range was adopted consistently by the former National Registry of Childhood Tumours as well as in the European wide Automated Childhood Cancer Information System (ACCIS) and EUROCARE projects [14, 17, 40]. Although this definition is used in cancer epidemiology, it has been criticised as merely an arbitrary cut off which does not relate to clinical aspects of cancer [41] and does not reflect the legal upper age boundary for children which includes anyone up to the age of 16, or the World Health Organization (WHO) definition which considers a child as an individual under 18 years of age. Age boundaries and definitions are even less clearly defined for TYAs with cancer, and there is a lot of variation in age limits between study groups within the UK, Europe and the U.S.. TYAs were defined as those aged between 15-24 by large scale projects such as the European wide EUROCARE-4 study [17], the NCIN report on survival of TYAs with

cancer in the UK [42] and the NICE guidance on improving outcomes for children and young people with cancer [43]. This age boundary is said to reflect clinical services and bridges the gap between childhood and adult services, a TYA specific principal treatment centres include patients up to the age of 25. However, many of the cancers common to TYAs are still observed amongst older TYAs aged between 25 and 29, and as such the upper age boundary could be feasibly extended to young adults as old as 29 [42, 43]. Several UK based studies, including Birch et al. [44] and Geraci et al. [45], define TYAs as between the ages of 13 and 24, and the same authors also published data on TYAs with cancer aged between 13 and 29 [46]. Individuals up to the age of 30 have also been included in studies of cancer epidemiology in the U.S. [47] and those up to the age of 39 have been included in the definition of TYAs for two other U.S. studies [48, 49] as well as in the UK TYA Cancer Survivor Study. Although justifications for age boundaries have been provided, it can be argued that most of these age limits are arbitrary cut off points as with childhood cancer definitions. The main data source for this thesis is the Yorkshire register which contains data on cases of cancer up to the age of 30, and as such, TYAs for the purposes of the analysis within this thesis will include those diagnosed between the ages of 15 and 29 in line with several previous studies as highlighted above. In addition, this age range provides equal 15-year age intervals for comparing children (0-14 years) to TYAs (15-29 years).

The primary reason for CYAs with cancer being distinct from adults with cancer is the distribution and type of cancers observed amongst CYAs. The most common childhood cancers within the UK, in terms of their incidence rate per million person years (IR/mpyears) are leukaemia (41/mpyears), CNS tumours (28/mpyears) and lymphomas (11/mpyears) [10, 50]. For TYAs in England, lymphoma (45/mpyears), carcinoma (31/mpyears) and germ cell tumours (GCTs) (25/mpyears) are most common [51]. The percentage of cases for all diagnostic groups for children and TYAs are shown in Figure 2.2. Secondly, the incidence of CYA cancer is very rare compared to that of adults, as childhood cancers account for less than 0.5% of cancers in the UK [9, 10] and TYA cancer makes up less than 1% of cancers at all ages in the UK [5]. Despite being rare, cancer is the leading cause of death for children with almost 21% of deaths in this age group being attributed to cancer [4]. For TYAs cancer is the leading cause of death from disease, contributing to 9% and 15% of deaths for 15-24 year old males and females respectively [5]. Nonetheless, survival rates amongst CYAs are much higher in general compared to that seen in adults, with almost 80% and over 80% of children and TYAs respectively becoming long term survivors compared to only 54% of adults surviving cancer more than 5 years [4, 5, 14] (§2.3.2 provides detailed information about CYA cancer survival). Finally, CYA cancers also differ from adult cancers in terms of their clinical behaviour as CYA cancers tend to grow much more rapidly, but also respond better to chemotherapy

treatment when compared to adult tumours [9].



(a) (b)

Figure 2.2: Percentage of cancer diagnoses for (a) 0-14 year olds by diagnostic group for diagnoses between 2001 and 2005 (Data Source: [2]) and (b) 15-24 year olds by diagnostic group for diagnoses between 2005 and 2008 (Data Source: [3])

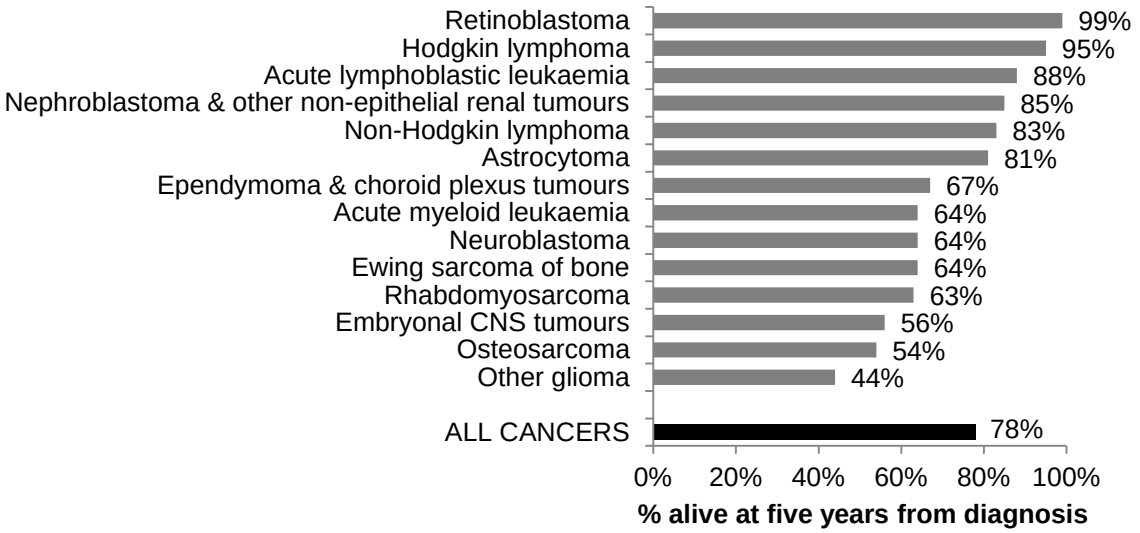
2.3.2 Survival

The literature on survival of CYA cancer in the UK between 1980 and 2014 was critically reviewed to determine what is already known about variation in cancer survival amongst CYAs. Full details of the search terms used are given in Appendix C. The search returned 1276 papers up until 2014 and a review of titles and abstracts resulted in final number of 18 papers assessing survival of CYA cancer in the UK, Britain or England [42, 44, 45, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66]. Variation in survival patterns according to diagnostic group, age at diagnosis, sex, temporal trends, stage at diagnosis, deprivation and ethnicity were evaluated focusing primarily on information contained within the 18 papers identified, however, additional reference was made to books including Little [9], Stiller [14] and Estlin et al. [67], in addition to grey literature from websites containing national statistics on CYA cancer which have not always been published in peer reviewed journals [2, 4, 5]. In addition, publications on adult cancers and studies from outside of the UK have been cited where appropriate for comparison.

2.3.2.1 Diagnostic Group

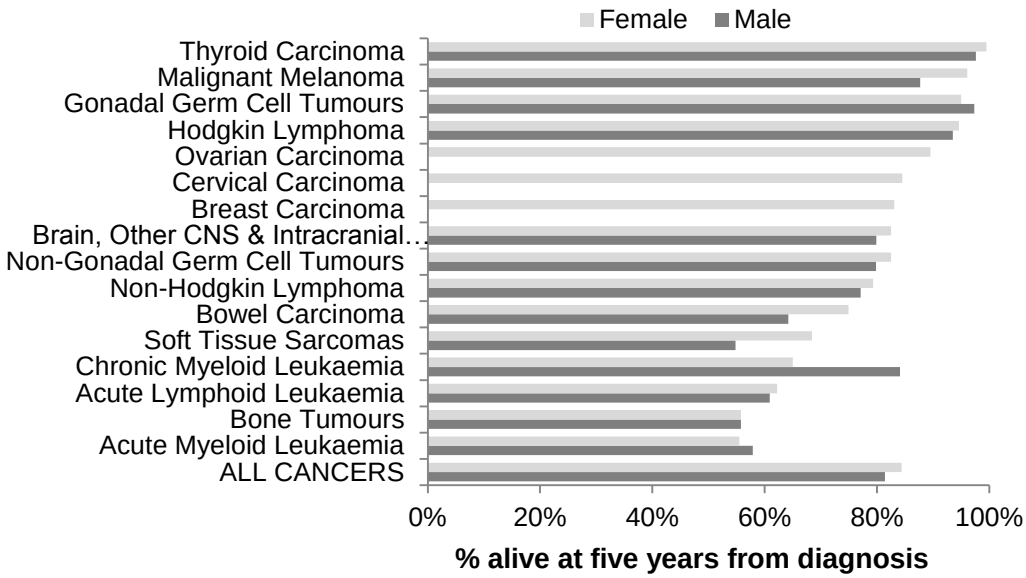
Changes and improvements in diagnostic tools and treatment protocols have led to overall improvements in survival of CYA cancer (see §2.3.2.4 for further details). Despite these improvements, survival varies significantly according to diagnostic group and subtype.

For children, survival from embryonal CNS tumours (56%), osteosarcoma (54%) and other glioma (44%) were substantially poorer than average (Figure 2.3). For TYAs, survival for males with STS (55%), males and females with bone tumours (56%) and males and females with acute myeloid leukaemia (AML) (58%) had a poorer prognosis than average for males and females of the same age (Figure 2.4). These data were obtained from the cancer research UK website and were based on national cancer registry data for children and TYAs respectively [4, 5]. The data provide a useful overview of survival for children and TYAs with cancer in the UK. However, the survival estimates are based on univariable data and as such do not take into account other potentially important factors, such as year of diagnosis, disease severity, ethnicity or deprivation. Furthermore, the data covered only a selection of childhood and TYA cancers and were four years out of date including only diagnoses up to 2005 with a maximum follow up period up until 2010. The available literature focusing on survival differences among cancer subtypes for each main ICCC diagnostic group is reviewed below.



Data source: NRCT/CCRG, http://www.ccrg.ox.ac.uk/datasets/survivalrates.htm

Figure 2.3: Five-year survival rates for selected childhood cancers, Great Britain, diagnosed during 2001-2005. Data Source: [4]



Data source: NWCIS, http://www.cancerresearchuk.org/cancer-info/cancerstats/teenage-and-young-adult-cancer/survival/#source1

Figure 2.4: Five-year relative survival rates for teenage and young adult cancers by diagnostic group and sex, ages 15-24, UK, 2001-2005. Data Source: [5]

Leukaemia

Leukaemias are cancerous blood cells which originate in the bone marrow and move into the blood stream. The type of blood forming cell that leukaemia develops from can be lymphoid or myeloid and it is these cells that determine the type of leukaemia. ALL (ICCC Ia) is the most common form of childhood leukaemia, making up over 75% of all leukaemias diagnosed in England [2]. Other subtypes of leukaemia include AML (ICCC Ib) and chronic myeloproliferative diseases (ICCC Ic). Approximately 90% of chronic myeloproliferative diseases are chronic myeloid leukaemias (CML), and the remaining 10% are chronic myelomonocytic leukaemias [68]. For childhood leukaemia, ALL is associated with the highest survival with almost 90% surviving at least 5-years [4]. Survival for AML is much poorer compared to ALL amongst children and TYAs, although the gap is much smaller for the latter due to poorer survival rates of ALL for TYAs compared to children [5, 14]. Very little literature exists on survival from childhood CML due to its extreme rarity of 8 cases diagnosed per year on average in England [10], however, Gatta et al. [68] indicates that 1 and 5-year survival from CML amongst children is 63% and 24% respectively. The large drop off in survival between 1 and 5-years from diagnosis is due to CML developing much more slowly compared to other leukaemias such as ALL and AML [68].

Lymphoma

Lymphoma is a type of cancer which evolves from abnormal white blood cells

(WBC) within the lymph nodes (known as lymphocytes). Two main classifications of lymphoma are HL and non-Hodgkin's lymphoma (NHL), these classifications are made based on the cell types within the tumour. Survival from lymphoma is high for both HL and NHL amongst children (97% and 79% 5-year survival respectively) and TYAs (93% and 72% 5-year survival respectively) [44, 63]. For childhood lymphoma, survival rates of HL and NHL remain high even at 15-year survival (96% and 75% respectively) [63]. Similar figures for TYAs at 10 and 15-years from diagnosis were not available.

Central Nervous System Tumours

The most common type of CNS tumour amongst children is astrocytoma, accounting for approximately 40% of cases. CNS tumour survival varies by diagnostic subgroup, such that children with astrocytomas, have 81% 5-year survival compared to 67% for ependymomas, 56% for intracranial and intraspinal embryonal tumours and 44% for other gliomas [14].

Neuroblastoma

Neuroblastoma is the most common cancer in infancy, with a median age of diagnosis of 18 months [69]. The oncogene MYCN has been identified as a major independent prognostic factor [70], with 5-year survival at 85% in infants without the MYCN amplification, compared to 25% for those with the MYCN amplification [69]. Only 3% of total neuroblastoma cases are diagnosed in children over the age of 10 [69]. Neuroblastoma in older cases has much slower progression and very poor survival compared to other childhood tumours. Despite poor survival, the MYCN amplification rarely occurs in tumours diagnosed within this age group. 5-, 10- and 15- year survival for childhood neuroblastoma was 54%, 52% and 51% respectively for diagnoses between 1991 and 1996 [63]. More recent results for diagnoses between 2001-2005 show 5-year survival has increased to approximately 65% [2, 64].

Retinoblastoma

Retinoblastoma is cancer of the eye, and is predominantly observed in children under the age of 5. Children diagnosed with retinoblastoma have the best survival chance compared to any other childhood cancer in England, with 5-year survival as high as 99% between 2001 and 2005 [2]. Retinoblastoma is extremely rare in TYAs or adults with only 26 reported cases across the world between 1919 and 2013 [71].

Renal Tumours

The 5-year survival rate for renal tumours amongst children is 84% overall. The Wilm's tumour subgroup (also known as nephroblastoma), which accounts for approximately 90% of childhood renal tumours, has 85% 5-year survival [2, 72]. 77% of childhood

Wilm's tumours are diagnosed before the age of 5, and 15% are diagnosed before the age of 1, however, there is no evidence that age is a prognostic factor for Wilm's tumours [14, 72]. Survival amongst the Rhabdoid renal tumour subgroup was much lower compared to Wilm's tumour, with 10-year survival being 25% [14].

Hepatic Tumours

Hepatic tumours are relatively rare in childhood, contributing to approximately 1.3% of all childhood cancers. Due to its rarity, very few outcome studies exist focusing specifically on hepatic tumours. Survival at 5-years for hepatic tumours diagnosed between 2001 and 2005 was 66% [2].

Malignant Bone Tumours

In adults, bone tumours commonly arise from other tumours which have metastasized and primary malignant bone tumours are rare amongst this age group. However, amongst children they are the sixth most common cancer and the third most common amongst TYAs. Osteosarcoma and Ewing sarcoma are the two most common types of primary malignant bone tumours amongst CYAs [73]. 5-year survival for Osteosarcoma and Ewing sarcoma in children was 54% and 64% overall respectively, however, for tumours which had metastasized, survival was less than 25% [73].

Soft Tissue Sarcomas

STS are a group of cancers which originate from cells in soft tissues such as muscle, fat, nerves or blood vessels. The most common form of STS amongst CYAs is rhabdomyosarcoma, which has 5-year survival of just below 65% [2]. Approximately 20% of CYAs diagnosed with STS present with advanced stage tumours which have metastasized, and survival for this group is less than 25% [74]. The histological subtype of rhabdomyosarcoma has also been identified as an important prognostic factor, with alveolar histology being a marker for poorer survival compared to embryonal histology (73% and 39% 10-year survival respectively) [14, 74].

Germ Cell Tumours and Neoplasms of Gonads

GCTs occur amongst children as well as TYAs, however, it is much more common for TYAs compared to children [75]. Survival at 5-years for GCTs in Britain is 86%, however, this high survival was driven by the malignant gonadal GCT subgroup with a 5-year survival of 96%. 5-year survival was considerably lower for the intracranial and intraspinal GCT subgroup and the malignant extracranial and extragonadal GCT subgroup at 79% each [14].

Carcinomas

Carcinomas are cancers which develop from epithelial tissue, it is the most common form of cancer in adults, however, it is extremely rare amongst children (see Figure

2.2). This is reflected in the classification of carcinomas in ICCC group XI - 'Other malignant epithelial neoplasms and malignant melanomas'. Although, still rare, carcinomas are more common amongst TYAs than in children. Birch et al. [44] document 5-year survival for TYAs with carcinomas according to site. The survival rates were 97% for thyroid carcinoma, 83% for head and neck carcinomas excluding thyroid carcinoma, 87% for ovarian carcinoma and 80% for cervical carcinoma, 70% for colorectal carcinoma and 67% for lung carcinoma.

2.3.2.2 Age at Diagnosis

Age has been identified as an important prognostic factor in some but not all CYA tumours, with varying patterns as described in Table 2.3. Marked improvements in survival over time have been observed for children (see §2.3.2.4), however, despite some improvements for TYAs, the survival benefit for this older age group have not been observed to the same extent. Several papers have assessed variation in survival according to age for CYAs, however, the literature largely focuses on survival amongst children [14, 58, 63, 64, 65] separately to that amongst TYAs [44] with no data in the UK comparing the survival differences between children and TYAs until a recent report by the NCIN in 2012 [42].

Table 2.3 indicates that increasing age was associated with poorer prognosis for ALL, AML, HL, NHL, astrocytoma, bone tumours, STS, extragonadal GCTs and breast cancer. Poorer survival was observed between childhood and TYA age groups (ALL, AML, NHL, bone tumours, STS and extra gonadal GCTs) as well as within the TYA age range (NHL and STS). Furthermore, for some diagnostic groups, the evidence of poorer prognosis with increasing age was restricted to within the TYA age range (astrocytoma, Ewing sarcoma and breast cancer) and the effect of age for these tumours among children had not been studied in detail. For ALL, differences according to age are thought to reflect differences in disease biology, for example, the $t[12, 21]$ translocation seen in 25% of children with B-cell ALL is much rarer amongst adults and only seen in 3% of cases [76], and TYAs have a higher ratio of T-cell compared to B-cell ALL compared to children [77, 78]. Despite improved survival from ALL for children overall compared to TYAs, the age pattern of prognosis amongst children with ALL is not linear, as those diagnosed under the age of 1 and over the age of 10 have poorer survival compared to those aged 1-9 years [53, 64, 65].

In contrast, survival rates have also been shown to improve with age for some diagnostic groups (intracranial embryonal tumours, GCTs, intracranial GCTs, CNS tumours and melanoma) (Table 2.3). For CNS tumours and melanoma, survival for TYAs was significantly better when compared to children, however, there was no evidence of an

effect of age within the childhood or TYA age range for either CNS tumours or melanoma. For GCTs overall, younger TYAs had poorer survival than older TYAs and evidence for children was limited to intracranial GCTs which showed that those under the age of 5 had poorer survival compared to 5-14 year olds. However, in contrast, for extragonadal GCTs, TYAs had poorer survival compared to children as highlighted in the previous paragraph. Furthermore, age was not a significant predictor of survival for the ovarian GCT subgroup.

Non-linear relationships with age have been observed for childhood ALL as indicated above. Further non-linearities were observed for rhabdomyosarcoma, the most common form of STS amongst children, with those under the age of 1 and over the age of 10 having poorer survival compared to those aged 1-9. In addition, for neuroblastoma, there was a sharp decline in survival for 1-4 year olds compared to those diagnosed under the age of 1, and a further, but smaller, decline for those aged 5-14 years old.

Finally, there was evidence to suggest that age was not a prognostic factor for some diagnostic groups (retinoblastoma, renal tumours and hepatic tumours) obtained from two national studies of children in Britain [14] and children in the UK [63]. The effect of age for these tumours is unknown for TYAs. In addition, age was not a prognostic factor within childhood AML, NHL, CNS tumours and bone tumours despite being identified as a prognostic factor amongst TYAs and between children and TYAs (Table 2.3).

The age patterns in relation to prognosis were consistent between studies in general, despite different time periods of study and differences between adjustment for confounding factors. However, there were some conflicting results for HL. O'Hara et al. [42] suggested there was no evidence of TYAs having poorer survival compared to children for HL diagnosed between 2001 and 2005, however, data from two separate studies indicated that TYAs with HL had a 5-year survival rate of 88% [44], which was lower than that observed amongst children (98% for 1-9 year olds and 93% for 10-14) [14]. Although the study by O'Hara et al. [42] only adjusted for age and sex, and did not include other potentially important confounding factors (for example year of diagnosis, deprivation, ethnicity or stage), the evidence comparing childhood and TYA survival for HL was stronger than the individual evidence of childhood survival compared to another study of TYA survival for several reasons. Firstly, the results from Stiller [14] and [44] were not directly comparable as they cover different, although overlapping, time periods (1991-2000 and 1979-2001). Finally, the results from Stiller [14] were unadjusted for confounders whereas those from Birch et al. [44] were adjusted for age, sex, deprivation and calendar year, thus again limiting their comparability.

Table 2.3: Evidence of age as a prognostic factor for children and young adults with cancer in England

Poorer prognosis with older age	**Age**	**Survival Summary**	**Reference(s)**
Acute lymphoblastic leukaemia	0-14	88% 5yrS	[44]
	13-16	50% 5yrS	[44]
	17-20	44% 5yrS	[44]
	21-24	37% 5yrS	[44]
	0-14 vs 15-24	HR = 0.22 (0.18-0.28)	[42]
Acute myeloid leukaemia	0-14 vs 15-24	HR = 0.71 (0.55-0.92)	[42]
Hodgkin's lymphoma	1-9	98% 5yrS	[14]
	10-14	93% 5yrS	[14]
	17-24	88% 5yrS	[44]
Non-Hodgkin's lymphoma	13-16	70% 5yrS	[44]
	17-20	65% 5yrS	[44]
	21-24	66% 5yrS	[44]
	0-14 vs 15-24	HR = 0.66 (0.0.49-0.89)	[42]
Astrocytoma	13-16	71% 5yrS	[44]
	17-20	61% 5yrS	[44]
	21-24	47% 5yrS	[44]
Bone Tumours	0-14 vs 15-24	HR = 0.79 (0.64-0.97)	[42]
Ewing Sarcoma	13-16	44% 5yrS	[44]
	17-20	30% 5yrS	[44]
	21-24	38% 5yrS	[44]
Soft tissue sarcoma	0-14 vs 15-24	HR = 0.79 (0.62-0.99)	[42]
	17-20	30% 5yrS	[44]
	21-24	25% 5yrS	[44]
Extra gonadal germ cell tumours	0-14 vs 15-24	HR = 0.46 (0.23-0.94)	[42]
Breast cancer	13-16	70% 5yrS	[44]
	21-24	58% 5yrS	[44]
Improved prognosis with older age			
Intracranial embryonal tumours	<4	21% 5yrS	[14, 58]
	4-14	63% 5yrS	[14, 58]
Germ cell tumours	13-16	80% 5yrS	[44]
	17-20	87% 5yrS	[44]
	21-24	90% 5yrS	[44]
Intracranial germ cell tumours	<5	61% 5yrS	[14]
	5-14	84% 5yrS	[14]
Central nervous system tumours	0-14 vs 15-24	HR = 1.49 (1.28-1.77)	[42]
Melanoma	0-14 vs 15-24	HR = 2.99 (1.61-5.54)	[42]
Non-linear age patterns			
Acute lymphoblastic leukaemia	<1 vs 1-9	HR = 4.33, P <0.001	[53]
	<1 vs 10-14	HR = 2.46, P <0.001	[53]

	10-14 vs 0-9	HR = 1.76 (1.35-2.30)	[64]
	<1	62.1% deaths	[65]
	1-9	21.0 % deaths	[65]
	10-14	40.6% deaths	[65]
Rhabdomyosarcoma	<1	53%, 66% 5yrS	[14, 58]
	1-9	72%, 70% 5yrS	[14, 58]
	10-14	51%, 48% 5yrS	[14, 58]
Neuroblastoma	<1	83% 5yrS	[14]
	1-4	46% 5yrS	[14]
	5-14	37% 5yrS	[14]
No effects of age observed			
Acute myeloid leukaemia	0-14	-	[14, 63]
Retinoblastoma	0-14	-	[14, 63]
Renal tumours	0-14	-	[14, 63]
Hepatic tumours	0-14	-	[14, 63]
Hodgkin's lymphoma	0-14 vs 15-24	HR = 0.83 (0.50-1.37)	[42]
Non-Hodgkin's lymphoma	0-14	-	[14, 63]
Ovarian germ cell tumours	0-14 vs 15-24	HR = 0.30 (0.03-3.38)	[42]
Central nervous system tumours	0-14	-	[14, 63]
Bone tumours	0-14	-	[14, 63]

2.3.2.3 Sex

As with adult cancers, survival from cancer amongst CYAs is generally poorer for males compared to females, however, associations vary according to diagnostic group. For childhood ALL, survival was approximately 25% poorer amongst boys compared to girls, however, this pattern was said to emerge 3 years post diagnosis [63, 65]. Stiller [14] showed that survival was only poorer for males compared to females with precursor-cell ALL and not for other immunophenotypes. However, Stiller [14] did not assess the difference in sex at different time periods. Limited evidence from a European wide study showed poorer survival for males compared to females with acute non-lymphocytic leukaemia (ANLL), however this result was not significant [68]. Furthermore, Stiller [14] and Johnston et al. [63] showed no evidence of a difference in survival between males and females with AML, which make up 80% of ANLL cases.

Male TYAs with HL have poorer survival compared to females [44], however, there was no evidence of difference by sex for childhood HL or for NHL amongst children or TYAs.

Amongst children, males with CNS tumours had poorer survival compared to females [60], however, this effect was reversed for the intracranial embryonal tumour subgroup,

in which females had a 37% increased risk of death compared to males [63]. Specifically, Birch et al. [44] have shown that gender does not affect prognosis of CNS tumours amongst TYAs. For children with neuroblastoma, survival overall was poor but significantly better for females compared to males (59% and 52% 5-year survival respectively) [14]. Female TYAs with osteosarcoma had improved survival compared to males [44], however, no sex differences have been shown for any bone tumour subgroup amongst children. Male TYAs with head and neck carcinoma and colorectal carcinoma had poorer survival compared to females, however, no sex differences were observed for thyroid or lung carcinomas [44].

There was no evidence of an effect of sex on survival for CYA renal or hepatic tumours in the UK literature.

2.3.2.4 Temporal Trends

Survival of cancer for children improved significantly over time, with overall 5-year survival for children in Britain having increased from 28% in the late 1960s to 77% in the period 1996-2000 [14]. For TYAs, improvements over time have not been as great as seen amongst children, with 5-year relative survival increasing from 63% in the late 1970s compared to 77% in the period 1996-2001 [44, 59]. However, data for TYAs did not date back as far as the 1960s, therefore potential earlier sharp increases in survival could not be determined.

National data showed improved survival over time for all diagnostic groups amongst children, and for all cancers except STS amongst TYAs [14, 44, 45]. Although regional evidence shows improved survival for TYAs with STS in the north of England (33% in 1968 to 67% in 2008) [64], and in the West Midlands (47% in 1971 to 69% in 2001) [45]. Furthermore, a study from the south East of England was the only study to show a decrease in survival for young adults with bone tumours in England (69% between 1968-1972 to 36.6% between 1998-2002) [61]. This compared to an increase in survival for children with bone tumours (23% in 1960s to 64% in 1990s) in Britain [14] and a 5% increase in the North of England (1991-2001) [62]. For TYAs in England, there was a modest survival improvement for those with osteosarcoma (41% in 1979 to 49% in 2001) and a larger improvement (29% in 1979 to 46% in 2001) for those with Ewing's sarcoma [45].

For childhood leukaemia, survival improved from as low as 9% in the 1960s to 79% in 2000 [14]. The rate of increase for AML was similar to that of leukaemias overall (6% to 65%), and although the rate of increase for ALL was slower, overall survival of ALL was higher in the latest period (83%) compared to AML. More recent data from the Cancer

Research UK website shows that survival from ALL increased further to 88% between 2001 and 2005, but for AML latest survival rates remained at 64% [4]. Improvements in survival for ALL were likely due to the introduction of more intensive treatment protocols after the UK medical research council trial in the 1980s [79]. For TYAs, survival from ALL in England has improved from 41% in the late 1970s to 55% in the period 1996-2001. For TYAs with AML, survival improved from 30% to 50% in the same period [45]. However, data from Birch et al. [44] showed that improvements were restricted to earlier time periods (1979-1984 and 1985-1989) with no further survival improvements beyond then. Despite improvements for both the childhood and TYA age ranges, survival for TYAs with leukeamia remains poorer compared to children (see §2.3.2.2).

Sharp increases in survival were observed for HL (20% to 90% between 1968 and 1988), however, little improvement has been observed after this period [64]. For NHL, a similarly sharp increase in survival was observed (23% to 65% between 1978 and 1997), with a further increase to 83% between 1998 and 2005 [64]. Large improvements in survival of lymphoma were a result of improved treatment regimens and increased understanding of tumour biology allowing for treatments to be more specifically targeted to specific tumour subgroups. For NHL, treatment moved from involved field radiation and combined chemotherapy which caused acute side effects, to a chemotherapy regimen derived from a leukaemia treatment protocol [80]. For HL, initial improvements in survival were a result of the introduction of extended field radiotherapy which involved radiotherapy being administered to surrounding areas of the affected lymph nodes. For TYAs, survival from HL was already high at 85% between 1979 and 1984 but improved further to 93% for the period 1996-2001 [45]. For NHL, survival for TYAs improved from 55% to 71% in the same respective periods [45]. Similar to trends for leukaemia, Birch et al. [44] showed that improvements were restricted to earlier time periods (late 1970s and early 1980s) without further improvements beyond then.

Survival for children with CNS tumours improved significantly from 37% to 71% (1969-2000) [14]. The most recently available data from the period 2001-2005 showed that survival from CNS tumours amongst children had remained the same at 71% [4]. For TYAs, survival had improved from 68% to 74% (1979-2001) [44, 45].

For GCTs, survival amongst TYAs has been consistently high, but despite already high rates, survival has continued improving from 84% in 1979-1984 to 96% between 1996-2001 [44, 45]. For childhood GCTs, survival was considerably poorer at 45% in the mid 1970s [4] but latest figures show it to be comparable to TYA survival (92% between 2001-2005) [4].

Large improvements in survival have been made for hepatic tumours amongst children, which have had historically poor survival (9%, 14% and 30% 5-year survival in the late

1960s and mid 1970s and 1980s respectively) [14]. Survival increased to its current rate of 66% in the early 1990's and has remained the same since then [2]. Treatment for hepatic tumours historically involved complete resection (removal of the tumour), however, advances in treatment have improved outcomes as chemotherapy treatment was introduced to shrink the tumour prior to its removal providing considerably better prognosis [81, 82, 83].

2.3.2.5 Stage at Diagnosis

By definition, the stage of cancer defines the extent and severity of cancer and therefore has strong bearing on the patients prognosis. Despite the importance of stage of disease or other measures of disease severity at diagnosis, many epidemiological studies do not include such measures within their survival analysis. Of the 18 key references in the UK describing survival patterns of CYA cancers, the majority did not include measures of disease severity within their analysis [14, 44, 45, 53, 54, 59, 61, 62, 63, 64, 65, 66]. Tseng et al. [60] adjusted for WHO Grade as a measure of disease severity for CNS tumours and Joshi et al. [58] adjusted for stage in their analysis of rhabdomyosarcoma. In addition, studies by Oakhill and Mann [52], McKinney et al. [55], Powell et al. [57] and Stiller et al. [56] included WBC count as a measure of disease severity for leukaemia, but the latter three of these included a range of other childhood cancers in addition to leukaemia for which disease severity was not studied. Lack of adjustment for the disease severity implies that results showing differences between age groups, ethnic groups and other variables could be confounded by casemix, therefore not showing the true effects of these covariates.

2.3.2.6 Ethnicity

Data on ethnic differences for children and TYAs in England is sparse, with only four studies of childhood cancer [52, 55, 56, 57] and just one recently published study looking at breast cancer differences amongst those aged less than 40 years [84]. Amongst the sparse childhood literature, few consistent patterns of ethnic differences in survival exist. Powell et al. [57] showed evidence of poorer survival for south Asians with ALL compared to white children which was similar to earlier findings of a small study (n=60) in the 1980s using data from UK ALL clinical trials [52]. However, two further studies [55, 56] showed no significant differences for south Asians with ALL compared to non-south Asians. Although Stiller et al. [56] did observe a higher risk of death for non-white children compared to white children with ALL. In the U.S., the overall 5-year cancer survival rate of Hispanic children (72%) was found to be lower than for white children

(84%), whilst black children with ALL had poorer survival than white children (75% vs. 85%) [24, 25]. There was some evidence from a UK wide study of poorer survival for black children diagnosed with neuroblastoma compared to white children, with a hazard ratio (HR) of 1.79 (P=0.047) [56], however, this analysis did not adjust for confounding factors such as deprivation.

The four UK based papers were indicative of possible inequalities in survival between ethnic groups, however the evidence was not convincing or consistent and was largely restricted to ALL. In particular, comparability between studies was difficult because of different methods of identifying ethnic groups. The study by Oakhill and Mann [52] was based on clinical trial data, in which south Asian ethnicity seemed to based on country of birth (India and Pakistan), although the method of identifying the country of birth was not clear. Stiller et al. [56] used self reported ethnicity, whereas McKinney et al. [55] used name analysis to identify south Asian vs. non-south Asian cases. Powell et al. [57] did not mention the origin of ethnicity data. These differences lead to lack of comparability as there could be discrepancies in the assignment or classification of ethnicity between studies. In particular, the name analysis programme was used to identify those of Indian, Pakistani or Bangladeshi origin, however, it is possible that non-south Asian people have south Asian names through marriage for example, or vice versa. In addition, there was no mention of any missing data for Oakhill and Mann [52] and Powell et al. [57], and it is therefore assumed their analysis was a complete case only analysis, which could have resulted in biased estimates (see Chapter 3 for further details). Both Stiller et al. [56] and McKinney et al. [55] treated missing values as separate categories, which is considered an inadequate method of handling missing data as described in Chapter 3, and the use of multiple imputation could have improved estimates and may have resulted in different outcomes.

2.3.2.7 Deprivation

Lightfoot et al. [65] document evidence of a survival gap in ALL according to socioeconomic status (HR=1.29, 95% CI 1.05-1.57, P=0.034) for the most deprived compared to least deprived cases. Importantly, this gap has been shown to widen around 6-9 months post diagnosis, which coincides with the timing of treatment regimens transferring to less intensive home-administered maintenance chemotherapy. This key point highlights that the deprivation gap is not as a result of access to health care, and perhaps explains why differences according to levels of deprivation for childhood cancer survival have not been observed in earlier studies, [54], unlike those seen in the adult population [85]. There is little evidence to suggest that survival of AML for children differs according to socioeconomic status, however, a study in Yorkshire showed that

deprivation does play a role in AML relapses, such that AML cases from the most deprived areas were significantly less likely to relapse (OR=0.54, 95% CI=0.32-0.93) [86].

For TYAs with leukaemia, Birch et al. [44] showed borderline evidence of poorer survival for those in the most deprived group compared to least deprived (41% vs 45%; P=0.048). The subgroup analysis revealed this difference was driven by ALL (42% vs 48%; P=0.066), with no differences in survival according to deprivation observed for AML or CML subgroups. However, the evidence of a deprivation gap in survival for TYAs with ALL is very weak compared to the evidence amongst children, as the former difference was small with borderline significance, compared to a larger 30% significant increase risk of death for more deprived children with ALL.

For lymphoma, there was some evidence that TYAs have poorer survival as deprivation increased by 25% (95% CI 1.08-1.43) on average [66], however, this study was limited to the north of England and similar effects were not observed for a larger study of TYAs across England [44].

For CNS tumours, Tseng et al. [60] found no differences in survival according to deprivation or region of residence in children in England, unlike amongst the adult group for which higher socioeconomic status and living in southern England was associated with better survival. This is in contrast to a study from the Yorkshire region, in which the authors showed that children with CNS tumours from middle affluence were 62% more likely to die than those from the most affluent area (P=0.020) [55]. This finding was confirmed by a later study in the same region, again showing CNS tumour survival for children from most affluent areas (quintile 1) was significantly worse compared to survival from those in the 2nd and 4th quintiles of deprivation [12]. In addition, a study from the North of England provided evidence of poorer survival for children with CNS tumours from more affluent areas compared with more deprived areas [64]. Overall, the evidence suggests that those from more affluent areas have poorer survival. However, the difference has not been consistently shown between the most deprived and most affluent areas, but instead, the significant difference appears to be between the most affluent and the middle affluent areas. Furthermore, the evidence to date appears to be limited to the Yorkshire and north of England region, with no effects observed nationally to date.

For neuroblastoma, a study from Yorkshire by McKinney et al. [55] show a reduced risk of death from childhood neuroblastoma in the second most affluent deprivation quintile compared to the most affluent area (HR=0.40, P=0.020). There were no national studies looking at survival from neuroblastoma in relation to deprivation, thus it is unclear whether or not this effect is consistently observed amongst other areas of the UK.

For TYAs, Birch et al. [44] provides evidence of poorer survival for carcinoma cases in more deprived areas (69% vs. 73% in the most vs. least deprived areas respectively; P=0.008). In particular, survival from head and neck carcinomas and colorectal carcinomas was poorer in more deprived areas [44]. Survival for carcinoma of the thyroid, lung, ovary and cervix did not vary according to deprivation level [44].

2.3.2.8 Geographical Variation

Differences in survival patterns have been observed within England and the UK and regional differences in temporal changes in survival for bone tumours have been described earlier in §2.3.2.4. For children with ALL, some evidence of regional variation in survival trends have been observed in a national study by Schillinger et al. [54] which covers diagnoses of ALL in the period 1971-1990. The authors describe that the North and West NHS region have 5-year survival rates of ALL which are 4% and 7% lower than the national average respectively. The authors do however make the overall conclusion that there is no strong evidence of a north/south divide in terms of survival of ALL. For TYAs, a study across all common cancer groups in England revealed that there was significant variation in survival by government office region for GCTs, with 5-year survival ranging from 87% in the south west compared to 95% survival in London [45]. In addition, significant survival differences by region for colorectal cancer were observed in the same study, with 5-year survival as low as 41% in the Yorkshire and Humber region compared to 77% in the east of England. For both GCTs and colorectal cancer, the regional effects remained significant after adjusting for socioeconomic deprivation per region.

2.3.2.9 Summary of variation in Cancer Survival Rates

Overall, childhood survival in each diagnostic group is well described within the literature and most studies consistently use the 0-14 year age range making results comparable in general. The literature portrays a positive message in terms of improvements in survival over time. Despite certain subgroups having experienced significant improvements historically, more recent figures show little improvement in the last 5 to 10 years. For example, while NHL survival continues to improve, HL survival improved drastically in the mid 1970s to 5-year survival of 90% with little or no improvement documented since the late 1980s [64]. Large increases in survival were mainly a result of the development of new treatment regimens and a better understanding of specific diagnostic groups, however, perhaps the reason for no further improvements beyond those was because improvements and refinements in treatment protocols were harder to make when survival was already very high. Nevertheless, there were other diagnostic groups for which prognosis remained

poor. For example, survival from bone tumours amongst CYAs was between 50 and 60%, and that of STS was 65% overall but less than 25% for those diagnosed with advanced stage tumours.

Despite positive improvements overall, there are still gaps in survival according to gender, age, deprivation and ethnicity. CYA cancer is more commonly diagnosed amongst males than females, but survival rates are also poorer in many diagnostic groups for males compared to females, including leukeamia, gliomas, neuroblastomas and osteosarcomas. In terms of age, children under the age of 1 have significantly poorer survival of leukaemia compared to children between 1 and 10 years of age, and the youngest astrocytoma cases also have a very poor prognosis. Additionally, in general TYAs have poorer survival for certain tumour groups compared to children. Reasons for TYAs having poorer outcomes were unclear, however, it could be a result of later diagnosis of cancer amongst this age group or due to differences in the precise types of tumours which occur in this age group. In addition, access to effective treatments and clinical trials are also known to be poorer for TYAs compared to children (19% and 51% of TYAs and children with cancer respectively were enrolled onto clinical trials in 2006 [87]), which could result in poorer survival for this age group.

There is strong evidence showing survival differences according to deprivation amongst adults with cancer across the world [88], with many studies showing poorer survival for people from more deprived areas for example for breast cancer [89, 90, 91], bowel cancer [92, 93] and cervical cancer [94]. Similar evidence of differences in survival by ethnic group [95, 96, 97] have been shown, and the effects of ethnicity and deprivation on health care were often thought to be highly correlated. Moreover, ethnicity and deprivation have both been shown to affect the severity of disease at presentation amongst adult cancers [95, 96, 98, 99, 100, 101]. Section 2.3.2.5 discusses the severity of cancer at diagnosis and its importance as a prognostic factor, however, inequalities in disease severity amongst CYAs with cancer are unclear despite evidence of such inequalities amongst adults. For example, for adult cancers in the U.S., African-American women with breast cancer and black and Hispanic patients overall have been shown to have a significantly higher proportion of advanced stage cancers compared to other ethnic groups [102, 103, 104]. However, there has been some suggestion that these differences by ethnic group can be largely explained by deprivation [105]. In Scotland, poorer survival of breast cancer in women from more deprived areas was attributed to these women presenting with more advanced tumours compared to women from less deprived areas [106]. However, a study in East Anglia concluded that stage only partly accounted for breast cancer survival differences [98]. A recent study in the U.S. attempted to disentangle the effects of ethnicity and deprivation on the effect of stage at diagnosis and concluded that ethnicity was a stronger predictor of stage than socioeconomic status [107]. However, importantly,

their analysis simply categorised missing data of stage at diagnosis as a separate level which does not adequately account for missing data and could lead to biased results (as discussed in Chapter 3).

Amongst CYAs with cancer, inequalities in survival according to deprivation or ethnic group are less clearly described than for adults, however, results tended to differ between studies which adjusted for stage or severity of disease at presentation and those that did not. Tseng et al. [60] showed no difference in survival of childhood CNS tumours according to deprivation after adjusting for WHO grade, whereas McKinney et al. [55] showed that children from middle affluent areas in Yorkshire were at an increased risk of death compared to those from affluent areas, however, data on WHO grade was not included in this analysis. For leukaemia, Lightfoot et al. [65] showed that children from more deprived areas had poorer survival with no adjustment for severity of disease, whereas Schillinger et al. [54] did not observe significant differences according to socioeconomic status for children with leukaemia. Although Schillinger et al. [54] also did not adjust for disease severity at presentation, they excluded 10% of cases using listwise deletion due to missing data, therefore calling into question the accuracy of these results. In particular, they excluded all cases who were diagnosed on the same day as their recorded date of death, which may have led to biased results as these were potentially the cases with most advanced disease at presentation.

In terms of ethnicity, two previous studies have shown no difference in survival of leukaemia amongst children according to ethnic group [55, 56] whereas Powell et al. [57] showed that south Asian children with leukaemia had poorer survival compared to non-south Asian children with the same diagnosis. The evidence of ethnic differences was restricted to childhood cancers, predominantly focused on leukaemia and was out of date (only including diagnosis up to the mid 1990s). Furthermore, missing data was not handled adequately in any of the papers assessing childhood survival by ethnic group as discussed in §2.3.2.6. The evidence base of ethnic differences for childhood cancers in the UK is very weak, and for TYAs is limited to one study of breast cancer cases including TYAs and older adults up to the age of 40 [84].

2.3.3 Long Term Effects Amongst Survivors of Cancer

The previous section on survival of CYA cancer in the UK shows that survival has improved substantially since the 1960s and 70s and is high in general for children as well as TYAs. This improvement in survival over time means there is a growing cohort of long term survivors of childhood and TYA cancers within the UK, and a study of all cancers in the UK at all ages shows increasing numbers of survivors by a rate of 3%

per year on average [108]. Due to this increase in survivorship, focus is shifting from simply improving overall survival towards the quality of that survival and minimising occurrences of long term side effects of treatment. This is also reflected by the NHS outcomes framework, which emphasizes the need to focus on reducing the potential years of life lost [21]. In addition, the improving outcomes guidance for children and young people [43] highlights that although many CYAs survive cancer, there is a substantial risk of late effects (LEs) for survivors and therefore an increasing need for this group to be monitored in long term follow up clinics.

There is a wealth of data on the long term effects amongst survivors of childhood cancers, which have primarily arisen from two large scale retrospective cohort studies; the North American Childhood Cancer Survivorship Study (CCSS) and the British Childhood Cancer Survivorship Study (BCCSS). The CCSS includes cases diagnosed under the age of 21 with cancer between 1970 and 1986, and the BCCSS includes cases of diagnosed under the age of 16 between 1940 and 1991. Thus far, 346 publications since the 1980s have emerged using CCSS and BCCSS data to describe the LEs amongst survivors of childhood cancer which include an increased risk of second malignant neoplasms, neurocognitive impairment, cardiotoxicity, fertility problems and psychological effects [109, 110, 111, 112, 113, 114, 115, 116, 117, 118]. Oeffinger et al. [16] provides evidence that approximately two-thirds of survivors of childhood cancer are expected to develop at least one chronic health condition, and almost a third developing a severe, life threatening or disabling condition. Reasons for this increased risk have largely been attributed to the intensive treatment protocols which have helped improve overall survival rates of cancer for CYAs but are also partially thought to be due to specific genetic predisposition [119]. Some specific examples of LEs are described below.

Survivors of childhood leukaemia are said to be at an increased risk of secondary cancers, cardiovascular conditions, infertility and growth failure [120, 121, 122]. Mody et al. [122] report that survivors of childhood leukaemia are at a 3-fold increased risk of developing multiple chronic medical conditions compared to siblings and a 4-fold increased risk of developing severe and life threatening conditions compared to siblings. The cumulative incidence of developing a chronic medical condition for survivors of childhood leukaemia was 13% at 25 years from diagnosis, but was as high as 21% for those who also received radiation therapy as part of their treatment [122]. Increased exposure to anthracylines (specific types of chemotherapy drugs) as well as treatment at a younger age are known risk factors for developing cardiovascular LEs for survivors of childhood leukaemia [123, 124].

For HL, initial improvements in survival were a result of the introduction of extended field radiotherapy which involves radiotherapy being administered to surrounding areas of the

affected lymph nodes. LEs as a result of these intensive treatments include weakening of the immune system, sterility, secondary tumours or cardiovascular complications [125, 126]. The excess risk of cardiac death amongst children and adolescents who have received intensive treatment for HL compared to the general population is 17.1 per 10,000 person years [126].

For childhood survivors of brain tumours, the relative risk for stroke is approximately 43 compared to healthy siblings, the relative risk for blood clots is almost 6-fold and the relative risk for angina-like symptoms is two-fold [127].

For retinoblastoma survivors, the risk of developing secondary non-ocular tumours (tumours not related to the eye) as late as 50 years post diagnosis is almost 50% (95% CI 38-60%) for heritable retinoblastoma, however, this risk is dramatically lower for non-heritable retinoblastoma cases (5%; 95% CI 2-12%) [128]. The most common types of cancer following hereditary retinoblastoma are sarcomas, carcinomas, brain tumours and melanoma [128, 129].

LEs following malignant bone tumours include cardiac complications due to anthracycline treatment, as well as a risk of developing secondary tumours such as leukaemia and lung cancer approximately 7.5 years after the end of treatment [130].

For rhabdomyosarcoma, effective treatment includes a combination of chemotherapy, surgery and radiotherapy. However, the use of all three of these treatments can cause adverse long term effects and there is an ongoing debate in the medical community around whether radiotherapy is really required in all patients [131]. Recurrences or secondary malignant neoplasms occur in approximately 9% of rhabdomyosarcoma cases 5-years beyond treatment, and the risk of having a late event after treatment is highest in those whose original diagnoses was for an advanced stage tumour, which is caused by the increased level of therapy required in these patients [132].

In general, cardiovascular related events have been identified as one of the most important LEs for survivors of CYA cancer due to the long term morbidity associated with it in addition to the potential for early mortality in survivors when compared to the general population [113, 114, 118]. The risk of cardiovascular related LEs are said to increase with the dose of anthracyclines, ranging from approximately four- to 28-fold compared to those not receiving anthracyclines and a 5-fold increased risk of cardiac disease is found after radiation to the heart [112, 116]. Although high dose anthracyclines increase the risk of cardiovascular LEs, cardiotoxicity is also observed in survivors who received low dose anthracycline treatment [116]. The risk of cardiovascular LEs is said to be greatest for patients receiving radiotherapy combined with chemotherapy [113, 118, 127].

2.3.4 Key Gaps in the Knowledge

Continued research into the epidemiology and aetiology of all diagnostic groups and subgroups is required with the aim of improving survival for all CYA cancers. The main challenge that exists in this field of research is the small number of cases within each diagnostic group and subgroup due to the rarity of cancer in children and TYAs. This becomes even more of an issue in terms of statistical power when researchers are interested in looking at survival within subgroups as well as by age, sex and other demographic and medical related variables.

Despite 11 studies of childhood cancer survival and 7 studies of TYA cancer survival across the UK, only 1 recent publication explicitly compares survival between childhood and TYA cancer [42]. The results of this study were only adjusted for age and sex, and although the research is broad and includes many diagnostic groups, the authors have not included year of diagnosis, deprivation, ethnicity or stage, which are all potentially important predictors of survival amongst children and TYAs. The authors do acknowledge that stage is an important prognostic factor, but declare they were unable to assess this due to the the high level of missing data. It is unclear whether variation in survival observed in the UK for CYA cancer is due to differences in stage at diagnosis as very little research exists for CYA survival which takes disease severity into account (see §2.3.2.5). In addition, survival differences between ethnic groups for children in the UK remain unclear, with few studies showing consistent results. There was only one study focusing on differences by ethnic group amongst TYAs, however, this study only looked at breast cancer diagnosis and their age range was not specific to TYAs as it included all cases under the age of 40 [84]. One of the key issues for lack of data on the survival of CYA cancer according to disease severity and ethnicity is missing data; the National Audit Office report on delivering the CRS as well as the improving outcomes for CYAs with cancer report recognises that stage at diagnosis remains a key gap in cancer intelligence [1, 22].

This study aims to address the paucity of survival data in the UK across the CYA age range which is adjusted for the risk factors identified from the current literature, including age, sex, deprivation, ethnicity, year of diagnosis and stage or disease severity at diagnosis. Adequate handling of missing data, through the use of multiple imputation, will allow for partially observed variables to be included within the analysis as well as minimising the chance of obtaining biased estimates (further details discussed in Chapter 3).

It remains unclear whether possible differences in survival by age, ethnicity or deprivation are related to inequalities in disease severity at diagnosis for CYAs in the same manner as for adults. The study by Lightfoot et al. [65] showed that the survival gap in deprivation

arose around the same time as a change from hospital based chemotherapy administration to home based chemotherapy treatment, thereby suggesting that the deprivation gap in survival for leukaemia was more likely a result of differences in treatment adherence according to deprivation rather than a result of delayed access to healthcare. To date, no research in the UK has focused on the effects of ethnic group or deprivation on the severity of cancer at diagnosis for CYAs. Furthermore, it is unclear whether there are differences in disease severity between children and TYAs. In order to address this, the effect of ethnic group, deprivation and age group on the severity of disease at diagnosis was studied using multiply imputed data for partially observed variables (Chapter 7).

Despite a growing understanding of the risks of LEs amongst survivors of childhood cancer, most studies rely on the CCSS and BCCSS for data. Despite these studies providing a huge resource and long term follow up data, they only include cases diagnosed up until 1986 and 1999. This means the long term effects of more modern treatment modalities are not yet explored, in particular for the CCSS data on which the majority of LEs papers are based. In addition, the BCCSS only includes data on childhood cancer survivors up to the age of 16 at diagnosis, and although the CCSS includes cases up to the age of 21 at diagnosis, this covers only a part of the TYA age range. The review by Woodward et al. [133] highlights the paucity of information on the long term effects of survivors of TYA cancer, despite hypothesised concerns over the risk of LEs amongst this age group. The development of long term follow up services for TYA survivors with cancer is under way in the UK [133], despite a lack of evidence about the burden of LEs amongst TYA survivors. Further limitations of the CCSS and BCCSS data are that they are retrospective cohort studies which rely on self-reported outcomes. Self-reported outcomes can produce recall bias as, for example, people may be more likely to report they have suffered from a heart problem, knowing that they are taking part in a survey about possible LEs of their cancer treatment. Population based data on survivors of CYA cancer are required to study LEs in an objective manor in order to improve intelligence on the subject [134]. Hawkins [135] and Zhang et al. [136] identify the need for data linkage of routine datasets in order to study LEs amongst survivors to provide timely and objective data on the LEs experienced amongst CYA survivors of cancer. Many LEs have been identified within the current literature, and it is outside of the scope of this thesis to study them all in detail. As described in §2.3.3, cardiovascular LEs have been identified as one of the most important co-morbidities amongst survivors of childhood cancer. To date, there is limited data on cardiovascular disease sufficient to warrant admission to hospital for childhood cancer survivors, for TYAs the paucity of information on cardiovascular LEs is acknowledged [117, 133]. This study aims to use linked routinely collected population-based cancer registry data with national administrative data on hospital admissions (HES data), to quantify the incidence and risk of cardiovascular

LEs amongst survivors of CYAs with cancer. The results of this analysis are provided in Chapter 8.

Chapter 3

Review of Missing Data Methodology

3.1 Introduction

This chapter provides a detailed evaluation of the effect missing data can have upon statistical analyses, a review of the techniques currently available to handle missing data problems as well as a critical evaluation of the use and accurate implementation of these techniques within medical research.

Missing data arises frequently in medical research and could occur in the outcome of interest or within the explanatory variables. Data can be missing as a result of never being collected (by design or otherwise), being lost after collection or through being incorrectly collected and therefore deleted. However, despite being common, researchers are often not aware of the impact missing data can have upon their analysis. Furthermore, many statistical techniques and packages are designed for complete data, and as such analyses are often performed under the assumption that the data are complete [137].

Ignoring missing data could lead to inaccurate and biased estimates, as well as lack of power and efficiency; missing data problems are discussed in detail in §3.4. However, prior to this, an overview of the terminology commonly used within the missing data literature is given §3.2.

There are a range of statistical methods which can be used to analyse incomplete data, including ad-hoc methods (deletion methods and single imputation techniques) as well as more advanced techniques (maximum likelihood estimation (MLE), multiple imputation (MI) and inverse probably weighting (IPW)). Ad-hoc techniques are explored in §3.5. More advanced techniques, including MLE, MI and IPW are explored in more detail in §3.5.3, §3.5.4 and §3.5.5 as these are recognised as established and advanced imputation techniques by key authors in the field of missing data techniques

[138, 139, 140, 141, 142, 143, 144] and offer considerable advantages over the ad-hoc techniques as is detailed in this Chapter. Furthermore, MLE, MI and IPW are all methods which can be applied to population based datasets as used within this thesis. Other advanced techniques for handling missing data include multiple imputation by latent class analysis [145, 146] and non-parametric Bayesian multiple imputation [147, 148]. Multiple imputation by latent class analysis and non-parametric Bayesian multiple imputation methods were developed for high dimensional categorical datasets (of the order of 80 or more categorical variables) and were therefore not applicable to this study which included a small number of categorical and continuous variables.

The level of use of each of MLE, MI and IPW within the medical literature was assessed by performing a critical review of the current literature. Searches were performed for English language articles only within MEDLINE and Web of Science dating back to the earliest known reference for each of the respective techniques to the present day. The search strategies excluded subject categories from Web of Science that were not medical related topics (for example economics and geography), however, mathematical and statistical papers were retained within the search. Details of the exact search strategies for imputation by MLE, MI and IPW are given in Appendices E, G and F respectively.

Finally, the chapter concludes with a detailed evaluation of the advanced missing data techniques in order to determine which of these is most appropriate for the analysis within this thesis (§3.6).

Missing data notation used throughout this chapter follows that as in Bartlett et al. [149].

3.2 Missing Data Mechanisms

There are several ways in which data can be missing, and in the presence of each of these mechanisms a different technique for handling the missing data may be required [150]. Missing data mechanisms, and their importance in determining the appropriate missing data techniques, were generally overlooked until Rubin [151] in 1976 formalised the concept. Rubin [151] developed the following three classifications of missing data; missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). In order to describe these three mechanisms, the following definitions are required:

Fully Observed Outcome (Y)

Y is the fully observed outcome variable of interest within the analysis.

Partially Observed Covariates (X)

The vector X contains p partially observed covariates, such that $X = (X_1, \ldots, X_p)$. X^{mis} and X^{obs} denote the missing and observed components of X for a given subject, such that $X = X^{mis} + X^{obs}$.

Fully Observed Covariates (Z)

The vector Z contains q fully observed covariates, such that $Z = (Z_1, \ldots, Z_q)$.

Auxiliary Variables

Auxiliary variables are a subset of fully and partially observed variables within the data which could aid the prediction of partially observed variables, but are not themselves of primary interest in the analysis.

Complete Data

The complete data refers to all elements of the data that are either observed or were intended to be observed, and thus includes Y, X and Z.

Missing Data Indicator R

The missing data indicator can be presented by a matrix of the same size as the completed data which contains a value of zero when the corresponding value of X is observed and a value of one when the corresponding value of X is missing.

3.2.1 Missing Completely at Random (MCAR)

Data are considered MCAR when the missing data are unrelated to either X^{obs} or X^{mis}, conditionally on the covariates of interest. In other words there are no systematic differences between the observed and missing data and information contained within other variables in the dataset is unable to aid prediction of the missing data. MCAR is represented by the following equation:

$$f(R|Y, X, Z, \Phi) = f(R|\Phi) \text{ for all } Y, X, Z, \Phi,$$

where Φ denotes unknown parameters.

For example, data are said to be MCAR if a laboratory sample is accidentally dropped. If the data are MCAR, then they are a simple random sample of the complete data. This implies that the analysis of a dataset with missingness under the MCAR assumption will not result in biased estimates, provided the analysis of the data in the scenario in which it was complete would also not result in bias [152]. However, a complete case analysis under MCAR would result in a loss in precision.

3.2.2 Missing at Random (MAR)

Data are MAR when the probability that the data are missing is independent of the missing data itself conditionally on the observed data and the covariates of interest. This is represented by the following equation:

$$f(R|Y, X, Z, \Phi) = f(R|Y, X^{obs}, Z, \Phi) \text{ for all } Y, X^{mis}, Z, \Phi.$$

In other words if any systematic differences between the observed and missing data can be explained by differences in the observed data, then the data are considered MAR. Therefore, the collection of as many covariates which can improve prediction of the missing values as possible increases the plausibility of the MAR assumption [153, 154, 155, 156]. These variables are known as the auxiliary variables. Missing values of blood pressure data can be considered MAR if, for example, older patients were more likely to have their blood pressure recorded. In this case, blood pressure would be MAR with the missingness conditioned on the observed ages of the individuals as missing blood pressure could be predicted from age.

MCAR is a special case of MAR, with the latter being an assumption which requires the missing data to be a random sample of the observed data only and not of the complete data [153]. Therefore, analyses under an MAR assumption would still be valid if the data were MCAR [148].

3.2.3 Missing Not at Random (MNAR)

Data are MNAR when the probability of the data being missing is dependent upon the missing values themselves after conditioning on the observed data and the covariates of interest. This implies that when the data are MNAR, the missing data are related to unobserved values of the partially observed variable. MNAR is also referred to as non-ignorable or informative missing data [148]. For example, if blood pressure was less likely to be recorded for those people who had low blood pressure and none of the other observed variables could fully explain this relationship, then any missing values of blood pressure would be MNAR. This implies that missing blood pressure could not be predicted by any other observed data; unlike under the MAR assumption.

3.3 How to Determine the Missing Data Mechanism

There are no definitive ways to test whether data are MCAR, MAR or MNAR. However, it is possible to use the observed data in order to make informed decisions about the missing data mechanism. Evidence against MCAR can be found by determining whether the missing data pattern differs according to a variable within the data that is fully observed. There is evidence against MCAR if for example age was related to whether or not blood pressure was recorded. However, if no such relations are observed, it does not imply that the data were necessarily MCAR, they could be MNAR. Despite this caveat, a test for MCAR was developed by Jamshidian and Jalal [157] and has recently been implemented in the R package MissMech [158]. The test determines whether subsets of data have identical missing data patterns or not, by testing for homoscedasticity of the covariance matrices of the missing data patterns using techniques described by Hawkins [159]. Although the MissMech test is useful for complex missing data problems, for simpler scenarios with only one or two partially observed variables, logistic regression models can be easily implemented to determine whether relationships between observed data and the missing data patterns exist. In order to determine whether an MAR assumption is plausible, a sensitivity analysis can be performed by analysing the data under an MNAR assumption [160] (see Chapter 4 for further details of sensitivity analysis methods).

3.4 Missing Data Implications

As discussed in the introduction of this chapter, missing data is often overlooked by researchers due to a lack of awareness or understanding of the impact missing data can have upon their analysis. Missing data could undermine the validity of research as it could lead to lack of power and inefficiency of the analysis as well as inaccurate and biased estimates unless it is accounted for using advanced missing data techniques (see §3.5 for further details).

Examples of ways in which missing data could lead to inaccurate results include excluding a variable which has partially observed data or excluding all cases for which any observations are missing (complete case analysis (CCA). These methods are discussed in more detail in §3.5.1.1, however, are introduced here to enable a discussion of the implications missing data can have on an analysis. If a partially observed variable is a confounding variable within an analysis (that is to say, a variable which correlates with the dependent and independent variables), then simply excluding this variable from the analysis could generate misleading results. The need to adjust for confounding variables is well known in statistical analysis in general to reduce the risk of false positive (Type

I) errors. Moreover, if the missing data occurs in the outcome of interest, then excluding this variable from the analysis is not a viable option.

More commonly, when researchers ignore missing data (knowingly or not), they tend to perform a CCA in which all cases with any missing data are excluded from the analysis. This has several implications on the analysis, including a loss of power and precision as well as the potential to produce biased results. Power and efficiency of the analysis are lost because CCA reduces the overall sample size included in the analysis, and although each partially observed variable may only have a small percentage of missing data, a large number of cases could be excluded from the analysis if there are many variables with missing data [142]. Moreover, CCA does not only exclude the missing observations themselves, but also excludes valuable recorded information in other variables for the deleted cases. For example, if a person's age was missing from the dataset, but all other data were recorded, the whole record would be excluded from analysis. The resulting estimates may be unable to detect statistical significance due to reduced power and therefore important associations in the data could be missed, or the analysis may lead to erroneous inferences due to increased standard errors by disregarding some of the recorded data. As discussed earlier (§3.2), CCA will not lead to biased estimates if the data are MCAR. However, if this assumption does not hold, then biased estimates could occur. This is because the observed data are no longer a simple random sample of the intended complete data. This implies there could be under-representation of certain subgroups (selection bias) [161]. For example, if data are more likely to be missing in rare subtypes of cancer then by excluding missing data, these rare subtypes would be under-represented in the analysis. If in addition the survival rates of this particular cancer were poorer compared to other more common subtypes, then the overall survival patterns of the cohort would be overestimated in a CCA. Furthermore, if the subtype is already rare, and data are also more likely to be missing for this subtype, a CCA could mean it is not possible to estimate survival for this particular group as the potential of excluding 100% or near 100% of these cases is high. In general, the presence of bias means that the results may have limited validity and limited generalisability [162]. In medicine, presenting inaccurate and invalid results could have serious consequences as the presented evidence may be used to change clinical practice [163].

3.5 Techniques for Handling Missing Data

There are several ad-hoc methods for handling missing data including deletion methods and a range of single imputation methods. These methods are described in this section alongside their advantages and disadvantages. More advanced methods such

as handling missing data by maximum likelihood estimation, multiple imputation and inverse probability weighting are given towards the end of this section. These methods are compared and contrasted and their use within the medical literature is reviewed.

3.5.1 Deletion Methods

3.5.1.1 Listwise Deletion

Listwise deletion (also referred to as CCA), is a method in which subjects with missing data in any of the variables of interest are excluded. CCA is represented diagrammatically in Figure 3.1.

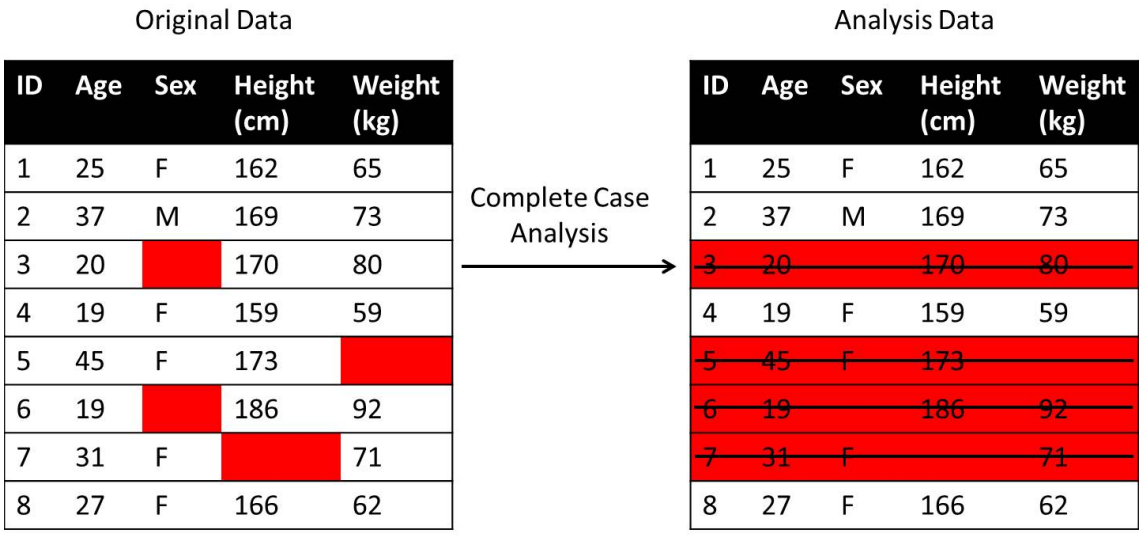


Figure 3.1: Diagrammatic representation of complete case analysis.

This technique has the advantage of being a simple method of handling missing data as no additional data manipulations are required prior to analysis. In a CCA the sample of subjects remains the same for all variables which means that for several univariable analyses on one dataset the results are directly comparable. In many statistical packages, the default setting of statistical analysis such as regression analysis is to discard subjects with any missing data, making it easy for the researcher to perform a CCA without necessarily realising that in doing so they are making several assumptions about the data which could lead to potential bias. Despite CCA being the default setting in many statistical software packages, the technique is not accepted as an appropriate method for handling missing data due to several disadvantages [161]. Unless the data are MCAR, a CCA will produce biased results, increased standard errors and loss of power.

Nevertheless, a CCA can be justified if the data are deemed to be MCAR or if the amount of missing data is relatively small. There are no specific rules for when a CCA is justified,

and therefore the researcher has to make an informed decision contrasting the added benefit of being able to complete a straightforward analysis to the loss of precision in the final results [161]. Despite no general rules being available, Schafer [153] suggests that when missing data occurs in less than 5% of cases, a CCA may be a reasonable approach.

There were several examples of the use of CCA within the medical literature; however, these techniques are often not fully justified within the journal articles. For example, a paper published in the British Medical Journal reporting results from a randomised control trial looking at the effect of acupuncture compared to standard treatment for lower back pain used a CCA [164]. Data on participants of the trial were collected at 3, 12 and 24 months after commencement of treatment and missing data occurred in 24% of cases at the 24 month time period. The authors state that in addition to the CCA, they also performed a sensitivity analysis by analysing the data using the 'last observation carried forward' (LOCF) technique (see §3.5.2.1 for details of this method). They concluded that there were no differences between the results between the CCA and the LOCF analysis. Although the authors explored two options of handling missing data, and concluded that there was no difference in the findings between the two methods, they did not acknowledge that the two methods chosen could have led to biased results and increased standard errors caused by the loss of information. The authors did not make any statements with regards to the structure of the missing data, and as discussed above, unless the missing data were MCAR, the CCA would have led to biased inferences.

3.5.1.2 Pairwise Deletion

Pairwise deletion (also referred to as available-case analysis) excludes cases for which the variables in a particular sub analysis contain missing data as opposed to the cases which have missing data in any variables within the whole dataset. For instance, if the researcher has a dataset with age, gender, height and weight and aims to analyse the relationship between height and weight, then only those cases for which height and weight data were missing would be excluded, regardless of missing data in the age or gender variables (see Figure 3.2).

In situations where the analysis is simple, and only performed on small subsets of the data, this approach could be useful as there is less data loss in comparison to listwise deletion. However, once regression analysis is required with even a small number of variables this method can soon result in a similar amount of data loss as would be the case with listwise deletion. A further disadvantage of pairwise deletion is that results on subsets of the data will have different sample sizes and will therefore not be comparable [161]. As with all

ID	Age	Sex	Height (cm)	Weight (kg)
1	25	F	162	65
2	37	M	169	73
3	20		170	80
4	19	F	159	59
5	45	F	173	
6	19		186	92
7	31	F		71
8	27	F	166	62

ID	Age	Sex	Height (cm)	Weight (kg)
1	25	F	162	65
2	37	M	169	73
3	20		170	80
4	19	F	159	59
~~5~~	~~45~~	~~F~~	~~173~~	
6	19		186	92
~~7~~	~~31~~	~~F~~		~~71~~
8	27	F	166	62

Figure 3.2: Diagrammatic representation of pairwise deletion analysis: example in which the variables required for analysis are height and weight.

deletion methods, subgroups of the study population could be underrepresented within each of the subset analyses, leading to selection bias and invalid estimates.

3.5.2 Single Imputation

Imputation is a method of filling in the missing data with alternative values, and then analysing it as if it were the true complete data. In single imputation the missing data are imputed with a single value. There are many types of single imputation, including hot deck imputation, unconditional mean imputation, regression imputation and stochastic imputation. These are discussed in more detail below. Single imputation has the advantage of producing a complete dataset which can then be analysed using standard statistical techniques and is superior to deletion methods as it does not exclude any observed data. The drawbacks of single imputation are that, except for stochastic imputation, the estimates are biased even when the data are MCAR. Furthermore, single imputation leads to an underestimation of the variance as the uncertainty of the imputed estimates is not taken into account, and therefore multiple imputation methods (§3.5.4) supersede the more simplistic single imputation techniques.

3.5.2.1 Last observation carried forward

The LOCF method is commonly used for missing data within longitudinal studies throughout the medical literature [165, 166]. The method simply replaces the missing value at time $t+1$, $X^{mis}_{(t+1)}$, with the previously observed value, $X^{obs}_{(t)}$, at time t.

The main flaw of this method is that it makes the assumption that the value of a certain variable remains unchanged after dropout which is a strong and unrealistic assumption in most situations [167]. Despite this known flaw, it is a method regularly used in longitudinal studies, for example, in papers by Ginsberg and Lindefors [165] and Tariot et al. [166].

3.5.2.2 Mean Imputation

The most basic form of mean imputation is referred to as unconditional mean imputation. This is a method of single imputation in which missing values are replaced by the mean of the observed values for that variable. Mean imputation dates back to 1932 [168], and has been used throughout medical research since then despite some clear disadvantages. Examples of mean imputation used within the literature include a recent paper published by the Hammill Institute on Disabilities looking at the expectations employers have of individuals with and without disabilities [169] and a study assessing major adverse cardiac events in relation to anxiety and depression [170]. Ju et al. [169] excludes 20 cases due to incomplete data collection forms, with a further two cases for which "the conventional data imputation method of substituting means for missing values" was used. Despite mean imputation being implemented for only two cases, the authors show a lack of understanding of missing data and the potential negative influence may have upon their results by also excluding 20 incomplete cases from the analysis. Frasure-Smith and Lesperance [170] use mean imputation to impute missing values of blood pressure at baseline for 11 cases out of approximately 800. The authors also perform regression imputation (see next section), but conclude that there was no difference in the results between the two methods.

Unconditional mean imputation has been described as the worst method of imputation which should always be avoided [141, 144], reasons for which are described below. The papers published in recent years by Frasure-Smith and Lesperance [170] and Ju et al. [169] indicate either a lack of understanding amongst the medical research community of the implications incorrect imputation can have or a lack of willingness to implement more complicated methods. Despite having identified only two such papers, there may be more within the literature that were hard to identify as papers may have used this technique of imputation without being fully aware of the fact that they were imputing data, and therefore would not have used the terminology for mean imputation or explicitly stated that they had missing data.

Mean imputation results in a distribution of $\mathbf{X}$ which has a spike at the mean and is therefore more tightly centred on the mean than the true distribution of $\mathbf{X}$. This implies

that the variance of the true distribution of $\mathbf{X}$ is underestimated. Mathematically, this is represented as follows:

If s^2 is the sample variance of the observed available data, then the sample variance of the imputed data combined with the observed data (s^2_{imp}) is described by the following equation:

$$s^2_{imp} = s^2 \frac{(n_{obs} - 1)}{(n - 1)},$$

where n_{obs} is the number of subjects with observed data and n is the total number of subjects. Under the MCAR assumption, s is an estimate of the true variance, so the variance of the filled in data is underestimated by a factor of $(n_{obs} - 1)/(n - 1)$. If subsequent analysis is performed under the assumption that the data are complete, resulting estimates such as variances or percentiles will not be accurate, as the distribution of the imputed and observed data is distorted from that of the true complete distribution.

Conditional mean imputation is a somewhat more refined approach of mean imputation in which mean values are conditioned upon observed data items that are subsequently used to replace missing values. For example in survey data, conditional mean imputation involves splitting the data into respondents and non-respondents based on various categories (referred to as classes in this context). The mean for the respondents in each of the classes is then used as the value for the non-respondents in the same class.

3.5.2.3 Regression Imputation

Regression imputation is a method by which observed data is used to predict missing data estimates. The method works by regressing all the observed cases of a partially observed variable on other variables within the data set. This model is then used to generate predictions for the missing values. For example, a regression line of the variable X_2 on X_1 will allow predictions to be made for the values of X_2. This method can be applied using a combination of variables within the data including continuous and categorical variables. Regression imputation reduces down to conditional mean imputation in the situation where the observed variables used in the prediction are dummy variables representing a categorical variable. This is because the resulting predictions are equal to the mean values within each class.

The disadvantage of regression imputation is that the imputed data values are highly correlated as the imputed values lie directly on a straight line if the imputation model is univariable or a flat surface if the imputation model is multivariable. This feature also implies that the imputed values have less variability compared to if the data would have

been fully observed. The attenuation of variances are not however to the same extent as those resulting from unconditional mean imputation. Beale and Little [171] and Buck [172] have developed adjustments that can be made to the results to ensure estimates are unbiased, however, this is only valid if the data are MCAR. Regression imputation can be unreliable if the missing data occurs for values of X_1 which lie outside of the range of X_1 values corresponding to the observed X_2 values. Furthermore, regression imputation does not allow for missing data in the predictors of the variable being imputed. Therefore, if such a situation arises, variables would be imputed based on a complete case analysis of the other variables in the data and therefore could result in biased estimates as with any complete case analysis.

3.5.2.4 Stochastic Regression Imputation

Stochastic regression imputation is similar to regression imputation as it imputes missing values based upon predictions from a regression model. However, it has the additional feature of including a normally distributed residual error. Instead of imputing the conditional mean as described above, the imputation is based upon a draw out of the set of observed values conditional upon the observed data (referred to as a conditional draw). This draw includes a residual term to reflect the uncertainty of predicted values, and is therefore an improvement upon regression imputation. Figure 3.3 gives a simple diagram of how regression and stochastic imputation work for imputing missing values of height based on weight.

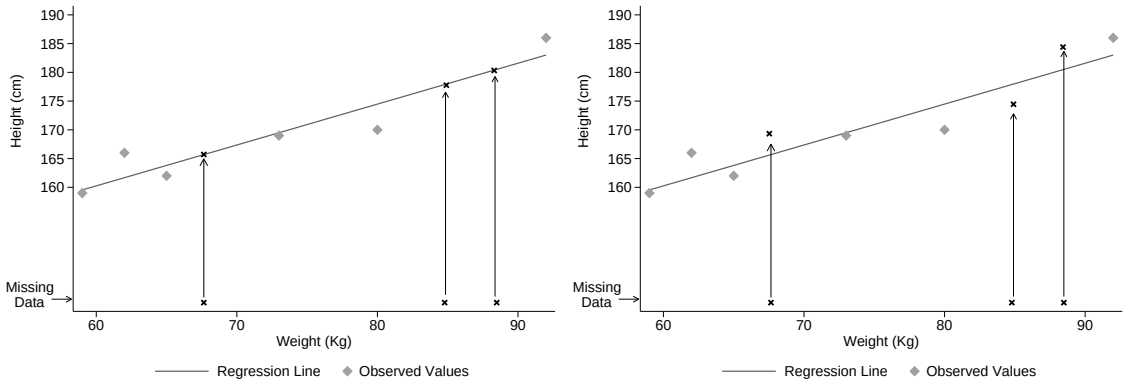


Figure 3.3: Diagrammatic representation of (a) regression imputation and (b) stochastic regression imputation: example of imputing height based on weight

The inclusion of a residual error term eliminates the bias which standard regression imputation suffers from and it is the only single imputation method which produces unbiased estimates under the MAR assumption [144]. Despite these advantages,

stochastic regression is still not recommended over multiple imputation or maximum likelihood methods discussed in the following section. This is because the imputed data are treated as the fully observed data when being analysed as with all single imputation techniques, and therefore still retains the problem of attenuated standard errors. A bootstrap sampling technique can be used to overcome this problem [167], however, maximum likelihood and multiple imputation techniques are superior techniques which negate the need to go through the effort of adding this extra step to the analysis.

3.5.2.5 Hot Deck Imputation

Hot deck imputation (also referred to as donor imputation) is a common type of imputation used in survey data in which imputed values are obtained from survey respondents with similar characteristics to those of the non-respondents. The method of defining similarity varies from simply selecting respondents of the same age with the same stage disease to complex and elaborate schemes involving a long list of variables. If the data does not contain any donor respondents, then the search is repeated using a less restrictive scheme (i.e. by trying to match fewer variables). Rao and Shao [173] stated that the main advantage of hot deck imputation is that the distribution of the data is preserved. However, it is the distribution of the available data that is maintained, and this is not necessarily the same as the distribution of the data had it been complete. Nonetheless, it is preferable to other single imputation techniques such as mean imputation in which the distribution of the data is distorted with a spike at the mean. Furthermore, results obtained from different analysis of the imputed data are comparable to each other as they would be based on the same sample, unlike in an available-case analysis as described in §3.5.1 [173]. As with all single imputation methods, the true variance is underestimated as a direct result of treating the imputed values as the truly observed values and any uncertainty introduced by the imputation is ignored.

3.5.3 Maximum Likelihood Estimation, Expectation-Maximization (EM) Algorithm

Maximum likelihood estimation (MLE) is a common mathematical technique of estimating unknown parameters given the data. Likelihood based methods can be used to handle missing data problems by considering the missing data matrix, $\mathbf{R}$, as an explicit part of the data. If the data are MAR, then the missing data can be treated like unknown random variables which are removed from the likelihood via summation or integration [154]. Dempster et al. [174] give a detailed description of how to use a technique known

as the Expectation-Maximization (EM) algorithm to find the MLE estimates when the data are incomplete.

The likelihood function, $L(\theta|x)$ is a function of an unknown parameter θ given the data x. For mathematical convenience, the log of the likelihood function, $l(\theta|x)$ is often used to find the MLE instead of the likelihood function itself. The MLE is found by maximising $l(\theta|x)$, and in the case of missing data, the EM algorithm is used to achieve this in an iterative process. The expectation step (E-step) estimates the missing values based on parameter estimates from the previous step by using the conditional expectation. The maximisation step (M-step) is subsequently used to re-estimate the parameters by maximising the log likelihood of θ given the observed data x^{obs} as well as the estimated values of x^{mis}. Starting values for θ are required for the first step of the algorithm. The likelihood is increased at each iteration of the algorithm, thus convergence is guaranteed [174]. Despite this feature, one of the main disadvantages of the EM algorithm is that convergence can be slow especially in situations where the amount of missing data is large [167]. Furthermore, the asymptotic variance-covariance matrix is not directly available from the EM algorithm for determining parameters such as standard errors. However, an extension to the EM algorithm which allows for standard errors to be calculated has been proposed by Meng and Rubin [175] as mentioned within the following section.

3.5.3.1 Applications of the EM Algorithm in the Medical Literature

Although the EM algorithm is well known within the mathematical and statistical community, it is not commonly used within the medical literature. This is likely to be due to its inaccessibility and the requirement of a good mathematical and statistical background in order to fully understand and implement the method.

A literature search of the EM algorithm combined with missing data initially produced 179 results. After removing duplicates and irrelevant articles, 98 remained. Details of the full search strategy can be found in Appendix E.

Most journal articles cite the EM algorithm back to Dempster et al. [174] in 1977, and although this paper can be credited with formalising the algorithm, the earliest known reference to the idea is that of McKendrick [176] in 1925. Further work exploring some of the theory of the algorithm was done by Orchard and Woodbury [177] and Sundberg [178] and specific examples were seen in Hartley [179] and Baum et al. [180].

The literature search returned many papers published within methodological journals such as Statistics in Medicine, Biometrics and Lifetime Data Analysis, with only a third of papers in various medical and epidemiological journals. This key point highlights the limited uptake of the algorithm by medical researchers who may not necessarily

have a good mathematical or statistical background. Many of the papers published within statistical journals propose extensions to original EM algorithm and give specific examples of how these methods can be applied to medical research. One key extension to the EM algorithm is that given by Meng and Rubin [175]. The authors propose a method named the supplemented EM algorithm (SEM) through which the asymptotic sample variance-covariance matrices for point estimates such as standard errors can be obtained. The SEM algorithm obtains the additional variability in the data caused by the fraction of missing information and adds this to complete-data variance-covariance matrix to obtain the sample variance-covariance matrix. This method builds on the simpler case for one parameter in which the rate of conversion of the EM algorithm, r, connects the sample data variance (i.e. the observed data variance), V, to the complete-data variance, V_c, by the following relationship:

$$V = \frac{V_c}{(1-r)}.$$

The same authors go on to propose another extension named the Expectation/Conditional Maximization (ECM) algorithm in which the normal M-step of the algorithm is replaced by several computationally simpler CM-steps based on the complete data conditional maximum likelihood instead of the complete data maximum likelihood [181]. Enders [182] describes an approach for calculating Cronbachs alpha with missing data (internal consistency reliability estimates for scales) using the EM algorithm, and Claeskens and Consentino [183] provide a method for determining the missing data Akaike's information criterion (AIC) for model selection after the EM algorithm.

Ibrahim [184] introduces a weighting method to handle data in which a categorical covariate is missing. The method gives weights to the complete data log likelihood within the E-step of the algorithm. The authors extend this weighting method further to parametric survival models [185] as well as Cox proportional hazards models [186]. Further work by Ibrahim et al. [187] describe a method of applying the EM algorithm to generalized linear models when the missing data are said to be non-ignorable. Non-ignorable missing data refers to the scenario when non-response is related to the values of the missing data itself [167]. The paper by Ibrahim et al. [187] in 1999 includes theoretical discussions of complex situations in which data are missing in multiple variables which can include a mixture of categorical and continuous variables. An additional paper in the same year [138] gives an applied example using quality of life data in breast cancer patients. The paper was published with the aim of aiding understanding of the more complicated material presented in Ibrahim et al. [187], and thus includes data which contained missing values in one covariate, the occurrence of which is rare in real life data. Horton and Laird [139] describe the method of weights by Ibrahim [184], including

its extensions, in detail and conclude that when the missing covariates are categorical the method is straightforward, however, becomes much more complex when the number of covariates is large.

Despite the use of specific examples within the methodological papers, the extension methods rely on a good mathematical understanding and the sheer number of extensions published highlight that the EM algorithm is not a simple approach that can be applied in the same manner to each missing data problem. This makes the method unsuitable to studies in which several different analyses of the same dataset are required, as individually tailored techniques for imputation would be required each time.

Despite the complex theoretical background, the EM algorithm method of handling missing data for a variety of data structures can be easily implemented in a range of statistical software packages, details of which are provided in Collins et al. [154]. The statistical package SPSS handles missing data problems via the EM algorithm, which is a program widely used within the medical research community due to its simple user interface. SPSS includes automatic settings such that the methods become incredibly easy to implement, however, this can mean that researchers do not give adequate thought to the assumptions required to produce unbiased results. Two papers published by Jenkinson et al. in 2006 and 2007 [188, 189] use the EM algorithm to handle missing data in questionnaires relating to Parkinsons disease and Amyotrophic Lateral Sclerosis respectively. In both articles, the authors aimed to assess whether the EM algorithm was a suitable method for handling missing data and if the results were reliable, the authors do so by deliberately removing data from a complete set of data. The authors conclude that the method produced satisfactory results in both cases, however, details of how the algorithm was performed, except for a statement that states the analysis are performed in SPSS, were not given. Nevertheless, the papers do discuss missing data patterns and question the use of the methods if the missing data were not MAR which was indicative of their understanding of the imputation process.

3.5.4 Multiple Imputation

Multiple imputation dates back to an idea developed by Rubin [190] in 1978, and is documented in more detail in Rubin [191] in 1987. The technique is an extension to stochastic regression imputation, however, instead of producing one set of imputed values, the process is repeated several times to create M sets of imputed data. This process ensures that the error variance lost with single imputation is accounted for. Once M sets of imputed data have been obtained, each from the same regression model (referred to as the imputation model), each set of data are then analysed individually according

to standard statistical techniques. The parameter estimates from each of these analyses are then combined to produce one final set of parameter estimates. The estimates are combined according to Rubins rules [191], represented by the equation below:

$$\bar{\theta}_M = \frac{1}{M}\sum_{m=1}^{M}\hat{\theta}_m,$$

where $\hat{\theta}_m$ is a scalar estimate of interest, for example a regression coefficient, obtained from imputed data set m (for $m = 1, 2, \ldots, M$).

There are two types of variance associated with this estimate; the within-imputation variance

$$\bar{W}_M = \frac{1}{M}\sum_{m=1}^{M}W_m,$$

where W_m is the variance associated with the estimate $\hat{\theta}_m$, and the between-imputation variance

$$B_M = \frac{1}{M-1}\sum_{m=1}^{M}(\hat{\theta}_m - \bar{\theta}_M)^2.$$

The total variance, T_M can be calculated as follows

$$T_M = \bar{W}_M + \frac{M+1}{M}B_M.$$

The process of multiple imputation can be represented by the following flow diagram (Figure 3.4).

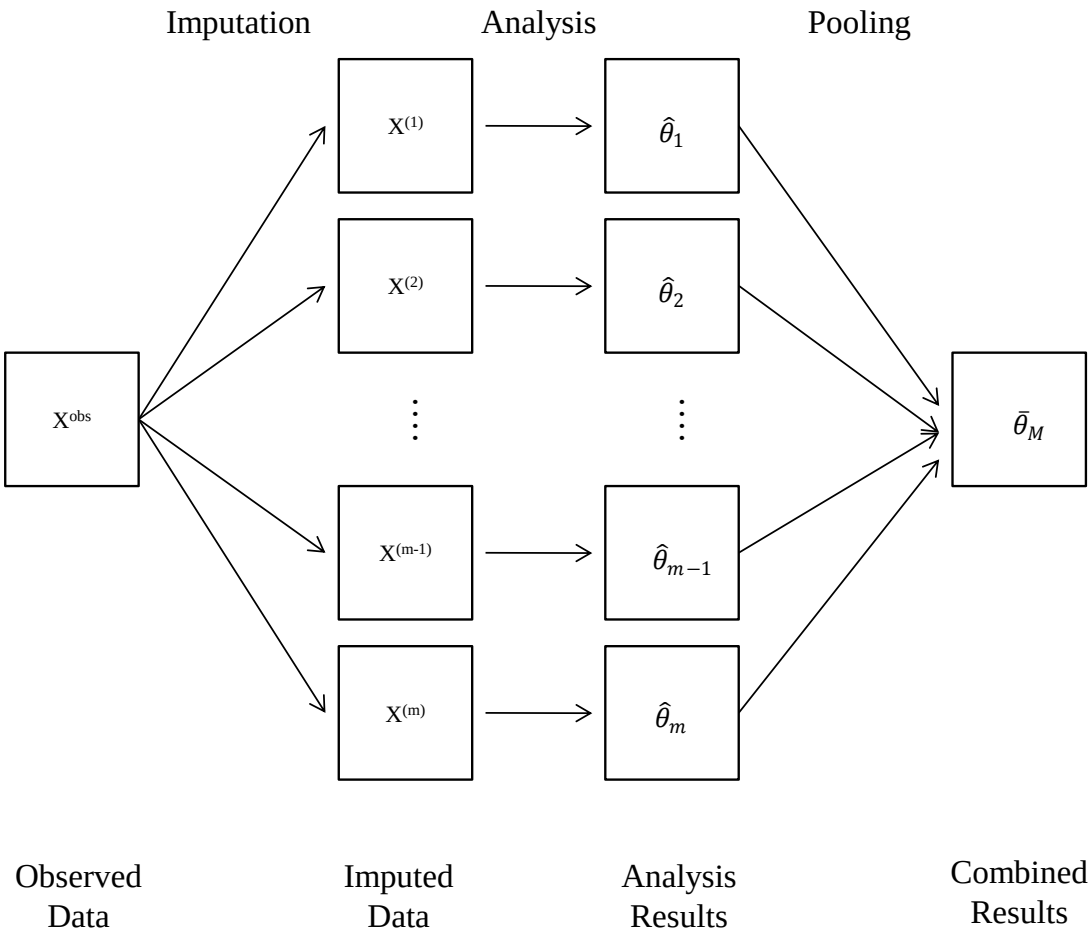


Figure 3.4: Diagrammatic Representation of the Multiple Imputation Process. Figure adapted from short course materials on multiple imputation delivered by the MRC Biostatistics Unit, Cambridge, 2011

3.5.4.1 Multiple Imputation for Multivariable Data

When the missing data occurs in more than one variable, there are a range of additional complexities which can occur [192, 193]. Imputation of multivariable data was not covered in the original literature by Rubin [191], however several methods have been suggested since then [153, 193, 194, 195]. These methods fall into two broad categories under the headings of joint modelling (JM) and fully conditional specification (FCS).

The JM approach, as proposed by Li [194], Rubin and Schafer [195] and Schafer [153] involves specifying a joint distribution $P(X, Z, R)$, where $X = (X_1, \ldots, X_p)$ is a set of p partially observed variables, $Z = (Z_1, \ldots, Z_q)$ is a set of q fully observed variables on the same subjects and R is the missing data indicator matrix. The joint model $P(X, Z, R)$ encompasses both the analysis model and the imputation model, and in reality this model is difficult to specify. Several problems are encountered during imputation by JM, which

are listed below:

- The same type of modelling process may not be appropriate for each variable which is to be imputed [193]. The variables in one data set can differ in terms of their type (i.e. continuous, binary or categorical), therefore specifying one joint model could end up creating unrealistic assumptions such as assuming normality of a binary variable [193].
- The imputation models could be complicated if the relationship between the variable which is to be imputed and the predictors is non-linear, contains interactions or is dependent on censoring.
- Imputation by JM could create non-valid combinations such as 'pregnant fathers' [143].
- The data set could contain derived variables, and the JM approach does not necessarily ensure that variables which are summed or transformed remain consistent with their original parts.

The problems of imputation by JM can be overcome by specifying individual conditional models for each variable with missing data (X_j) and treating this as a separate modelling process from the final analysis model. In this case, each conditional model can be represented by the conditional model in the form $P(X_j^{mis}|Y, Z, X_{-j}, R)$, for each X_j , $j = 1, \ldots, q$. Where X_j^{mis} is the missing part of X_j and X_{-j} refers to the set of X variables excluding X_j. This process of specifying a conditional distribution for each variable which is to be imputed is known as multiple imputation by fully conditional specification (FCS). The FCS method consists of two parts, firstly, the set of conditional models specified above, which are referred to as the imputation model, and secondly, the analysis model (also known as the substantive model). Each model can be represented as follows:

Imputation Model

$f(X_j^{mis}|Y, Z, X_{-j}, R, \omega)$, for parameter ω ($\omega \in \Omega$).

Analysis Model

$f(Y|X, Z, \theta)$, for parameter θ ($\theta \in \Theta$).

A widely used method of imputation by FCS is proposed by van Buuren et al. [160], and is more commonly known as multiple imputation by chained equations (MICE). Other terminologies used to refer to this process include variable by variable imputation,

regression switching and sequential regressions [160, 196]. The main benefit of FCS over a JM approach is that for categorical variables which contain missing data, a multinomial logistic regression model can be applied whilst a linear regression model can be used for imputing a continuous variable in the same dataset. Additional benefits include the possibility of maintaining specific features of the data such as upper and lower boundaries for specific variables, as well as the possibility of including constraints to ensure the imputation process does not impute non-valid combinations [143].

MICE can be implemented in a range of statistical software. The MICE package by van Buuren and Groothuis-Oudshoorn [192] is available in both S-Plus and R, the SPSS module MVA is available in version 17 onwards and allows the user to perform multiple imputation however it does not appear to be as flexible as the programs in S-Plus and R which are command line driven. Royston [140] introduced the 'ice' software package to Stata, implementing the same method as that in S-Plus and R. This package has since been updated with additional flexibility around categorical variables [197].

3.5.4.2 Substantive model compatible fully conditional specification (SMC-FCS)

Despite MICE being a widely applied method in a range of medical applications, the validity of the statistical properties underlying MICE are comparatively understudied [145, 198]. The main question over the validity of the MICE method is the possibility of specifying a set of imputation models in the case of multiple partially observed variables which are not mutually compatible [198]. Two models are said to be incompatible if there is no joint distribution for the corresponding set of conditional distributions. Furthermore, there is concern that the imputation model is not always compatible with the analysis model. The imputation and analysis models specified in the previous section are said to be compatible if there exists a joint model:

$$f(Y, X_j | X_{-j}, Z, \psi)$$

Recent work by Bartlett et al. [149] in 2014 indicated that incompatibility of the imputation models and analysis models is particularly a problem when the analysis model is non-linear, such as the Cox proportional hazards (PH) model. The Cox model is a non-linear model as it does not belong to the family of generalized linear models (GLMs), which are formed from a random component, a linear predictor and a link function. Furthermore, Cox PH models cannot be evaluated using standard maximum likelihood estimation as with GLMs, and instead rely on the partial likelihood. Although the log hazard function is a linear function in the covariates, Cox PH models are considered non-linear in relation to how partially observed variables are imputed using MICE. As

discussed in more detail in §3.5.4.4, the outcome variable should be included in the imputation model so that covariate-outcome associations are not diluted. For example, for a continuous partially observed variable (X), the time to event variable (T; or more precisely, the Nelson-Aalen estimate of the hazard function of T), is used to generate imputations of X using a normal linear imputation model. The imputation model therefore assumes linearity between X and the hazard function of T, which is not compatible with the analysis model in which the relationship between the log hazard function of T is linear with X [149]. The authors show that if the imputation and analysis models are incompatible, then imputations will be drawn from the incorrect distribution, and have therefore proposed an extension to the MICE method which ensures compatibility between the analysis model and the conditional imputation models. The method is known as Substantive Model Compatible Fully Conditional Specification (SMC-FCS).

The MICE algorithm works by specifying a non-informative prior distribution for the imputation model parameters ($f(\omega)$). At the t^{th} iteration, missing values for the $t + 1^{th}$ iteration are imputed using parameters drawn from the posterior distribution (the product of the prior and the likelihood of all observed data and the imputed values up to the t^{th} iteration). The SMC-FCS, instead specifies a non-informative prior distribution based on the imputation model parameters ($f(\omega)$) as well as the analysis model ($f(\theta)$). At the t^{th} iteration, parameters are first drawn from the posterior distribution based on the analysis model prior and likelihood, these parameters are then used to impute missing data by drawing parameters from the posterior of the imputation model prior and likelihood. Thus the $t + 1^{th}$ iteration is imputed from a distribution which is proportional to the product of the imputation and analysis models, which by definition is compatible with the analysis model.

3.5.4.3 Number of Imputations

Early papers describing MI suggest that very few imputations (between 5 and 10) are required to produce reliable results [160, 199]. Rubin [191] shows that the number of imputations depends on the fraction of missing information, λ, and gives the following equation for the relative efficiency of m imputations compared to 1 based on an infinite number of them as:

$$(1 + \frac{\lambda}{m})^{-1}.$$

Schafer [199] uses this equation to show that by performing 5 imputations on a dataset with 50% missing information, the standard deviation of the estimates is only

approximately 5% wider than it would be if using an infinite number of imputations, since $(1 + \frac{0.5}{5})^{-1} = 1.049$. Schafer [199] therefore concludes that there is no added practical benefit of performing more than 5 to 10 imputations.

Since these early methods have been implemented, several researchers have questioned whether such a small number of imputations are sufficient, and simulation studies have been performed to assess the optimum value for m, weighing up the reproducibility of the results and the computational time required [200, 201].

Graham et al. [201] recommends the use of m = 20,20,40,100 and > 100 for true $\lambda =$ 0.10, 0.30, 0.50, 0.70 and 0.90 respectively as the simulations have shown a considerable amount of loss of statistical power for detecting a small effect size when performing fewer imputations than this. Bodner [200] also concludes that although the previous suggestion of 5 to 10 imputations do not result in a great loss in efficiency compared to using infinite imputations, the small number of imputations repeated independently on the same dataset can result in substantial variability in important statistics such as P-values and confidence intervals (CIs). This could have the important implication that several researchers analysing the same data set could come to different conclusions. This variability in results caused by the use of a finite value for m is referred to as the imputation variance, as it is a measure of the variability obtained if the multiple imputation process was repeated several times. The Monte Carlo (MC) error can be used to measure this variability. The MC error is the standard deviation between several repetitions of the same imputation model on the same dataset, this error tends to zero as the number of imputations increases [202]. White et al. [202] also agree that reproducibility of the analysis is important, and come to the same conclusion as Bodner [200] that this will be achieved by choosing the number of imputations as approximately equal to the percentage of missing data. MC errors can be calculated for all estimates obtained from multiple imputed data, including parameter estimates, standard errors and P-values. Therefore, the reproducibility of the analysis can be assessed after the analysis by checking whether the MC error is sufficiently small so that the interpretation of the results remain the same at the extreme ends of the MC error (i.e. by adding or subtracting the value of the MC error to the parameter estimates and P-values).

3.5.4.4 Specifying the Imputation Model

Another important aspect of multiple imputation is correctly specifying the imputation model, as an incorrectly specified model can result in inaccurate imputations. Several papers describe methods for specifying the imputation model. The following general principles apply;

MAR Assumption

Including as many imputation predictors in the model as possible will make the MAR assumption more plausible [153, 154, 155, 156], in general, MI will have minimal bias if all available information is used [160]. Including all covariates from data sets is not feasible within many medical research settings, as some data sets can contain over 100 variables. This would not only be computationally extensive, but would also create problems with multicollinearity. van Buuren et al. [160] recommend that no more that 15 to 25 variables are to be included in the imputation model, as beyond this, an increase in explained variance within the regression model becomes minor.

Congeniality

All variables which are to appear in the analysis model, including the outcome variable, should also be included within the imputation model. An imputation model containing at least the same variables as the analysis model is said to be congenial. If variables which are used within the analysis are not included in the imputation model, then predictive relationships within the imputed data will be incorrectly diluted [203]. Often, the outcome variable is overlooked in the imputation process, however, the importance of its inclusion was highlighted by Sterne et al. [142], who discuss a study published in the British Medical Journal in which the authors claimed no association between cholesterol level and cardiovascular risk [204]. The lack of association was a surprising result, and the reason for this result was later found to be the exclusion of cardiovascular risk from the imputation model for cholesterol level, which led to an underestimation of the association between cholesterol and cardiovascular risk. Simulation studies have shown that in the case of survival data, the outcome variable should be included in the form of the Nelson-Aalen estimate of the hazard function in addition to including the censoring indicator [205]. Previous to this, the log of survival time was commonly used in imputation of survival data, however, White and Royston [205] and White et al. [202] show that this tends to underestimate the covariate-outcome association.

Interaction Terms and Non-Linear Transformations

For the same reasons as described above, interaction terms and non-linear transformations which may be of interest within the analysis should also be included as predictors in the imputation model so any predictive relationships within the data are not falsely diluted.

Auxiliary Variables

Any additional variables available which may help predict the missing data, but are not necessarily of interest in the main analysis should also be included in the imputation model. Such variables are referred to as auxiliary variables.

Normality Assumption and Face Validity

Multiple imputation by chained equations assumes normality of any continuous variables

to be imputed, and therefore it is important that this assumption is checked prior to the imputation process. If a non-Normal continuous variable is imputed under the assumption that the variable follows a Normal distribution, then the resulting set of imputations will be closer to a Normal distribution than the original distribution of the observed data, this is referred to as a lack of face validity [202]. For example, if the variable x followed a bimodal distribution and was imputed under a Normal assumption, then the resulting imputations are likely to contain values between the two modes of the data, thus creating a distribution which does not represent the original bimodal shape. There are several options for dealing with non-Normally distributed continuous variables, these are transformation of the variable, the use of predictive mean matching or multiple imputation by splitting. Non-Normally distributed variables can often be transformed to an approximately Normal distribution prior to the imputation process. The imputed values are then transformed back to their original scale prior to using them within the analysis model. There are two main families of transformation, include the log transformation, or more precisely, a shifted-log transformation, and a Box-Cox transformation. Transformations of these types for a non-Normal variable x are represented by the following formulae:

Shifted-Log Transformation

$$f(x) = ln(\pm x - c)$$

Where the sign of x is set to positive or negative if the distribution of x is positively or negatively skewed respectively. The constant c is chosen such that $(\pm x - c) > 0$, to ensure that the logarithm can be calculated.

Box-Cox Transformation

$$f(x) = (x^{\lambda} - 1)/\lambda$$

Where the parameter λ can be estimated using maximum likelihood.

An alternative method for dealing with non-Normally distributed variables is predictive mean matching. This method imputes values which are sampled from the observed values of the variable that is to be imputed, this results in a set of imputed values which closely follows the distribution of the observed data. However, the main downfall of this method occurs when the observed values only form a small subset of the true distribution of that variable, therefore the distribution of the imputed values would also be restricted to this subset of the distribution.

Finally, when the variable to be imputed has a bimodal distribution it could be possible to split this data into two Normally distributed variables according to some other variable. For example, if height had a bimodal distribution, and this was because all males in the dataset were taller than females, then imputation for males and females separately would

allow height to be imputed under a Normal distribution in each case. This option avoids complications of finding a transformation, and does not restrict the sample of imputed values to those in the observed data, however, it depends heavily on whether or not such an explanatory variable of the two modes is available within the dataset.

Imputation of the Outcome

Although multiple imputation is a modern method to produce unbiased and efficient results in the presence of missing data, many applied researchers worry that the method entails simply guessing missing data values. This concern is even more apparent, even among those familiar with imputation, when it comes to imputing the dependent variable of interest [206]. Due to this concern, a popular approach to handling missing outcome data is a mixture of the CCA and MICE methods, in which all cases with missing outcome data are excluded from the imputation and analysis, and then MICE is applied to all independent variables. This method has been shown to produce unbiased estimates in the situation where data are MCAR (as is the case for any CCA under MCAR) or if there are no missing data in the independent variables (in which case no imputations would be required) [207]. However, similarly to a CCA of independent variables, a CCA of the outcome variable could result in underrepresentation of certain subgroups as well as a lack of statistical power and efficiency. Moreover, there are many situations where a dataset is to be used for several different research questions and it is not uncommon in this scenario for a variable to be an independent variable for one analysis and a dependent variable for a subsequent analysis. In this scenario it is not clear whether or not the variable of interest should be imputed or not. Young and Johnson [206] argue that in fact there is no difference between missing data within the independent or dependent variable and both should be handled using multiple imputation techniques. Despite this, an alternative method for imputing the outcome variable has been proposed, the multiple imputation then deletion (MID) method. Using this terminology, the argument by Young and Johnson [206] can be thought of as the multiple imputation and retention (MIR) approach as discussed below.

MID was proposed in 2007 by von Hippel [208] and involves including the outcome variable in the imputation model, but then only retaining imputed values for the independent variables for the subsequent analyses and deleting imputed values of the outcome variable. This method ensures congeniality between the imputation and analysis model. However, it removes the imputed values for the outcome variable, which are argued to be simply adding random noise to the dataset. The MIR method, although used commonly according to Young and Johnson [206], has not been labelled as a specific ‘new method’, but instead is simply considered as an approach which treats the outcome variable as any other variable in the imputation process. Therefore, the imputed outcome values are retained within the subsequent analysis as would be the case for imputation of independent variables. This ensures congeniality of the imputation and analysis models as

well as allowing variables to be used as outcome or independent variables across differing analysis. Both Young and Johnson [206] and von Hippel [208] agree that there is little difference between the MID and MIR analysis as the number of imputations increases beyond 5. White et al. [202] in 2011 compared the MIR approach to a CCA for the outcome variable and showed that MC errors were substantially reduced using the CCA analysis. However, this was for a limited example which did not include any auxiliary variables in the imputation process. If the imputation process were to include auxiliary variables, especially those which are highly correlated with the outcome variable, then imputing the outcome would result in an analysis which contains additional information and would be preferable to an analysis which excluded this additional information [202, 206].

3.5.4.5 Applications of Multiple Imputation in the Medical Literature

There have been several reviews of the use of MI in the medical literature including work by Klebanoff and Cole [209], Sterne et al. [142] and Wood et al. [210]. These reviews were used in combination with a full literature search as they only included searches on restricted time periods and a restricted selection of medical journals. This search was subsequently refined to applications of MI used within cancer survival research. Full details of the search strategy used can be found in Appendix G.

The initial literature search returned 622 results, and there was a steady increase in the use of MI within the medical literature in the last 10 years. The original publications by Rubin in 1978 and 1987 [190, 191] resulted in fewer than 10 publications per year until around 1996 when a few key publications were made. In 1995, Greenland and Finkle [211] published a paper called ‘A Critical Look at Methods for Handling Missing Covariates in Epidemiological Regression Analyses’. They reiterate the importance of using more sophisticated methods instead of ad-hoc methods such as deletion methods and single imputation techniques. Despite their recommendations, they acknowledged that the reason these methods had not taken off within the medical literature was because of the lack of understanding of the techniques and the lack of statistical software packages to easily implement the more statistically sound methods for handling missing data. Rubin [212] attempted to address the issues people have in their understanding of the method by reviewing the important aspects of MI without going into much of the technical details described in his earlier work in 1978 and 1987. However, again, concluding remarks were that statistical software packages were urgently required and the authors concluded that once these become available, MI would likely become the standard method of handling missing data problems [212].

Schafer [153] is often credited with popularising the method of MI because his book published in 1997 describes the methods and their implementation in a much more accessible manner compared to earlier work [212]. In combination with the publication of this book, the availability of statistical software packages for the implementation of MI, namely van Buuren's MICE software in R and S-Plus first introduced in 1999 [192] and the ice package for Stata by Royston [140] amongst others, are likely to have led to the increase in the use of MI in recent years.

Papers by Klebanoff and Cole [209] and Sterne et al. [142] reviewed the use of MI within the medical literature; the first focusing specifically on epidemiological journals and the latter on major general medical journals. Both papers conclude that the use of MI is sparse. More importantly, both papers highlight the variation in the reporting of MI and attempt to provide guidelines for future use. These include;

- Reporting the number of missing values for each variable alongside possible reasons for their missingness
- Providing details regarding potential differences between cases with complete and incomplete data
- Describing the precise methods used for imputation (including software used, variables included in the imputation model, the number of imputations, the use of transformations of non-normal variables, any interactions used)
- Discussion of assumptions that are required for inferences to be valid
- Comparison of complete case analysis with MI analysis

Despite many methodological papers and several recent review papers, MI methods are still uncommon within medical research and more specifically cancer research. A literature search focusing specifically on MI within cancer research resulted in 60 publications between 1997 and 2014 (Appendix G). Although the number of publications per year increased around 2008 (Figure 3.5), the use of multiple imputation remains relatively uncommon. Moreover, those papers which did use imputation, the methods were often not adequately described and implemented, with the most common problems occurring around not accurately specifying the imputation model by excluding the outcome variable and interaction terms as well as not performing an adequate number of imputations for the amount of missing data.

Clark et al. [213] uses MI to impute missing values of prognostic factors of ovarian cancer (including stage of cancer, grade of cancer, performance status, ascites and debulking) in a data set of 1189 cases of ovarian cancer. The authors described in detail the pattern of

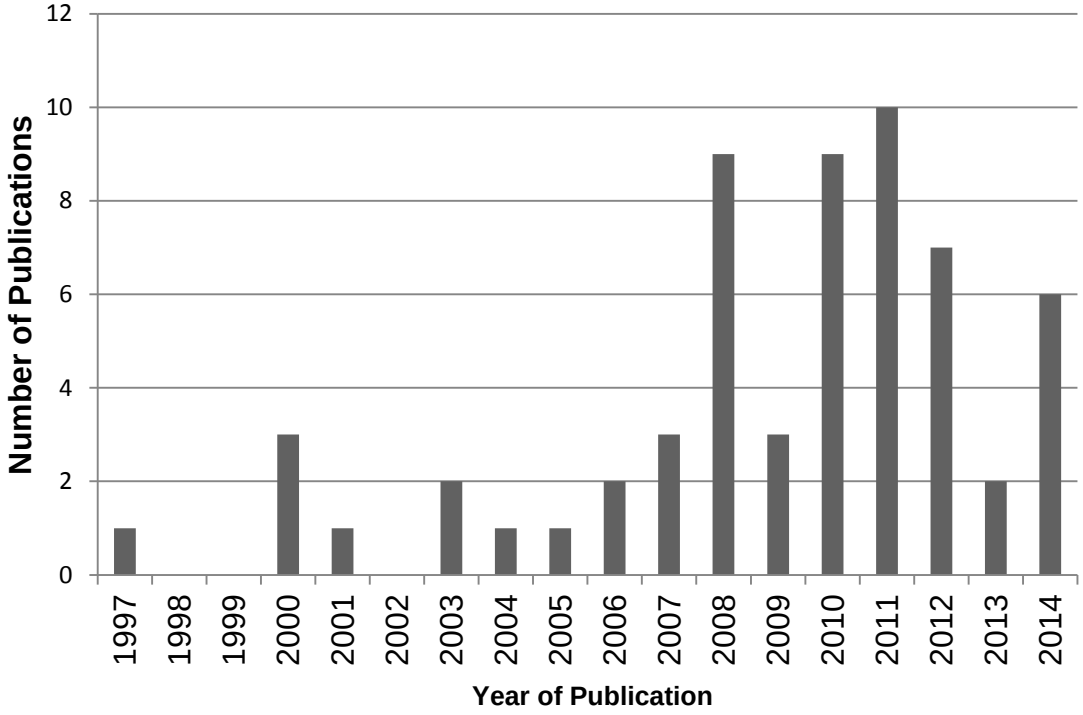


Figure 3.5: Number of publications of multiple imputation per year within cancer survival research

missing data and the level of missing data (17.2% missing in 69% of cases) and discussed the plausibility of the MAR assumption. The authors used 10 imputations, but provide no further discussion as to how they have arrived at this number. In light of recent papers discussing the number of imputations, the recommended number of imputations for this analysis would have been 20; approximately equal to the amount of missing data. Their results show non-significant effects of interactions between several of their covariates (stage and the tumour marker CA125, stage and grade and stage and histology), however, it was not clear from the paper whether or not these interactions were included within the imputation model. If they were excluded, the association between these variables may have been falsely weakened as discussed in the previous section. The same authors go on to publish a methodological paper using the same data as their case study [214]. This paper included step by step details on how the imputation was performed pointing out important theoretical ideas along the way, in an attempt to demystify the imputation process for other medical researchers. However, a few key points were not covered by the authors. Namely, the authors referenced van Buuren et al. [160] and quoted: "all variables that appear in a model constructed using complete cases should be included in the imputation models." Yet the authors appeared to fail to include interaction terms within their imputation model, despite testing for interactions within their final analysis model. The authors used 10 imputations in their analysis, again referencing van Buuren

et al. [160]. As discussed in the previous section, the arbitrary figure of between 5 and 20 imputations is outdated, and papers by Graham et al. [201] and Bodner [200] indicate that the number of imputations should be considerably higher than this, especially in the case of large amounts of missing information. It is therefore important for the continued publications of papers such as Clark and Altman [214] so that medical researchers who rely on them to understand the MI methodologies are receiving the best possible and most up to data advice as to their implementation.

More recent applications of MI within cancer survival research include Ferguson et al. [215], Nur et al. [216] and Mandleblatt et al. [217]. Ferguson et al. [215] and Nur et al. [216] both described the imputation models in detail and discussed the appropriateness of the MAR assumption; Mandleblatt et al. [217] did not mention either of these points. A small number of imputations was used in each of these papers; Ferguson et al. [215] used 5 imputations on data which was missing in 32% in one variable and 17% in another, Nur et al. [216] cite the use of 10 imputations quoting approximately 10% missing data and Mandleblatt et al. [217] used 10 imputations quoting 5% of missing data in most variables and two other variables with 17% missing data. The number of imputations were too few for studies by Ferguson et al. [215] and Mandleblatt et al. [217] based on the amount of missing information. The number of imputations used within Nur et al. [216] were appropriate in relation to the amount of missing data, however, it was difficult to determine whether or not this was a conscious decision or just by chance.

Although the use of MI within cancer survival research is not common practice, several researchers have begun using the methods within their work over recent years. Despite many methodological papers and several recent review papers the methods are not adequately described and the most common pitfall is failing to specify the imputation model adequately by including outcome variables and interactions in addition to the choice of the number of imputations.

3.5.5 Inverse Probability Weighting

IPW is another technique for handling missing data which works very differently from MI. It is similar to a CCA in that the analysis model is fitted only to the complete cases. However, unlike in a CCA, the complete cases are weighted by the inverse of their probability of being a complete case. IPW can remove the bias that results from a CCA when the data are not MCAR [218]. The IPW analysis model estimates $\hat{\theta}$ as the solution to the following score equation:

$$\sum_{i=1}^{n} R_i w_i \mathbf{U_i}(\theta) = 0$$

Where $\mathbf{U_i}(\theta)$ is the first derivative with respect to θ of the log likelihood function and w_i is the weight given to individual i. This equation reduces down to a CCA in the case when $w_i = 1$ for all i.

As w_i is unknown, it needs to be estimated, this is achieved by fitting a missingness model. The missingness model is a logistic regression model using the missing data indicator $\mathbf{R}$ as the outcome and variables from the set $\{X, Y, Z\}$ as predictors (where X and Y are the covariates and dependent variable from the analysis model respectively and Z represents additional variables which are observed but not used in the analysis model) [218, 219, 220]. The missingness model can be fitted provided the predictors in the model are all fully observed, or if the missing data has a monotone pattern [218]. A monotone missing pattern occurs when, for a set of variables $X = (X_1, \ldots, X_k)$, X_k is only observed for an individual if X_{k-1} is observed for each $k = 2, \ldots, K$. The Markov randomised monotone missingness model can be used in the case of non-fully observed variables [221], however, this model is more complicated to fit and is not described further here as this would invalidate the main advantage IPW has over MI, which is its simplicity as discussed in §3.6.

3.5.5.1 Applications of Inverse Probability Weighting in the Medical Literature

A literature search of IPW returned a total of 76 papers published between 2000 and 2014. However, the majority of these papers used IPW for the design and analysis of surveys, without its application to dealing with missing data. In a refined literature search combining IPW with missing data terminology, only 8 papers remained. Details of the search strategy can be found in Appendix F. Applications of IPW included analyses of missing data in surveys, clinical trials and longitudinal cohort studies. Reasons for the lack of uptake of this method for handling missing data could be because IPW estimates are known to be inefficient compared to likelihood based methods [222] as well as the estimates being sensitive to the form of the probability response model [167]. The lack of use of IPW as a method for handling missing data within the medical literature does not seem to be an issue related to inaccessibility of the method to medical researchers. The advantages and disadvantages of IPW as a missing data method compared to likelihood based approaches and MI are discussed in §3.6.

3.6 Comparison of Likelihood Based Approaches, Multiple Imputation and Inverse Probability Weighting

If the predictors in the missingness model are all fully observed, or if their missingness pattern is monotone, IPW has the advantage of being easier to implement as there is less technical sophistication compared to likelihood based approaches and MI. This makes the method more accessible to researchers who are not necessarily experts in statistical techniques. The increased availability of MI software has also made the use of MI more accessible to researchers, however, methods are still largely inadequately handled as shown earlier in §3.5.4.5. Many of the common pitfalls documented by Sterne et al. [142] in 2009, including failure to specify the correct imputation model or using a sufficient number of imputations, are still apparent in current applications of MI. Likelihood based methods for handling missing data continue to be unpopular within applied medical research as discussed in §3.5.3.1. This is largely due to the high level of mathematical knowledge required to gain a detailed understanding of the method, as well as its many extensions.

The desired imputation method depends not only on the ability of the researcher, but also on the design of the study [161, 218, 223]. IPW is commonly used to handle survey non-response or longitudinal studies, as missing data in these scenarios can often result in large blocks of data being missing for each individual. The use of MI or likelihood based methods to impute missing values for entire questionnaires, or for data collected at a hospital visit which was not attended, are not recommended [218, 220]. However, for registry based or routine population based data, MI and likelihood methods are preferred as they are able to incorporate all of the available information within the datasets to minimise bias and improve efficiency.

Although likelihood based methods, MI and IPW can all reduce bias compared to ad-hoc methods and CCA, MI and likelihood based methods have a clear advantage over IPW. Namely their efficiency, as likelihood based methods and MI do not solely rely on information from complete cases, but additionally include partially observed cases thereby allowing more information to be used within the final analysis [224, 225]. In addition, imputation by both the EM algorithm and MI allow the researcher to account for a degree of uncertainty within the final analysis [226], which ensures appropriate estimates of standard errors and P-values are obtained [227].

MI has the additional advantage over IPW of allowing many statistical techniques to be used and applied after the imputation process, provided the imputation process is

completed in a thorough manner to ensure congeniality between the imputation process and subsequent analysis. With the exception of methods for handling hierarchical data structures after MICE; methodologies for which are still open for debate and in development [223]. The data used within this study did not warrant the use of multilevel modelling as, although patients were nested within hospitals, the data were not robust enough in terms of their recorded principle treatment centres and therefore a clear hierarchical structure could not be identified; methods used for imputation of multilevel data were therefore not further explored in this thesis. As discussed in §3.5.3.1, there are many extensions to the EM algorithm which are required for a range of different missing data problems. The implication of this is that the method is not efficient for scenarios in which multiple analyses are required from the same dataset, unlike with MI. A further disadvantage of the EM algorithm is that with a large number of partially observed variables and large amounts of missing data, convergence of the algorithm can be slow and difficult to achieve [153].

Despite the clear advantages of MI over IPW and likelihood based approaches, MICE (the most popular implementation of MI), is predominantly based on observational and experimental evidence and lacks a strong theoretical basis [202, 228]. Another concern raised over MICE is the possibility of incompatibility between the imputation and analysis model which could result in imputations which are derived from the incorrect distribution, in particular when the analysis model is non-linear [149]. van Buuren et al. [229] and White et al. [202] agree there is little evidence to suggest that incompatibility is a concern as it does not tend to have a real effect upon the analysis. Despite this, an extension to the MICE algorithm, SMC-FCS, has recently been proposed by Bartlett et al. [149] in 2014 which recommends its use for non-linear models such as Cox PH modelling.

The analysis within this thesis is based on population based cancer registry data with a moderate to high level of missing data (see Chapter 5 for a detailed description of missing data specific to this thesis). The cancer registry dataset is used for many different analyses, in particular, within this thesis, two analyses were completed based on multiply imputed data. Firstly, a survival analysis based on imputed data of stage and ethnicity (Chapter 6) and secondly, an analysis determining predictors of late stage diagnoses based on imputed data of stage (as the outcome variable) and ethnicity (Chapter 7). Due to the nature of the study design as well as the requirement of multiple analyses based on imputed data, MI was the preferred imputation method within this thesis. MICE was used for both analyses, however, results were also compared to SMC-FCS for the survival analysis which was based on Cox PH modelling.

3.7 Conclusion

The key implications of missing data are well understood in the statistical community. Despite this, many applied health researchers fail to realise that the lack of an appropriate statistical method to handle missing data does not only reduce the statistical power and thereby the ability to identify potentially important factors, it can also result in invalid and imprecise estimates. This can have serious consequences if the results are subsequently used to determine clinical practice.

Despite the clear disadvantages of ad-hoc missing data methodologies, the uptake of appropriate techniques within the medical literature remain limited. In particular, for research focusing on survival of cancer including studies conducted worldwide covering all age groups and cancer types, there were just 60 publications between 1997 and 2014 which implemented MICE. Moreover, the majority of the identified papers failed to satisfy the full requirements for sound implementation of MICE through misspecification of the imputation models and using an insufficient number of imputations to ensure validity of the results.

As discussed in Chapter 2, missing data of stage and ethnicity is a common problem within cancer registry data. In addition to the paucity of research on childhood and TYA cancer survival which is adequately adjusted for stage of disease (Chapter 2), cancer survival studies have not, in general, implemented adequate methods for handling missing data. In particular, previous studies of survival of childhood leukaemia in the UK tended to use either a CCA [52, 57] or the indicator method [55, 56] for missing data which produces incorrect and biased estimates.

MICE and SMC-FCS were implemented in this thesis to produce efficient and reliable results of CYA cancer survival following a detailed evaluation of the available methods and discussion of the advantages and disadvantages of IPW, likelihood based approached and MI (§3.6). Details of the specific methodology used throughout this thesis, including a description of the imputation model specifications, are provided in Chapter 4, followed by a description of the missing data patterns for the Yorkshire register in Chapter 5.

Chapter 4

Methods

4.1 Introduction

This chapter includes a description of the datasets and methods used throughout the thesis. The chapter begins with specific details of the Yorkshire Specialist Register of Cancer in Children and Young People (Yorkshire register from here on in), which was the main data source for this thesis. Additionally, inpatient hospital episodes statistics (HES) data was used for analysing hospital activity amongst long term survivors of cancer as well as allocating ethnicity codes. Several different types of HES data exist, which are described within this chapter, before providing reasons for focusing specifically on inpatient HES data. Both data sources contained detailed and identifiable information on a person-by-person basis and these were therefore of a highly sensitive nature. The ethical implications and required ethical approvals are discussed in §4.3. Following an overview of the data sources, §4.4 focuses on the statistical methods used throughout the thesis by initially describing the methods used for descriptive data analysis, followed by the multiple imputation methods used within the thesis. Subsequently, methods for the following three main analyses were described in turn;

i) Variation in Cancer Survival (results presented in Chapter 6)

ii) Inequalities in Disease Severity at Presentation (results presented in Chapter 7)

iii) Long Term Effects amongst Survivors of Cancer (results presented in Chapter 8)

4.2 Data Sources

The importance of high quality cancer registration data in the use of large scale epidemiological studies, as well as the structure of cancer registration in England has been described in §2.2. This section focuses on the specific datasets used throughout the thesis, namely the Yorkshire register and HES data. The Yorkshire register is appropriate for research into both short and long term outcomes of CYA cancer due to the historical nature of the database as well as its ongoing follow up of cases every two years beyond their date of diagnosis. The dataset was used alongside HES data, which enabled research into outcomes other than survival, such as late effects resulting from cancer treatment. The Yorkshire register was linked to inpatient and outpatient HES data so that individual cases could be identified in both datasets. Details of the data linkage process are given in §4.2.4, following a general description of the Yorkshire register and HES datasets. The percentage of linked cases as well as differences between linked and non-linked cases are explored in Chapter 5.

4.2.1 Cancer Registry Data

The Yorkshire register is a regional population based dataset containing detailed demographic and clinical information on CYAs aged 0-29 years inclusive at diagnosis. Since 1974, cases of cancer were registered for children diagnosed under the age of 15 within the former Yorkshire regional health authority. Two expansions of the register have occurred since then, the first was an extension of the age limit to include young adults up to the age of 30 for diagnoses from 1990 onwards. This data was initially collected retrospectively from 1999 and prospectively from that point forward. The second expansion occurred in 2006 to align with the new definition of the Yorkshire Strategic Health Authority, which consists of south Yorkshire in addition to the former Yorkshire region. Data collection was extended to include cases diagnosed in south Yorkshire from 1998 onwards.

Notifications of CYA cancer cases in the region were received from the Northern and Yorkshire Knowledge and Intelligence Team (NYKIT) and East Midlands Knowledge and Intelligence Team (EMKIT) (the regional cancer registry teams part of the National Cancer Registration Service in the UK covering the Yorkshire Strategic Health Authority). These notifications include data on the date of diagnosis as well as the morphology and site of the tumour. Detailed data, including treatment information for each of these cases is then obtained by a data collection officer via the medical records at relevant hospitals in the area, and annual follow up of all cases takes place to ascertain data on any relapses

or deaths.

Within this thesis, cases diagnosed under the age of 30 in the former Yorkshire regional health authority (which excludes south Yorkshire) between 1990 and 2009 were included. This time frame was chosen so that there was a consistent age range across the study period. Although data collection is ongoing, the process of manually collecting data for each case in the area is time consuming and thus at the time of analysis, complete data was only available up to the end of 2009.

South Yorkshire data was not included within this analysis to avoid causing bias in any results by a change in the inclusion criteria during the study period of interest. Although survival analysis can account for differences in the entry point to the study (the date of diagnosis in this case), a change in the underlying cohort over time could introduce unknown biases. For example, there could be differences in the types of cancers which are diagnosed in south Yorkshire compared to the former Yorkshire region due to differences in the expertise or specialist services offered by these two areas. If survival was generally higher or lower in south Yorkshire compared to the former Yorkshire region, then including this cohort part way through the study period (from 1998 onwards) would mean that survival rates would appear to be improving or worsening over time due to the structure of the data. Without detailed research into any potential differences between these two areas in terms of the population demographics, the clinical practices and clinical coding procedures it was not possible to determine whether such a bias would exist, what size this may be and in which direction it would occur.

Additionally, the south Yorkshire region is covered by EMKIT, whereas the former Yorkshire region is covered by NYKIT. Differences in the procedures for notifying the Yorkshire register as well as the process of data collection could lead to differences in case ascertainment within these areas, and therefore has the potential to introduce information or ascertainment bias. Ascertainment bias can lead to biased estimation of survival rates, the bias will be towards improved survival rates should the cases which were missed in data collection have poorer survival than those that have been collected and towards poorer survival rates if those that were missed had better survival. There will be no bias if the missing cases are MCAR as the collected cases will be a random sample of the total cases, however, without in depth analysis and examination of the potential differences between the two regions it is not possible to determine the potential size of the bias introduced.

One possibility to avoid the types of biases described above would be to only include data from 1998 onwards covering both the south Yorkshire region and the former Yorkshire region. Any potential differences between these two areas could then be adjusted for within the analysis by including a variable to indicate which region the

data is from. However, restricting the analysis to diagnosis from 1998 onwards has several disadvantages. Firstly, the ability to analyse longer term outcomes would be lost as there would be a reduction in the maximum follow up time of cases by 8 years. Secondly, although the size of the study region would be increased, south Yorkshire is geographically much smaller than the former Yorkshire region, thus excluding 8 years of data from this larger region in favour of including data from an additional small region would inevitably lead to an overall reduction in the number of cases within the study and therefore result in a loss of statistical power.

4.2.2 Hospital Episode Statistics Data

HES is the national statistical data warehouse for England containing data on all national health service (NHS) care in England and is managed by the Health and Social Care Information Centre (HSCIC). Data on private patients treated within NHS hospitals are also included (approximately 11% of the UK population have private health insurance). There are six datasets within the HES data warehouse, including inpatient, outpatient, accident and emergency (A&E), patient reported outcome measures (PROMs), adult critical care and mortality data.

4.2.2.1 Inpatient Data

Of the six datasets that comprise the full HES data warehouse, the most established is HES inpatient data (also referred to as admitted patient data). Inpatient data dates back to the 1st of April 1989, however, the level of completeness and quality of data was not considered adequate for large scale statistical analysis until 1st April 1996, and the Pseudo HES ID field used to link records from individual cases across the HES data warehouse was not introduced until after this point [230]. Inpatient HES data contains details of all admissions to NHS hospitals in England; additionally it includes private patients treated in NHS hospitals. Within inpatient HES, data are recorded by individual episodes, referred to as a finished consultant episode (FCE). This term often gets confused with a hospital admission, however, one hospital admission can comprise one or more FCEs if a patient is cared for by more than one consultant during their admission. An admission can therefore be defined as a continuous inpatient spell (CIPS) which can be constructed for each patient. A detailed algorithm for the construction of these CIPS is given in Lakhani et al. [231], the implementation of which is described in Chapter 5. As well as having a data structure in which multiple FCEs make up a single CIPS, each FCE can also contain up to 20 diagnoses and up to 24 operation codes which are classified according to the International Classification of Diseases (ICD-10) and the Office of Population Censuses

and Surveys Classification of Interventions and Procedures (OPCS-4.5) [30, 232]. A summary of the inpatient HES data used within this thesis is provided in Chapter 5.

4.2.2.2 Outpatient Data

Outpatient data contains individual level data on all outpatient hospital appointments in England since the 1st of April 2003. This dataset is less well established and in its data dictionary, the introduction describes this data as 'experimental due to known problems'. The main problem with outpatient HES data was the incompleteness of clinical codes, since collection of this data was not mandatory for outpatient appointments, unlike for inpatient admissions. However, the specialty of the consultant under which the patient was seen was recorded consistently within outpatient HES data. A summary of an outpatient data extract is given in Chapter 5.

4.2.2.3 Accident and Emergency Data

In the 2007/08 financial year, HES expanded its database to include an A&E dataset containing diagnosis, investigation and treatment codes on all A&E hospital visits in England. As with outpatient data, the A&E dataset is listed as an experimental one, with many data items not being mandatory. Although the inpatient and outpatient datasets contain admission method codes so that patients who came into hospital via A&E could be identified, any activity which occurred prior to referral to a specific department was recorded within the A&E dataset, and was not available from inpatient data.

4.2.2.4 Mortality Data

Although inpatient HES data contains mortality data for those who died during an inpatient admission, this does not cover those who died outside of hospital or additional data such as the underlying cause of death. The mortality dataset was constructed using the Office for National Statistics (ONS) mortality data linked to any patient within the HES data warehouse. This linkage provides data on those who died in and out of hospital as well as providing their cause of death. However, mortality data was only available for patients who have had at least one hospital record in one of the three main datasets mentioned above.

4.2.2.5 Patient Reported Outcomes Measures (PROMs) Data

The PROMs data covers patient reported outcomes for four elective surgical procedures, including groin hernia operations, hip replacements, knee replacements and varicose vein operations for the whole of England since 2009.

4.2.2.6 Adult Critical Care Data

The critical care dataset comprises all records for adults who have had critical care hospital stays. Individuals may be associated with more than one critical care stay which could either be for the same or a different condition, and may be during the same or a different time period.

4.2.3 HES data and Cancer Epidemiology

Each of the HES datasets as described above have complicated structures, their own limitations and cover different time periods. Therefore careful consideration into which of the datasets was most useful in supplementing the Yorkshire register data was required. Inpatient HES data could not be used as a stand alone dataset for the purposes of cancer epidemiology. Although cancer diagnoses codes were recorded within inpatient data, the data was not sufficient to accurately determine the number of cases of cancer in England, as only those in hospital care at the time of their diagnosis would be included. Therefore, a cohort of cancer cases obtained solely from HES data could be biased towards a more 'unhealthy' cohort. Furthermore, some non-specific cancer diagnoses codes could be recorded within an episode before a definitive diagnosis had been made. There was also the possibility of duplication of diagnostic codes in which a cancer diagnosis was recorded each time a person with cancer had a hospital admission regardless of it being the admission in which the diagnosis was made. However, inpatient data linked to cancer registry data can provide a powerful tool for cancer epidemiology analysis, with examples including Pollock and Vickers [233], Morris et al. [234] and Maddams et al. [235]. Detailed information on the definitive diagnosis of cancer and related specialist oncology data such as treatment and staging can be supplemented with detailed hospital admissions and all possible clinical diagnoses both before and after the diagnosis of cancer.

Within this thesis, linked cancer registry and inpatient HES data was used to obtain additional information about the patients' demographics, namely, their ethnicity, as well as to describe and quantify late effects of cancer treatment such as cardiovascular disease within this cohort. As described in the literature review in Chapter 2, there was a paucity

of objective data of this kind to accurately quantify the incidence of cardiovascular late effects within long term survivors of CYA cancer. The details of this analysis can be found in Chapter 8.

Inpatient data, however, only portrays a subset of a patients hospital activity and a more complete picture would be seen when also considering outpatient appointments. In the example of cardiovascular late effects, the incidence of such late effects could be underestimated by simply assessing inpatient appointments as a proportion of heart problems could be dealt with entirely within outpatient appointments. As mentioned within the description of outpatient data, the dataset was not sufficient for this type of analysis due to the lack of clinical diagnostic coding within the data. Although specialty codes were included within the outpatient data, this information would not be sufficient to determine the frequency of cardiovascular late effects within outpatient appointments. For example, you cannot assume that someone who was seen under the 'Cardiology' specialty had a heart problem, as they may have been referred there by an oncologist as a precaution.

The other HES datasets mentioned will not be used within this thesis due to their irrelevance to CYA cancer (Adult Critical Care and PROMs data) and their limited time frame (A&E data). Mortality data for the cohort was available via the Yorkshire register, obtained through the follow up of patients and notifications received via NYKIT. Both the mortality dataset held by the HES data warehouse and the data held by NYKIT were obtained from death certificates. There was therefore no need to obtain duplicate mortality data via HES.

4.2.4 Data Linkage Methods

Inpatient HES data between 1996 and 2011 was linked to specialist cancer registry data including diagnosis from 1974 to present. Only those cases who were alive in 1996 were eligible for linkage, as those who died before this time would not be present within the inpatient dataset. NHS number, date of birth, sex and postcode at diagnosis for all eligible cases were extracted from the Yorkshire register and sent to the HSCIC for linkage, alongside a detailed application form containing all required HES data fields. The linkage process was performed by the HSCIC trusted data linkage service, who performed deterministic matching (also known as exact matching) in combination with fuzzy matching using the above identifiers. Deterministic matching requires all identifiers to match exactly. In fuzzy matching, exact matching is performed on part of the identifier (for example, matching exactly on month and year of birth, instead of full date of birth). The matching process was completed in several stages (referred to as passes) as outlined

below.

1st Pass Matching on NHS number, sex, date of birth and postcode

2nd Pass Matching on NHS number, sex and date of birth

3rd Pass Matching on NHS number, sex, postcode and partial date of birth

4th Pass Matching on NHS number, sex, and partial date of birth

5th Pass Matching on NHS number, postcode

6th Pass Matching on sex, date of birth and postcode, where NHS number does not contradict the match, date of birth is not the 1st January and postcode is not on the ignore list (a list of postcodes that have a dense population such as Army Bases)

7th Pass Matching on sex, date of birth and postcode, where NHS number does not contradict the match and date of birth is not the 1st January

Details of the number of cases which successfully linked to the register, as well as differences between linked and non-linked cases are explored in §5.2.1.

4.3 Ethical Approval and Data Security

The University of Leeds requires ethical review of any research which involves human participants - this includes the use of their data or their tissue. For example, the collection and use of data obtained via interviews or questionnaires, as well as via observations or testing. Any research involving human participants from the NHS (including their personal data or tissue) requires ethical review by the National Research Ethics Service. A research project which has obtained NHS ethical approval does not require University of Leeds approval in addition. Throughout this thesis, personal data on NHS patients as collected by the Yorkshire register as well as HES data was used. Ethical approval for the collection and use of this data for research purposes has been granted by the Northern and Yorkshire Research Ethics Committee, reference: MREC/00/3/001. This ethical approval covers the work set out within the study protocol for the Yorkshire register, of which this thesis is one element. Original approval was obtained in May 2000, with the most recent update being approved in April 2014.

Most research involving individuals requires informed consent from each participant involved in the study according to the Data Protection Act 1998, the Human Rights Act 1998 and the common law. However, any member of the UKACR, which included the

Yorkshire register, was exempt from having to obtain informed patient consent for the holding of personal data. This exemption was granted under Section 251 of the NHS Act 2006 to all members of UKACR, and was conditional upon annual reports to the National Information Governance Board (NIGB) outlining the fulfillment of the conditions on the retention and disposal of personal information. As of 2009, the NIGB became responsible for issuing Section 251 approval and following national cancer registration structural change, the Yorkshire register was required to obtain individual Section 251 approval through the Confidentiality Advisory Group (CAG). The application was approved in August 2014, reference: CAG 1-07(b)/2014.

The data held by the Yorkshire register is of a highly sensitive nature as it identifies all individuals in Yorkshire diagnosed with a malignancy under the age of 30. The data held are fully identifiable, including names, dates of birth, addresses and unique NHS number alongside detailed clinical data. This data is only accessed by those covered under the ethical approval, and data security is therefore very important. As stated within the registers protocol, all data is held in a secure room on a University of Leeds secure area network, which can only be accessed by authorised members of staff via password authentication. The following additional conditions of the data are also in place;

- No information is ever published in which individuals can be identified.
- No individuals on the Yorkshire register are ever approached directly.
- Data are only released according to the requirements of the security policy which specifies the circumstances for data release.

All work within this thesis will abide by the ethical and data security conditions as outlined here. In particular, the publication of any data from the Yorkshire register will not contain identifiable patient data (such as NHS number), or potentially identifiable data (such as the publication of fewer than five individuals in any one group).

4.4 Statistical Analysis

The main analysis within this thesis was performed in Stata, version 13 [236]. R was used to perform the sensitivity analysis for multiple imputation [237] (§6.6.1).

4.4.1 Descriptive Data Analysis

A detailed description of the cohort is given in Chapter 5, including the number of cases by age group and gender for each diagnostic group, exploration of cancer registry and HES data linkage rates and an overview of missing data patterns.

To determine diagnostic groups, morphology and site codes for each diagnosis were initially allocated to ICCC classification as well as Birch TYA classification schemes (see §2.2.2), and results compared. The cohort covered two age groups for which separate classification schemes were devised, however, using two classification schemes in the same analysis would make comparisons by age group difficult. The classification scheme which minimised the number of cases allocated to the 'other' categories was used throughout the analysis (see Chapter 5). For ICCC, groups XI and XII contain 'other malignant epithelial neoplasms' and 'other and unspecified neoplasms' respectively, whereas for Birch TYA, groups 9 and 10 contain 'miscellaneous neoplasms not elsewhere classified' and 'unspecified malignant neoplasms not elsewhere classified' respectively. Each scheme also resulted in a number of cases which could not be classified, such cases were described and labelled as 'unknown.' Unknown classifications could have resulted from poor data quality, for example, not all site codes are valid for all morphology codes.

Further exploratory analysis included a detailed description of the number of cancer registry cases which successfully linked to at least one HES record, and differences between linked and non-linked cases were explored using the appropriate statistical test for demographic and clinical variables (including Pearson χ^2 test, McNemar's test, t-test or Mann-Whitney U-test).

Following detailed descriptions of the cohort, the level of missing data was explored per variable in which missingness occurred (stage/disease severity and ethnicity). This was completed on a diagnostic group basis, which was of particular importance for the stage or disease severity measure due to the fact that each diagnostic group has their own measure of stage or disease severity. Once the missingness for each variable was explored in detail, missing data patterns by diagnostic group were assessed.

The diagnostic groups taken forward for imputation and detailed analysis were determined based on the overall amount of missing data discovered in the descriptive data analysis. There are no standard rules or cut off points for how much missing data is too much missing data, with very little research addressing this point specifically. Statements on the topic of the amount of missing data in research papers are vague and include for example "if the extent of missing data is not too great" [238], "the potential for bias usually increases with the extent of missing data" [239], and "70% missing [data] may require more imputations" [240]. However, a paper by Barzi and Woodward [241] looking

at different imputation methods at different levels of missing data stated that multiple imputation was acceptable for studies with up to 60% missing data, after which none of the tested imputation methods (including multiple imputation and the EM algorithm) produced satisfactory results. Although indicative of a level at which missing data may become problematic, the research by Barzi and Woodward [241] was a single paper comparing a specific set of datasets in which cholesterol level was imputed. The results found by Barzi and Woodward [241] could be specific to the exact missing data mechanism of cholesterol level and was not necessarily generalisable to other studies. The general consensus between researchers appeared to be that with large amounts of missing data (60/70/80% or higher), any errors made within the imputation process would lead to greater inaccuracies in the results compared to situations in which the level of missing data was small. Extra thought, therefore, needs to be put into the exact methods of imputation, the number of imputations and correctly specifying the imputation model(s) in situations where the amount of missing data is large.

Based on the available information, only diagnostic groups for which the overall level of missing data was below 60% were considered for imputation and subsequent survival analysis. These were central nervous system (CNS) tumours, leukaemia and germ cell tumours (GCT). WHO Grade (four level ordered categorical variable), white blood cell (WBC) count (continuous measure) and stage (four level ordered categorical variable) were the disease severity measures to be imputed for each diagnostic group respectively and ethnicity (three level categorical variable) was imputed for all three diagnostic groups (see Chapter 5 for a detailed description of the missing data for all diagnostic groups).

4.4.2 Multiple Imputation

Missing data for disease severity and ethnic group was imputed to allow for the inclusion of these variables in the 'Variation in Cancer Survival' analysis (Chapter 6) and 'Inequalities in Disease Severity at Diagnosis' analysis (Chapter 7). This section contains specific methods used for imputation purposes, §4.4.3, §4.4.4 and §4.4.5 contain details of the main analysis methods. As discussed in Chapter 3, multiple imputation was the preferred method of handling missing data within this thesis. Two multiple imputation techniques, MICE and SMC-FCS, were compared to a complete case analysis (CCA) in Chapter 6. The SMC-FCS method was only applicable to survival analysis due to the use of Cox PH modelling which is a non-linear model. For the access to healthcare analysis, which is based on logistic regression modelling (a form of generalised linear modelling), only MICE was used for imputation. The final analysis in Chapter 8 did not include any imputation, thus all methods relating to that chapter are provided in §4.4.5.

Prior to any imputation process, the missingness mechanism was explored. As described in Chapter 3, there is no definitive way to determine from the observed data whether the missing data are MCAR, MAR or MNAR. A distinction between MCAR and MAR can be made by assessing whether any of the non-missing variables within the observed dataset can predict missingness. This was determined with the use of logistic regression models, where a binary indicator variable for missingness of a particular variable was treated as the outcome measure, and the non-missing variables within the dataset were entered into the model as independent variables. One such model was produced for each variable which was to be imputed (i.e. disease severity and ethnicity for CNS tumours, leukaemia and GCT). The variables included in these missingness predictor models included all variables of interest in the main survival analysis (see §4.4.3 for variable choice), as well as any additional variables available within the Yorkshire register data, namely three binary variables for initial treatment modalities (chemotherapy, radiotherapy or surgery) and a binary variable to indicate whether a relapse had occurred. Furthermore, Kaplan-Meier survival curves for cases with missing data compared to cases without missing data were plotted in order to further explore the missing data mechanism. The presence of significant differences between cases with missing data compared to cases without missing data implies that the data are not MCAR, however, the absence of such differences is not sufficient to conclude that the data are MCAR. Evidence for MNAR cannot be obtained from the observed data alone, however, a sensitivity analysis can be performed (see §4.4.2.2).

4.4.2.1 Imputation Model Specification

Imputation models were specified for CNS tumours, leukaemia and GCTs individually due to their individual measures of disease severity. The imputation models included all variables of interest in the main survival analysis, including interaction terms (see §4.4.3), the Nelson-Aalen survival estimate and the censoring indicator. Any variables which were shown to be significant predictors of missingness based on the results of the missingness predictor models described above were included in the imputation models as auxiliary variables. Any continuous variables which could be categorised, such as age, year of diagnosis and deprivation were all included as continuous variables within the imputation models. Tests for linearity (see §4.4.3) for each of these variables were subsequently performed at the analysis stage.

The same imputation model specifications were used for the 'Inequalities in Disease Severity at Diagnosis' analysis. The outcome variables for survival analysis (Nelson-Aalen estimate and censoring indicator) were also retained in the imputation models as disease severity and survival were highly correlated as demonstrated in the results section

of Chapter 6. The outcome measure for the inequalities analysis was disease severity (see §4.4.4 for details), which was also a variable to be imputed. Chapter 3 contains a discussion of the literature surrounding imputation of the outcome variable. Based on this discussion the outcome variable was imputed and observed and imputed data were included within the analysis as firstly, auxiliary variables were used within the imputation process thereby providing additional information in the imputed outcome variable and secondly, the power of the analysis was substantially increased using this method due to the amount of missing disease severity data.

Collinearity of variables was checked using Spearman correlation coefficients between variable pairs. All imputation models for ethnicity were based on multinomial logistic regression models as this was an unordered categorical variable. Imputation of WHO grade for CNS tumours and stage for GCTs was based on ordered logistic regression due to the natural ordering of the variables (high grade/stage represents more severe disease at presentation). For leukaemia, WBC count was imputed using linear regression. As described in §3.5.4.4, it is important to check the normality assumption for any continuous variable which is to be imputed. The distribution of WBC count was checked graphically using a histogram with a Normal density curve overlay. WBC count was transformed using a natural logarithmic transformation due to evidence of non-Normality (§6.3.2).

4.4.2.2 Imputation Assessment and Sensitivity Analysis

In order to determine the quality of the imputations, stacked bar charts were plotted for the observed cases compared to each imputation number for all imputed categorical variables. For WBC count, a histogram showing observed and imputed data was produced. These graphics provided a visual method of assessing the general pattern of imputed values. Additionally, pooled parameter estimates obtained from the multiple imputation analysis models were compared to those obtained via a CCA as suggested by good practice guidelines [142]. In order to determine the number of imputation iterations required, the method suggested by Royston [242] of running one imputation with a large number of iterations and assessing the point at which the imputations became stationary was used. The point at which imputations became stationary was determined visually using trace plots. For categorical variables, convergence was difficult to assess using trace plots of the mean, however, the literature provides no alternative suggestions and were therefore still used. The recommended number of cycles used for multiple imputation was between 10 and 20 according to van Buuren and Groothuis-Oudshoorn [192], thus 20 cycles were used for tumour groups in which there was no evidence of later convergence.

For imputation models in which there was reason to believe the MAR assumption could

be violated, the data was analysed under plausible MNAR assumptions by allowing the underlying missingness probability to differ systematically. For example, by assuming that those with missing stage data had poorer survival compared to those with completed stage data. For continuous variables, the method is described in detail by van Buuren and Oudshoorn [193], in which a constant, referred to as the $\delta-$adjustment, is added to the imputed values. For categorical variables, the method is more complicated as it requires adjustment of the probability that the missing data takes a specific value, rather than adjustments of the values themselves. This method can be implemented in the `sens.mice` package in R, where parameters of θ are specified to allow a percentage increase or decrease in the missingness probability in relation to the outcome. The resulting imputations are analysed using the same techniques as for imputations under the MAR assumption and results compared. The data are deemed to be robust to deviations from the MAR assumption if the results obtained under the MNAR assumption do not differ substantially from those obtained under an MAR assumption (i.e. the magnitude, direction and significance of effect sizes are similar). van Buuren [243] states that sensitivity analysis should only be performed in situations where there is a good idea of different plausible missing data scenarios, and a careful imputation procedure is considered more important than conducting a poorly informed sensitivity analysis. Furthermore, van Buuren [243] state that it is sufficient to make adjustments to one imputation variable within the model, rather than all imputed variables. The reason for this is that the effect of changing the imputed values of one variable will change the imputations of the other imputed variables due to the cyclical nature of imputation by chained equations.

A sensitivity analysis using the $\delta-$adjustment method was performed for the missing WBC variable within the leukaemia diagnostic group analysis. For CNS tumours and GCTs, sensitivity analysis for WHO grade and stage were performed using the `sens.mice` package. The missingness probability adjustments were set at 50% and 80% respectively.

Further sensitivity analysis to determine whether or not the number of imputations was sufficient for repeatability of the analysis was conducted by assessing Monte Carlo errors (as described in §3.5.4.3). The number of imputations were deemed to be sufficient if the Monte Carlo error was sufficiently small so that the addition or subtraction of this error to the parameter estimates and P-values of the survival analysis did not change their interpretation. Initial choices for the number of imputations were set to be approximately equal to the amount of missing data, thus these are specified in Chapter 5.

4.4.3 Variation in Cancer Survival

Survival patterns of the cohort were initially described in a univariable analysis, and were summarised by one, three and five-year survival estimates alongside 95% confidence intervals. For imputed variables, it was initially thought that pooled estimates of the survivor function and the standard error could be obtained by manually calculating these according to Rubin's rules. However, after initial calculations, it became clear that although the pooled estimates were sensible, the associated standard errors taking into account the within and between imputation variance (see §3.5.4) produced confidence intervals outside of the range zero to one. The combining of estimates and their related standard errors across imputed dataset relied on the assumption that the estimates follow a normal distribution, which was not the case for the survivor function. Although transformations of the data could have been applied, research by Marshall et al. [244] suggests that despite such transformations it is unclear how to produce sensible pooled standard errors for measures which are bound by zero and one, and instead the authors suggest that providing the full range of estimates across all imputed datasets provides a more appropriate representation of the variability within the data.

In order to represent the univariable survival estimates for imputed variables, Kaplan-Meier (K-M) plots of the pooled estimates were overlayed onto a K-M plot showing the range of survival estimates at each time point over all imputations. However, several complications arose during this process. K-M curves could be obtained for each single imputation, however, these tended to overlap with each other and therefore simply selecting the minimum and maximum lines out of all the imputations was not possible. Instead, initially, the maximum and minimum value at each time point was selected. However, by doing so, the sequence of increasing time corresponding to a decrease in the proportion of people survived was not necessarily maintained. This was because for each single imputation of stage for example, one person could be assigned a different stage to the subsequent imputation(s). This implies that each K-M curve associated with each imputation contained different people and therefore different times at which events (deaths) occurred. Therefore, in order to obtain sensible K-M curves which maintained the standard pattern of decreasing proportion survived over time, whilst highlighting the variability across all imputations, individual K-M curves for each imputation were plotted in a light background colour, with the average over all imputations overlayed on to this plot in a dark foreground colour. As Stata has a maximum number of plot options (70), lines could not be coloured via the command line. Therefore, they were instead coloured using schemes written via the 'record' option in the Stata Graphics editor. Details of the Stata code required to implement this technique are given in Appendix H.

4.4.3.1 Model Selection Process

A multivariable Cox PH model was implemented for each imputed dataset for CNS tumours, leukaemia and GCTs and the resulting parameter estimates were combined using Rubin's rules [191], which were implemented in Stata by use of the `mi estimate, hr:` prefix. The model selection procedure for the Cox PH model followed the method described in Collett [245], pages 80-89, however, instead of testing significant improvements to model fit using a likelihood ratio test, model improvements were checked using the Wald test, which is asymptotically equivalent to the likelihood ratio test [246]. Model comparisons using an approximation of likelihood ratio testing [247] after imputation is possible for models in which the log-likelihood function is fully specified, and the package `milrtest` in Stata can be used for such tests in linear and logistic regression models. However, likelihood ratio testing after multiple imputation was not possible for Cox PH models as these models use the partial log-likelihood rather than the full log-likelihood function due to the baseline hazard ratio not being specified. Furthermore, White et al. [202] suggests that although likelihood ratio tests are often used to compare models for CCA, the Wald test is generally recommended for comparing models after imputation as there is no evidence to suggest that the approximation of the likelihood ratio test for multiple imputed estimates offers an advantage over the Wald test.

The model selection procedure recommended by Collett [245] involved identifying a set of explanatory variables for potential inclusion in the model alongside any potential interaction terms of interest based on the literature review provided in Chapter 2 and clinical input from Dr DP Stark. Subsequently, each explanatory variable was fitted individually in a univariable Cox PH model and compared to the null model to determine which variables had individual predictive power. All variables with individual predictive power were then sequentially added to the model, and only those that continued to improve model fit were retained in the model in addition to those of specific clinical interest in the study. Explanatory variables that were not individually predictive were then added to the model one at a time, as they may become important in combination with other variables. Again, only those that improved the model were retained. Once all main effects in the model were determined, interactions were added to the main effects model and individually tested against the nested main effects model. Main effects which formed part of interaction terms of interest were retained in the model even if they did not improve model fit so that the corresponding interaction could be tested and the hierarchic principle remained intact. If subsequently, the interaction term was not significant, then this was removed from the model alongside the non-significant main effect term. Linearity was checked for continuous variables which could be categorised (specified below) and were tested individually by adding a categorical variable to the full model containing the

corresponding continuous variable, a significant improvement in model fit gave evidence against linearity and would result in the categorical variable being chosen in favour of the continuous form.

Collett [245] highlights the importance of using clinical knowledge alongside sensible model checking procedures and therefore recommends a lenient significant value of 10% while selecting variables for inclusion or exclusion from the model. Alternative approaches to the procedure outlined above include automated stepwise procedures or the use of Directed Acyclic Graphs (DAGs). The reason for not choosing an automated forward, backward or stepwise model selection procedure is that the final model obtained can depend on the procedure used (for example, forward elimination could result in a different final model than backward elimination) and they depend on a pre-specified significance level which does not allow for clinical knowledge, or variables of key interest in a study to be included. DAGs can be helpful in determining the intricate causal pathways between explanatory variables and any sources of bias within a multivariable analysis. However, the use of DAGs to inform model selection, is more suited to an analysis in which there is one main exposure for a given outcome. Although it may be possible to select one variable as the main exposure, the purpose of this research was primarily to discover where variations in survival may arise due to a number of demographic and clinical factors, with no single most important exposure. Therefore, the analysis models presented in Chapters 6 and 7 are mutually adjusted multivariable models, which can be used to predict the survival for a particular person with characteristics of all the variables included in the model (for example a south Asian person, aged 25 diagnosed with a certain type of cancer in a particular year). Without the use of DAG, care must be taken to not make any causal inferences based on the analysis results as the effect of each variable includes both direct and indirect effects. Nevertheless, the approach of using mutually adjusted variables is arguably more applicable to a clinical scenario in which a doctor can determine the prognosis of a patient according to a full range of patient and clinical characteristics in combination.

Variables considered for inclusion in the Cox PH models were based upon those which were previously used within the CYA cancer survival literature. These were diagnostic subgroup, age, sex, deprivation, stage or severity of disease, ethnicity and year of diagnosis [40, 64, 248]. Deprivation was measured using the 2007 Index of Multiple Deprivation (IMD) [249], assigned to the lower super output area associated with the postcode at diagnosis. IMD 2007 is a measure of seven domains, which include income, employment, health and disability, education, skills and training, barriers to housing and services, living environment and crime. This data was obtained from GeoConvert, which is a web based tool run by the UK Data Service Census Support and allows extraction of geographical data based on postcodes. The IMD is an area based measure of deprivation

and is routinely used within epidemiological research in cases where individual based deprivation data is not available. Lower super output areas (LSOA) are confined to a population of approximately 1500 people, and there are a total 32,482 of these areas in England. The deprivation fifth for each lower super output area in England has been shown to remain relatively stable over a 25 year study period [250]. Despite IMD being the measure of socioeconomic status of choice by the UK government and its common use for public health research, the inclusion of a health domain has been criticised as it could lead to a statistical phenomenon known as mathematical coupling [251]. Mathematical coupling occurs when one variable contains all or part of another variable, and therefore when analysing relationships between these variables they will inevitably be highly correlated [252], thus potentially causing spurious results. However, Adams and White [253] assessed the impact of analysing health outcomes using IMD by comparing such an analysis to one in which the health domain was removed. Quintiles based on the full IMD score and the score without the health domain had a high level of agreement (92%). Furthermore, although the study showed that there were some differences in the relationship between health outcomes and the two IMD measures (with and without the health domain), these differences were very small and the authors concluded the inclusion of the health domain was of little practical importance. The IMD score was chosen in favour of other more historical area based deprivation scores (including the Townsend Deprivation Score [254] and Carstairs Index [255]) as it was updated using routine data collected in inter-census years, rather than being confined solely to census data. Furthermore, IMD assigns weights to each of its domains, compared to the equal weighting used in the Townsend deprivation score. The latest version of the 2007 IMD score includes postcodes from the most up to date National Statistics Postcode Products at time of analysis (2010), which is a directory of all postcodes in England. IMD 2007 is a contemporary measure of deprivation which falls within the period of diagnosis covered in this thesis.

The linearity of age, deprivation and year of diagnosis were all tested, if there were retained as explanatory variables in the modelling selection procedure, using the categorisation method. The problems of using categorisation within epidemiological studies are well described [256, 257, 258, 259] as discussed below, however, this method was adopted in favour of more advanced methods to account for non-linear relationships such as splines, fractional polynomials and generalized additive models (GAMs) [260, 261, 262] due to ease of interpretation of results by a clinical, non-statistical, audience. The disadvantages of categorisation include the assumption of homogeneity between the categorised groups of data and the introduction of multiple hypotheses testing through pairwise comparisons [257]. Furthermore, comparing results across studies can become problematic if categorisation is done based on data driven cut

off points such as quantiles [256]. Although other methods of testing for linearity are available (such as splines, fractional polynomials and GAMs), these methods are prone to over fitting which could lead to spurious findings and are known to hinder communication of results to a non-statistical audience [256, 259]. The results of this study were intended for a clinical audience and to be published in clinical journals, therefore, the ease of interpretation by clinical audiences and the translation of the results into practice was key. Moreover, the use of this approach compared with other methods allowed for straight forward quantification of the effect of imputation on the coefficients. Whereas quantifying the effect of imputation on coefficients of spline functions or fractional polynomials is not immediately obvious, although these could be described, rather than quantified, graphically.

For age at diagnosis, data was considered in 5-year age bands (0-4, 5-9, 10-14, 15-19, 20-24 and 25-29) as well as two main age groups (0-14 and 15-29). For leukaemia, the following age grouping was additionally checked (< 1, 1-10, 11-14, 15-29 years). For year of diagnosis, 5-year periods (1990-1994, 1995-1999, 2000-2004, 2005-2009) were considered and linearity of deprivation was checked against deprivation fifths, which were generated using the following quintiles of LSOA: 6496.4, 12992.8, 19489.2 and 25985.6. These quintiles were based on splitting the total number of LSOA in England into fifths, rather than by splitting the deprivation values for the cohort into fifths to ensure groups were relative to the whole of England rather than Yorkshire to ensure generalisability of results across England.

Age and sex were known to have an impact upon survival for certain diagnostic groups, for example, TYAs with acute lymphoblastic leukaemia (ALL) had poorer survival compared to children with the same diagnosis. The effect of age on survival could vary by sex [263]. In particular, as the cohort in this thesis spanned two main age groups (children < 15 years of age and TYAs aged 15-29), gender differences in survival could exist for childhood tumours for example, but not for TYAs. Therefore, interaction terms between age and sex as well as between age and diagnostic subgroup were considered for analysis. Furthermore, year of diagnosis could have a direct effect upon survival through changes in treatment regimens and protocols over time, and improvements in survival of CYA cancer over time are well documented (see §2.3.1). In order to determine whether potential improvements over time were restricted to specific diagnostic subgroups, an interaction between year of diagnosis and diagnostic subgroup was also considered for analysis.

Interactions were only included within the imputation and survival analysis if there were a sufficiently large number of events per interaction level. A minimum of 10 events per variable (EPV) have been shown to be required to ensure sufficient power in any survival analysis [264, 265, 266], thus for each tumour group, the number of deaths per interaction

level were summarised (§6.2.1) to determine whether interactions were to be included in the analysis. Further work by Vittinghoff and McCulloch [267] in 2007 highlights that the use of 10 EPV may be too conservative, therefore, 10 EPV was simply used as a guide to inform whether further subgroup analysis and interactions were to be studied, but was not used at the expense of including important confounding effects and variables of key interest in the analysis (such as ethnicity and disease severity). In cases where there was sufficient data, interactions were included within the imputation models in addition to the analysis models so that any potential interaction effects were not diluted by the imputation process as described in §3.5.4.4. Interaction terms were specified by creating dummy variables for each interaction level. The interactions of interest (as mentioned above) only included fully observed variables, thus special methods for interactions such as the 'just another variable' method (whereby a new variable representing the interaction is created and it is included in the imputation model as an ordinary main effects variable) or passive imputation (whereby the main effects within the interaction are imputed individually, and the interaction term is derived post-imputation analysis based on the imputed main effects) were not required. If including interactions with partially observed variables, the 'just another variable' method would be preferred, as this has been shown to perform better than passive imputation [207, 268].

4.4.3.2 Model Assessment

The PH assumption was checked graphically using log cumulative hazard plots for each variable in the final model of each diagnostic group across all imputations. Continuous variables were categorised as follows prior to plotting: year of diagnosis - 1990–1998 and 1999–2009 and WBC count - standard risk ($< 50,000\mu/L$) and high risk ($\geq 50,000\mu/L$). Goodness of fit (GOF) was assessed by plotting deviance residuals against risk scores (predicted values) for all imputations. Deviance residuals were chosen in favour of Martingale residuals as deviance residuals have been shown to detect residuals at the extreme negative (observed deaths occurred later than predicted by the model) as well as the extreme positive (observed deaths occurred earlier than predicted by the model) end of the scale, whereas Martingale residuals are not sensitive to extreme positive residuals [269, 270]. Therefore, lack of model fit was indicated by extreme outlying residuals on the deviance residual plots. Harrell's C-index of discrimination was used to assess the predictive performance of each model [271], which was obtained via Stata's post estimation command `estat concordance`. Harrell's C-index of discrimination cannot be combined using Rubin's rules [244], and therefore the range of values was described over all imputation models rather than providing one pooled estimate and CI.

4.4.4 Inequalities in Disease Severity at Diagnosis

In order to assess inequalities in disease severity at diagnosis, the analysis focused around three main factors of interest as identified in the literature (Chapter 2), these were ethnicity, deprivation and age. Missing data for disease severity and ethnicity were imputed using MICE (see §4.4.2 above for detailed imputation methods). In order to describe the imputed data, the average percentage over all imputations of cases by age group (0-14 vs. 15-29), ethnicity (White, Asian, other) and deprivation (1-5, where 1 was least deprived and 5 was most deprived) were summarised.

The effects of ethnicity and deprivation on disease severity were modeled using ordinal logistic regression for WHO grade of tumour for CNS tumours (grade I-IV) and stage for GCTs (stage I-IV). For leukaemia, the effects of ethnicity and deprivation on disease severity were modelled using logistic regression according to standard risk and high risk leukaemia based on WBC count measures of $< 50,000\mu/L$ and $\geq 50,000\mu/L$.

4.4.4.1 Model Selection Process

Initially, the effects of age, ethnicity and deprivation on disease severity were modelled individually using ordinal logistic regression for WHO grade of tumour for CNS tumours (grade I-IV) and stage for GCTs (stage I-IV) whilst adjusting for sex and year of diagnosis. Similarly, for leukaemia, the effects of age, ethnicity and deprivation on disease severity were modelled using logistic regression according to standard risk and high risk leukaemia based on WBC count measures of $< 50,000\mu/L$ and $\geq 50,000\mu/L$ respectively, again, adjusting for sex and year of diagnosis. After assessing the independent effects of age, ethnicity and deprivation, a multivariable model was fitted which included mutual adjustments for age, ethnicity and deprivation in addition to sex and year of diagnosis.

4.4.4.2 Model Assessment

The GOF for the logistic regression model (i.e. the leukaemia analysis) was tested using the Hosmer-Lemeshow (HL) test for each imputation [272]. The HL-GOF test was chosen in favour of assessing the Pearson or Deviance residuals as continuous variables were included in the model (year of diagnosis), implying that there could be as many covariate patterns as subjects. This condition has been shown to lead to incorrect P-values for model fit tested using Pearson or Deviance residuals [273], which the HL-GOF test overcomes by grouping data according to values of the estimated probabilities (i.e. the estimated probability of having high risk compared to low risk leukaemia based on

values of the independent variables). The HL-GOF tests the null hypothesis that the observed probability of the event (high risk leukaemia) and expected probability of the event obtained from the model are the same, therefore, a non-significant result provides no evidence against model fit. The recommended number of 10 subgroups was used for all HL-GOF tests. The discriminatory power (i.e. the sensitivity versus 1 minus specificity) for logistic regression models can be assessed using a receiver operator (ROC) curve, however, as the discriminatory power needed to be assessed for all imputations models, the area under the ROC curve (AUROC) values for each model were calculated and the mean and 95% CI of these values were summarised. This process was performed in Stata using the following commands, as described in the `mim:` Stata help file:

```
mim: logit x1 x2 x3, or
mim: predict xb
mim, cat(combine) byvar est(r(area)) se(r(se)) : roctab y xb
```

where `x1,x2` and `x3` are independent variables and `y` is the dependent variable.

In order to test the model fit of the ordered logistic regression models for the CNS tumour and GCT analysis, a series of logistic regression models were tested using the HL-GOF and AUROC methods described above per imputation. Although methods have been developed based on extensions of the HL-GOF test [274], these are not readily available or easily applied in current software, therefore Hosmer and Lemeshow [273] recommends use of the individual logistic regression method by Begg and Gray [275]. Furthermore, this method could be easily repeated for each imputation model as described above. The outcomes WHO grade and stage each have four ordered categories, therefore, model fit and predictive power was assessed for the set of logistic regression models with the following binary outcomes, whilst retaining the ordered structure of the data:

i) Grade/stage I compared to grade/stage II-IV,

ii) Grade/stage I-II compared to grade/stage III-IV,

iii) Grade/stage I-III compared to grade/stage IV.

4.4.5 Long Term Effects amongst Survivors of Cancer

In order to describe the risk of cardiovascular LEs amongst long term survivors of CYA cancer, a subset of the linked cancer registry and inpatient HES data was studied. A long term survivor of cancer was defined as anyone having survived a minimum of 5

years post diagnosis of cancer and a LE was defined as a co-morbidity (in this case a cardiovascular hospital event; see detailed definition below) occurring a minimum of 5 years post diagnosis. As described earlier in this chapter, inpatient HES data was available from 1996 to 2011, therefore, cases diagnosed between 1991 and 2006 who survived a minimum of 5 years were included in the subset of cancer registry linked HES data. This ensured that inpatient HES data for identification of late effects was available for all linked cases from 5 years of diagnosis. A summary of the linkage rates for this subset of the cohort are provided in §5.2.1. In order to determine whether the rate of cardiovascular LEs amongst survivors of CYA cancer differed to the rate of cardiovascular hospital events amongst the background population, record level inpatient HES data were obtained for the whole population resident within the former Yorkshire Regional Health Authority between 1996 and 2011 and matched on age and sex. The cohort of survivors were aged between 0 and 29 years at diagnosis and were diagnosed between 1991 and 2006. To ensure comparability of admission rates amongst the survivor cohort and the general population, the age at hospital admission was used to match the survivor cohort to the general population cohort. In 1996, the first year of available HES data, the possible age range at hospital admission for the survivor cohort was 5-34 years old, in 1997, the possible age range was 5 to 35 years old and so forth up until 5 to 49 year olds in 2011. Table 4.1 shows the age structure by diagnosis and hospital admission year. Any admissions for the general population outside of the highlighted age ranges in each year of hospital admission were excluded from the general population based HES data to ensure the overall age range included in the general population matched that of the cancer survivor cohort.

Table 4.1: Age structure at hospital admission by year of diagnosis (1991-2006) and year of hospital admission (1996-2011). Highlighted age ranges for each year of hospital admission were selected from the general population HES data as comparator data.

Year of Diagnosis	**Year of Hospital Admission**								
	1996	**1997**	**1998**	**1999**	…	**2008**	**2009**	**2010**	**2011**
1991	5-34	6-35	7-36	8-37	…	17-46	18-47	19-48	20-49
1992	4-33	5-34	6-35	7-36	…	16-45	17-46	18-47	19-48
1993	3-32	4-33	5-34	6-35	…	15-44	16-45	17-46	18-47
1994	2-31	3-32	4-33	5-34	…	14-34	15-44	16-45	17-46
1995	1-30	2-31	3-32	4-33	…	13-42	14-34	15-44	16-45
1996	0-29	1-30	2-31	3-32	…	12-41	13-42	14-34	15-44
1997	0-28	0-29	1-30	2-31	…	11-40	12-41	13-42	14-34
1998	0-27	0-28	0-29	1-30	…	10-39	11-40	12-41	13-42
1999	0-26	0-27	0-28	0-29	…	9-38	10-39	11-40	12-41
2000	0-25	0-26	0-27	0-28	…	8-37	9-38	10-39	11-40
2001	0-24	0-25	0-26	0-27	…	7-36	8-37	9-38	10-39
2002	0-23	0-24	0-25	0-26	…	6-35	7-36	8-37	9-38
2003	0-22	0-23	0-24	0-25	…	5-34	6-35	7-36	8-37
2004	0-21	0-22	0-23	0-24	…	4-33	5-34	6-35	7-36
2005	0-20	0-21	0-22	0-23	…	3-32	4-33	5-34	6-35
2006	0-19	0-20	0-21	0-22	…	2-31	3-32	4-33	5-34

Cardiovascular LEs were defined using inpatient diagnoses and operations fields as recorded in HES and grouped as follows: hypertension, cardiomyopathy and heart failure, coronary artery disease, pulmonary heart disease, pericardial and endocardial disease, valvular heart disease, conduction disorders, cerebrovascular disease and operations & procedures requiring hospitalisation. Expert advice from a cardiologist (Dr Chris Gale) was sought to form these groups and identify specific codes used to categorise the diagnoses codes and operations and procedure codes (classified according to ICD-10 and OPCS-4.5 respectively; see §4.4.5) and specific codes in each group are given in Table 4.2. A patient was identified as having a cardiovascular LE if they experienced at least one hospital admission containing any diagnoses or procedure in the above categories occurring exclusively five or more years after the diagnosis of cancer. Those who had either no reported cardiovascular LEs, or who had any cardiovascular hospital admission prior to, or within five years of, cancer diagnosis were defined as having had no cardiovascular LE as the latter could be an indication that the cardiovascular admission was related to a pre-existing or underlying condition, and not as a result of their cancer. The number of cases with a cardiovascular admission prior to or within five years of diagnosis was small and occurred in less than 1% of the cohort. Events were identified from all diagnosis and procedure fields, however, only the first occurrence of a particular event was included in the analysis so that ongoing conditions, which could be recorded multiple times, were not duplicated. It was important to include the primary diagnoses and all subsequently recorded diagnoses codes within the HES data to compare to the general population, as for the cohort of survivors, the majority of primary diagnoses were recorded as a neoplasm (see §5.2.2 for a data summary). Again, ongoing conditions, such as cancer, are sometimes recorded within diagnoses fields despite not being the diagnoses related to that specific hospital episode. Therefore, by only assessing primary diagnoses codes compared to the general population, the true rate of cardiovascular LEs could be underestimated.

Table 4.2: Diagnoses and operations classification codes (ICD-10 and OPCS-4.5 respectively) used to identify cardiovascular late effects and grouped into 9 categories

Cardiovascular Category and ICD-10 Code	**Description**
Hypertension	
I10	Essential (primary) hypertension
I11	Hypertensive heart disease
I12	Hypertensive renal disease
I13	Hypertensive heart and renal disease
I15	Secondary hypertension
Cardiomyopathy and Heart Failure	
I42	Cardiomyopathy
I43	Cardiomyopathy in diseases classified elsewhere

I50	Heart Failure
I51	Complications and illdefined descriptions of heart disease
I52	Other heart disorders in diseases classified elsewhere
Coronary Artery Disease	
I20	Angina pectoris
I21	Acute myocardial infarction
I22	Subsequent myocardial infarction
I23	Certain current complications following acute myocardial infarction
I24	Other acute ischaemic heart diseases
I25	Chronic ischaemic heart disease
Pulmonary Heart Disease	
I26	Pulmonary embolism
I27	Other pulmonary heart diseases
I28	Other diseases of pulmonary vessels
Pericardial disease	
I30	Acute pericarditis
I31	Other diseases of pericardium
I32	Pericarditis in diseases classified elsewhere
Valvular Heart Disease	
I08	Multiple valve diseases
I34	Nonrheumatic mitral valve disorders
I35	Nonrheumatic aortic valve disorders
I36	Nonrheumatic tricuspid valve disorders
I37	Pulmonary valve disorders
I33	Acute and subacute endocarditis
I38	Endocarditis, valve unspecified
I39	Endocarditis and heart valve disorders in diseases classified elsewhere
Conduction Disorders	
I44	Atrioventricular and left bundle-branch block
I45	Other conduction disorders
I46	Cardiac arrest
I47	Paroxysmal tachycardia
I48	Atrial fibrillation and flutter
I49	Other cardiac arrhythmias
Cerebrovascular Disease	
I60	Subarachnoid haemorrhage
I61	Intracerebral haemorrhage
I62	Other non traumatic intracranial haemorrhage
I63	Cerebral Infarction
I64	Stroke, not specificied as haemorrhage or infarction
I65	Occlusion and stenosis of precerebral arteries, not resulting in cerebral infarction
I66	Occlusion and stenosis of cerebral arteries, not resulting in cerebral infarction

I67	Other cerebrovascular diseases
I68	Cerebrovascular disorders in diseases classified elsewhere
I69	Sequelae of cerebrovascular disease
G45	Transient cerebral ischeamic attacks and related syndromes

Operations and Procedures[1]

K02	Other Transplantation of heart
K13	Transluminal repair of defect of septum
K14	Other open operations on septum of heart
K16	Other therapeutic transluminal operations on septum of heart
K23	Other operations of wall of heart
K25	Plastic repair of mitral valve
K26	Plastic repair of aortic valve
K27	Plastic repair of tricuspid valve
K28	Plastic repair of pulmonary valve
K29	Plastic repair of unspecified valve of hear
K30	Revision of plastic repair of valve of heart
K31	Open incision of valve of heart
K32	Closed incision of valve of heart
K33	Operations on aortic root
K34	Other open operations on valve of heart
K35	Therapeutic transluminal operations on valve of heart
K36	Excision of valve of heart
K38	Other operations on structure adjacent to valve of heart
K40	Saphenous vein graft replacement of coronary artery
K41	Other autograft replacement of coronary artery
K42	Allograft replacement of coronary artery
K43	Prosthetic replacement of coronary artery
K44	Other replacement of coronary artery
K45	Connection of thoracic artery to coronary artery
K46	Other bypass of coronary artery
K47	Repair of coronary artery
K48	Other open operations on coronary artery
K49	Transluminal balloon angioplasty of coronary artery
K50	Other therapeutic transluminal operations on coronary artery
K51	Diagnostic transluminal operations on coronary artery
K52	Open operations on conducting system of heart
K53	Other incision of heart
K54	Open heart assist operations
K55	Other open operations on heart
K56	Transluminal heart assist operations
K57	Other therapeutic transluminal operations on heart
K58	Diagnostic transluminal operations on heart

K59	Cardioverter defibrillator introduced through vein
K60	Cardiac pacemaker system introduced through vein
K61	Other cardiac pacemaker system
K62	Therapeutic transluminal operations on heart
K63	Contrast radiology of heart
K65	Catherisation of heart
K67	Excision of pericardium
K68	Drainage of pericardium
K69	Incision of pericardium
K71	Other operations on pericardium
K75	Percutaneous transluminal ballon angioplasty and insertion of stent into coronary
K77	Transluminal drainage of pericardium
K78	Transluminal operations on internal mammary artery side branch

[1]OPCS-4.5 Coding

4.4.5.1 Statistical Methods for Long Term Effects amongst Survivors of Cancer

The number of cardiovascular LEs amongst the cancer survivor cohort was summarised by age group at cancer diagnosis (0-14 year olds and 15-29 year olds) and the number of distinct cardiovascular LEs per survivor was described. Furthermore, the time to first cardiovascular hospital admission per survivor was described using the median and interquartile range (IQR). The cumulative incidence for cardiovascular LEs was estimated as a function of years since diagnosis whilst death without experiencing a cardiovascular LE was treated as a competing risk.

In order to compare the survivor cohort to the general population, the number and crude incidence (per 10,000 person-years) of cardiovascular hospital admissions amongst the survivor cohort and the general population were summarised by cardiovascular diagnosis and age group. For the cancer survivor cohort, total person-years (PYs) was calculated by summing up the exposure time of all cancer survivors, where exposure time was taken as the time to first cardiovascular event for all cases classified as having had a cardiovascular LE. For those who did not experience a cardiovascular LE, PYs was taken as the time until the end of the study period (31st December 2011) or the date of death. Time zero was taken as 5 years beyond cancer diagnosis, the point at which the patient became classified as a long term survivor. For the general population, the PYs were calculated as the sum of the total age-sex matched population of Yorkshire following the age and year structure highlighted in Table 4.1.

The excess risk of cardiovascular hospitalisation amongst the survivor cohort compared to the general population was described by calculating hospitalisation rate ratios (HRR) overall and by age group using the indirect standardisation method [276]. This method compares the rate of cardiovascular hospital admissions in the survivor cohort (study population) to the rate of cardiovascular admissions in the general population (reference population). HRRs were standardised to the general population by single year of attained age (age at hospital admission), year of event and sex.

To determine the risk factors for cardiovascular LEs amongst the survivor cohort Royston-Parmar relative survival modelling was implemented, which allowed for adjustment of the risk of a cardiovascular event in the general population by attained age, year of event and sex (Royston & Lambert, 2006). Explanatory variables included gender, age and year at diagnosis, diagnostic group, deprivation [249] and initial treatment type. In order to further assess the risk factors for cardiovascular LEs, two further models were fitted to the following subgroups:

i) Cases who received chemotherapy to examine the effect of the number of different anthracycline drugs administered (in the absence of accurate dose information).

ii) Cases who received radiotherapy (excluding cerebrovascular LEs) to examine the effect of radiation to the chest.

Models were fitted for all cancers combined and no further subgroup analyses for specific diagnoses was possible due to the low number of cardiovascular events in the cohort (n=119). As diagnostic groups were not studied individually, the impact of stage or disease severity was not assessed as disease severity was specific to each individual diagnostic group and individual analyses would be required.

4.4.5.2 Model Assessment for Long Term Effects amongst Survivors of Cancer

As the flexible parametric RP survival models rely on cubic splines to estimate the baseline hazard function, the scale and the complexity of the splines need to be determined [277]. Models were compared using the AIC and BIC for the proportional hazards (PH) model, proportional odds (PO) model and probit model between 1 and 5 degrees of freedom (df). The optimal (lowest) AIC and BIC values determined which combination of model scales and df provided the most appropriate model fit to the data. The PH assumption for PH models was checked using the log cumulative hazard plot for each variable included in the final model. Finally, sensitivity to the choice of model was assessed by plotting relative survival curves and 95% CI's at varying df for each of the PH, PO and probit models.

Chapter 5

Descriptive Data Analysis

5.1 Introduction

This chapter includes general descriptive statistics of the Yorkshire register and HES datasets. The cohort is described according to diagnostic group, age group (children and TYAs) and gender. Prior to a detailed description of the cohort, the ICCC and TYA Birch Classification schemes were applied to the data and compared to determine the suitability of each scheme for the analysis within this thesis. The number of cases which successfully linked to HES data are described in detail, and any differences between linked and non-linked cases were also explored. A summary of inpatient and outpatient hospital episodes is provided, alongside summary statistics for the cohort subgroup used within the analysis of cardiovascular LEs amongst survivors of CYA cancer. In addition, this chapter focuses on the level of missing data of stage (disease severity) and ethnicity. However, specific missing data mechanisms (described in §3.2) were considered in Chapter 6 alongside the main imputation analysis. The chapter concludes with a strategy for imputation within the thesis (§5.4).

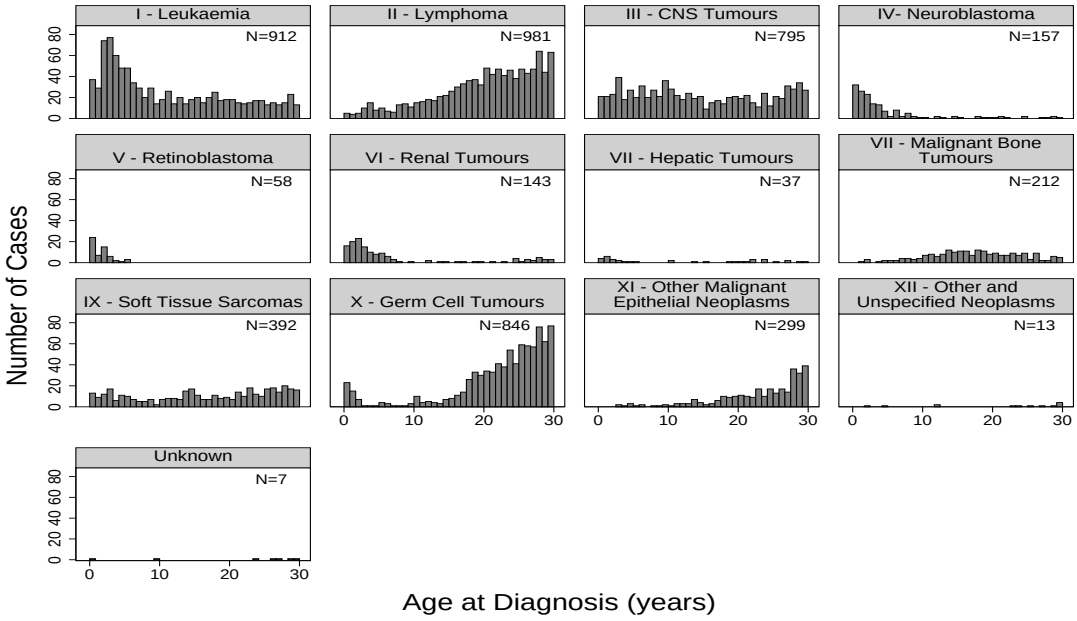
5.2 Study Population

There were a total of 4852 cases diagnosed between 1990 and 2009 inclusive in the former Yorkshire regional health authority and registered on the specialist register. After applying the ICCC and Birch TYA classification schemes to the morphology and topography data within the Yorkshire register, 0.1% (n=7) and 2.0% (n=95) of cases were classified as unknown for each scheme respectively (Figure 5.1). Furthermore, the ICCC resulted in fewer allocations to the 'other' categories (ICCC XI - other malignant epithelial neoplasms, ICCC XII - other and unspecified neoplasms) compared to the Birch TYA

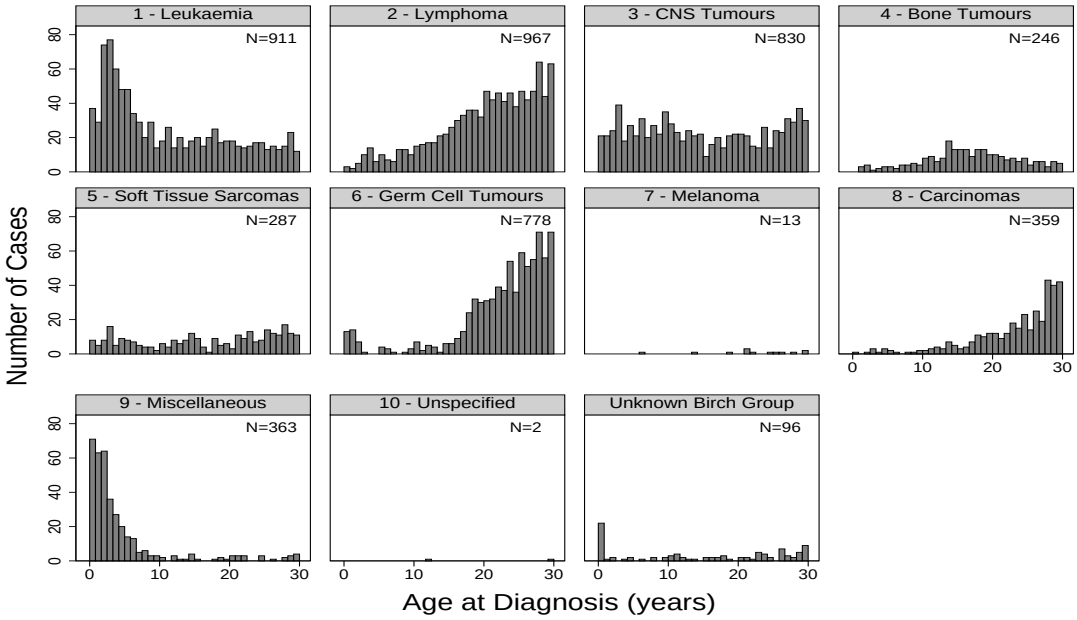
scheme (Birch Group 9 - Miscellaneous neoplasms not elsewhere classified and Birch Group 10 unspecified malignant neoplasms not elsewhere classified), with 6.4% (n=312) and 7.5% (n=365) for ICCC and Birch respectively. The ICCC scheme minimised the number of cases allocated to the 'other' and unknown categories, thus resulting in a larger number of cancers being allocated to a detailed classification. Therefore, the ICCC was used throughout the thesis in favour of the Birch TYA scheme.

Table 5.1 shows that amongst the cohort, the most commonly diagnosed cancers were leukaemia, lymphoma and GCTs. There were approximately one and a half times as many males as females in the cohort and the largest number of cases were diagnosed aged 0-4 years, 20-24 years and 25-29 years. Diagnoses of neuroblastoma, hepatic tumours and other and unspecified malignant neoplasms had poorest 5-year survival (54%, 57% and 54% respectively), compared to retinoblastoma and GCTs with much better survival at 5-years (98% and 94% respectively). There was little variation in 5-year survival according to gender or age group, averaging at 77%.

Cases of leukaemia, neuroblastoma, retinoblastoma and renal tumours tended to be diagnosed in younger children, with a peak in the number of leukaemia cases diagnosed around the age of 2 and 6 (Figure 5.1). Lymphomas, GCTs and other malignant epithelial neoplasms more commonly occurred amongst older TYAs. However, there was an incidence peak in GCTs for those diagnosed under the age of 1. The number of cases diagnosed with 'other specified malignant neoplasms' increased with age, which reflects the choice of a childhood specific classification scheme. This diagnostic group predominantly included tumours more commonly diagnosed amongst TYAs, consisting of 55% 'other and unspecified carcinomas' and 35% 'thyroid carcinomas'. CNS tumours and STS were diagnosed across the whole 0-29 year age range in roughly equal numbers, and malignant bone tumours was most common amongst 10-19 year olds and less common in very young children.



(a)



(b)

Figure 5.1: Age at diagnosis by (a) the international classification of childhood cancer (ICCC) and (b) the Birch teenage and young adult (TYA) diagnostic group classification for cases of cancer diagnosed in Yorkshire between 1990 and 2009

Table 5.1: Cases of childhood and young adult cancer diagnosed in the former Yorkshire regional health authority, 1990-2009

Variable	Cases (N)	Percentage	5-year survival (%)
Diagnostic group			
Leukaemia	912	18.8	72
Lymphoma	981	20.2	85
CNS tumours	795	16.4	69
Neuroblastoma	157	3.2	54
Retinoblastoma	58	1.2	98
Renal tumours	143	2.9	86
Hepatic tumours	37	0.8	57
Bone tumours	212	4.4	64
Soft tissue sarcomas	392	8.1	66
Germ cell tumours	846	17.4	94
Other malignant epithelial neoplasms	299	6.2	69
Other unspecified malignant neoplasms	13	0.3	54
Unknown ICCC group	7	0.1	86
Gender			
Male	2905	59.9	77
Female	1947	40.1	76
Age group (years)			
0-4	932	19.2	76
5-9	510	10.5	79
10-14	533	11.0	76
15-19	692	14.3	76
20-24	918	18.9	77
25-29	1267	26.1	78
Total	4852	-	77

5.2.1 Summary of Data Linkage

Of the 4852 cases diagnosed between 1990 and 2009, 382 cases died before 1st of April 1996 and were therefore not eligible for linkage to HES (which was available from this date forward). A total of 4113 (92%) of the cases eligible for linkage were successfully linked to at least one inpatient HES record (Figure 5.2).

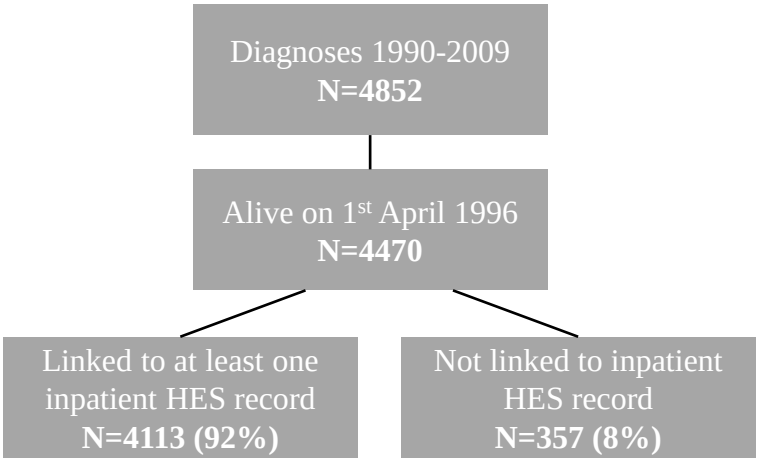


Figure 5.2: Number of linked and non-linked cases of cancer registry patients diagnosed between 1990 and 2009 to inpatient hospital episode statistics (HES) data for admissions between 1996 and 2011

Of the 4113 successful matches, 82% matched at the 1st pass, 16% were matched after the 2nd pass and fewer than 1% were matched at each of the remaining passes (Table 5.2).

Table 5.2: Match Rank for Linked Cases

Pass	**Number of Cases**	**Percentage**
1st Pass	3378	82%
2nd Pass	641	16%
3rd Pass	38	1%
4th Pass	5	<1%
5th Pass	12	<1%
6th Pass	0	0%
7th Pass	39	1%
Total	4113	100%

Outpatient data was available for 89% (n=3669) of the linked HES cases for outpatient appointments between 2003 and 2011. Outpatient data was initially obtained in addition to inpatient HES data to assess the number of cardiovascular LEs experienced by long

term survivors of cancer. For reasons outlined in §5.2.2.1 below, outpatient data were not used in any analysis.

5.2.2 Summary of Inpatient HES Data Admissions

There were a total of 83,614 FCEs for the 4,113 Yorkshire register cases who successfully linked to one or more inpatient HES record. Of these, the majority were single episode CIPS (n=71,653, 91%). Overall, the median number of admissions per case was 11, with an IQR of 4-27 (Table 5.3). Leukaemia patients had the highest median number of admissions per person (33, IQR=10-52), which could be a reflection of the number of day case admissions which are also recorded within inpatient HES data. The majority of primary HES diagnosis codes were neoplasms (67%), followed by 'symptoms, signs and abnormal clinical and laboratory findings' (6.0%) and 'factors influencing health status' (5.9%) (Table 5.4), although the latter were much more commonly seen amongst secondary diagnoses (14.3% and 27.6%). Other relatively frequent primary diagnoses included 'diseases of the blood and blood-forming organs' (2.9%), 'diseases of the digestive system'(2.7%) and 'injury, poisoning and certain consequences of external causes' (2.9%). The majority of neoplasms were recorded as the primary diagnosis, a relatively large proportion of neoplasms were recorded in secondary (9.9%), tertiary (6.6%) and quaternary (2.9%) diagnoses fields. The percentage of missing data increased rapidly for subsequent diagnoses fields, with 25%, 64% and 82% blank diagnoses fields for secondary, tertiary and quaternary diagnoses. This was an expected finding, as not every admitted patient is expected to have multiple diagnoses.

Table 5.3: Median and interquartile range (IQR) for the number of admissions[a] by main diagnostic group

Diagnostic Group[b]	**Median**	**IQR**
Leukaemia	33	10-52
Lymphoma	15	5-25
CNS Tumours	7	3-17
Other Solid Tumours	8	3-18
Overall	11	4-27

[a]Admissions were defined as continuous inpatient spells (CIPS)
[b]Based on the international classification of childhood cancer groups I, II, III and IV-XI respectively.

Table 5.4: Number and percentage of finished consultant episodes (FCEs) recorded in inpatient hospital episode statistics (HES) data between 1996 and 2011 grouped according to ICD-10 chapters for cases linked to the Yorkshire register diagnosed between 1990 and 2009

ICD-10 Chapters	**Diagnosis 1**		**Diagnosis 2**		**Diagnosis 3**		**Diagnosis 4**	
	N	**%**	**N**	**%**	**N**	**%**	**N**	**%**
A00-B99 Certain infectious and parasitic diseases	1,308	1.6	1,135	1.4	1106	1.3	644	< 1
C00-D48 Neoplasms	56,171	67.2	8,288	9.9	5,520	6.6	2,416	2.9
D50-D89 Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism	2,425	2.9	2,105	2.5	1,681	2.0	927	1.1
E00-E90 Endocrine, nutritional and metabolic diseases	574	< 1	716	< 1	801	1.0	614	< 1
F00-F99 Mental and behavioural disorders	132	< 1	272	< 1	327	< 1	204	< 1
G00-G99 Diseases of the nervous system	740	< 1	1,052	1.3	944	1.1	473	< 1
H00-H59 Diseases of the eye and adnexa	360	< 1	293	< 1	252	< 1	148	< 1
H60-H95 Diseases of the ear and mastoid process	306	< 1	184	< 1	179	< 1	106	< 1
I00-I99 Diseases of the circulatory system	501	< 1	586	< 1	508	< 1	355	< 1
J00-J99 Diseases of the respiratory system	1,888	2.3	1,216	1.5	1,164	1.4	525	< 1
K00-K93 Diseases of the digestive system	2,225	2.7	1,125	1.3	930	1.1	682	< 1
L00-L99 Diseases of the skin and subcutaneous tissue	635	< 1	237	< 1	183	< 1	140	< 1
M00-M99 Diseases of the musculoskeletal system and connective tissue	933	1.1	596	< 1	372	< 1	225	< 1
N00-N99 Diseases of the genitourinary system	1,064	1.3	514	< 1	455	< 1	297	< 1
O00-O99 Pregnancy, childbirth and the puerperium	1,671	2.0	155	< 1	255	< 1	107	< 1
P00-P96 Certain conditions originating in the perinatal period	122	< 1	50	< 1	34	< 1	25	< 1
Q00-Q99 Congenital malformations, deformations and chromosomal abnormalities	386	< 1	609	< 1	483	< 1	258	< 1
R00-R99 Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified	4,998	6.0	11,982	14.3	2,604	3.1	1,374	1.6
S00-T98 Injury, poisoning and certain other consequences of external causes	2,465	2.9	610	< 1	602	< 1	274	< 1
V01-Y98 External causes of morbidity and mortality	0	0.0	2,258	2.7	709	< 1	658	< 1
Z00-Z99 Factors influencing health status and contact with health services	4,702	5.6	23,108	27.6	10,522	12.6	4,053	4.8
Incorrectly recorded as a morphology code (not in ICD-10 format)	8	0.0	5,966	7.1	549	< 1	415	< 1
Blank	0	0.0	20,557	24.6	53,434	63.9	68,694	82.2
Total	83,614	100	83,614	100	83,614	100	83,614	100

5.2.2.1 Cardiovascular Late Effects Dataset

As described in §4.4.5 of Chapter 4, a subset of the cohort was used to analyse cardiovascular LEs among survivors of cancer (see Chapter 8 for main results). The linkage chart applicable to this subset is given in Figure 5.3. There were 3939 diagnoses of cancer between 1991 and 2006 registered on the Yorkshire register. Of these, 3,306 had survived a minimum of 5-years and were alive on the 1st of April 1996. A total of 3247 (98%) of cases linked to at least one inpatient HES record. Outpatient data were available for 74% (n=2412) of those successfully linked to inpatient HES data. In order to successfully use outpatient data to identify cardiovascular LEs, diagnoses code data were required. However, 99.5% of primary diagnoses codes in the available outpatient data were in the 'Symptoms, signs and abnormal clinical and laboratory findings not elsewhere classified' ICD-10 chapter, almost all of which were the specific code for 'unknown and unspecified causes of morbidity' (Table 5.5). The remaining 0.5% of outpatient records contained more specific diagnoses information from a range of ICD-10 chapters. For subsequent diagnoses, the outpatient data became even more sparse with just over 1% of secondary diagnoses codes being recorded as 'unknown and unspecified causes of morbidity' and 99% of missing data. For tertiary and quaternary diagnoses, there was virtually 100% missing data. In addition to the diagnoses fields for outpatient appointments, the main specialty under which the patient was seen was also recorded. This data showed that the majority of outpatient appointments for long term survivors of CYA cancer were with paediatric (21%) and medical oncology (14%) specialists. In total, 0.9% of appointments were recorded as being with 'cardiology' or 'paediatric cardiology' specialists. However, these data could not be used reliably to determine whether a survivor of cancer had experienced a cardiovascular LE or not, as the person could be referred to a cardiologist as a precautionary measure. Without data on the diagnosis made during an outpatient appointment it is not possible to confidently attribute a cardiovascular LE to someone who had attended an outpatient appointment under a cardiovascular specialty code. Outpatient data were therefore not used to analyse cardiovascular LEs, and the main analysis focused instead on cardiovascular LEs experienced in inpatient admissions (Chapter 8).

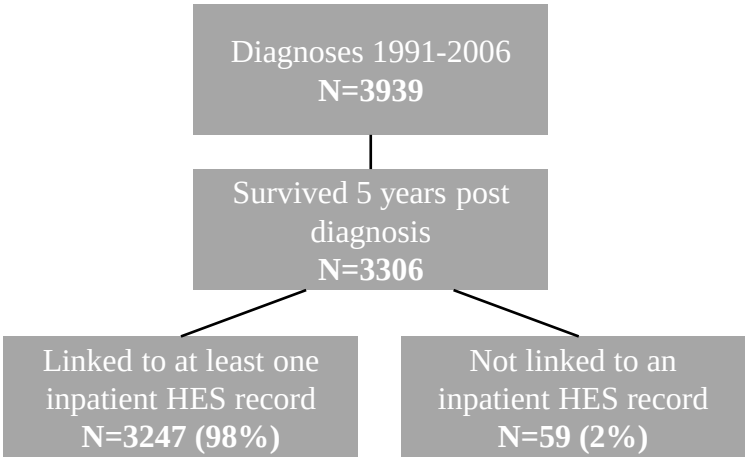


Figure 5.3: Number of linkages and non-linkages of cancer registry patients diagnosed between 1991 and 2006 to inpatient hospital episode statistics (HES) data for admissions between 1996 and 2011

Table 5.5: Number and percentage of finished consultant episodes (FCEs) recorded in outpatient hospital episode statistics (HES) data between 2003 and 2011 grouped according to ICD-10 chapters for cases linked to the Yorkshire register diagnosed between 1990 and 2009

ICD-10 Chapters	**Diagnosis 1**		**Diagnosis 2**		**Diagnosis 3**		**Diagnosis 4**	
	N	**%**	**N**	**%**	**N**	**%**	**N**	**%**
A00-B99 Certain infectious and parasitic diseases	< 5	< 1	5	< 1	5	< 1	0	0
C00-D48 Neoplasms	345	< 1	10	< 1	0	0	0	0
D50-D89 Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism	< 5	< 1	< 5	< 1	0	0	0	0
E00-E90 Endocrine, nutritional and metabolic diseases	19	< 1	0	0	< 5	< 1	0	0
F00-F99 Mental and behavioural disorders	40	< 1	0	0	0	0	0	0
G00-G99 Diseases of the nervous system	< 5	< 1	0	0	0	0	0	0
H00-H59 Diseases of the eye and adnexa	< 5	< 1	0	0	0	0	0	0
H60-H95 Diseases of the ear and mastoid process	< 5	< 1	0	0	0	0	0	0
I00-I99 Diseases of the circulatory system	< 5	< 1	0	0	0	0	0	0
J00-J99 Diseases of the respiratory system	7	< 1	0	0	0	0	0	0
K00-K93 Diseases of the digestive system	6	< 1	0	0	0	0	0	0
L00-L99 Diseases of the skin and subcutaneous tissue	5	< 1	0	0	0	0	0	0
M00-M99 Diseases of the musculoskeletal system and connective tissue	8	< 1	0	0	0	0	0	0
N00-N99 Diseases of the genitourinary system	8	< 1	0	0	0	0	0	0
O00-O99 Pregnancy, childbirth and the puerperium	14	< 1	0	0	0	0	0	0
P00-P96 Certain conditions originating in the perinatal period	0	0	0	0	0	0	0	0
Q00-Q99 Congenital malformations, deformations and chromosomal abnormalities	0	0	0	0	< 5	< 1	0	0
R00-R99 Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified	156,340	99.7	1,733	1.1	0	0	0	0
R69X - Unknown and unspecified causes of morbidity	156,324	99.7	1,732	1.1	0	0	0	0
S00-T98 Injury, poisoning and certain other consequences of external causes	0	0	0	0	0	0	0	0
V01-Y98 External causes of morbidity and mortality	0	0	0	0	0	0	0	0
Z00-Z99 Factors influencing health status and contact with health services	42	< 1	38	< 1	0	0	< 5	< 1
Missing	0	0	155,057	98.9	156,837	99.9	156,843	99.9
Total	156,844	100	156,844	100	156,844	100	156,844	100

5.2.3 Comparison of Linked and Non-Linked Cases

There were more cases of leukaemia and fewer cases of other solid tumours in those that linked compared to those that did not, there was a 3:2 male to female ratio compared to a 7:3 ratio for linked and non-linked cases respectively and the median age at diagnosis was significantly lower in linked cases (18.2 years) compared to non-linked cases (22.2 years) (Table 5.6).

Table 5.6: Distribution of cases for linked and non-linked Yorkshire registry and hospital episode statistics (HES) data

Variable	**Linked**		**Not Linked**		**Total**	**P-value**
	N	**Col %**	**N**	**Col %**	**N**	
Diagnostic group[a]						
Leukaemia	774	18.8	48	13.4	822	
Lymphoma	838	20.4	83	23.2	921	0.019^b
CNS Tumours	667	16.2	48	13.4	715	
Other solid tumours	1834	44.6	178	49.8	2012	
Gender						
Male	2424	58.9	257	72.0	2681	$< 0.001^b$
Female	1689	41.1	100	28.0	1789	
	Median	**IQR**	**Median**	**IQR**		
Age at diagnosis (years)	18.2	17.6	22.2	17.4		$< 0.001^c$
Total (N)	4113	-	357	-	4470	

[a]Broad diagnostic groupings were used to avoid small numbers
[b]Based on Pearson χ^2 test
[c]Based on Mann-Whitney U-test due to non-normal distribution of age at diagnosis

5.3 Missing Data

This section assesses missing data structures of stage and disease severity from the Yorkshire register and ethnicity data from the linked HES data. Missing data was described for each variable separately initially, after which missing data patterns across both variables were described for the whole dataset as well as by year of diagnosis to determine whether there were any changes over time in the level of completeness. Missing data was assessed overall as well as by diagnostic group and a strategy for imputation is given at the end of this chapter (§5.4).

5.3.1 Stage and Disease Severity

Table 5.7 summarises the available stage data by each of the 12 ICCC diagnostic groups. The table includes all recorded and raw values within the stage field of the Yorkshire register, which is a single field for all diagnostic groups combined. The data includes many values which appear to be invalid, as they are outside of the standard staging range of 1 to 4. However, validity cannot be assessed in detail for all cancers combined, as different staging mechanisms apply for each diagnostic group (discussed below). Nonetheless, the high number of cases for which a '9' was recorded suggests this value may have been entered to indicate missing data, despite the correct procedure for missingness in the Yorkshire register being to leave the field blank. Counting missing data as only those cases for which the stage was blank shows that other malignant epithelial neoplasms and other and unspecified neoplasms have the highest levels of missing data (93.3% and 100% respectively), whereas GCTs have the lowest level of missing data (40%). However, when additionally including '9' as a missing value, the level of missingness was very high (over 90%) for the majority of diagnostic groups, including leukaemia, retinoblastoma, hepatic tumours, bone tumours, other malignant epithelial neoplasms and other and unspecified neoplasms.

Table 5.7: Summary of recorded values of stage for cases of cancer diagnosed between 1990 and 2009 according to the international classification of childhood cancer (ICCC). Sensitive data on fewer than 5 cases were replaced by #

Recorded value for stage	**ICCC**[a]													**Total**
	I	**II**	**III**	**IV**	**V**	**VI**	**VII**	**VIII**	**IX**	**X**	**XI**	**XII**	**Missing ICCC**	
1	10	23	160	#	0	7	0	#	8	282	#	0	0	495
1A	0	31	12	0	0	0	0	0	0	8	0	0	0	51
1B	0	5	0	0	0	0	0	0	0	2	#	0	0	9
1C	0	0	0	0	0	0	0	0	0	#	0	0	0	#
1E	0	#	0	0	0	0	0	0	0	0	0	0	0	#
1M	0	0	0	0	0	0	0	0	0	8	0	0	0	8
1V	0	0	0	0	0	0	0	0	0	#	0	0	0	#
2	#	34	98	#	#	9	0	#	5	60	#	0	#	217
2A	0	60	#	0	0	0	0	0	0	9	0	0	0	72
2B	0	59	0	0	0	0	0	0	0	20	#	0	0	80
2C	0	0	0	0	0	0	0	0	0	6	0	0	0	6
2D	0	0	0	0	0	0	0	0	0	#	0	0	0	#
3	55	12	41	#	0	#	0	0	0	20	#	0	0	136
3A	#	18	#	0	0	0	0	#	0	0	0	0	0	27
3B	0	17	0	0	0	0	0	0	0	0	0	0	0	17
3C	0	0	0	0	0	0	0	0	0	#	0	0	0	#
4	8	19	135	34	0	7	0	#	#	44	#	0	0	257
4A	0	#	0	0	0	0	0	0	#	0	0	0	0	#
4B	0	15	0	0	#	0	0	0	0	#	0	0	0	18
4H	0	0	0	0	0	0	0	0	0	#	0	0	0	#
4L	0	0	0	0	0	0	0	0	0	#	0	0	0	#
4S	0	0	0	#	0	0	0	0	0	0	0	0	0	#
6	#	0	0	0	0	0	0	0	0	0	0	0	0	#
B	#	0	0	0	0	0	0	0	0	0	0	0	0	#
9	115	43	49	27	11	32	#	22	52	59	7	0	0	420
Blank	719	640	290	86	45	85	34	184	322	317	279	13	#	3020
Total cases	912	981	795	157	58	143	37	212	392	846	299	13	7	4,852
Missing stage														
Percentage blank	78.8	65.2	36.5	54.8	77.6	59.4	91.9	86.8	82.1	37.5	93.3	100.0	85.7	62.2
Percentage blank or '9'	91.4	69.6	42.6	72.0	96.6	81.8	100.0	97.2	95.4	44.4	95.7	100.0	85.7	70.9

[a] I - leukaemia, II - lymphoma, III - central nervous system tumours, IV - neuroblastoma, V - retinoblastoma, VI - renal tumours, VII - hepatic tumours, VIII - bone tumours, IX - soft tissue sarcomas, X - germ cell tumours, XI - other epithelial, XII - other and unspecified neoplasms.

The following sections explore the level of missing stage data per diagnostic group in detail by taking into consideration the validity of recorded values.

5.3.1.1 I - Leukaemia

Leukaemia is a haematological tumour which means that a stage, in terms of its traditional definition which includes a measure of the physical size of the tumour, does not apply. The AJCC does not include a TNM stage for leukaemia, nor does it mention another staging mechanism for this cancer. The French-American-British (FAB) classification is sometimes considered to stage leukaemia, as it classifies leukaemias in terms of cell type, with classifications ranging from M1 to M7 for myeloid leukaemia and L1 to L3 for lymphocytic leukaemia [278]. The values recorded in the stage field for leukaemia were likely to be based on this system (Table 5.7). However, the FAB classification is not related to the prognosis or severity of disease and will therefore not be used within this thesis as a staging mechanism. Many studies use WBC count as a prognostic factor for leukaemia, with high levels of white cells in the blood indicating poorer outcomes [279, 280, 281]. This measure was used throughout this thesis to measure disease severity at diagnosis for leukaemia.

Just over half (51%) of WBCs were missing for leukaemia overall, with a higher than average level of missing data for AML (64%) (Table 5.8). Despite there being roughly 50% fewer leukaemia cases amongst TYAs compared to children, TYAs had a much higher level of missing WBC data (69% compared to 42%).

WBC is commonly measured for cases diagnosed with leukaemia, and should therefore be available within the medical notes. Issues relating to missing data of stage due to its complexity therefore do not apply. Due to the unexpected high level of missing data for WBC amongst cases of leukaemia, efforts were made to retrieve this data from paper records held by the register as well as electronic records held by the paediatric oncology department at the Leeds General Infirmary, where the majority of childhood cancers in the region were treated. The level of missingness amongst children was minimised to 3% overall (Table 5.9). However, the level of missing WBC was still high (58%) for 15-29 year olds. The reason for this was that data on WBC was electronically stored for childhood leukaemias at the paediatric oncology department, however, data for TYAs around the region was not available in this format, therefore despite retrieving some of the data, there was still a high level of missingness of WBC for the TYA cohort.

Table 5.8: Summary of missing white blood cell (WBC) count for Leukaemia by age group at diagnosis. Sensitive data on fewer than 5 cases were replaced by #

Diagnostic Subgroup	Recorded WBC	Missing WBC	Total
Children (0-14 year olds)			
a) Lymphoid leukaemias	290 (60%)	195 (40%)	485
b) Acute myeloid leukaemias	46 (46%)	55 (55%)	101
c) Chronic myeloproliferative diseases	7 (54%)	6 (46%)	13
d) Myelodysplastic syndrome	# (75%)	# (25%)	#
e) Unspecified and other leukaemias	# (83 %)	# (17%)	#
a-e) All leukaemias	351 (58%)	258 (42%)	609
Teenagers and Young Adults (15-29 year olds)			
a) Lymphoid leukaemias	40 (36%)	70 (64%)	110
b) Acute myeloid leukaemias	42 (29%)	104 (71%)	146
c) Chronic myeloproliferative diseases	9 (24%)	28 (76%)	37
d) Myelodysplastic syndrome	# (40%)	# (60%)	#
e) Unspecified and other leukaemias	# (0%)	# (100%)	#
a-e) All leukaemias	93 (31%)	210 (69%)	303
All Ages (0-29 year olds)			
a) Lymphoid leukaemias	330 (55%)	265 (45%)	595
b) Acute myeloid leukaemias	88 (36%)	159 (64%)	247
c) Chronic myeloproliferative diseases	16 (32%)	34 (68%)	50
d) Myelodysplastic syndrome	# (56%)	# (44%)	#
e) Unspecified and other leukaemias	# (45%)	# (55%)	#
a-e) All leukaemias	444 (49%)	468 (51%)	912

Table 5.9: Summary of white blood cell count (WBC) data for leukaemia after retrieval of additional data from medical notes. Sensitive data on fewer than 5 cases were replaced by #

Diagnostic Subgroup	Recorded WBC	Missing WBC	Total
Children (0-14 year olds)			
a) Lymphoid leukaemias	472 (97%)	13 (3%)	485
b) Acute myeloid leukaemias	96 (95%)	5 (5%)	101
c) Chronic myeloproliferative diseases	# (92%)	# (8%)	#
d) Myelodysplastic syndrome	# (100%)	# (0%)	#
e) Unspecified and other leukaemias	# (100%)	# (0%)	#
a-e) All leukaemias	590 (97%)	19 (3%)	609
Teenagers and Young Adults (15-29 year olds)			
a) Lymphoid leukaemias	51 (46%)	59 (54%)	110
b) Acute myeloid leukaemias	60 (41%)	86 (59%)	146
c) Chronic myeloproliferative diseases	14 (38%)	23 (62%)	37
d) Myelodysplastic syndrome	# (60%)	# (40%)	#
e) Unspecified and other leukaemias	# (0%)	# (100%)	#
a-e) All leukaemias	128 (42%)	175 (58%)	303
All Ages (0-29 year olds)			
a) Lymphoid leukaemias	523 (88%)	72 (12%)	595
b) Acute myeloid leukaemias	156 (63%)	91 (37%)	247
c) Chronic myeloproliferative diseases	26 (52%)	24 (48%)	50
d) Myelodysplastic syndrome	# (78%)	# (22%)	#
e) Unspecified and other leukaemias	# (55%)	# (45%)	#
a-e) All leukaemias	718 (79%)	194 (22%)	912

5.3.1.2 II - Lymphoma

Similarly to leukaemia, lymphoma is a haematological tumour and cannot be staged according to the TNM system. However, the AJCC Staging Manual suggests the use of the Ann Arbor Staging System for HL and NHL to determine the disease severity at diagnosis. The Ann Arbor Staging System classifies lymphoma similarly to a grouped TNM stage, with levels I-IV (Table 5.10). Each stage is further accompanied with an A or B classification which refers to the absence or presence respectively of the following three symptoms; unexplained fever, night sweats and unexplained weight loss. Additionally, optional staging information based on the presence of extra nodal involvement can be recorded, which is indicated by an 'E'. The recorded values of stage for lymphoma cases show inconsistencies with the Ann Arbor staging system (Table 5.11). The value '9' is not valid, and stages recorded as '1', '2', '3' or '4' should be accompanied by an A or B. As discussed earlier, we can assume that the value of '9' indicates missingness. The symptoms (A or B codes) and extranodality (E code) provide additional information to main staging levels I-IV, and can therefore be considered as separate data items. Missingness is described in further detail by counting any values of I-IV, regardless of an additional A, B or E code, as complete data.

Overall, there was 70% missing stage data, which did not vary by age (76% and 68% for childhood and TYA cases respectively) (Table 5.12). Amongst the two largest subgroups (HL and NHL), there was a higher level of missingness for NHL compared to HL in both age groups (75% vs. 67% and 84% vs. 61% for childhood and TYA cases respectively).

Table 5.10: The Ann Arbor Classification System

Stage	Description
I	Involvement of a single lymphatic site
II	Involvement of two or more lymph node regions on the same side of the diaphragm
III	Involvement of lymph node regions on both sides of the diaphragm
IV	Diffuse or disseminated involvement of one or more extralymphatic organs[1]

[1]Extralymphatic organs are those that are outside of the lymphatic system, such as the liver, lungs or brain.

Table 5.11: Recorded values of stage for lymphoma cases diagnosed between 1990 and 2009. Sensitive data on fewer than 5 cases were replaced by #

Recorded value	**Cases**	
	N	**Percentage**
1	23	6.7
1A	31	9.1
1B	5	1.5
1E	#	#
2	34	10.0
2A	60	17.6
2B	59	17.3
3	12	3.5
3A	18	5.3
3B	17	5.0
4	19	5.6
4A	#	#
4B	15	4.4
9	43	12.6
Total	**341**	**100**

Table 5.12: Summary of missing stage for lymphoma by age group at diagnosis. Sensitive data on fewer than 5 cases were replaced by #

Diagnostic Subgroup	Recorded Stage	Missing Stage	Total
Children (0-14 year olds)			
a) Hodgkin lymphomas	31 (33%)	62 (67%)	93
b) Non-Hodgkin lymphomas	16 (25%)	48 (75%)	64
c) Burkitt lymphoma	# (3%)	# (97%)	#
d) Miscellaneous lymphoreticular neoplasms	0 (0%)	10 (100%)	10
e) Unspecified lymphomas	# (26%)	# (74%)	#
a-e) All lymphomas	53 (24%)	164 (76%)	217
Teenagers and Young Adults (15-29 year olds)			
a) Hodgkin lymphomas	213 (39%)	332 (61%)	545
b) Non-Hodgkin lymphomas	21 (16%)	114 (84%)	135
c) Burkitt lymphoma	# (0%)	# (100%)	#
d) Miscellaneous lymphoreticular neoplasms	# (0%)	# (100%)	#
e) Unspecified lymphomas	11 (16%)	56 (84%)	67
a-e) All lymphomas	245 (32%)	519 (68%)	764
All Ages (0-29 year olds)			
a) Hodgkin lymphomas	244 (38%)	394 (62%)	638
b) Non-Hodgkin lymphomas	37 (19%)	162 (81%)	199
c) Burkitt lymphoma	# (2%)	# (98%)	#
d) Miscellaneous lymphoreticular neoplasms	# (0%)	# (100%)	#
e) Unspecified lymphomas	16 (19%)	70 (81%)	86
a-e) All lymphomas	298 (30%)	683 (70%)	981

5.3.1.3 III - Central nervous system tumours

There is no TNM stage for CNS tumours, as the key factor relating to prognosis of CNS tumours is the growth of the tumour, which can damage brain functioning due to pressure on other parts of the brain. Furthermore, the N and M parts of the staging mechanism do not apply to CNS tumours as there are no lymph nodes within the brain or spinal cord, and CNS tumours tend to move around the CNS but do not tend to metastasize to other parts of the body [32]. The AJCC recommends the use of the WHO Grading of Tumours of the Central Nervous System scheme ([6], see Appendix D). This classification scheme assigns certain tumour morphologies to a specific grade (I to IV). The grading system does not align to the ICCC classification scheme, as for example, ICCC group III(b) is 'Astroyctoma' which includes WHO grade I tumours such as pilocytic astrocytoma and subependymal giant cell astrocytomas, as well as WHO grade IV tumours such as glioblastoma and gliosarcoma. The WHO grading system is a different type of grade to that mentioned in the TNM staging of STS in the example in Chapter 2, Table 2.1, where grade was used to supplement the stage, and refers to the level of differentiation of cells.

Table 5.13 gives the values for grade recorded by the Yorkshire register for this cohort. Those with a '9' and a number followed by the letter 'A' do not correspond with valid values for the WHO grading system. Both instances were treated as missing.

There was 46% missing data on the grade of CNS tumours overall, which did not vary by age (43% and 49% for children and TYAs respectively) (Table 5.14). The level of missing data varied according to diagnostic subgroup. Amongst children, the level of missing data was much lower for ependymomas, astrocytomas and intracranial and intraspinal embryonal tumours (34%, 38% and 24% respectively) compared to other gliomas, other specified and unspecified intacranial and intraspinal neoplasms (80%, 65% and 100% respectively). The variation in the level of missingness did not appear to be related to how rare the cancer was. Ependymomas made up less than 10% of all childhood CNS tumours, but had a similar level of missingness to astrocytomas, which account for almost half of CNS tumours. For the TYA age range the overall level of missingness was similar compared to children, however, there was less variation in the level of missing grade data according to diagnostic subgroup. Nevertheless, the pattern was similar to that of children in that other gliomas and unspecified intacranial and intraspinal neoplasms had the highest level of missingness (62% and 100% respectively).

Table 5.13: Recorded grade values for central nervous system tumours diagnosed between 1990 and 2009. Sensitive data on fewer than 5 cases were replaced by #

Recorded value	**Cases**	
	N	**Percentage**
1	160	31.7
1A	12	2.4
2	98	19.4
2A	#	#
3	41	8.1
3A	#	#
4	135	26.7
9	49	9.7
Total	**505**	**100**

Table 5.14: Summary of missing grade for central nervous system tumours by age group at diagnosis. Sensitive data on fewer than 5 cases were replaced by #

Diagnostic Subgroup	Recorded Grade	Missing Grade	Total
Children (0-14 year olds)			
a) Ependymomas and choroid plexus tumour	25 (66%)	13 (34%)	38
b) Astrocytoma	119 (62%)	74 (38%)	193
c) Intracranial and intraspinal embryonal tumours	78 (76%)	24 (24%)	102
d) Other gliomas	9 (20%)	37 (80%)	46
e) Other specified intracranial and intraspinal neoplasms	18 (35%)	34 (65%)	52
f) Unspecified intracranial and intraspinal neoplasms	0 (0%)	5 (100%)	5
a-f) All CNS tumours	249 (57%)	187 (43%)	436
Teenagers and Young Adults (15-29 year olds)			
a) Ependymomas and choroid plexus tumour	15 (52%)	14 (48%)	29
b) Astrocytoma	110 (58%)	81 (42%)	191
c) Intracranial and intraspinal embryonal tumours	# (41%)	# (59%)	#
d) Other gliomas	20 (38%)	32 (62%)	52
e) Other specified intracranial and intraspinal neoplasms	28 (48%)	30 (52%)	58
f) Unspecified intracranial and intraspinal neoplasms	# (0%)	# (100%)	#
a-f) All CNS tumours	184 (51%)	175 (49%)	359
All Ages (0-29 year olds)			
a) Ependymomas and choroid plexus tumour	40 (60%)	27 (40%)	67
b) Astrocytoma	229 (60%)	155 (40%)	384
c) Intracranial and intraspinal embryonal tumours	89 (69%)	40 (31%)	129
d) Other gliomas	29 (30%)	69 (70%)	98
e) Other specified intracranial and intraspinal neoplasms	46 (42%)	64 (58%)	110
f) Unspecified intracranial and intraspinal neoplasms	0 (0%)	7 (100%)	7
a-f) All CNS tumours	433 (54%)	362 (46%)	795

Although the level of missing WHO grade data was 46%, the WHO grading system can be used to infer the grade of tumour where it is missing, as the grading system is based upon the exact morphology of a tumour. Table 5.15 gives the level of missing data by subgroup after WHO grades were assigned based on morphology. This process was only completed for cases in which grade was missing, as it was assumed that data available via the data collection procedures had a higher degree of accuracy as additional information other than the morphology code may have helped determine the grade in those instances. After this process, 9% of missing grade data remained (Table 5.15). The reason for this was that some morphology codes were associated with multiple grades, and therefore the morphology code alone did not provide sufficient information to assign a unique grade. In these instances, the grade could not be inferred from the morphology code and were therefore imputed (Chapter 6).

Table 5.15: Missing grade for central nervous system tumours after assigning grades based on morphology using the WHO grading scheme [6]. Sensitive data on fewer than 5 cases were replaced by #

Diagnostic Subgroup	Recorded Grade	Missing Grade	Total
Children (0-14 year olds)			
a) Ependymomas and choroid plexus tumour	38 (100%)	0 (0%)	38
b) Astrocytoma	177 (92%)	16 (8%)	193
c) Intracranial and intraspinal embryonal tumours	102 (100%)	0 (0%)	102
d) Other gliomas	15 (33%)	31 (67%)	46
e) Other specified intracranial and intraspinal neoplasms	# (94%)	# (6%)	#
f) Unspecified intracranial and intraspinal neoplasms	# (0%)	# (100%)	#
a-f) All CNS tumours	372 (85%)	64 (15%)	436
Teenagers and Young Adults (15-29 year olds)			
a) Ependymomas and choroid plexus tumour	# (93%)	# (7%)	#
b) Astrocytoma	# (99%)	# (1%)	#
c) Intracranial and intraspinal embryonal tumours	27 (100%)	0 (0%)	27
d) Other gliomas	44 (85%)	8 (15%)	52
e) Other specified intracranial and intraspinal neoplasms	# (93%)	# (7%)	#
f) Unspecified intracranial and intraspinal neoplasms	# (50%)	# (50%)	#
a-f) All CNS tumours	338 (94%)	21 (6%)	359
All Ages (0-29 year olds)			
a) Ependymomas and choroid plexus tumour	# (97%)	# (3%)	#
b) Astrocytoma	367 (96%)	17 (4%)	384
c) Intracranial and intraspinal embryonal tumours	1129 (100%)	0 (0%)	129
d) Other gliomas	59 (60%)	39 (40%)	98
e) Other specified intracranial and intraspinal neoplasms	103 (94%)	7 (6%)	110
f) Unspecified intracranial and intraspinal neoplasms	# (14%)	# (86%)	#
a-f) All CNS tumours	710 (89%)	71 (9%)	795

5.3.1.4 IV - Neuroblastoma

Neuroblastoma is a disease which predominantly occurs amongst children, and as such, there was no TNM staging mechanism for neuroblastoma. There were two alternative staging systems in place for neuroblastoma, the International Neuroblastoma Staging

System (INSS) [282] and The International Neuroblastoma Risk Group (INRG) staging system [283]. The former was established in 1986, however, staging according to this system is dependent on surgery. Table 5.16 gives a full description of the INSS staging system. Dependency on surgery means that comparisons of staging between different centres is difficult as the same tumour could be Stage I or Stage III depending on the extent of surgical excision, which could vary between surgeons [284, 285]. These issues have led to the development of a pre-treatment staging system by the INRG, published by Cohn et al. [283]. The cancer register used the original INSS mechanism for staging tumours as the INRG system was relatively new, however, a recommendation to adopt the use of the new INRG staging system was made. This will depend on the uptake of the system by clinicians and upon which system is recorded within the medical notes. In addition to the INRG staging system, the same authors also developed an INRG risk group which is a pre-treatment risk group system in which patients were identified as very low, low, intermediate and high risk [283]. These data are not currently collected by the cancer register, however, it was recommended for future data collection.

Table 5.17 gives the values for stage contained within the cancer register for this cohort. All values except for the 27 cases of '9' were valid. The level of missing stage data was high (72%) overall, despite the data being relatively clean compared to stage data in other diagnostic groups (Table 5.18). Neuroblastoma is extremely rare in TYAs compared to amongst children, making up just over 10% of cases. All stage data for these cases was missing, possibly a result of its rarity. Particularly, if patients in this age range were treated within an adult clinic environment (as opposed to receiving paediatric care) the medical team were not necessarily aware of how to stage this particular tumour given that it does not have a TNM staging mechanism. Amongst children, stage was missing in 68% of the more common subgroup 'neuroblastoma and ganglioneuroblastoma', and missingness was as high as 78% for the less common 'other peripheral nervous cell tumours' subgroup.

Table 5.16: The International Neuroblastoma Staging System (INSS)

Stage	Description
I	The tumour can be removed completely during surgery. Lymph nodes removed during surgery may or may not contain cancer, but other lymph nodes near the tumour do not.
II-A	The tumour is located only in the area it started and cannot be completely removed during surgery. Nearby lymph nodes do not contain cancer.
II-B	The tumour is located only in the area where it started and may or may not be completely removed during surgery, but nearby lymph nodes do contain cancer.
III	The tumour cannot be removed with surgery. It has spread to regional lymph nodes (lymph nodes near the tumour) or other areas near the tumour, but not to other parts of the body.
IV	The original tumour has spread to distant lymph nodes (lymph nodes in other parts of the body), bones, bone marrow, liver, skin, and/or other organs (except for those listed in stage 4S, below).
IV-S	The original tumour is located only where it started (as in stage I, IIA, or IIB), and it has spread only to the skin, liver, and/or bone marrow (in infants younger than one). The spread to the bone marrow is minimal (usually less than 10% of cells examined show cancer).

Table 5.17: Recorded stage values for neuroblastoma cases diagnosed between 1990 and 2009. Sensitive data on fewer than 5 cases were replaced by #

Recorded value	Cases	
	N	Percentage
1	#	#
2	#	#
3	#	#
4	34	47.9
4S	#	#
9	27	38.0
Total	**71**	**100**

Table 5.18: Summary of missing stage for neuroblastoma by age group at diagnosis. Sensitive data on fewer than 5 cases were replaced by #

Diagnostic Subgroup	Recorded Stage	Missing Stage	Total
Children (0-14 year olds)			
a) Neuroblastoma and ganglioneuroblastoma	# (32%)	# (68%)	#
b) Other peripheral nervous cell tumours	# (22%)	# (78%)	#
a-b) All Neuroblastomas	44 (31%)	97 (69%)	141
Teenagers and Young Adults (15-29 year olds)			
a) Neuroblastoma and ganglioneuroblastoma	0 (0%)	5 (100%)	5
b) Other peripheral nervous cell tumours	0 (0%)	11 (100%)	11
a-b) All Neuroblastomas	0 (0%)	16 (100%)	16
All Ages (0-29 year olds)			
a) Neuroblastoma and ganglioneuroblastoma	# (31%)	# (69%)	#
b) Other peripheral nervous cell tumours	# (20%)	# (90%)	#
a-b) All Neuroblastomas	44 (28%)	113 (72%)	157

5.3.1.5 V - Retinoblastoma

Survival for retinoblastoma is very high compared to other childhood cancers, with 5-year survival reaching 99% in England (§2.3.1). As such, the main priority in treating retinoblastoma is that of saving the sight of the patient. There are two staging systems in place, both on a scale of 1 to 5. These are the Reese-Ellsworth staging system [286] and the International Classification for Intraocular Retinoblastoma [287]. Both staging mechanisms were designed for intraocular tumours (tumours within the eye), as most retinoblastoma tumours are diagnosed before they spread outside of the eye. As staging of retinoblastoma concerns saving of the patients eye sight rather than being related to the patients prognosis, staging of retinoblastoma was not included in the analysis for this thesis which focuses on survival, and therefore, imputation for this variable or a description of missingness was not completed.

5.3.1.6 VI - Renal Tumours

Renal tumours can be staged using the TNM Staging Mechanism [32]. The grouped stage contains levels I, II, III and IV and does not include any additional subgrouping. The only non-valid value recorded on the register for renal tumours was '9', which was treated as missing (Table 5.19). The level of missingness by diagnostic subgroup and age group was given in Table 5.20. Despite renal tumours having a relatively simple staging mechanism, there was still a high level of missing data (82% overall). Table 5.20 highlights differences between children and TYAs in terms of which type of renal tumour they were more likely to have (nephroblastoma and renal carcinomas for children and TYAs respectively), however, the pattern of missingness did not appear to differ by age group.

Table 5.19: Recorded stage values for renal tumours diagnosed between 1990 and 2009. Sensitive data on fewer than 5 cases were replaced by #

Recorded value	**Cases**	
	N	**Percentage**
1	7	12.1
2	9	15.5
3	#	#
4	#	#
9	32	55.2
Total	**58**	**100**

Table 5.20: Summary of missing stage for renal tumours by age group at diagnosis. Sensitive data on fewer than 5 cases were replaced by #

Diagnostic Subgroup	Recorded Stage	Missing Stage	Total
Children (0-14 year olds)			
a) Nephroblastoma and other nonepithelial renal tumours	21 (19%)	92 (81%)	113
b) Renal carcinomas	# (67%)	# (33%)	#
c) Unspecified malignant renal tumours	# (#)	# (#)	#
a-c) All renal tumours	23 (20%)	93 (80%)	116
Teenagers and Young Adults (15-29 year olds)			
a) Nephroblastoma and other nonepithelial renal tumours	# (0%)	# (100%)	#
b) Renal carcinomas	# (12%)	# (88%)	#
c) Unspecified malignant renal tumours	0 (-)	0 (-)	0
a-c) All renal tumours	# (11%)	# (89%)	#
All Ages (0-29 year olds)			
a) Nephroblastoma and other nonepithelial renal tumours	21 (18%)	93 (82%)	114
b) Renal carcinomas	5 (17%)	24 (83%)	29
c) Unspecified malignant renal tumours	0 (-)	0 (-)	0
a-c) All renal tumours	26 (18%)	117 (82%)	143

5.3.1.7 VII - Hepatic Tumours

Although a TNM staging system exists for liver tumours amongst adults, this only includes hepatic carcinomas and not hepatoblastoma. For paediatric hepatic tumours, the PRETEXT Staging System for Hepatoblastoma and Hepatocellular Carcinomas exists (Table 5.21).

There were a total of 37 hepatic tumours within this cohort (hepatoblastoma - n=16, hepatic carcinomas and unspecified malignant hepatic tumours - n=21), of these, <5 cases had a recorded value of ‘9’, which were treated as missing. No other data on stage was available for these hepatic tumours, thus stage was missing in 100% of cases.

Table 5.21: The PRETEXT staging system for Hepatoblastoma and Hepatocellular Carcinomas

Stage	Description
PRETEXT 1	Tumour involves only one liver sector; three adjoining liver sectors are free of tumour.
PRETEXT 2	Tumour involves one or two liver sectors; two adjoining liver sectors are free of tumour.
PRETEXT 3	Tumour involves three liver sectors and one liver sector is free of tumour or tumour involves two liver sectors and two non-adjoining liver sectors are free of tumour.
PRETEXT 4	Tumour involves all four liver sectors; there is no liver sector free of tumour.

5.3.1.8 VIII - Malignant Bone Tumours

Malignant bone tumours can be staged using the TNM staging mechanism, with additional subgrouping for stages I, II and IV (Tables 5.22 and 5.23).

The majority of recorded values were '9', which was invalid and treated as missing (Table 5.24). There were 6 further recorded values out of 212 bone tumour cases, however, these were not consistent with the TNM staging system as the A and B codes which supplement stages I, II and IV were not present. As with lymphoma, these were treated as separate data items and therefore, stages I, II and IV were counted as complete data. Any recorded values of '3A' were invalid and counted as missing as it was unclear whether the recorded stage should have been III, or any of IA, IIA and IVA. Overall bone tumour staging was missing in 98% cases. By age, there was 100% missing data for children with bone tumours. Table 5.25 gives the breakdown of recorded and missing stage by diagnostic subgroup (age group breakdown is not provided due to small numbers). There was no indication of any pattern of missing data, however, this could be because of the limited number of cases which had a recorded stage.

Table 5.22: Prognostic groups based on TNM staging for bone tumours

Grouped Stage	Primary Tumour (T)	Regional Lymph Nodes (N)	Distant Metastasis (M)	Histologic Grade (G)
Stage IA	T1	N0	M0	G1,2 GX
Stage IB	T2	N0	M0	G1,2 GX
	T3	N0	M0	G1,2 GX
Stage IIA	T1	N0	M0	G3, G4
Stage IIB	T2	N0	M0	G3, G4
Stage III	T3	N0	M0	G3, G4
Stage IVA	Any T	N0	M1a	Any G
Stage IVB	Any T	N1	Any M	Any G
	Any T	Any N	M1b	Any G

Table 5.23: TNM staging definitions for bone tumours

Primary Tumour (T)	
TX	Primary tumour cannot be assessed
T0	No evidence of primary tumour
T1	Tumour 8cm or less in greatest dimension
T2	Tumour more than 8cm in greatest dimension
T3	Discontinuous tumours in the primary bone site
Regional Lymph Nodes (N)	
NX	Regional lymph nodes cannot be assessed
N0	No regional lymph node metastasis
N1	Regional lymph node metastasis
Distant Metastasis (M)	
M0	No distant metastasis
M1	Distant metastasis
M1a	Lung
M1b	Other distant sites
Histologic Grade (G)	
GX	Grade cannot be assessed
G1	Well differentiated
G2	Moderately differentiated
G3	Poorly differentiated
G4	Undifferentiated

Table 5.24: Recorded values of stage for malignant bone tumours diagnosed between 1990 and 2009. Sensitive data on fewer than 5 cases were replaced by #

Recorded value	Cases	
	N	Percentage
1	#	#
2	#	#
3A	#	#
4	#	#
9	22	78.6
Total	**28**	**100**

Table 5.25: Summary of missing stage for malignant bone tumours by age group at diagnosis. Sensitive data on fewer than 5 cases were replaced by #

Diagnostic Subgroup	Recorded Stage	Missing Stage	Total
All Ages (0-29 year olds)			
a) Osteosarcomas	0 (0%)	103 (100%)	103
b) Chondrosarcomas	# (10%)	# (80%)	#
c) Ewing tumour and related sarcomas of bone	# (3%)	# (97%)	#
d) Other specified malignant bone tumours	0 (0%)	20 (100%)	20
e) Unspecified malignant bone tumours	# (25%)	# (75%)	#
a-e) All soft tissue sarcomas	5 (2%)	207 (98%)	212

5.3.1.9 IX - Soft Tissue and Other Extraosseuous Sarcomas

STS can be staged using the TNM staging mechanism, with additional subgroups 'A' and 'B' for stages I and II (See Tables 2.1 and 2.2, Chapter 2). There were 52 cases with the value of '9' recorded, all of which were classed as missing (Table 5.26). There was one value of '4A' recorded which is inconsistent with the staging mechanism in Table 2.2 and was treated as missing. The overall level of missingness of stage was very high at 96%. The high level of missingness occurred consistently across diagnostic subgroups as well as by age group (Table 5.27).

Table 5.26: Recorded stage values for soft tissue sarcomas. Sensitive data on fewer than 5 cases were replaced by #

Recorded value	**Cases**	
	N	**Percentage**
1	8	11.4
2	5	7.1
4	#	#
4A	#	#
9	52	74.3
Total	**70**	**100**

Table 5.27: Summary of missing stage for soft tissue sarcomas by age group at diagnosis. Sensitive data on fewer than 5 cases were replaced by #

Diagnostic Subgroup	Recorded Stage	Missing Stage	Total
Children (0-14 year olds)			
a) Rhabdomyosarcomas	# (3%)	# (97%)	#
b) Fibrosarcomas, peripheral nerve sheath tumours, and other fibrous neoplasms	0 (0%)	23 (100%)	23
c) Kaposi sarcoma	# (0%)	# (100%)	#
d) Other specified soft tissue sarcomas	# (7%)	# (93%)	#
e) Unspecified soft tissue sarcomas	# (17%)	# (83%)	#
a-e) All soft tissue sarcomas	7 (4%)	159 (94%)	166
Teenagers and Young Adults (15-29 year olds)			
a) Rhabdomyosarcomas	# (6%)	# (94%)	#
b) Fibrosarcomas, peripheral nerve sheath tumours, and other fibrous neoplasms	6 (7%)	75 (93%)	81
c) Kaposi sarcoma	0 (0%)	9 (100%)	9
d) Other specified soft tissue sarcomas	# (3%)	# (97%)	#
e) Unspecified soft tissue sarcomas	0 (0%)	26 (100%)	26
a-e) All soft tissue sarcomas	10 (4%)	216 (96%)	226
All Ages (0-29 year olds)			
a) Rhabdomyosarcomas	# (3%)	# (97%)	#
b) Fibrosarcomas, peripheral nerve sheath tumours, and other fibrous neoplasms	6 (6%)	98 (94%)	104
c) Kaposi sarcoma	0 (0%)	10 (100%)	10
d) Other specified soft tissue sarcomas	7 (5%)	142 (95%)	149
e) Unspecified soft tissue sarcomas	# (3%)	# (97%)	#
a-e) All soft tissue sarcomas	17 (4%)	375 (96%)	392

5.3.1.10 X - Germ Cell Tumours

Staging of GCTs is complex as different staging mechanisms exist depending on which subgroup the tumour belongs to. However, these subgroups do not align to the ICCC GCT subgroups. The ICCC subgroups for GCTs include the following categories;

Xa) Intracranial and intraspinal germ cell tumours

Xb) Malignant extracranial and extragonadal germ cell tumours

Xc) Malignant gonadal germ cell tumours

Xd) Gonadal carcinomas

Xe) Other and unspecified malignant gonadal tumours

The most commonly occurring GCT in this cohort were malignant gonadal GCTs (Xc). This diagnostic subgroup includes both ovarian and testicular GCTs, each of which have their own TNM staging system, which also apply to subgroups Xd and Xe for the same topographies. In addition to the TNM staging system, ovarian GCTs can be staged according to the International Federation of Obstetricians and Gynaecologists (FIGO) staging system. The FIGO and TNM staging systems have the same number of groups and subgroups (see Table 5.28), however, there are differences in terms of how these are defined [32]. There was also an additional staging mechanism for testicular GCTs, known as the Royal Marsden staging system [288].

Malignant extracranial and extragonadal GCTs (Xb) were staged using the International Germ Cell Cancer Collaborative Group (IGCCCG) Classification, using a basic four stage system. The more complex GCTs classified in group Xa do not have a universally accepted staging mechanism [289], however, they are sometimes staged according to the Chang TM staging system for medulloblastoma, which is a CNS tumour [290, 291].

Table 5.28 shows the main staging systems in place for different types of GCTs alongside an indication of which ICCC subgroups align to these systems. In total, there are five staging systems for GCTs covering most of the diagnostic subgroups. Each staging system has their own subgroups but are broadly speaking on a four level scale. Additionally, there was no applicable staging mechanism for intracranial and intraspinal GCTs. The implications on the analysis caused by these complications are discussed in Section 5.4.

The Yorkshire register does not record the staging mechanism alongside the stage value, therefore, it was difficult to ascertain which values were valid or not. As a consequence, any value of stage which matched to any of the possible stages indicated in Table 5.28 was deemed valid. Blank or recorded values of ‘9’ were classified as missing.

Missingness is presented by age and diagnostic subgroup, excluding group Xa for which no staging mechanism applied (Table 5.29). Despite GCTs being a complex group of tumours to stage, the level of missingness was 43% overall, which was much lower compared to other diagnostic groups. The overall level of missingness was lower amongst the TYA age group (36% missing) which made up the biggest proportion of GCTs (91% of GCTs in this cohort were diagnosed amongst TYAs). Overall and amongst TYAs, the group which had the lowest level of missing stage data was Xc (<40% missing) which was associated with clearly defined TNM staging systems for ovarian and testicular tumours.

However, the same pattern was not seen amongst children, which despite being subject to the same TNM staging mechanisms, had 82% missing stage for group Xc.

Table 5.28: Germ cell tumour staging systems

Name	Stage Values	ICCC Subgroup
TNM Staging for Ovarian Germ Cell Tumours	I, IA, IB, IC, II, IIA, IIB, IIC, III, IIIA, IIIB, IIIC, IV	Xc, Xd, Xe (provided site is ovary)
FIGO Staging for Ovarian Germ Cell Tumours	I, IA, IB, IC, II, IIA, IIB, IIC, III, IIIA, IIIB, IIIC, IV	Xc, Xd, Xe (provided site is ovary)
TNM Staging for Testicular Germ Cell Tumours	0, I, IA, IB, IS, II, IIA, IIB, IIC, III, IIIA, IIIB, IIIC	Xc, Xd, Xe (provided site is testis)
The Royal Marsden System for Testicular Germ Cell Tumours		Xc, Xd, Xe (provided site is testis)
IGCCCG Classification for Extracranial and Extragonadal Germ Cell Tumours	I, II, III, IV	Xb (any site)

Table 5.29: Summary of missing stage for germ cell tumours by age group at diagnosis. Sensitive data on fewer than 5 cases were replaced by #

Diagnostic Subgroup	Recorded Stage	Missing Stage	Total
Children (0-14 year olds)			
b) Malignant extracranial and extragonadal germ cell tumours	# (6%)	# (94%)	#
c) Malignant gonadal germ cell tumours	7 (18%)	32 (82%)	39
d) Gonadal carcinomas	# (0%)	# (100%)	#
e) Other and unspecified malignant gonadal tumours	# (0%)	# (100%)	#
b-e) Germ Cell Tumours (exlcuding Intracranial and Intraspinal GCTs)	9 (12%)	66 (88%)	75
Teenagers and Young Adults (15-29 year olds)			
b) Malignant extracranial and extragonadal germ cell tumours	7 (30%)	16 (70%)	23
c) Malignant gonadal germ cell tumours	435 (64%)	241 (36%)	676
d) Gonadal carcinomas	# (28%)	# (72%)	#
e) Other and unspecified malignant gonadal tumours	# (25%)	# (75%)	#
b-e) Germ Cell Tumours (exlcuding Intracranial and Intraspinal GCTs)	451 (62%)	281 (38%)	732
All Ages (0-29 year olds)			
b) Malignant extracranial and extragonadal germ cell tumours	9 (16%)	46 (84%)	55
c) Malignant gonadal germ cell tumours	442 (62%)	273 (38%)	715
d) Gonadal carcinomas	# (27%)	# (73%)	#
e) Other and unspecified malignant gonadal tumours	# (18%)	# (82%)	#
b-e) Germ Cell Tumours (exlcuding Intracranial and Intraspinal GCTs)	460 (57%)	347 (43%)	807

5.3.1.11 XI - Other Malignant Epithelial Neoplasms

Age structures for each diagnostic group were explored at the beginning of §5.2, which highlighted that most tumours within ICCC group XI occurred primarily in the older TYA age range. This implied that the types of tumours classified within this 'other' category, tended to be those more commonly seen in older adults and could be seen according to the subgroups of XI which were predominantly 'other and unspecified carcinomas' and 'thyroid carcinomas'. Staging mechanisms are difficult to define for a diagnostic group which contains a range of different tumours, especially when the majority of those were 'other and unspecified carcinomas'. Therefore the validity of stage data was difficult to determine. Values of '9' were counted as missing and the remaining values I, IB, II, IIB, III and IV were all considered complete (Table 5.30). Stage was missing for all 44

childhood cases, as well as for all cases in diagnostic subgroups XIa, XIb, XIc and XId. For XIf, the level of missing data was high (100% and 92%) amongst 0-14 and 15-29 year olds respectively. Overall, stage was missing in 96% of cases.

Table 5.30: Summary of missing stage for other malignant epithelial neoplasms by age group at diagnosis. Sensitive data on fewer than 5 cases were replaced by #

Diagnostic Subgroup	Recorded Stage	Missing Stage	Total
Children (0-14 year olds)			
a) Adrenocortical carcinomas	0 (0%)	14 (100%)	14
b) Thyroid carcinomas	0 (0%)	18 (100%)	18
c) Nasopharyngeal carcinomas	0 (0%)	# (100%)	#
d) Malignant Melanomas	0 (0%)	# (100%)	#
f) Other and unspecified malignant neoplasms	0 (0%)	6 (100%)	6
a-d, f) Other Malignant Neoplasms[a]	0 (0%)	44 (100%)	44
Teenagers and Young Adults (15-29 year olds)			
a) Adrenocortical carcinomas	0 (0%)	# (100%)	#
b) Thyroid carcinomas	0 (0%)	87 (100%)	87
c) Nasopharyngeal carcinomas	0 (0%)	16 (100%)	16
d) Malignant Melanomas	0 (0%)	# (100%)	#
f) Other and unspecified malignant neoplasms	13 (8%)	145 (92%)	158
a-d, f) Other Malignant Neoplasms[a]	13 (5%)	252 (95%)	265
All Ages (0-29 year olds)			
a) Adrenocortical carcinomas	0 (0%)	7 (100%)	7
b) Thyroid carcinomas	0 (0%)	105 (100%)	105
c) Nasopharyngeal carcinomas	0 (0%)	10 (100%)	10
d) Malignant Melanomas	0 (0%)	13 (100%)	13
f) Other and unspecified malignant neoplasms	13 (8%)	151 (92%)	164
a-d, f) Other Malignant Neoplasms[a]	13 (4%)	286 (96%)	299

[a]Category e) 'Skin Carcinomas' was excluded as tumours in this category were not registered on the Yorkshire register.

5.3.1.12 XII - Other and Unspecified Malignant Neoplasms

It is not possible to stage tumours which were of an unknown or unspecified nature, and therefore no stage data were available for these tumours.

5.3.1.13 Summary of Missing Stage Data

The above detailed descriptions of stage per tumour group highlight the problems which arise for staging of CYA cancers. Firstly, there were a number of different mechanisms for staging across diagnostic groups as well as across subgroups. Furthermore, stage does not relate to prognosis in all cases (for example, leukaemia and retinoblastoma) and no staging mechanism existed for some tumour types (for example, malignant extracranial and extragonadal GCTs). For the purposes of this thesis, the term stage and disease severity will be used interchangeably to refer to the extent and severity of disease at diagnosis. The term stage will be used generally throughout this thesis to refer to a measure of disease severity. For clarity, Table 5.31 provides the measures of disease severity used for each diagnostic group.

Table 5.31: Stage and disease severity by diagnostic group for cancer in children and young people

Diagnostic Group	**Measure of Stage/Disease Severity**
I - Leukaemia	White blood cell (WBC) count (continuous, $10^3\mu/L$)
II - Lymphoma	Stage (Ann Arbor Staging Classification, I-IV)
III - CNS Tumours	Grade (WHO Grade, I-IV)
IV - Neuroblastoma	Stage (INSS, I-IV)
V - Retinoblastoma	Not Applicable
VI - Renal Tumours	Stage (TNM, I-IV)
VII - Hepatic Tumours	Stage (PRETEXT, I-IV)
VIII - Malignant Bone Tumours	Stage (TNM, I-IV)
IX - Soft Tissue Sarcomas	Stage (TNM, I-IV)
X - Germ Cell Tumours	Stage (TNM ovarian, I-IV; TNM testicular, 0-III; FIGO I-IV; Royal Marsden, I-IV and IGCCCG I-IV)
XI - Other Epithelial Neoplasms	Stage (unspecified systems, I-IV)
XII - Other and Unspecified Neoplasms	Not Applicable

The level of missing stage data was very high in most diagnostic groups (Figure 5.4). For leukaemia, CNS tumours and GCTs, the level of missing data was below 50%. However, for all other diagnostic groups, the level of missing stage data was above 70% and close to 100% in a few cases.

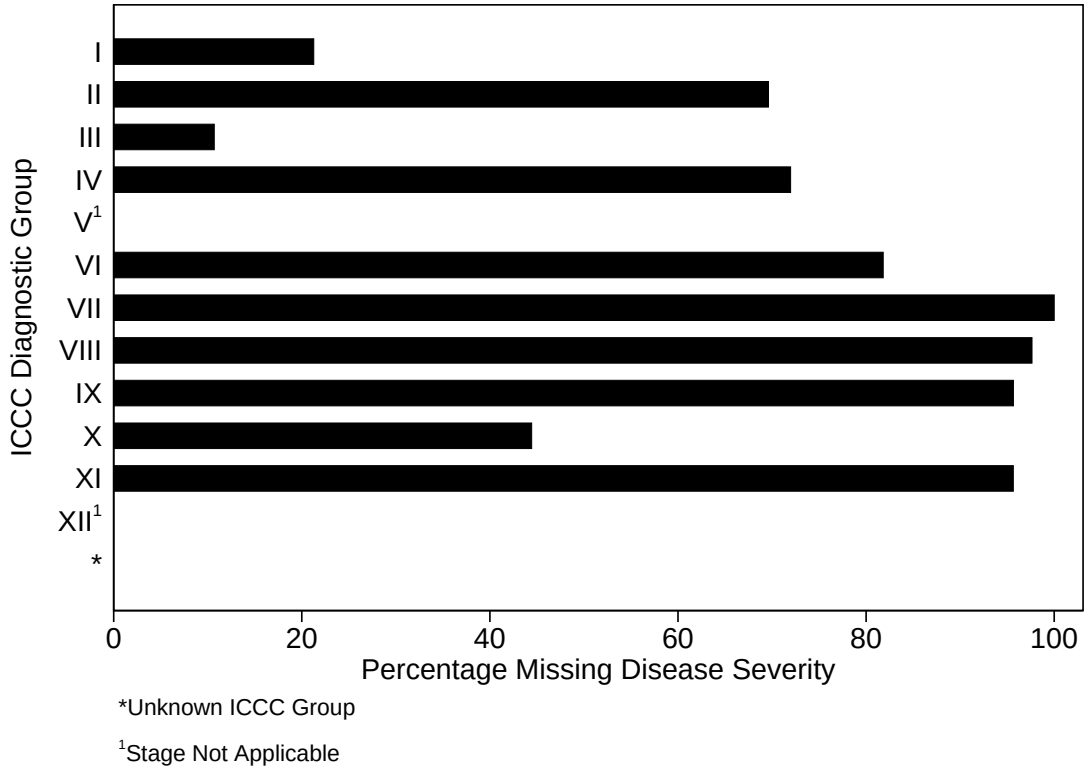


Figure 5.4: Level of missing disease severity by diagnostic group

The Yorkshire register database was unable to accurately capture staging data of this nature due to a single field for stage which was applicable for all tumours. This meant that there was no validation upon data entry of the values recorded, which has resulted in a large amount of missing or invalid recordings of stage. A detailed scheme for each diagnostic group has been developed (Appendix I) including all the information on stage and disease severity detailed in this chapter as well as additional data items in relation to prognosis for some tumour groups. This information formed part of the development of the Yorkshire register's electronic database system, which is currently ongoing. The implementation of accurate and up to date staging information on the data collection and storage system for the Yorkshire register is anticipated to improve the completeness and importantly, the quality of data on stage and disease severity.

5.3.2 Ethnicity

Detailed ethnicity data was obtained from linked HES records. These data were coded according to an 11 category system until April 2001, after which a more detailed 18 category system based on 2001 census classifications was adopted. In order to analyse these data, pre-2001 ethnicity codes and post-2001 ethnicity codes were mapped into one

system of broad ethnic groups, as shown in Table 5.32.

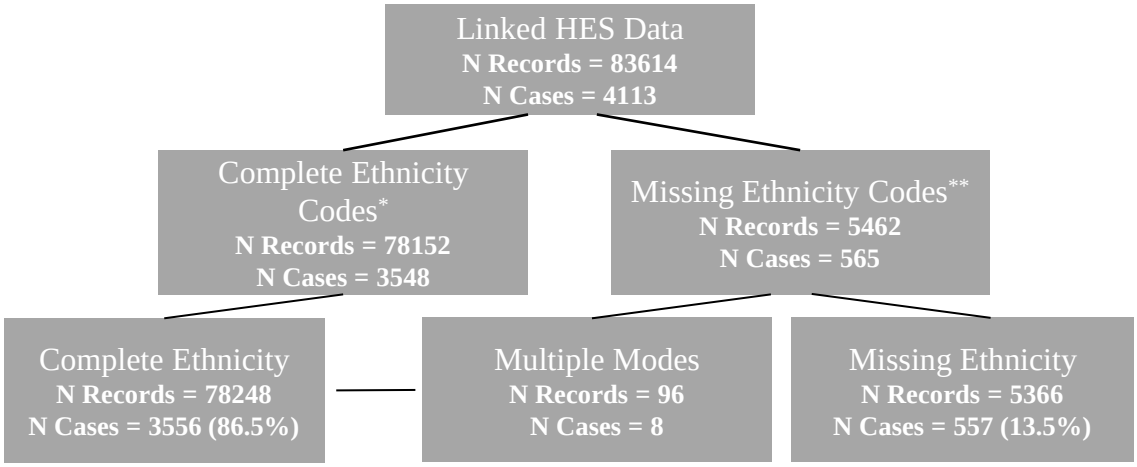
Table 5.32: Ethnicity coding schemes contained in hospital episode statistics (HES) data pre- and post-2001 alongside suggested broad grouped categories

Pre-2001 Census Coding	**2001 Census Coding**	**Grouped Categories**
0 = White	A = British (White) B = Irish (White) C = Any other White background	White
	D = White and Black Caribbean (Mixed) E = White and Black African (Mixed) F = White and Asian (Mixed) G = Any other Mixed background	Mixed
4 = Indian 5 = Pakistani 6 = Bangladeshi	H = Indian (Asian or Asian British) J = Pakistani (Asian or Asian British) K = Bangladeshi (Asian or Asian British) L = Any other Asian background	Asian
1 = Black - Caribbean 2 = Black - African 3 = Black - Other	M = Caribbean (Black or Black British) N = African (Black or Black British) P = Any other Black background	Black
7 = Chinese	R = Chinese (other ethnic group)	Chinese
8 = Any other ethnic group	S = Any other ethnic group	Other
X = Not known 9 = Not given	X = Not known Z = Not stated	Unknown

HES data can contain multiple episodes (FCEs) per patient, and patient demographic data is entered and recorded separately for each such FCE rather than being carried forward from a previous FCE. This has the implication that one patient can be associated with multiple ethnicity codes. Furthermore, some records associated with one person could have missing ethnicity codes, whilst other ethnicity codes for that person are recorded. In order to overcome this, an initial imputation scheme, selecting the most commonly recorded ethnicity code (the ethnicity mode) for each person, was implemented. Missing values were not included when determining the most common code. For example, one person with 10 FCEs of which three had been recorded as ‘White’ ethnicity and the

remaining 7 ethnicity codes were missing, was assigned to 'White' ethnicity. This was done to avoid the loss of valuable ethnicity information. In some cases, it was not possible to identify the most commonly recorded ethnicity value either because all were missing or because there were multiple codes recorded an equal number of times. The former scenario was treated as an ordinary missing data problem and accounted for using multiple imputation techniques (results in Chapter 6), the latter was determined using the method described below.

Figure 5.5 provides the missing data structure for ethnicity, which indicates that there were 8 individuals in the dataset with multiple modes of ethnicity. Each of these records was reviewed manually to determine the ethnicity. Another option would be to treat these cases as having missing ethnicity, however, that would disregard potentially useful data. The assignment of ethnicity was based on the record which contained the most detailed information. For example, if one of the modes was white, and the other mode was mixed then the latter was the assigned ethnicity as this was a more detailed ethnic category. In cases where the two modes were white and other, white was allocated as this was more detailed than 'other'. In cases where there was conflicting information such as white and black or white and Chinese, a cross check with the full name of the person was made where this was available. If the name was not available or did not provide further indications of ethnicity, the non-white categories were chosen as white could have been incorrectly entered as a default value. Lastly, if there were cases with conflicting information between categories, where neither category was white, and where no mixed category existed for those two ethnicities then this would have been assigned to missing, however, no such cases occurred in this data. There were no cases with more than two modes. Table 5.33 shows all combinations of multiple modes of ethnicity that were observed alongside their final assignments.



*Includes HES records with complete ethnicity data and cases with multiple ethnicity codes per person with on distinct mode.
**Include HES records for patients for which all records had missing ethnicity, and patients for which there were multiple modes.

Figure 5.5: Ethnicity Data Chart

Table 5.33: Ethnicity assignment for observed scenarios of multiple ethnicity modes

Scenario	Mode 1	Mode 2	Assigned Ethnicity
1	White	Mixed	Mixed
2	White	Other	White
3	White	Black	Black
4	White	Chinese	Chinese
5	Mixed	Asian	Mixed

After individual review of these 8 cases, there were a total of 3556 (86.5%) cases with recorded ethnicity and a remaining 557 (13.5%) with missing ethnicity data. However, there were an additional 739 cases which did not link to HES data and as such also had missing ethnicity. Therefore, ethnicity was missing in 1297 (26.7%) cases overall. The majority of the cohort was White (64.2%), and the second largest group were Asian (6.4%) (Figure 5.6). There was little variation in the level of missing ethnicity data by diagnostic groups (Figure 5.7), ranging from 20% to 40% on average. Missing data appeared high for 0-14 year olds in ICCC group XII (50%), however, this group was very small with fewer than 5 cases in total.

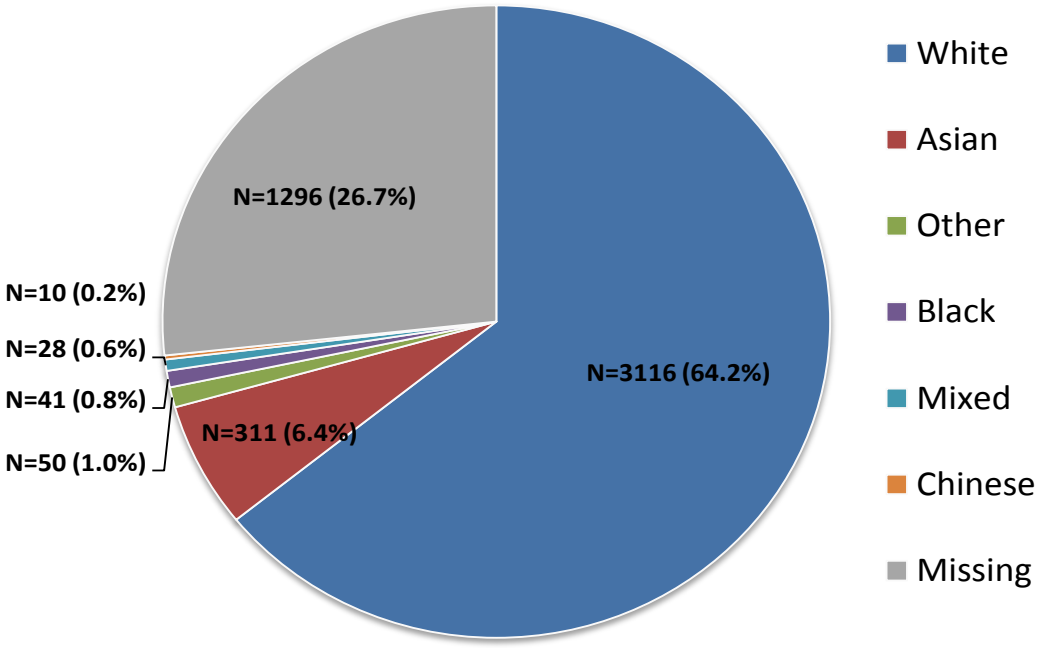


Figure 5.6: Distribution of ethnic groups amongst children and young adults with cancer in Yorkshire, 1990-2009

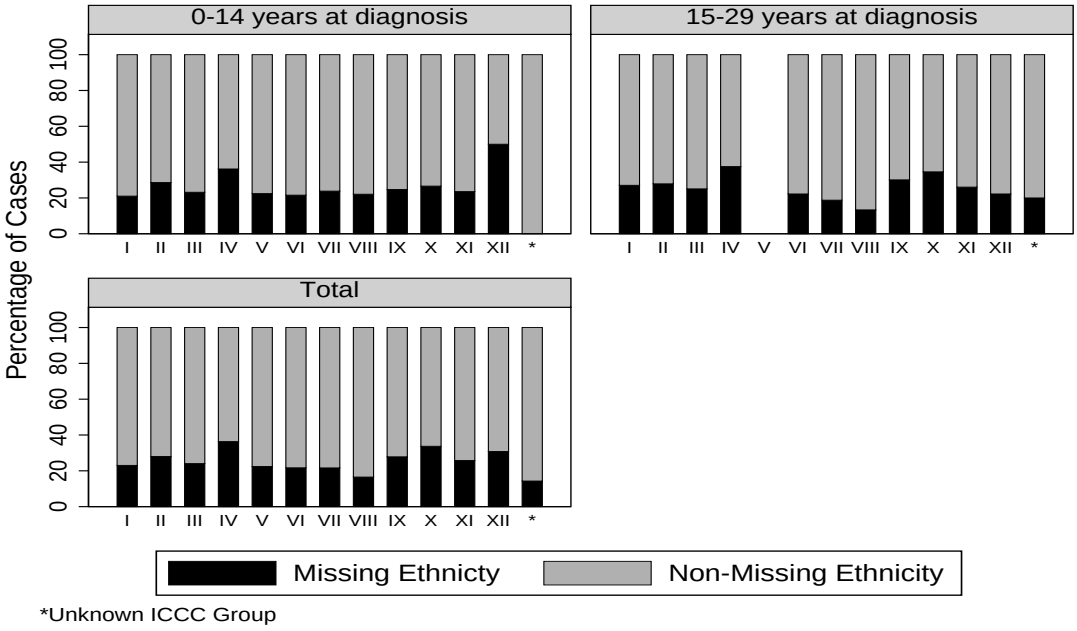


Figure 5.7: Percentage of Missing Ethnicity Data by Diagnostic Group and Age at Diagnosis

5.3.3 Missing Data Patterns

In order to fully understand the overall amount of missing data within the cohort, the missing data for individuals across multiple variables was assessed as well as missing

data trends over time. This section continues to focus on missing disease severity and ethnicity data as the remaining variables of interest within the Yorkshire register did not contain any missing data.

Figure 5.8 shows that complete data was available for over 60% of cases of leukaemia (I), CNS tumours (III), retinoblastoma (V), other and unspecified neoplasms (XII) and for those with unknown ICCC group. For GCTs (X), the level of complete data was approximately 40% and for the remaining diagnostic groups, complete data was available for less than 20% of cases. The level of completeness was high for retinoblastoma, other and unspecified neoplasms (XII) and for those with unknown ICCC group as stage was not applicable for these diagnostic groups, and thus only missing data for ethnicity was counted towards the overall percentage of missing data. For leukaemia, additional efforts were made to retrospectively collect WBC count and for CNS tumours, the WHO grading system was used to assign a grade according to morphological code which had resulted in higher levels of completeness for these tumour groups.

Figure 5.9 shows the percentage of missing disease severity and ethnicity data by year of diagnosis. For disease severity, the overall average level of missing data appears stable over the study period, between 40 and 60% missingness (univariable incidence rate ratio (IRR) = 0.99, 95% CI 0.99-1.01, P-value=0.630). For ethnicity, the percentage of missing data decreased significantly over time by 12% on average (univariable IRR=0.88 (95% CI 0.88-0.89, P-value$<$0.001). There was a large drop in missingness of ethnicity around the mid 1990s (Figure 5.9), which was due to linkage to HES data (available from 1996/7 onwards). After a small increase in the level of missing data, there was a further drop in the level of missing ethnicity between 2002 and 2009 which is indicative of improvements in HES data over time.

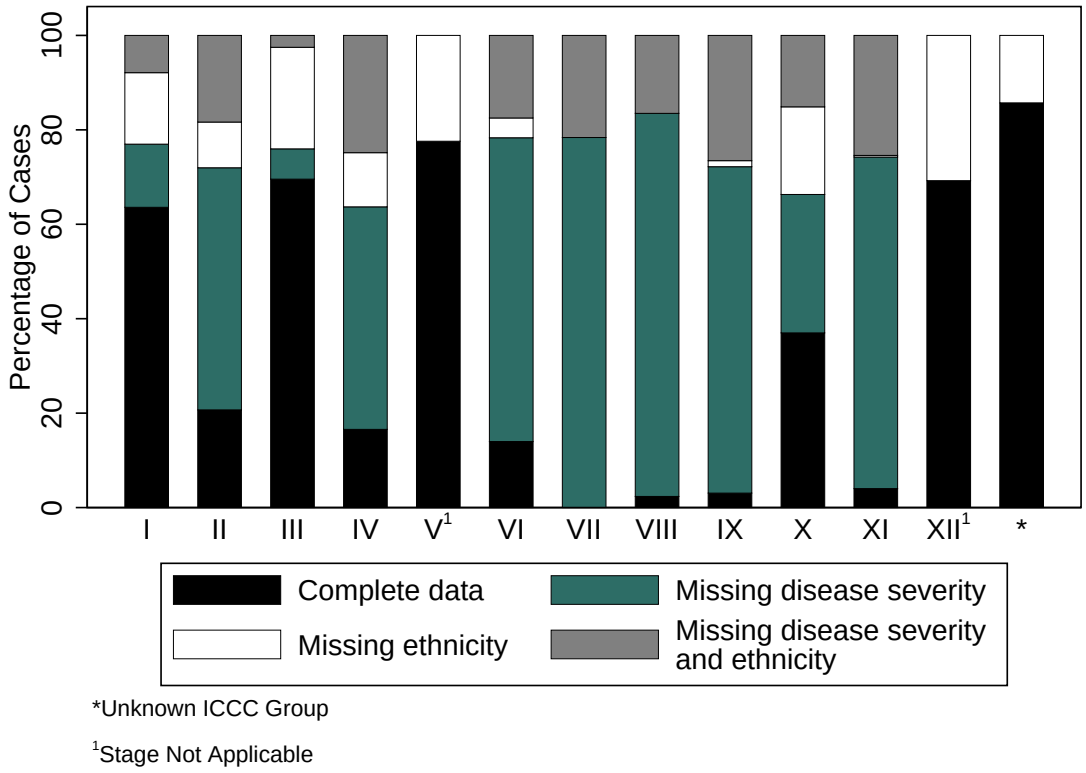


Figure 5.8: Percentage of missing disease severity and ethnicity by international classification of childhood cancer (ICCC) diagnostic group

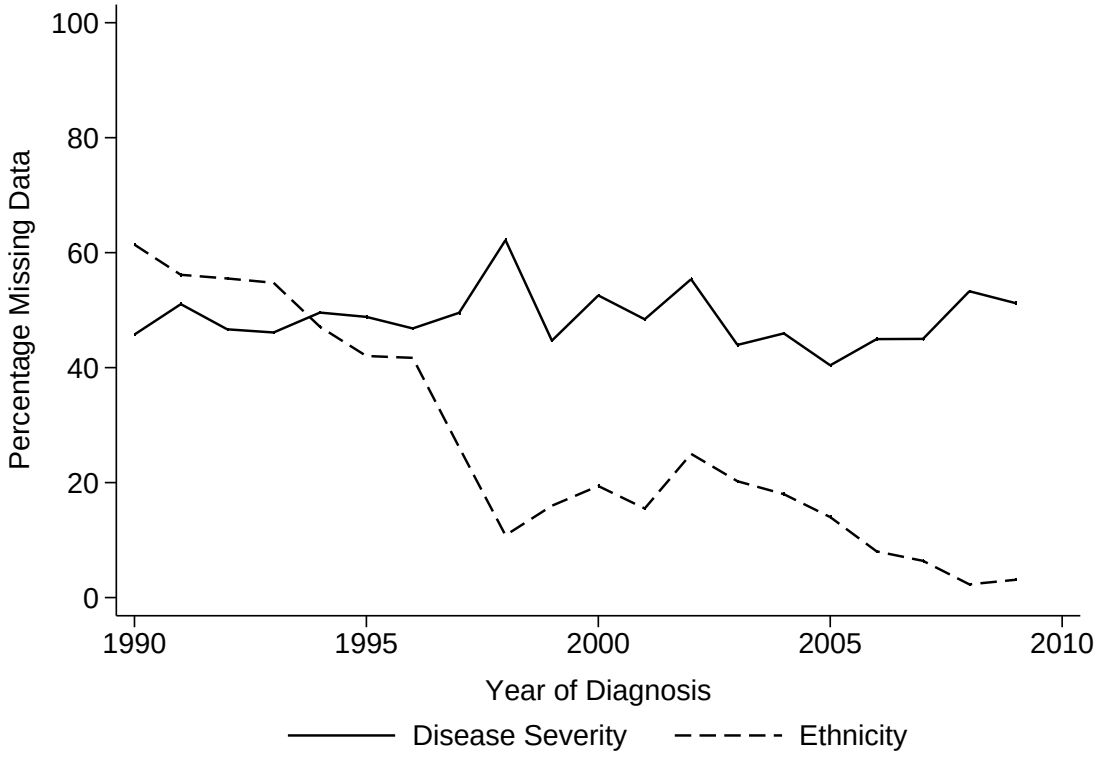


Figure 5.9: Percentage of missing disease severity and ethnicity by year of diagnosis for children and young adults with cancer in Yorkshire, 1990-2009

5.4 Imputation Strategy

The overall level of missing data within this cohort was very high (over 80%) in most diagnostic groups. Based on the results described within this chapter and the discussion of an acceptable level of missingness for imputation (§4.4.1), the diagnostic groups considered for imputation and further analysis were leukaemia, CNS tumours and GCTs as their overall levels of missingness was below 60%. Disease severity and ethnicity were imputed for each of these diagnostic groups, using individual imputation models specific to each diagnostic group due to differences in disease severity measures.

For leukaemia, WBC count was used as a proxy measure for stage at diagnosis as discussed in §5.3.1.1. WBC count is a continuous measure, therefore the imputation model was a linear regression model. However, WBC count cannot be negative, and it is possible that the imputation model could impute negative values. Therefore, the logarithm of WBC count was imputed instead to ensure only positive values of WBC were obtained.

For CNS tumours, missing values for WHO grade were imputed. The WHO classification scheme contained values on a four level scale, from grade I-IV, and therefore the imputation model was based on an ordered logistic regression model.

As described in §5.3.1, GCTs were staged using multiple systems per subgroup. In addition, there was no known staging mechanism for intracranial and intraspinal GCTs, which could be considered as missing by design. The staging mechanisms available for the other subgroups of GCTs all had four main levels, with additional subgroups in each category. For the purposes of imputing these values, only the four main stages were considered and subcategories were not imputed. Despite no known staging mechanism for intracranial and intraspinal GCTs, values of stage were still imputed on a scale of I to IV. The purpose of imputation was to model the distribution of disease severity, rather than to impute a specific value for a single persons disease severity. Therefore, imputing values of disease severity for this subgroup allowed stage to be adjusted for within the final survival analysis for all GCT subgroups. As suggested by von Hippel [292], the main focus of multiple imputation is to obtain a set of values which behave like the complete data if these values had not been missing rather than to obtain a set of values which look like those in the observed data. Using four main levels of stage gave a measure of increasing disease severity, whether including additional subcategories or not.

Detailed considerations of the assumptions required for multiple imputation as well as the missing data mechanisms and the imputation models are given in Chapter 6 alongside the analysis of variation in survival from CYA cancer.

5.5 Conclusion

This chapter provides a detailed scrutiny of the linked Yorkshire register and HES datasets used for analysis in Chapters 6, 7 and 8. Overall, there were 4852 diagnoses of CYA cancer in the former Yorkshire region between 1990 and 2009, of whom 92% (n=4113) successfully linked to at least one inpatient HES record. Specifically, for the sub-cohort used for assessing cardiovascular LEs amongst long term survivors of CYA cancer (results provided in Chapter 8), the linkage rate was 98%. This linkage enabled the capturing of inpatient hospital activity for the vast majority of the cohort. Despite available outpatient records for 3669 cases, these data were found to be unsuitable for identifying specific diagnoses or conditions due to 99% of data being recorded as 'unknown and unspecified causes of morbidity'.

The analysis revealed that both the quantity and quality of data on disease severity within the Yorkshire register was poor, with more than 50% missing data for most tumour groups. Based on the evidence given in §4.4.1, detailed survival analysis for CYA cancers within Yorkshire was performed for CNS tumours, leukaemia and GCTs in which missing data levels were deemed suitable for multiple imputation techniques.

Importantly, the level of missing stage data did not decrease over the 20 year study period. The results within this chapter were used to develop and compile a detailed staging scheme (Appendix I) for all CYA cancers to be implemented into the Yorkshire registers electronic database system. The successful implementation of this scheme is hoped to improve the quality of data through validation of data upon entry, implying that data loss due to recording errors would be minimised. Furthermore, an electronic data collection system which contains detailed information about staging mechanisms may provide a level of additional information to the data collection officer, which was previously inaccessible.

Linkage to a routine dataset, inpatient HES data, has enabled data on ethnicity for CYAs with cancer in Yorkshire to be identified in a more objective manor than the previous method of relying on name analysis programs. Despite a 92% overall linkage rate and improvement in data completeness over time, ethnicity was missing in 26.7% of cases overall and was therefore imputed for use within the analysis (Chapters 6 and 7).

The main results of the thesis are presented in three chapters focusing on variation in cancer survival (Chapter 6), inequalities in disease severity at diagnosis (Chapter 7) and long term effects amongst survivors of cancer (Chapter 8) followed by a detailed discussion of these results in Chapter 9.

Chapter 6

Variation in Cancer Survival

The following publications have arisen from the analysis and results in this chapter:

1 van Laar, M., P.A. McKinney, D.P. Stark, A. Glaser, S.E. Kinsey, I.J. Lewis, S.V. Picton, M. Richards, P.D. Norman, and R.G. Feltbower. (2012). Survival trends of cancer amongst the south Asian and non-south Asian population under 30 years of age in Yorkshire, UK. *Cancer Epidemiology* 36(1): e13-e18.

2 van Laar, M., D. Greenwood, D. Stark, R.G. Feltbower. (2014). Missing data and survival analysis of central nervous system tumours amongst children and adolescents in Yorkshire, UK, 1990-2009. *Pediatric Blood & Cancer*. 61: S293-S293. Conference Abstract.

3 van Laar, M., D. Greenwood, D. Stark, R.G. Feltbower. (2013). Missing data and survival analysis of central nervous system tumours amongst children and young people in Yorkshire, 1990-2009. *European Journal of Cancer*. 60:3-3. Conference Abstract.

4 van Laar, M., D. Greenwood, D. Stark, R.G. Feltbower. (2011). Multiple imputation and survival analysis: an example using cancer registry data. *Journal of Epidemiology and Community Health* 65:A395-A395. Conference Abstract.

6.1 Introduction

The analysis presented in this chapter focuses on survival patterns for CYAs with CNS tumours, leukaemia and GCTs. Missing data for disease severity (WHO grade, WBC count and stage respectively) and ethnicity (white, Asian and other) were imputed using the SMC-FCS method (see §4.4.2). Preliminary results focusing on survival analysis after multiple imputation for CYAs with cancer in Yorkshire were published in 2012 in van Laar et al. [263], however, the work in this chapter includes more detailed analysis for three main tumour groups using the more advanced SMC-FCS imputation method and extends the analysis from 2005 up to 2009. This chapter includes the imputation analysis results and survival analysis results in detail, and therefore focuses on two types of models:

1. Imputation models - including predictor, outcome and auxiliary variables to impute partially observed variables, and
2. Analysis models - including predictor variables after multiple imputation to model overall survival.

The structure of this chapter follows the analysis order as follows; proposed survival analysis models were specified prior to the imputation process (§6.2) to ensure congeniality between the analysis and imputation. Subsequently, §6.3 details the imputation analysis, including missing data mechanism exploration, imputation model specification and imputation results. Although §6.3 includes analysis model results, these were presented for the purpose of comparing CCA, MICE and SMC-FCS methods; full consideration of the survival models is given in §6.4 including 1, 3 and 5-year survival estimates, K-M plots for imputed data, multivariable model selection, and pooled Cox PH model results for the final analysis models. Survival model diagnostics are summarised in §6.5, and sensitivity to the missing data mechanism assumption is explored in §6.6.

6.2 Analysis Model Specification

Partially observed variables were imputed separately for each tumour group and analysed according to tumour specific analysis models, however, the initial model specification was the same for all tumour groups, with the exception of the disease severity variable. Table 6.1 shows the Spearman correlation coefficients for all variables to be included in the analysis (variables were identified in §4.4.3). The correlation coefficients between ethnicity and deprivation were higher than for any other variable pairs, however,

all correlation coefficients were $< |0.3|$, thus providing no concern for collinearity. Therefore, all variables were considered in the model selection process for the final analysis models (§6.4).

Table 6.1: Spearman Correlation Coefficients by Tumour Group

Central Nervous System Tumours							
Variable	WHO Grade	Ethnic Group	Sex	Age	Diagnostic Subgroup	Year of Diagnosis	Deprivation
WHO Grade	1						
Ethnic Group	-0.0061	1					
Sex	-0.0732	0.0179	1				
Age	-0.0538	0.0219	0.0115	1			
Diagnostic Subgroup	0.0257	0.0115	0.0193	-0.0086	1		
Year of Diagnosis	-0.0486	0.1505	0.0357	0.0694	0.0066	1	
Deprivation	-0.0494	0.2947	-0.0116	0.0001	0.0152	0.0452	1

Leukaemia							
Variable	WBC Count	Ethnic Group	Sex	Age	Diagnostic Subgroup	Year of Diagnosis	Deprivation
WBC Count	1						
Ethnic Group	-0.0052	1					
Sex	-0.0147	-0.0105	1				
Age	-0.0628	-0.0205	0.0345	1			
Diagnostic Subgroup	0.1848	-0.0417	0.0718	0.2081	1		
Year of Diagnosis	0.0144	-0.0946	-0.0038	0.1295	0.0841	1	
Deprivation	-0.0606	0.2402	-0.0529	-0.0497	0.0813	-0.0027	1

Germ Cell Tumours							
Variable	Stage	Ethnic Group	Sex	Age	Diagnostic Subgroup	Year of Diagnosis	Deprivation
Stage	1						
Ethnic Group	0.0778	1					
Sex	0.0956	0.0388	1				
Age	-0.1364	-0.0856	-0.1753	1			
Diagnostic Subgroup	-0.1610	0.0056	0.0592	0.0639	1		
Year of Diagnosis	-0.0443	0.1630	-0.0401	-0.0165	0.0376	1	
Deprivation	-0.0056	0.2115	-0.1017	0.1282	-0.0414	0.1163	1

6.2.1 Interactions

As discussed in §4.4.3, interactions were only included in the imputation and analysis models if there was sufficient power (greater than 10 deaths per interaction level) [264, 265, 266]. Interaction terms considered for analysis were age at diagnosis by sex, age at diagnosis by diagnostic subgroup and year of diagnosis by diagnostic subgroup (§4.4.3) and were summarised according to the number of deaths (Table 6.2). The variables to be considered for analyses and the overall events per variable (EPV) are given in Table 6.3.

For CNS tumours, fewer than 10 deaths occurred for 'other specified CNS' and 'other unspecified CNS' subgroups, which were therefore combined to allow for the year by diagnostic subgroup interaction to be considered for analysis. Despite combining these groups, there remained too few events for the age group by diagnostic subgroup interaction which was therefore excluded. The age group by sex interaction was included as the number of events was sufficient. Table 6.3 indicates that the overall EPV for the proposed CNS tumour analysis model was sufficiently large at 14.5. For leukaemia, the number of events was sufficiently large for the age group by sex interaction, however, there was not enough power to study the effects of a year- or age- by diagnostic subgroup interaction for leukaemia. The overall EPV for leukaemia was sufficiently large at 19.8. For GCTs, there were an insufficient number of events for all proposed interaction terms, therefore no interactions were included. Furthermore, the proposed analysis model, despite excluding interactions, had insufficient power with EPV of 4.8. The final analysis models for each diagnostic group were determined in §6.5 and the low EPV value for GCTs was addressed simultaneously.

Table 6.2: Number of deaths by tumour group for proposed interaction terms; diagnostic subgroup by year, diagnostic subgroup by age group and sex by age group. Sensitive data on fewer than 5 deaths were replaced by #

Central Nervous System Tumours			
	Year of Diagnosis	**Age at Diagnosis**	
		0-14 years	**15-29 years**
Diagnostic Subgroup			
Ependymoma	18	#	#
Astrocytoma	144	41	103
Embryonal	63	51	12
Other Gliomas	57	30	27
Other specified CNS	17	10	7
Other unspecified CNS	6	#	#
Sex			
Male	N/A	88	95
Female	N/A	61	61

Leukaemia			
	Year of Diagnosis	**Age at Diagnosis**	
		0-14 years	**15-29 years**
Diagnostic Subgroup			
Lymphoid Leukaemia	140	91	49
Acute Myeloid Leukaemia	114	38	76
Chronic Myeloproliferative disaeases	17	#	#
Meylodysplastic syndrome	#	#	#
Unspecified Leukaemia	#	#	#
Sex			
Male	N/A	80	73
Female	N/A	55	69

Germ Cell Tumours			
	Year of Diagnosis	**Age at Diagnosis**	
		0-14 years	**15-29 years**
Diagnostic Subgroup			
Malignant Gonadal GCTs	37	0	37
Intracranial and Intraspinal GCTs	9	#	#
Malignant extracranial and extragonadal GCTs	8	#	#
Gonadal Carcinomas	#	#	6
Other and unspecified GCTs	#	#	#
Sex			
Male	N/A	#	45
Female	N/A	#	11

Table 6.3: Proposed variables for analysis by tumour group including the number of levels for each variable, the number of dummy variables associated with each variable and the number of events per variable

Central Nervous System Tumours		
Variable	**Levels (N)**	**Dummy Variables (N)**
WHO grade	4	3
Ethnic group	3	2
Sex	2	1
Age at diagnosis	1	1
Diagnostic subgroup	5	5
Year of diagnosis	1	1
Deprivation	1	1
Year by diagnostic subgroup	5	4
Age by sex	4	3
Totals	**N**	
Total dummy variables	21	
Total cases	795	
Total deaths	305	
Events per variable[a]	14.5	

Leukaemia		
Variable	**Levels (N)**	**Dummy Variables (N)**
WBC count	1	1
Ethnic group	3	2
Sex	2	1
Age at diagnosis	1	1
Diagnostic subgroup	5	4
Year of diagnosis	1	1
Deprivation	1	1
Age by sex	4	3
Totals	**N**	
Total dummy variables	11	
Total cases	912	
Total deaths	277	
Events per variable[a]	19.8	

Germ Cell Tumours		
Variable	**Levels (N)**	**Dummy Variables (N)**
Stage	4	3
Ethnic group	3	2
Sex	2	1
Age at diagnosis	1	1
Diagnostic subgroup	5	4
Year of diagnosis	1	1
Deprivation	1	1
Totals	**N**	
Total dummy variables	13	
Total cases	846	
Total deaths	62	
Events per variable[a]	4.8	

[a]Events were deaths

6.3 Imputation Analysis

6.3.1 Missing Data Mechanism Assessment

There were significant differences between survival curves for cases with observed and missing disease severity and ethnicity for CNS tumours and leukaemia but not for GCTs (Figure 6.1).

For CNS tumours, grade was 7% less likely to be missing per single yearly increase in age at diagnosis and 95% less likely to be missing for those who received surgery compared to those who did not (Table 6.4). Cases diagnosed later in the study period as well as those who received chemotherapy were 10% and 3-fold more likely to have missing grade respectively. Grade was more likely to be missing for those with other gliomas and unspecified intracranial and intraspinal tumours compared to astrocytomas. Missing ethnicity was 44% more likely for males compared to females, however, missing ethnicity became less likely over the study period. Furthermore, those who received surgery and radiotherapy were less likely to have missing data on ethnicity.

For leukaemia, WBC count was more likely to be missing for older cases of leukaemia and those who had received radiotherapy. WBC count became less likely to be missing over the study period and was less likely to be missing for those who had relapsed or received chemotherapy. Missingness of ethnicity was also less likely over the study period as well as for those who received chemotherapy or radiotherapy. CYAs with AML were 80% more likely to have missing ethnicity data compared to those with ALL.

For GCTs, missingness of stage was 4% less likely to be missing per single yearly increase in age at diagnosis. Cases diagnosed with extracranial & extragonadal GCTs and gonadal carcinomas were over 3.5- and 3- times respectively more likely to have missing stage compared to those with gonadal GCTs. Cases who received any of the three treatment modalities (chemotherapy, radiotherapy or surgery) were all significantly less likely to have missing data on stage. Ethnicity was less likely to be missing by an average of 10% per year over the study period. There were no differences in the missingness of ethnicity between diagnostic subgroups, however, cases who experienced a relapse were less likely to have missing data on ethnicity.

For CNS tumours and leukaemias, it was evident from K-M survival curves that the data were not MCAR, due to evidence of a systematic difference in the outcome between missing and observed cases. For GCTs, although survival did not differ overall between those with missing and observed stage or ethnicity, there were systematic differences in missingness according to other clinical and demographic variables as described above and

therefore the MCAR assumption could also not be made for GCTs. All data were assumed to MAR, however, a sensitivity analysis to this assumption is presented in §6.6.1.

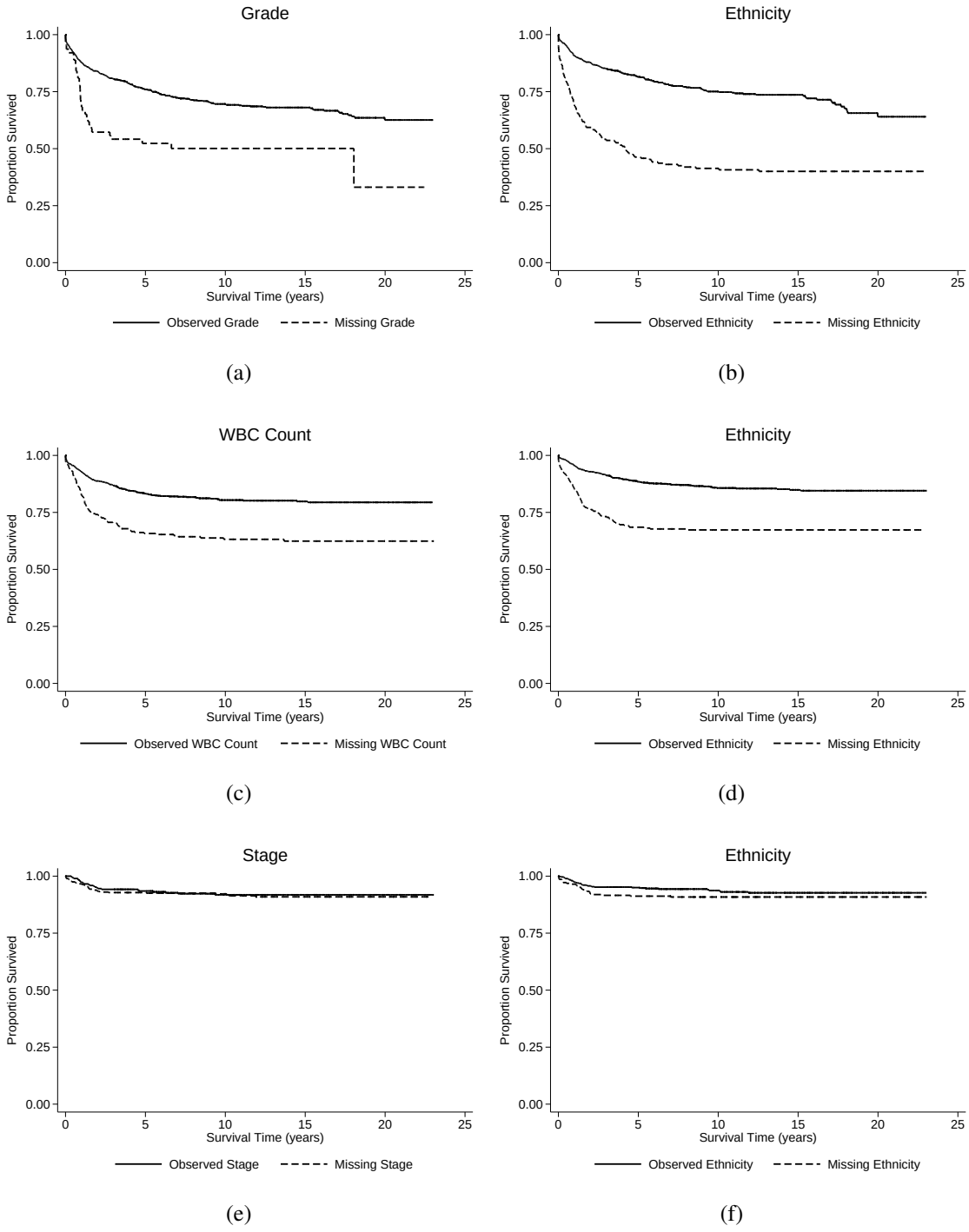


Figure 6.1: Kaplan-Meier curves for central nervous system (CNS) tumours by observed and missing grade (a) and ethnicity (b), for leukaemia by observed and missing white blood cell (WBC) count (c) and ethnicity (d) and for germ cell tumours by observed and missing stage (e) and ethnicity (f)

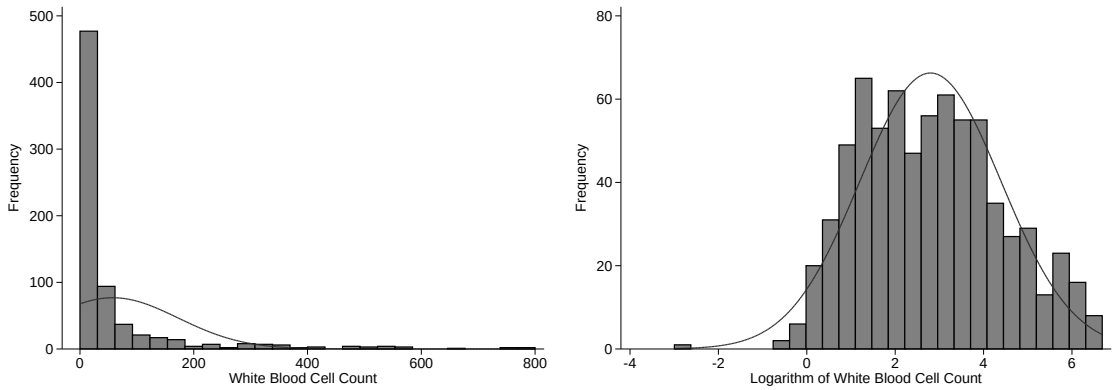
Table 6.4: Predictors of missingness for disease severity (WHO grade, white blood cell (WBC) count and stage) and ethnicity amongst children and young adults by tumour group

Central Nervous System Tumours		
Variable	**Odds Ratios**[a]	
	WHO Grade	**Ethnicity**
Age (years)	0.93**	1.02
Sex (Female vs. Male)	1.31	0.66*
Year of Diagnosis	1.10**	0.77**
Deprivation Score[b]	1.05	0.95
Diagnostic Subgroup		
Astrocytomas	1	1
Ependymomas	1.60	0.94
Intracranial and intraspinal embryonal tumours	1.00	1.00
Other gliomas	21.58**	1.15
Other specified intracranial and intraspinal neoplasms	2.92	1.07
Unspecified intracranial and intraspinal neoplasms	88.49**	1.94
Relapse (Yes vs. No)	0.42	0.79
Surgery (Yes vs. No)	0.05**	0.43**
Chemotherapy (Yes vs. No)	3.05**	1.01
Radiotherapy (Yes vs. No)	0.71	0.57*
Leukaemia		
	WBC Count	**Ethnicity**
Age (years)	1.22**	1.02
Sex (Female vs. Male)	0.87	0.71
Year of Diagnosis	0.87**	0.80**
Deprivation Score[b]	0.99	1.00
Diagnostic Subgroup		
Lymphoid leukaemias	1	1
Acute myeloid leukaemias	1.40	1.81**
Chronic myeloproliferative diseases	1.66	0.69
Myelodysplastic syndrome	0.38	3.35
Unspecified and other leukaemias	2.43	2.14
Relapse (Yes vs. No)	0.54*	0.83
Surgery (Yes vs. No)	1.23	0.81
Chemotherapy (Yes vs. No)	0.09**	0.50*
Radiotherapy (Yes vs. No)	2.69**	0.30**
Germ Cell Tumours		
	Stage	**Ethnicity**
Age (years)	0.96 **	1.00
Sex (Female vs. Male)	1.48	0.64
Year of Diagnosis	1.02	0.90 **
Deprivation Score[b]	1.01	1.00
Diagnostic Subgroup		
Malignant Gonadal GCTs	1	1
Malignant extracranial and extragonadal GCTs	3.65 **	0.84
Gonadal carcinomas	3.02 *	1.03
Other and unspecified malignant gonadal tumours	4.05	0.85
Relapse (Yes vs. No)	0.61	0.45 *
Surgery (Yes vs. No)	0.09 **	1.05
Chemotherapy (Yes vs. No)	0.58 **	0.78
Radiotherapy (Yes vs. No)	0.47 **	1.06

[a]Odds ratios were obtained from a multivariable logistic regression models for missing data by disease severity (WHO Grade, white blood cell (WBC) count and stage) and ethnicity
[b]2007 Index of Multiple Deprivation (IMD)
*Significant at 5% level, **Significant at 1% level.

6.3.2 Imputation Model Specification

WBC count was the only continuous variable to be imputed, and it was therefore checked for normality. Figure 6.2 shows that WBC count was highly positively skewed, and was transformed to an approximately normal distribution using a log transformation. The log transformation was also convenient in this case to restrict imputations to positive values, as negative values of WBC count are not possible. Table 6.5 contains the imputation model specifications for each tumour group based on the predictors for the proposed analysis model, outcome variables and interactions which were based on clinical relevance in addition to study power. Furthermore, the auxiliary variables identified within the previous section as predictors of missing values were also included. The number of imputations were $m = 40$ for CNS tumours and leukaemia, and $m = 60$ for GCTs, according to the overall level of missing data (justified in §3.5.4). Suitability of the number of imputations in each case was checked by assessing MC errors (see §6.3.3). The number of cycles used for each imputation model was 20 after checking for convergence using 500 cycles for $m = 1$ (Figure 6.3). There were no clear patterns of non-convergence, thus the default of 20 cycles was used for all analysis. The primary imputation method was SMC-FCS, however, CCA and MICE analyses are also presented for comparison.

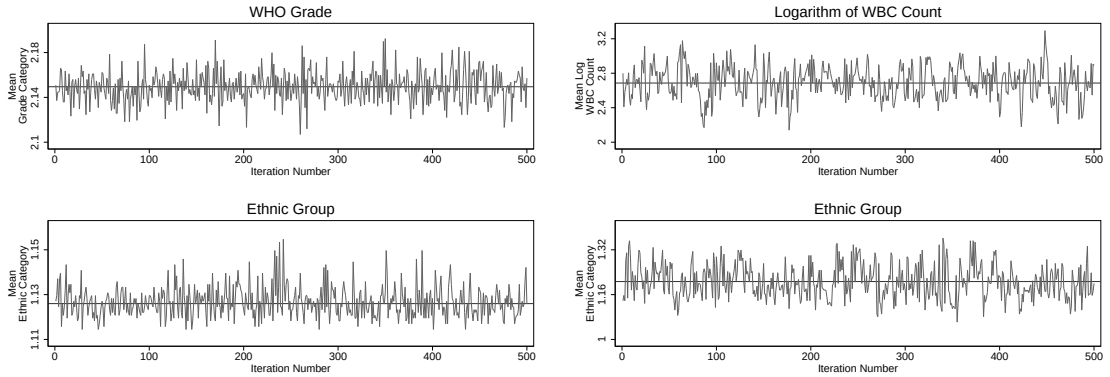


(a) (b)

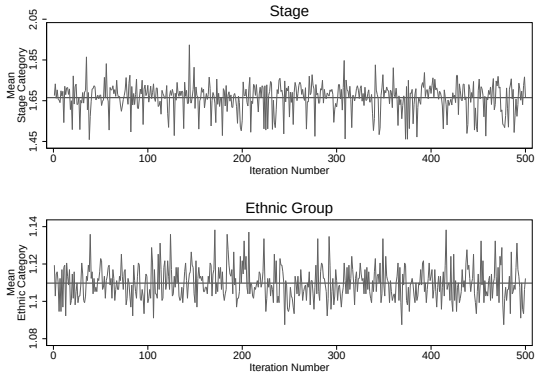
Figure 6.2: Histogram and Normal density curve for white blood cell count (a) and logarithm of white blood cell count (b)

Table 6.5: Imputation model specifications showing partially observed variables, fully observed predictors, interaction terms of interest, outcome variables and auxiliary variables by tumour type

	Partially Observed Variables	**Predictors**	**Interactions**	**Outcome Variables**	**Auxiliary Variables**
Central Nervous System Tumours	Grade (Ordered Logistic) Ethnic Group (Multinomial Logistic)	Age (Continuous) Sex (Binary) Deprivation (Continuous) Year (Continuous) Diagnostic Subgroup (Categorical)	Age by Sex Year by Diagnostic Subgroup	Nelson-Aalen Estimates Censoring Indicator	Chemotherapy (Binary) Surgery (Binary) Radiotherapy (Binary)
Leukaemia	Log WBC Count (Linear Regression) Ethnic Group (Multinomial Logistic)	Age (Continuous) Sex (Binary) Deprivation (Continuous) Year (Continuous) Diagnostic Subgroup (Categorical)	Age by Sex	Nelson-Aalen Estimates Censoring Indicator	Chemotherapy (Binary) Relapse (Binary) Radiotherapy (Binary)
Germ Cell Tumours	Stage (Ordered Logistic) Ethnic Group (Multinomial Logistic)	Age (Continuous) Sex (Binary) Deprivation (Continuous) Year (Continuous) Diagnostic Subgroup (Categorical)		Nelson-Aalen Estimates Censoring Indicator	Chemotherapy (Binary) Surgery (Binary) Radiotherapy (Binary) Relapse (Binary)



(a) (b)



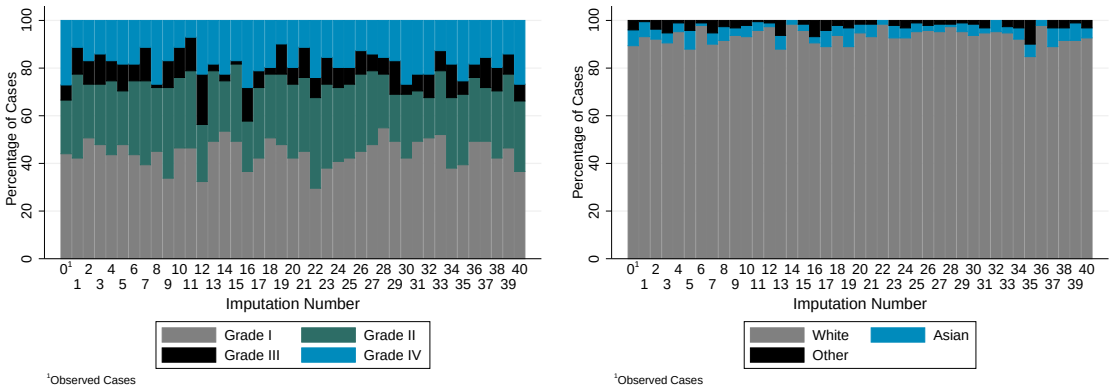
(c)

Figure 6.3: Trace plots of the mean value of partially observed variables over 500 iterations (imputation cycles) based on one imputation for central nervous system tumours (a), leukaemia (b) and germ cell tumours (c).

6.3.3 Imputation Results

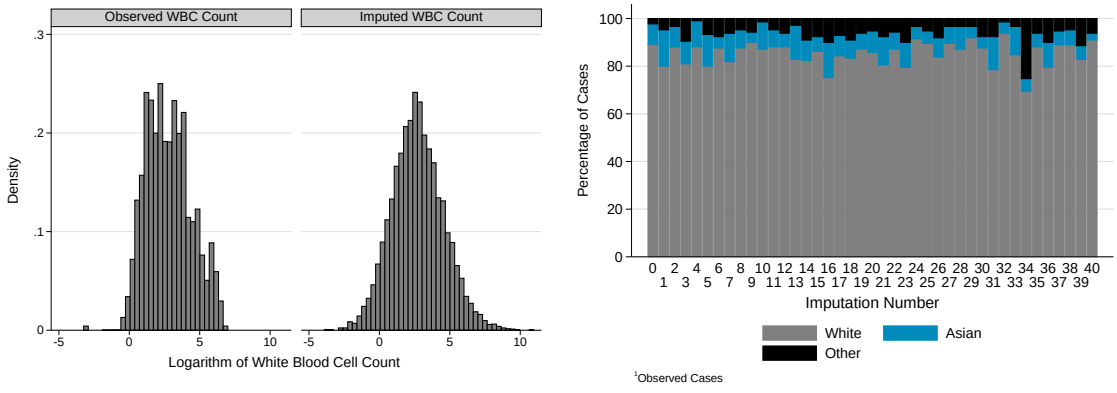
Imputations were generally stable with no obvious outliers, showing only expected variation between imputations for each tumour group (Figure 6.4 for CNS tumours, Figure 6.5 for leukaemia and Figure 6.6 for GCTs). Imputations of stage for GCTs varied more than for other imputation models by imputation number, this additional uncertainty between imputations was a reflection of the larger amount of missing data for this variable. Amongst CNS tumours, there was a slightly higher proportion of grade II tumours and lower proportion of grade IV tumours amongst the imputed cases compared to the observed cases, however, the overall distribution of completed data matched the observed data very closely (Figure 6.7). For leukaemia, it was evident that the complex pattern at the centre of the logarithm of WBC count distribution was not fully replicated by the imputation model, which had a normal distribution (Figure 6.7). Nonetheless, the

distribution of the imputed values was close to that of the observed values outside of this central area, and the observed and completed distributions were very similar. In addition, the imputation model for leukaemia predicted more cases into to the 'other' than 'white' ethnic category compared to the observed data, however the observed and completed data ethnic distributions were very similar. For GCTs, there were slightly more stage III and IV imputed values and fewer stage II values compared to observed data. The distribution of ethnic groups between the observed, imputed and completed data for GCTs did not vary (Figure 6.7). MC errors for each HR and associated P-value were sufficiently small so that neither the direction of significant variables nor the significance level of any variable were affected at the extreme boundaries of the error for all three tumour groups (Table 6.6).



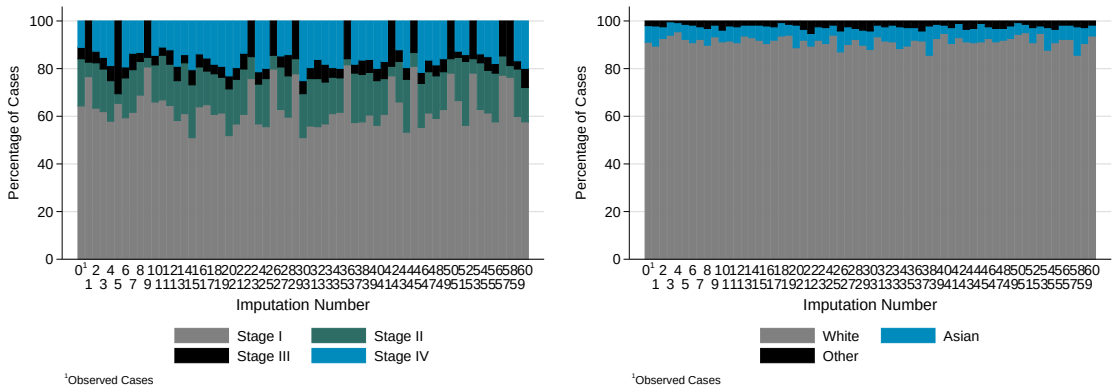
(a) (b)

Figure 6.4: The percentage of central nervous system tumour cases by grade (a) and ethnicity (b) within the observed data and each imputed dataset



(a) (b)

Figure 6.5: The distribution of the logarithm of white blood cell (WBC) count for leukaemia cases (a) and the percentage of leukaemia cases by ethnicity (b) for the observed data and each imputed dataset



(a) (b)

Figure 6.6: The percentage of germ cell tumour cases by stage (a) and ethnicity (b) within the observed data and each imputed dataset

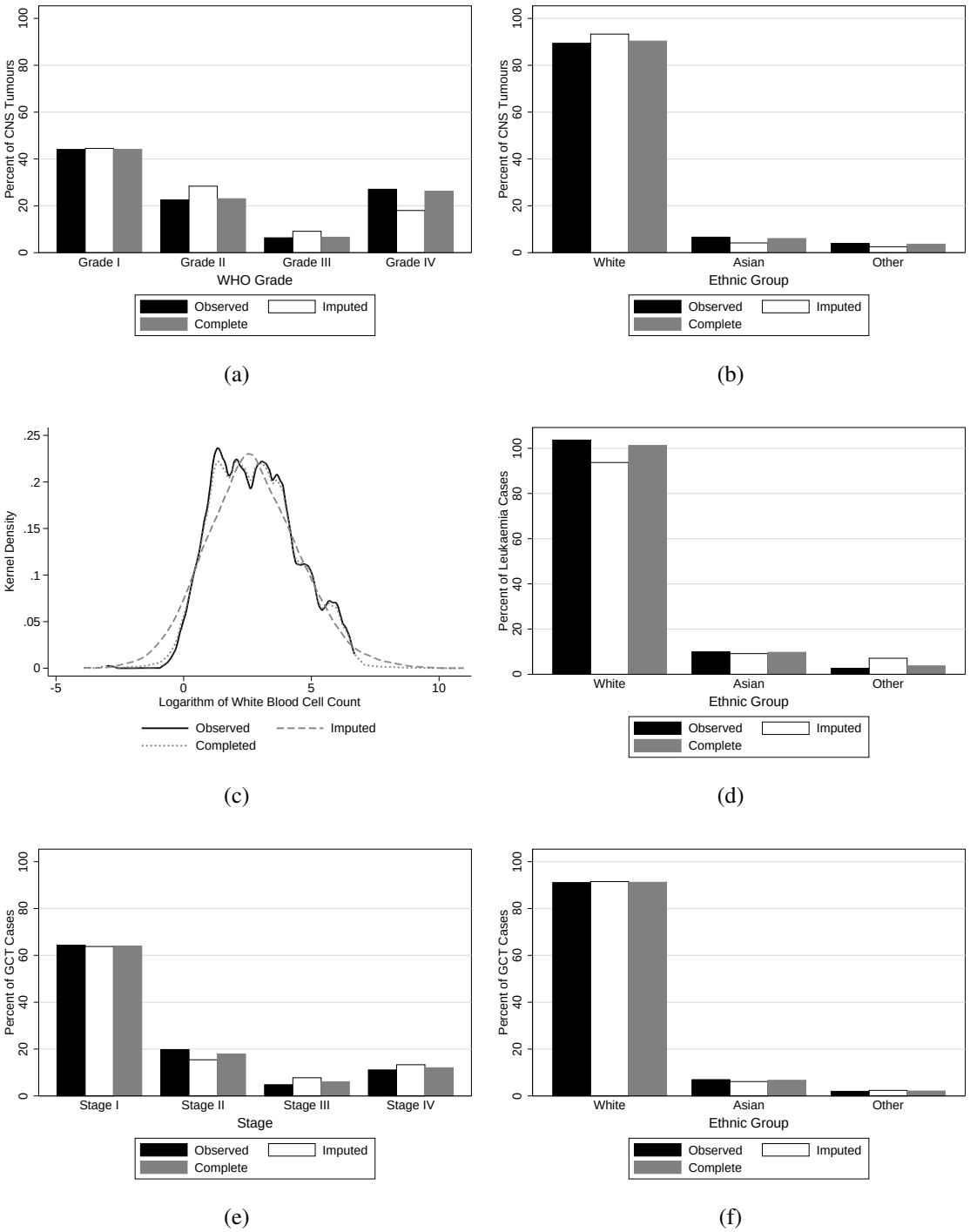


Figure 6.7: Distribution of observed, imputed and completed cases for central nervous system tumours by WHO grade (a) and ethnicity (b), leukaemia by logarithm of white blood cell (WBC) count (c) and ethnicity (d) and germ cell tumours by stage (e) and ethnicity (f)

Table 6.6: Monte Carlo (MC) errors of pooled hazard ratios (HRs) and $P-$values obtained from survival analysis models after multiple imputation for central nervous system tumours, leukaemia and germ cell tumours[a]

Variable	HR	MC Error	HR ± MC Error	P-value	MC Error	P-value ± MC Error
Central nervous system tumours						
Age	1.01	0.0002	1.0141-1.0146	0.036	0.0030	0.0330-0.0690
Sex (Female vs. Male)	0.84	0.0042	0.8322-0.8406	0.149	0.0100	0.1390-0.2880
Year	0.96	0.0004	0.9632-0.9639	0.001	< 0.0001	0.0010-0.0020
Deprivation	1.00	0.0002	1.0002-1.0005	0.912	0.0400	0.8720-1.7840
Diagnostic subgroup						
Astrocytoma	1					
Ependymoma	0.60	0.0035	0.5921-0.5990	0.047	0.0020	0.0450-0.0920
Embryonal	0.38	0.0033	0.3739-0.3805	< 0.001	< 0.0001	< 0.0001-< 0.0001
Other gliomas	1.51	0.0212	1.4929-1.5353	0.023	0.0070	0.0160-0.0390
Other CNS	0.80	0.0044	0.7916-0.8005	0.340	0.0120	0.3280-0.6680
Ethnicity						
White	1					
Asian	1.43	0.0358	1.3897-1.4614	0.216	0.0380	0.1780-0.3940
Other	2.22	0.0759	2.1443-2.2962	0.032	0.0100	0.0220-0.0540
Grade						
I	1					
II	3.64	0.0403	3.6027-3.6833	< 0.001	< 0.0001	< 0.0001-< 0.0001
III	6.31	0.0815	6.2259-6.3888	< 0.001	< 0.0001	< 0.0001-< 0.0001
IV	11.02	0.1194	10.903-11.142	< 0.001	< 0.0001	< 0.0001-< 0.0001
Leukaemia						
Age	1.04	0.0003	1.0428-1.0435	< 0.001	< 0.0001	< 0.0001-< 0.0001
Sex (Female vs. Male)	0.97	0.0041	0.9664-0.9745	0.811	0.0270	0.7840-0.8380
Year	0.94	0.0004	0.9400-0.9409	< 0.001	< 0.0001	< 0.0001-< 0.0001
Deprivation	1.00	0.0001	1.0034-1.0037	0.331	0.0180	0.3131-0.3491
Diagnostic subgroup[b]						
Ia	1					
Ib	1.79	0.0099	1.7843-1.8041	< 0.001	< 0.0001	0.0001-0.0001
Ic	0.65	0.0087	0.6424-0.6597	0.216	0.0050	0.2112-0.2212
Id	0.33	0.0075	0.3255-0.3406	0.280	0.0090	0.2713-0.2893
Ie	2.34	0.0420	2.2994-2.3835	0.067	0.0060	0.0611-0.0731
Ethnicity						
White	1					
Asian	1.25	0.0331	1.2157-1.2820	0.438	0.0520	0.3861-0.4901
Other	1.52	0.0733	1.4464-1.5929	0.347	0.0640	0.2828-0.4108
Log WBC Count	1.21	0.0060	1.2077-1.2197	< 0.001	< 0.0001	< 0.0001-< 0.0001
Germ Cell Tumours						
Age	1.04	0.0017	1.0355-1.0389	0.158	0.0190	0.1390-0.1770
Sex (Female vs. Male)	0.64	0.0251	0.6178-0.6680	0.408	0.0420	0.3660-0.4500
Year	1.01	0.0017	1.0063-1.0097	0.775	0.0450	0.7300-0.8200
Deprivation	1.01	0.0005	1.0069-1.0080	0.376	0.0370	0.3390-0.4130
Diagnostic subgroup[c]						
Xc						
Xa	18.24	1.6397	16.5976-19.8771	0.001	0.0010	< 0.0001-0.0020
Xb	1.72	0.0681	1.6489-1.7852	0.334	0.0380	0.2960-0.3720
Xd	13.31	0.9252	12.3875-14.2378	0.002	0.0010	0.0010-0.0030
Xe	15.20	1.2187	13.9813-16.4186	0.009	0.0040	0.0050-0.0130
Ethnicity						
White						
Asian	1.55	0.0662	1.4802-1.6126	0.441	0.0470	0.3940-0.4880
Other	2.65	0.2175	2.4294-2.8645	0.296	0.0440	0.2520-0.3400
Stage						
I						
II	5.27	0.2869	4.9781-5.5520	0.014	0.0040	0.0100-0.0180
III	18.87	1.0537	17.8181-19.9254	< 0.001	< 0.001	< 0.0001-< 0.0001
IV	32.77	2.2626	30.5098-35.0349	< 0.001	< 0.001	< 0.0001-< 0.0001

[a] Full model results including confidence intervals and standard errors are provided in Tables 6.7, 6.8 and 6.9

[b] Ia - Lymphoid leukaemias; Ib - Acute myeloid leukaemias; Ic - Chronic myeloproliferative diseases; Id - Myelodysplastic syndrome and other myeloproliferative diseases; Ie - Unspecified and other specified leukaemias

[c] Xa - Intracranial and Intraspinal GCTs; Xb - Malignant extracranial and extragonadal GCTs; Xc - Malignant Gonadal GCTs; Xd - Gonadal carcinomas; Xe - Other and unspecified malignant gonadal tumours

The CCA, MICE and SMC-FCS methods were compared by assessing estimates from the Cox PH model for CNS tumours, leukaemia and GCTs (Tables 6.7, 6.8 and 6.9). A basic analysis model including only main effects was used to compare results between CCA and imputation methods; an assessment of the most appropriate analysis model (including interactions and testing for linearity) is provided in the survival analysis section (§6.4). For GCTs, the proposed imputation model did not work using the SMC-FCS method, therefore imputations generated using the MICE method were used for further analysis of GCTs. A detailed description providing reasons for this is provided in the following subsection (§6.3.3.1).

6.3.3.1 Exploration of Germ Cell Tumour Substantive Model Compatible Fully Conditional Specification (SMC-FCS) Analysis

There were several complications when imputing and analysing stage and ethnic group for GCTs. The SMC-FCS imputation method for GCTs did not run using the proposed imputation model specification (Table 6.5), and the below paragraph contains a detailed description exploring the possible reasons. The SMC-FCS method required the imputation model and the analysis model to be specified at the imputation stage to ensure an imputation model that is compatible with the analysis model can be derived. This meant that any problems could arise in either the imputation or analysis part, and identifying the model which was not working was not immediately obvious from the Stata output. In terms of potential problems with the imputation model, the GCT analysis differed from other tumour groups due to the large amount of missing data (60% overall). Furthermore, as there was no known staging mechanism for the intracranial and intraspinal GCT subgroup, this subgroup had a 100% missing stage data. In addition, perfect prediction arose for some variables as all observed cases of intracranial and intraspinal GCTs were of white ethnicity, and all observed cases diagnosed between 5 and 9 years of age were of white ethnicity and had stage III tumours. In addition, there were only 9 cases of unspecified malignant gonadal tumours, 2 of which had a non-missing stage which were both stage I. In terms of the problems encountered for the survival analysis of GCTs, there was a much higher overall 5-year survival rate for GCTs (94% 5-year survival) compared to CNS tumours (59% 5-year survival) and leukaemia (72% 5-year survival). This meant that the EPV value for the proposed analysis model was too low (EPV = 4.7) to provide the required power for a Cox PH model. Initially, SMC-FCS using the full proposed imputation and analysis models were implemented, resulting in the following error messages “valid imputations have not been generated” and “convergence was not achieved”. The default number of imputation cycles (c=20) and the default rejection sampling limit (rjlimit=5000) were increased by 50 and 1000 respectively at a time, however, even at c=1000 and rjlimit=500000 respectively, convergence and valid imputations for each subject were not achieved. Subsequently, the imputation model was constructed in a stepwise

manner by adding a single variable at a time whilst keeping the basic analysis model constant (including only stage and ethnic group as explanatory variables). The algorithm converged and generated valid imputations for each subject based on a model including all proposed imputation variables (age, sex, year of diagnosis, deprivation, relapse status, surgery status, chemotherapy status and radiotherapy status) except diagnostic subgroup. Subsequently, the analysis model was constructed in a stepwise manner whilst maintaining the aforementioned imputation model constant throughout. The algorithm produced successful imputations for an analysis model containing all variables in the proposed analysis model (age, sex, year of diagnosis, stage, ethnicity and diagnostic subgroup) except for deprivation. The diagnostic subgroup variable had two main features (100% missing stage for one subgroup and perfect prediction) which were likely to lead to the SMC-FCS method failing to converge and produce valid imputations whilst including this variable. Being unable to include diagnostic subgroup in the imputation model led to uncertainty over the imputation validity and the strength of the MAR assumption as diagnostic subgroup was a significant predictor of missingness of stage as shown in Table 6.4. For this reason, further analysis of GCTs were performed using imputations produced by MICE for which the proposed imputation model was successfully implemented. It was unclear why inclusion of the deprivation variable in the analysis model caused the SMC-FCS algorithm to fail, however, SMC-FCS required more power than MICE due to its attempt to find a complex compatible imputation and analysis model, therefore it was likely to be a combination of the low EPV value for GCT tumours and the high level of missing data which were causing a problem. The remainder of §6.3.3 focuses on comparing the CCA, MICE and SMC-FCS methods for all three tumour groups, however, for GCTs, the MICE and SMC-FCS results compared were based on different imputation models as described here.

Tables 6.7, 6.8 and 6.9 contain multivariable survival estimates based on CCA, MICE and SMC-FCS methods. The results show, for the first time in CYA cancer survival, that ignoring missingness can lead to serious bias and incorrect inferences as discussed in detail for each tumour group below. The final analysis models and clinical interpretation of these are discussed later in §6.4.

For the CNS tumour analysis, despite very similar HRs, both the MICE and SMC-FCS analysis resulted in significant differences in survival for age, other ethnicity compared to white ethnicity as well as ependymoma and other gliomas compared to astrocytmas which were not observed amongst the CCA (Table 6.7). This showed that without the use of an imputation technique, the study was underpowered to detect these significant differences. The HR for year of diagnosis reversed direction, moving from an increased risk of death in the CCA (HR=1.02, 95% CI 0.99-1.05, P=0.294), compared to a decreased risk of death in the MICE and SMC-FCS analysis (HR=0.96, 95% CI 0.94-0.98, P=0.001 for MICE and SMC-FCS). Furthermore, the CCA indicated a significant difference in survival for the other CNS tumour subgroup compared to astrocytomas (HR=0.41, 95% CI 0.19-

0.86, P=0.019), whereas in the MICE and SMC-FCS methods there was no difference in the survival of this subgroup compared to astrocytomas (HR=0.79, 95% CI 0.50-1.27, P=0.335 and HR=0.80, 95% CI 0.50-1.27, P =0.340 for MICE and SMC-FCS respectively). The missing data, if ignored, would have caused incorrect inferences to be drawn about the 'other CNS tumour' subgroup, which appeared to have poorer survival rates in the CCA, however, survival was similar to that of astrocytomas after imputation. The availability of grade data for diagnoses which are not well defined, such as those in the 'other CNS tumour' group was poor (discussed in §5.3.1.3). However, imputation allowed for all information related to cases within this subgroup to be included within the analysis, regardless of whether grade was missing or not, thereby providing more accurate results. These important differences between CCA and imputation techniques indicate how ignoring missingness can cause bias in terms of the direction, size and significance of survival effects.

The significant difference in survival between embryonal tumours compared to astrocytomas as well as between grades II, III and IV compared to grade I were evident in all three analyses, and their effect sizes were comparable. The standard error was either reduced or of similar magnitude for all variables in the MICE and SMC-FCS analyses compared to the CCA except for those associated with ethnicity. Overall, imputation led to an increase in efficiency resulting from including all available data within the analysis.

For the leukaemia, the HR for year of diagnosis was not significant (HR=1.00, 95% CI 0.96-1.04, P=0.999) in the CCA, compared to a decreased risk of death observed using MICE and SMC-FCS methods (HR=0.94, 95% CI 0.92-0.96, P <0.001 for MICE and SMC-FCS) (Table 6.8). A similar change between HRs in the CCA and imputation analysis for year of diagnosis was also observed in the CNS tumour analysis. Estimates for diagnostic subgroup Id - 'Myelodysplastic syndrome and other myeloproliferative disease' were not available in the CCA as this group was small (n=6) and contained no deaths when using listwise deletion. The use of imputation techniques allowed estimates of survival for this subgroup to be obtained, although CIs around these estimates remained large. MICE and SMC-FCS resulted in an increase in HRs and borderline significant evidence of increased risk of death of subgroup Ie -'Unspecified and other specified leukaemias' compared to Ia - 'lymphoid leukaemias' compared to CCA.

HRs under CCA, MICE and SMC-FCS were very similar for those variables which displayed significant effects in the CCA, including HRs for age, diagnostic subgroup Ib and log WBC count. The HRs for other ethnicity compared to white ethnicity gave conflicting information between the MICE and SMC-FCS results in terms of the direction of effect (HR=0.90 and HR=1.52 respectively). However, both effects were not significant (P=0.814 and P=0.347 respectively), indicating that there was no certainty of

the direction of effect for this estimate. Furthermore, the percentage of cases and deaths within this category were very small (3.2% and 5.1% respectively) thereby highlighting the uncertainty and unstable nature of these results.

For the GCT analysis, the CCA showed a worrying significant increase in the risk of death by 18% on average per year of diagnosis (HR=1.18, 95% CI 1.04-1.33, P=0.008), however, after imputation using MICE and SMC-FCS, no significant increase in the risk of death was observed (HR=1.01, 95% CI 0.95-1.06, P=0.775 and HR=1.02, 95% CI 0.97-1.08, P=0.448 for MICE and SMC-FCS respectively) (Table 6.9). Estimates for diagnostic subgroup Xa - 'Intracranial and intraspinal GCTs' were not available as stage was missing in 100% of cases, causing the variable to be automatically excluded from the CCA through listwise deletion. The MICE and SMC-FCS methods both showed an increased risk of death for this subgroup compared to Xc - 'Malignant gonadal GCTs' (HR=18.24, 95% CI 3.33-99.99, P=0.001 and HR=7.51, 95% CI 2.70-20.88, P <0.001 for MICE and SMC-FCS respectively). In addition, estimates for Xe - 'Other and unspecified malignant gonadal GCTs' were not available in the CCA as there were no observed deaths in this subgroup after listwise deletion. MICE and SMC-FCS again showed a significant increased risk of death for this diagnostic subgroup compared to Xc - 'Malignant gonadal GCTs' (HR=15.20, 95% CI 1.97-117.09, P=0.009 and HR=7.24, 95% CI 1.26-41.68, P=0.027 for MICE and SMC-FCS respectively). For Xb - 'Malignant extracranial and extragonadal GCTs' CCA showed a significant 7-fold increased risk of death compared to Xc - 'Malignant gonadal GCTs', however, this effect was much smaller and non-significant (HR=1.72, 95% CI 0.57-5.15, P=0.334 and HR=2.27, 95% CI 0.82-6.28, P=0.113) using the MICE and SMC-FCS methods respectively. Although the increased risk of death for Xd - 'Gonadal carcinomas' was evident in CCA, MICE and SMC-FCS, the standard error of this estimate reduced from 85.14 to 17.07 and 6.45 for each method respectively. No significant effect of ethnicity on survival was observed in the CCA or MICE despite increased HRs for Asian and other ethnic groups compared to white ethnicity. However, there was a significant 5-fold increased risk of death for the other ethnic group compared to white ethnicity after using the SMC-FCS method. The risk of death increased significantly for stage II, III and IV tumours compared to stage I tumours across all analyses. The standard errors of these estimates reduced substantially by approximately 4- and 5-fold for MICE and SMC-FCS respectively. The SMC-FCS models differ from the MICE models as the former method excluded diagnostic subgroup from the imputation model (see §6.3.3.1). Standard errors for SMC-FCS were smaller on average than MICE, however, due to the inability to specify the proposed imputation model, the validity of the SMC-FCS results was uncertain.

Overall, important differences between CCA and each of the imputation methods were observed, however results between MICE and SMC-FCS were very similar. HRs were

identical or within < 0.1 of each other for each of the fully observed variables between the latter two methods for CNS tumours and leukaemia. For the partially observed variables, there was a slightly larger difference between HRs when comparing MICE and SMC-FCS for CNS tumours and leukaemia, however, the largest difference was still very small ($|0.11|$). For GCTs, the differences were larger (up to 10.7 difference in HRs of fully observed variables and up to 4.1 difference in HRs for partially observed variables), however, this was expected as the imputation models differed for each method. There was no change in the significance for any of the variables between the MICE and SMC-FCS methods. Standard errors were very similar for fully observed as well as partially observed variables between the MICE and SMC-FCS methods, except for the significant affect of other ethnicity compared to white ethnicity in the SMC-FCS analysis for GCTs. There was no consistent pattern in terms of the direction of the small differences observed between MICE and SMC-FCS, although on average, effect sizes were smaller using the SMC-FCS method compared to MICE (i.e. HRs tended to be closer to 1 using the SMC-FCS method compared to MICE), however this was not the case for all estimates.

Table 6.7: Cox proportional hazards model for central nervous system tumour survival: Complete Case Analysis (CCA) and two multiple imputation analyses (multiple imputation by chained equations (MICE) and substantive model compatible fully conditional specification (SMC-FCS))

	CCA (N=553)				MICE (N=795)				SMC-FCS (N=795)			
Variable	**HR**	**95% CI**	**SE**	P**-value**	**HR**	**95% CI**	**SE**	P**-value**	**HR**	**95% CI**	**SE**	P**-value**
Age	1.01	0.99-1.03	0.01	0.221	1.01	1.00-1.03	0.01	0.043	1.01	1.00-1.03	0.01	0.036
Sex												
Male	1				1				1			
Female	0.80	0.59-1.09	0.13	0.155	0.84	0.66-1.06	0.10	0.142	0.84	0.66-1.07	0.10	0.149
Year	1.02	0.99-1.05	0.02	0.294	0.96	0.94-0.98	0.01	0.001	0.96	0.94-0.98	0.01	0.001
Deprivation	1.00	0.99-1.01	0.00	0.507	1.00	0.99-1.01	0.00	0.897	1.00	0.99-1.01	0.00	0.912
Diag Subgroup												
Astrocytoma	1				1				1			
Ependymoma	0.59	0.30-1.14	0.20	0.117	0.60	0.36-1.00	0.16	0.051	0.60	0.36-0.99	0.16	0.047
Embryonal	0.30	0.18-0.49	0.08	< 0.001	0.38	0.25-0.56	0.08	< 0.001	0.38	0.25-0.56	0.08	< 0.001
Other gliomas	1.52	0.94-2.46	0.37	0.086	1.57	1.12-2.21	0.27	0.009	1.51	1.06-2.17	0.28	0.023
Other CNS	0.41	0.19-0.86	0.16	0.019	0.79	0.50-1.27	0.19	0.335	0.80	0.50-1.27	0.19	0.340
Ethnicity												
White	1				1				1			
Asian	1.33	0.76-2.33	0.38	0.324	1.47	0.83-2.60	0.43	0.184	1.43	0.81-2.50	0.41	0.216
Other	2.04	0.94-4.40	0.80	0.071	2.14	1.05-4.37	0.78	0.036	2.22	1.07-4.60	0.82	0.032
Grade												
I	1				1				1			
II	3.66	2.24-5.98	0.92	< 0.001	3.66	2.49-5.36	0.71	< 0.001	3.64	2.46-5.39	0.73	< 0.001
III	6.93	3.75-12.78	2.17	< 0.001	6.42	3.97-10.37	1.57	< 0.001	6.31	3.89-10.22	1.55	< 0.001
IV	13.06	7.72-22.07	3.50	< 0.001	11.11	7.41-16.66	2.30	< 0.001	11.02	7.35-16.53	2.28	< 0.001

Table 6.8: Cox proportional hazards model for leukaemia survival: Complete Case Analysis (CCA) and two multiple imputation analyses (multiple imputation by chained equations (MICE) and substantive model compatible fully conditional specification (SMC-FCS))

	CCA (N=577)				MICE (N=912)				SMC-FCS (N=912)			
Variable	**HR**	**95% CI**	**SE**	P**-value**	**HR**	**95% CI**	**SE**	P**-value**	**HR**	**95% CI**	**SE**	P**-value**
Age	1.04	1.01-1.07	0.01	0.003	1.05	1.03-1.06	0.01	< 0.001	1.04	1.03-1.06	0.01	< 0.001
Sex												
Male	1				1				1			
Female	0.79	0.54-1.17	0.16	0.237	0.97	0.76-1.23	0.12	0.807	0.97	0.76-1.24	0.12	0.811
Year	1.00	0.96-1.04	0.02	0.999	0.94	0.92-0.96	0.01	< 0.001	0.94	0.92-0.96	0.01	< 0.001
Deprivation	1.00	0.99-1.01	0.01	0.515	1.00	1.00-1.01	< 0.001	0.468	1.00	1.00-1.01	< 0.001	0.331
Diag subgroup[a]												
Ia	1				1				1			
Ib	1.67	1.05-2.64	0.39	0.029	1.84	1.39-2.43	0.26	< 0.001	1.79	1.35-2.39	0.26	< 0.001
Ic	0.71	0.29-1.76	0.33	0.459	0.55	0.30-1.02	0.17	0.059	0.65	0.33-1.29	0.17	0.216
Id[b]	-	-	-	-	0.38	0.05-2.71	0.38	0.332	0.33	0.05-2.45	0.38	0.280
Ie	1.97	0.27-14.31	1.99	0.503	2.48	1.01-6.09	1.14	0.048	2.34	0.94-5.82	1.14	0.067
Ethnicity												
White	1				1				1			
Asian	1.33	0.71-2.50	0.43	0.375	1.33	0.87-2.04	0.29	0.185	1.25	0.71-2.19	0.29	0.438
Other	1.30	0.41-4.16	0.77	0.660	0.90	0.37-2.19	0.41	0.814	1.52	0.63-3.65	0.41	0.347
Log WBC Count	1.21	1.08-1.37	0.07	0.002	1.21	1.10-1.32	0.06	< 0.001	1.21	1.10-1.34	0.06	< 0.001

[a]Ia - Lymphoid leukaemias; Ib - Acute myeloid leukaemias; Ic - Chronic myeloproliferative diseases;
Id - Myelodysplastic syndrome and other myeloproliferative diseases; Ie - Unspecified and other specified leukaemias
[b]For complete cases analysis, n=6 and no deaths were observed in this diagnostic subgroup, therefore HR was not estimated

Table 6.9: Cox proportional hazards model for germ cell tumour survival: Complete Case Analysis (CCA) and two multiple imputation analyses (multiple imputation by chained equations (MICE) and substantive model compatible fully conditional specification (SMC-FCS))

	CCA (N=305)				MICE (N=846)				SMC-FCS (N=846)[a]			
Variable	**HR**	**95% CI**	**SE**	P**-value**	**HR**	**95% CI**	**SE**	P**-value**	**HR**	**95% CI**	**SE**	P**-value**
Age	1.04	0.92-1.18	0.07	0.534	1.04	0.99-1.09	0.03	0.158	1.04	0.99-1.09	0.03	0.146
Sex												
Male	1				1				1			
Female	0.32	0.01-8.30	0.54	0.497	0.64	0.23-1.83	0.34	0.408	0.71	0.27-1.85	0.35	0.482
Year	1.18	1.04-1.33	0.07	0.008	1.01	0.95-1.06	0.03	0.775	1.02	0.97-1.08	0.03	0.448
Deprivation	1.00	0.96-1.03	0.02	0.835	1.01	0.99-1.02	0.01	0.376	N/A	N/A	N/A	N/A
Diagnostic subgroup[b]												
Xc	1				1				1			
Xa	-	-			18.24	3.33-99.99	15.68	0.001	7.51	2.70-20.88	3.91	0
Xb	6.73	1.04-43.75	6.43	0.046	1.72	0.57-5.15	0.96	0.334	2.27	0.82-6.28	1.18	0.113
Xd	54.96	2.64-1144.59	85.14	0.010	13.31	2.59-68.41	11.07	0.002	9.04	2.23-36.67	6.45	0.002
Xe	0.00	-			15.20	1.97-117.09	15.79	0.009	7.24	1.26-41.68	6.46	0.027
Ethnicity												
White	1				1				1			
Asian	1.98	0.49-8.03	1.41	0.340	1.55	0.51-4.70	0.87	0.441	1.66	0.54-5.10	0.95	0.378
Other	8.57	0.72-102.12	10.84	0.089	2.65	0.42-16.52	2.46	0.296	5.40	1.09-26.69	4.35	0.039
Stage												
I	1				1				1			
II	11.72	1.07-128.56	14.32	0.044	5.27	1.40-19.76	3.54	0.014	4.93	1.45-16.82	3.08	0.011
III	33.89	2.31-498.03	46.47	0.010	18.87	4.97-71.65	12.80	< 0.001	14.74	3.90-55.73	9.97	< 0.001
IV	84.50	9.33-765.11	94.99	< 0.001	32.77	8.25-130.19	22.88	< 0.001	31.17	9.18-105.77	19.31	< 0.001

[a] SMC-FCS imputation model excluded diagnostic subgroup and SMC-FCS analysis model excluded deprivation (see §6.3.3.1)

[b] Xa - Intracranial and Intraspinal GCTs; Xb - Malignant extracranial and extragonadal GCTs; Xc - Malignant Gonadal GCTs; Xd - Gonadal carcinomas; Xe - Other and unspecified malignant gonadal tumours

6.4 Survival Analysis

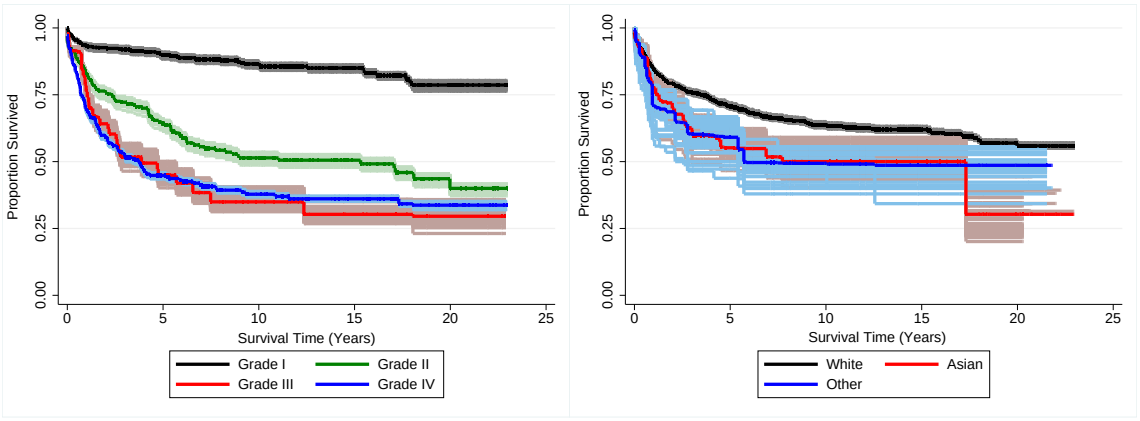
This section contains details of the survival analysis following multiple imputation of partially observed variables, including 1, 3 and 5-year survival estimates and multivariable Cox PH models. The analysis for CNS tumours, leukaemia and GCTs are summarised in §6.4.1, §6.4.2 and §6.4.3 respectively. An assessment of each analyses model is provided in §6.5.

6.4.1 Central Nervous System Tumours

For CNS tumours, 1 and 3-year survival was similar for children and TYAs, however, by 5-years, survival was poorer for TYAs (65%, 95% CI 60-70%) compared to children (73%, 95% CI 68-77%) although not significantly so (Table 6.10). A similar pattern was observed for males and females with CNS tumours, whereby there was little or no difference in 1 and 3-year survival, but by 5-years, males had poorer survival compared to females (66%, 95% CI 62-71% vs. 73%, 95%CI 68-77%). Across deprivation quintiles, survival decreased steadily between 1, 3 and 5-years, however, this decline was slightly steeper for the three most deprived fifths compared to the two least deprived fifths. Survival was highest for the 'other CNS' tumour subgroup and ependymomas and was poorest for 'other gliomas' and 'unspecified CNS tumours.' Of note was the particularly poor 5-year survival for those in the unspecified CNS tumour subgroup, with only 29% survival. K-M curves for grade of CNS tumour showed that those with grade I tumours had a much improved prognosis compared to any other grade of tumour (Figure 6.8). Of note, there was little difference between grade III and grade IV tumours, with a slight indication that survival for those with grade III tumours was poorer compared to those with grade IV tumours, particularly beyond approximately 6-years from diagnosis. K-M curves for ethnicity showed very little difference between survival amongst Asian and other ethnicities, whereas the white ethnic group indicated better survival compared to the Asian and other ethnic category.

Table 6.10: 1, 3 and 5-year survival estimates for central nervous system tumours (displayed as percentages of cases survived)

Variable	1 year (95% CI)	3 year (95% CI)	5 year (95% CI)
Age			
0-14	84 (81-87)	75 (70-79)	73 (68-77)
15-29	82 (78-86)	74 (69-78)	65 (60-70)
Sex			
Male	82 (78-85)	72 (67-76)	66 (62-71)
Female	85 (81-88)	77 (72-81)	73 (68-77)
Deprivation			
Least deprived (1)	82 (75-88)	75 (67-81)	71 (62-78)
2	80 (74-86)	70 (63-77)	67 (59-73)
3	88 (81-92)	80 (73-86)	73 (65-80)
4	86 (79-91)	77 (69-83)	71 (63-78)
Most deprived (5)	82 (76-86)	71 (65-77)	66 (59-72)
Diagnostic sugroups			
Astrocytoma	85 (81-89)	76 (72-80)	70 (65-74)
Ependymoma	91 (81-96)	82 (71-89)	77 (65-86)
Embryonal	79 (71-85)	65 (56-73)	61 (52-69)
Other gliomas	69 (59-77)	55 (45-64)	51 (41-60)
Other CNS	92 (85-96)	92 (85-96)	90 (82-94)
Unspec. CNS	43 (10-73)	43 (10-73)	29 (04-51)



(a) (b)

Figure 6.8: Kaplan-Meier (K-M) survival curves for central nervous system tumours for each imputation, superimposed with the mean K-M curve averaged over all imputations by grade (a) and ethnic group (b)

In order to determine a final analysis model, the method recommended by Collett [245] was implemented (discussed in §4.4.3.1 in detail) and results given in Table 6.11. Importantly, this method relies on clinical knowledge and considers variables of key interest (as identified from previous literature) rather than relying on an automated forward, backward or stepwise model selection procedure. A lenient significant cut off of 10% was used for guidance on individual explanatory power and evidence of interactions

and linearity while selecting variables for inclusion or exclusion from the model.

WHO grade and diagnostic subgroup both had individual explanatory power on survival which was significant at the 1% level. Age, sex and year of diagnosis all had explanatory power at the 5% significance level and ethnicity had explanatory power at the 10% significance level. The single effect of deprivation was not significant and was therefore excluded from the final model. Sex did not significantly improve the model containing WHO grade and diagnostic group, however, it was initially retained in the model to allow the age by sex interaction to be tested. There was no significant evidence against linearity of age (P=0.582 for 5-year age bands and P=0.313 for 15 year age bands), and year of diagnosis (P=0.148 for 5-year periods), therefore these variables were initially retained in the model as continuous variables. However, upon assessment of the PH assumption, a model with 15 year age groups was deemed more appropriate than one with continuous age (see §6.5 below). There was no evidence of an interaction between year and diagnostic subgroup (P=0.408) or age and sex (P=0.286), therefore these interactions were excluded from the final model. In addition, the non-significant main effect of sex was also excluded from the final model.

Table 6.11: Model selection process for central nervous system tumours using P-values of overall model fit obtained from Wald tests to guide selection of the analysis model

Variables in model	P**-value**
None	-
Age	0.0030
Sex	0.0263
Ethnicity	0.0941
Year	0.0016
Grade	< 0.001
Deprivation	0.6316
Diagnostic subgroup	< 0.001
Diagnostic subgroup + Grade	< 0.001
Diagnostic subgroup + Grade + Age	0.0924
Diagnostic subgroup + Grade + Sex	0.1037
Diagnostic subgroup + Grade + Year	0.0025
Diagnostic subgroup + Grade + Ethnicity	0.0765
Diagnostic subgroup + Grade + Year + Age	0.0339
Diagnostic subgroup + Grade + Year + Age + Sex	0.1559
Diagnostic subgroup + Grade + Year + Age + Sex + Ethnicity (Full Model)	0.0272
Linearity Tests	
Full model + Age Group (0-14, 15-29 years)	0.3131
Full model + Age Group (0-4, 5-9, 10-14, 15-19, 20-24, 25-29 years)	0.5818
Full model + Period (1990-1994, 1995-1999, 2000-2004, 2005-2009)	0.1475
Interactions	
Year x Diagnostic subgroup	0.4075
Age x Sex	0.2856
Final Model	
Diagnostic subgroup + Grade + Year + Age + Ethnicity	

6.4.1.1 Survival analysis results for central nervous system tumours

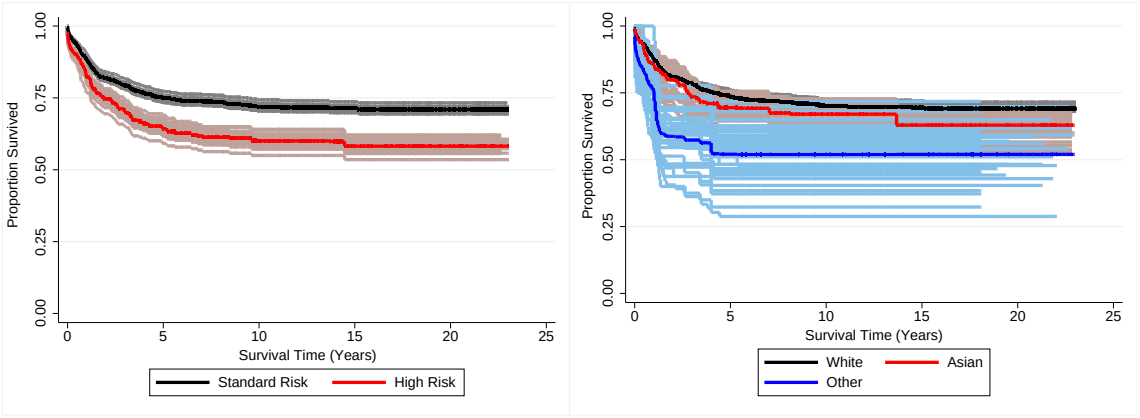
Survival for CNS tumours was significantly worse for older cases, with a 34% increase in the risk of death for those diagnosed aged 15-29 years compared to 0-14 year olds (Table 6.12). There was a two-fold increased risk of death for other ethnicity compared to white ethnicity, and an increased risk for those of Asian ethnicity compared to white ethnicity (HR=1.50) although the latter was not significant. Survival worsened significantly with increasing WHO grade at presentation by 3.5-fold, almost 6.5-fold and 10-fold for grade II, III and IV tumours respectively compared to grade I tumours. Over the study period, survival rates of CNS tumours improved by 4% on average per year, and survival for ependymomas and embryonal tumours was significantly better compared to astrocytomas. In the univariable analysis, survival of ependymomas was significantly worse compared to survival of astrocytomas (Table 6.10), however, after adjusting for WHO grade, this effect was reversed as all embryonal tumours were WHO grade IV, compared to only 1.8% of astrocytoma cases at WHO grade IV. Those with other gliomas had a 51% increased risk of death compared to astrocytomas and those with unspecified CNS tumours had a 3-fold increased risk of death compared to astrocytomas, as reflected by the univariable 1, 3 and 5-year survival rates.

Table 6.12: Pooled Cox proportional hazards model estimates based on 40 imputations for CNS tumours amongst children and young adults in Yorkshire, 1990-2009

Variable	**HR**	**95% CI**	**SE**	P**-value**
Age Group				
0-14 years	1			
15-29 years	1.34	1.04-1.71	0.169	0.022
Year	0.96	0.94-0.98	0.011	< 0.001
Diagnostic Subgroup				
Astrocytoma	1			
Ependymoma	0.58	0.35-0.96	0.150	0.035
Embryonal	0.40	0.27-0.61	0.083	< 0.001
Other gliomas	1.51	1.06-2.16	0.273	0.022
Other CNS	0.60	0.35-1.01	0.160	0.055
Unspecified CNS	2.97	1.17-7.50	1.401	0.021
Ethnicity				
White	1			
Asian	1.50	0.89-2.53	0.399	0.128
Other	2.25	1.12-4.52	0.797	0.023
Grade				
I	1			
II	3.53	2.41-5.18	0.690	< 0.001
III	6.41	3.95-10.41	1.585	< 0.001
IV	10.10	6.72-15.20	2.103	< 0.001

6.4.2 Leukaemia

For leukaemia, survival for those diagnosed under the age of 1 was very poor compared to other age groups and continued to decline from 55% to 37% and 29% at 1-, 3- and 5-years respectively (Table 6.13). Children diagnosed between the age of 1 and 14 experienced better survival compared to all other age groups, and although 1-year survival for TYAs aged 15-29 years was good, survival declined from 78% to 62% and 55% at 3 and 5-years respectively. Survival did not vary by gender or deprivation and remained stable over all three time points. Diagnostic subgroup Ib - Acute myeloid leukaemias displayed a similar survival pattern as that for TYAs, which was reflected by the fact that AML occurred more commonly amongst TYAs compared to children. Survival amongst diagnostic subgroup Ie - Unspecified and other specified leukaemias was poor at 55% survival by 1-year, but remained the same for 3 and 5-year survival. K-M curves for log WBC count were categorised according to standard and high risk, and these risk classifications accurately represented survival differences with clear separation between curves (Figure 6.9). The graph also showed there was less between-imputation variation for those with standard risk compared to high risk cases. K-M curves for ethnicity showed no difference between survival amongst white, Asian and other ethnic categories.



(a) (b)

Figure 6.9: Kaplan-Meier (K-M) survival curves for leukaemia for each imputation, superimposed with the mean K-M curve averaged over all imputations by standard ($< 50,000\mu/L$) and high risk ($\geq 50,000\mu/L$) WBC count (a) and ethnic group (b)

Table 6.13: 1, 3 and 5-year survival estimates for leukaemia (displayed as percentages of cases survived)

Variable	**1 year (95% CI)**	**3 year (95% CI)**	**5 year (95% CI)**
Age (years)			
< 1	51 (35-65)	37 (22-51)	29 (16-43)
1-10	95 (93-97)	90 (87-92)	87 (84-90)
11-14	89 (80-94)	78 (68-85)	75 (61-80)
15-29	78 (72-82)	62 (56-67)	55 (49-61)
Sex			
Male	88 (84-90)	78 (74-81)	73 (68-76)
Female	86 (82-89)	76 (72-80)	72 (68-76)
Deprivation Quintile			
Least deprived (1)	89 (82-93)	80 (72-86)	76 (68-83)
2	89 (83-93)	80 (73-85)	74 (67-80)
3	85 (78-89)	78 (70-83)	74 (67-80)
4	88 (82-92)	75 (68-81)	69 (61-76)
Most deprived (5)	85 (80-89)	74 (69-79)	71 (65-75)
Diagnostic subgroup[a]			
Ia	93 (90-94)	84 (81-87)	80 (76-83)
Ib	72 (66-77)	59 (53-65)	55 (49-61)
Ic	94 (83-98)	84 (71-92)	72 (57-82)
Id	100 (100-100)	89 (43-98)	89 (43-98)
Ie	55 (23-78)	55 (23-78)	55 (23-78)

[a]Ia - Lymphoid leukaemias; Ib - Acute myeloid leukaemias; Ic - Chronic myeloproliferative diseases; Id - Myelodysplastic syndrome and other myeloproliferative diseases; Ie - Unspecified and other specified leukaemias

Table 6.14 shows the model selection process for the Cox PH model for survival of leukaemia, following the same process as for CNS tumours. Age, year of diagnosis, logarithm of WBC count and diagnostic subgroup all had individual explanatory power on survival which was significant at the 1% level. The addition of each of these variables individually to the model continued to improve the model fit, therefore all were included in the final model. The single effects of sex, ethnicity and deprivation were not significant. Deprivation was excluded from the final model, however, ethnicity was retained within the model as it was of primary interest in the analysis. In addition, sex was initially retained in the model in order to study the interaction of age and sex. There was significant evidence against linearity of age when tested against children and TYA age groupings (0-14 and 15-29 years), 5-year age groupings (0-4, 5-9, 10-14, 15-19, 20-24, 25-29 years) as well as clinically relevant age groups (< 1, 1-10, 11-14, 15-29 years). The latter was chosen to be included in the model due to its clinical relevance in addition to the observed differences in survival by age groups in the univariable 1, 3 and 5-year survival estimates. The interaction between age group and sex was not significant (P <0.124), therefore the non-significant main effect of sex was also excluded from the final model.

Table 6.14: Model selection process for leukaemia using P-values of overall model fit obtained from Wald tests to guide selection of the analysis model

Variables in model	P**-value**
None	-
Age	< 0.001
Sex	0.875
Ethnicity	0.301
Year	< 0.001
log WBC Count	0.006
Deprivation	0.248
Diagnostic subgroup	< 0.001
Diagnostic subgroup + Ethnicity	0.4701
Diagnostic subgroup + Ethnicity + Sex	0.6118
Diagnostic subgroup + Ethnicity + Sex + Year	< 0.001
Diagnostic subgroup + Ethnicity + Sex + Year + Age	< 0.001
Diagnostic subgroup + Ethnicity + Sex + Year + Age + log WBC count (Full Model)	< 0.001
Linearity Tests	
Full model + Age Group (0-14, 15-29 years)	0.001
Full model + Age Group (0-4, 5-9,10-14,15-19,20-24,25-29 years)	0.001
Full model + Age Group (< 1, 1-10, 11-14, 15-29 years)	< 0.001
Full model + Period (1990-1994, 1995-1999, 2000-2004, 2005-2009)	0.773
Interactions	
Age Group (< 1, 1-10, 11-14, 15-29 years) by Sex	0.124
Final Model	
Diagnostic subgroup + Ethnicity + Year + Age Group (< 1, 1-10, 11-14, 15-29 years) + log WBC count	

6.4.2.1 Survival analysis results for leukaemia

Survival was significantly better for all age groups compared to those diagnosed under the age of 1 (Table 6.15). Furthermore, survival improved significantly over the study period by 7% on average per year (P <0.001). Survival of AML was 44% poorer compared to lymphoid leukaemias (P=0.009), and there was a 2.3-fold increased risk of death for unspecified and other specified leukaemias compared to lymphoid leukaemias, however, this effect was not significant (P=0.087). Those diagnosed with chronic myeloproliferative diseases had better survival compared to lymphoid leukaemias by 50% (P=0.013), and despite a low HR for those with myelodysplastic syndrome (HR=0.30, 95% CI 0.04-2.22) there was no significant difference in survival compared to lymphoid leukaemias. There was no significant difference for survival of leukaemia by ethnic group, despite large HRs of 1.34 and 1.38 for Asian and other ethnicity compared to white ethnicity, and survival worsened significantly by 18% on average per increase in the logarithm of WBC count (P=0.001).

Table 6.15: Pooled Cox proportional hazards model estimates based on 40 imputations for leukaemia amongst children and young adults in Yorkshire, 1990-2009

Variable	**HR**	**95% CI**	**SE**	P**-value**
Age Group				
<1 year	1			
1-10 years	0.16	0.10-0.25	0.04	<0.001
11-14 years	0.34	0.19-0.59	0.10	<0.001
15-29 years	0.63	0.41-0.97	0.14	0.037
Year	0.93	0.91-0.95	0.01	<0.001
Diagnostic Subgroup				
Lymphoid leukaemias	1			
Acute myeloid leukaemias	1.44	1.09-1.90	0.20	0.009
Chronic myeloproliferative diseases	0.49	0.28-0.86	0.14	0.013
Myelodysplastic syndrome	0.30	0.04-2.22	0.31	0.240
Unspecified and other specified leukaemias	2.26	0.89-5.73	1.07	0.087
Ethnicity				
White	1			
Asian	1.34	0.80-2.26	0.36	0.266
Other	1.38	0.56-3.34	0.63	0.481
Log white blood cell (WBC) count	1.18	1.07-1.30	0.06	0.001

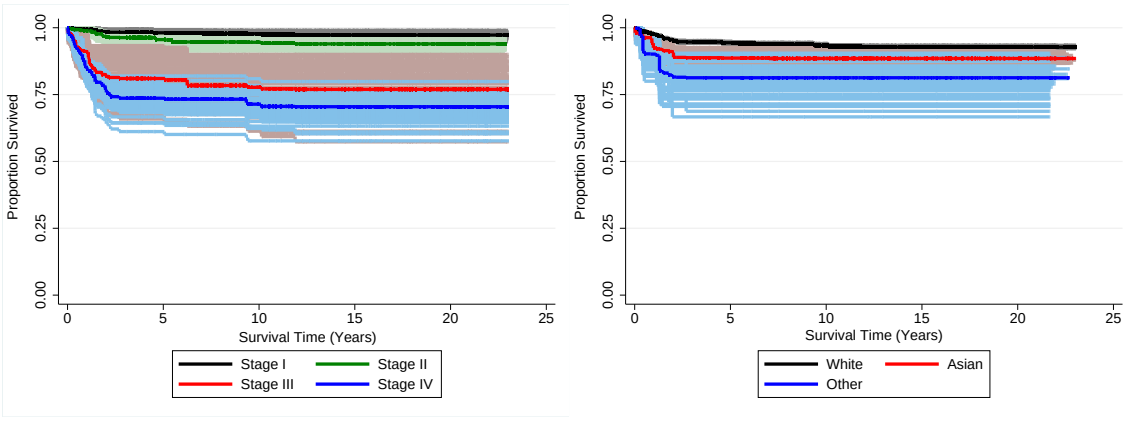
6.4.3 Germ Cell Tumours

For GCTs, survival was very high overall and in all cases remained high even at 5-years from diagnosis (Table 6.16). Changes between 1 and 3 year survival did occur amongst the smaller diagnostic subgroups, with intracranial and intraspinal GCTs decreasing from 87% to 82%, malignant extracranial and extragonadal GCTs from 91% to 85%, gonadal carcinomas from 96% to 85% and other and unspecified malignant gonadal tumours from 100% to 82%. However, no further deaths were observed between 3 and 5 years beyond diagnosis. The largest diagnostic subgroup, malignant gonadal GCTs, had a 1-year survival rate of 98%, which dropped only slightly to 96% at 3 and 5-years.

Table 6.16: 1, 3 and 5-year survival estimates for germ cell tumours (displayed as percentages of cases survived)

Variable	1 year (95% CI)	3 year (95% CI)	5 year (95% CI)
Age			
0-14	96 (89-98)	95 (88-98)	95 (88-98)
15-29	97 (96-98)	94 (92-95)	94 (92-95)
Sex			
Male	97 (96-98)	94 (92-96)	94 (92-96)
Female	97 (91-99)	92 (85-95)	92 (85-95)
Deprivation Quintile			
Least Deprived (1)	97 (92-99)	94 (88-97)	94 (88-97)
2	97 (93-99)	95 (91-98)	95 (90-97)
3	99 (96-100)	96 (92-98)	96 (92-98)
4	96 (92-98)	92 (87-95)	92 (86-95)
Most Deprived (5)	96 (93-98)	93 (89-96)	93 (89-96)
Diagnostic subgroup[a]			
Xa	87 (72-94)	82 (66-91)	82 (66-91)
Xb	91 (80-96)	85 (73-92)	85 (73-92)
Xc	98 (97-99)	96 (94-97)	96 (94-97)
Xd	96 (76-99)	85 (64-94)	85 (64-94)
Xe	100 (-)	82 (45-95)	82 (45-95)

[a]Xa - Intracranial and Intraspinal GCTs; Xb - Malignant extracranial and extragonadal GCTs; Xc - Malignant Gonadal GCTs; Xd - Gonadal carcinomas; Xe - Other and unspecified malignant gonadal tumours



(a) (b)

Figure 6.10: Kaplan-Meier (K-M) survival curves for germ cell tumours for each imputation, superimposed with the mean K-M curve averaged over all imputations by stage (a) and ethnic group (b)

Stage and diagnostic subgroup both had individual explanatory power on survival which was significant at the 1% level (Table 6.17). Sex had individual explanatory power at the 5% level, however, the addition of sex to a model containing stage and diagnostic subgroup did not provide a significant improvement in model fit. Despite age not having individual explanatory power, a model containing age group, stage and diagnostic

subgroup significantly improved the model. This was not observed when adding continuous age, or any other variable individually to the model.

Table 6.17: Model selection process for germ cell tumours using P-values of overall model fit obtained from Wald tests to guide selection of the analysis model

Variables in model	P**-value**
None	-
Age	0.720
Sex	0.049
Ethnicity	0.163
Year	0.646
Stage	< 0.001
Deprivation	0.181
Diagnostic subgroup	< 0.001
Diagnostic subgroup + Ethnicity	0.1841
Diagnostic subgroup + Ethnicity + Stage	< 0.001
Diagnostic subgroup + Ethnicity + Stage + Sex	0.1567
Diagnostic subgroup + Ethnicity + Stage + Age Group (0-14, 15-29 years)	0.0219
Final Model	
Diagnostic subgroup + Ethnicity + Stage + Age Group (0-14, 15-29 years)	

6.4.3.1 Survival analysis results for germ cell tumours

Table 6.18 shows estimates from the final Cox PH model for survival of GCTs. Those diagnosed aged 15 to 29 years had a significant 4-fold increased risk of death compared to children diagnosed aged 0-14 years (P=0.022). Survival was poorer for all diagnostic subgroups compared to gonadal GCTs, although the effect was not significant for extracranial and extragonadal GCTs. There was no significant difference by ethnic group despite the large HRs (HR=1.68, 95% CI=0.59-4.82 and HR=2.85, 95% CI 0.50-16.32 for Asian and other ethnicity respectively). This may be a result of the low EPV for the final model (EPV=6.2) in addition to the small number of cases diagnosed with GCT within these ethnic groups as reflected by the wide confidence intervals. Survival worsened significantly by 5-fold, 16-fold and 32-fold for stage II, III and IV tumours respectively.

Table 6.18: Pooled Cox proportional hazards model estimates based on 60 imputations for germ cell tumours amongst children and young adults in Yorkshire, 1990-2009

Variable	**HR**	**95% CI**	**SE**	P**-value**
Age Group				
0-14 years	1			
15-29 years	3.92	1.22-12.60	2.33	0.022
Diagnostic subgroup				
Malignant gonadal GCTs	1			
Intracranial and Intraspinal GCTs	19.88	4.19-94.43	15.67	< 0.001
Malignant extracranial and extragonadal GCTs	1.67	0.58-4.81	0.90	0.342
Gonadal carcinomas	10.09	2.55-39.97	7.05	0.001
Other and unspecified GCTs	15.21	1.84-125.93	16.34	0.012
Ethnicity				
White	1			
Asian	1.68	0.59-4.82	0.90	0.333
Other	2.85	0.50-16.32	2.53	0.238
Stage				
I	1			
II	5.18	1.38-19.41	3.48	0.015
III	15.63	4.34-56.27	10.18	< 0.001
IV	31.59	8.11-122.96	21.75	< 0.001

6.5 Analysis Model Assessment

Figures 6.11, 6.12 and 6.13 show log cumulative hazard plots for all variables included in the CNS tumour, leukaemia and GCT models. In cases where the data was sparse, most notably in the GCT analysis for diagnostic subgroup and ethnicity between -6 and 0 log survival time, the log cumulative hazard plots cross for certain subgroups (Xb - Malignant extracranial and extragonadal GCTs crosses Xa - Intracranial and Intraspinal GCTs and the plot for Asian ethnicity crosses the white ethnicity group). However, despite this, for all tumour groups and variables, the lines do not cross in the most concentrated area of the graph and therefore, there is no strong evidence against the PH assumption for the CNS tumour, leukaemia or GCT analysis.

The deviance residuals plots shown in Figures 6.14, 6.15 and 6.16 show that there were no extreme outlying residuals for any tumour group. There were two clear groups of residuals in each individual plot; the residuals associated with censored observations forming a dense scatter, and the residuals associated with the observed failure events forming a curved pattern. For all three tumour groups, the censored observations appeared to be randomly scattered and there were no extreme outliers for the observed failure events in any tumour group. The lack of extreme outliers indicated an overall adequate model fit as the model predicted events close to the time of observed events for all three tumour groups. Furthermore, the predictive power of all models was high, with Harrell's C statistic ranging between 74% and 76% over all imputations for CNS tumours, 74% and 75% for leukaemia and 81% to 92% for GCTs.

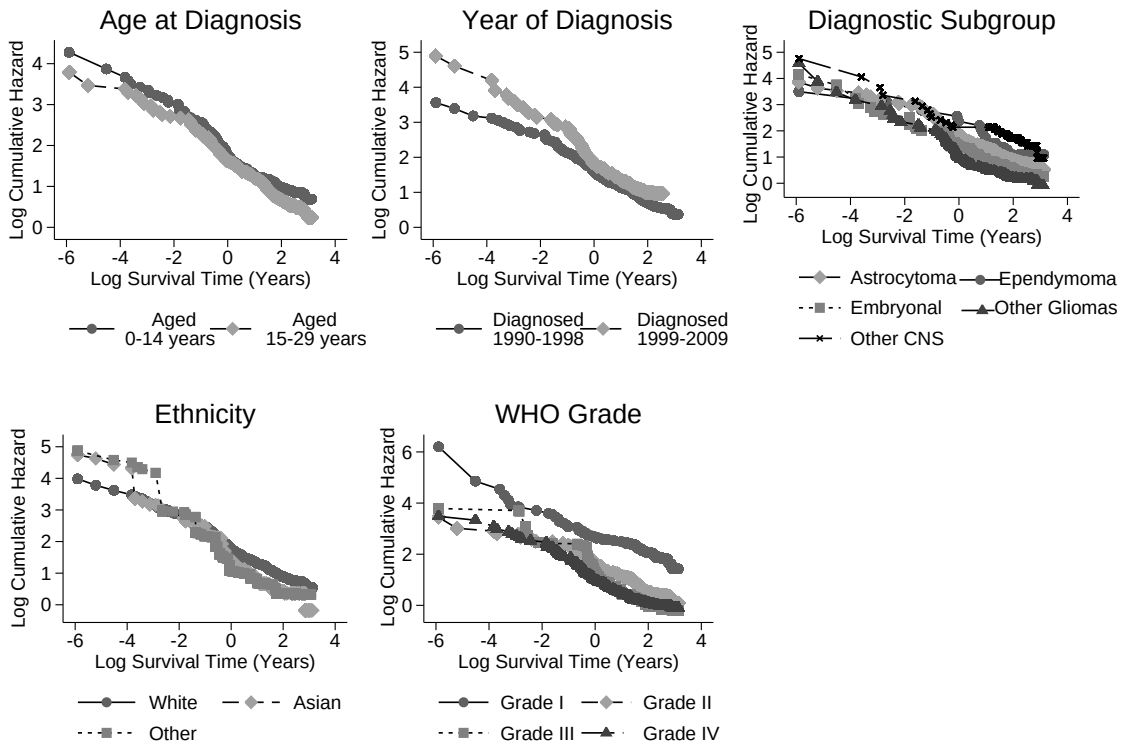


Figure 6.11: Log cumulative hazard plot for one imputation by age group at diagnosis, year of diagnosis, diagnostic subgroup, ethnicity and WHO grade for central nervous system tumours based on the final analysis model (see Table 6.12)

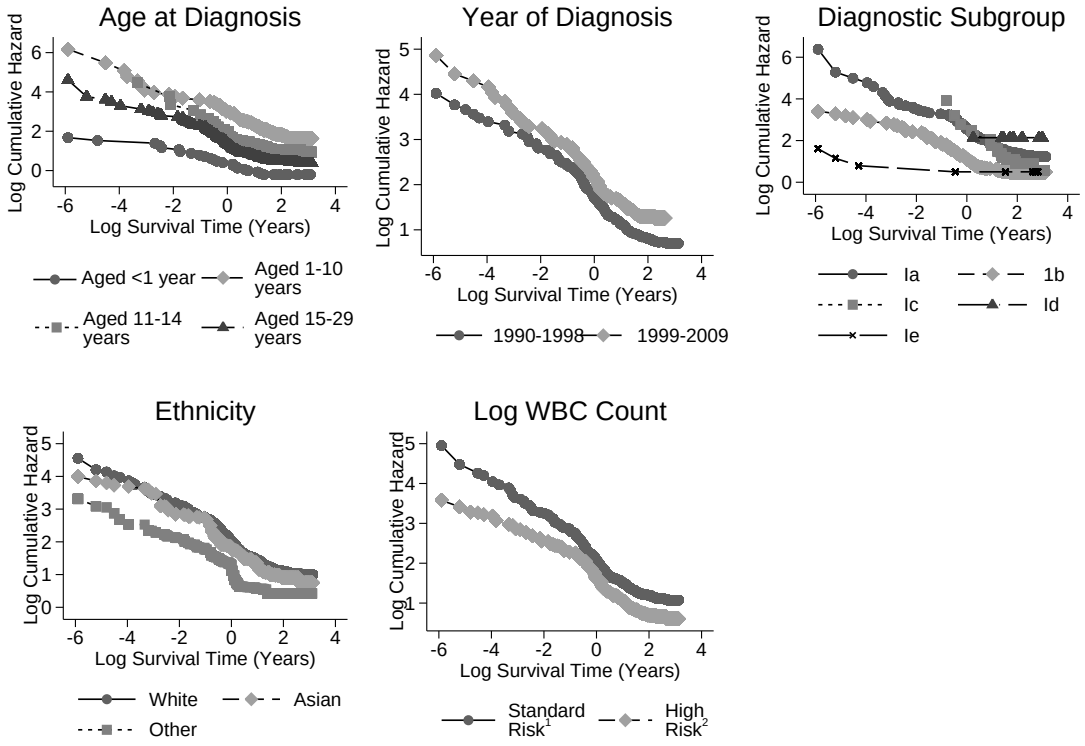


Figure 6.12: Log cumulative hazard plot for one imputation by age group at diagnosis, year of diagnosis, diagnostic subgroup, ethnicity and logarithm of white blood cell (WBC) count for leukaemia based on the final analysis model (see Table 6.15).
Diagnostic subgroups: Ia - Lymphoid leukaemias; Ib - Acute myeloid leukaemias; Ic - Chronic myeloproliferative diseases; Id - Myelodysplastic syndrome and other myeloproliferative diseases; Ie - Unspecified and other specified leukaemias. [1] Standard Risk:$< 50,000\mu/L$, [2] High Risk: $\geq 50,000\mu/L$

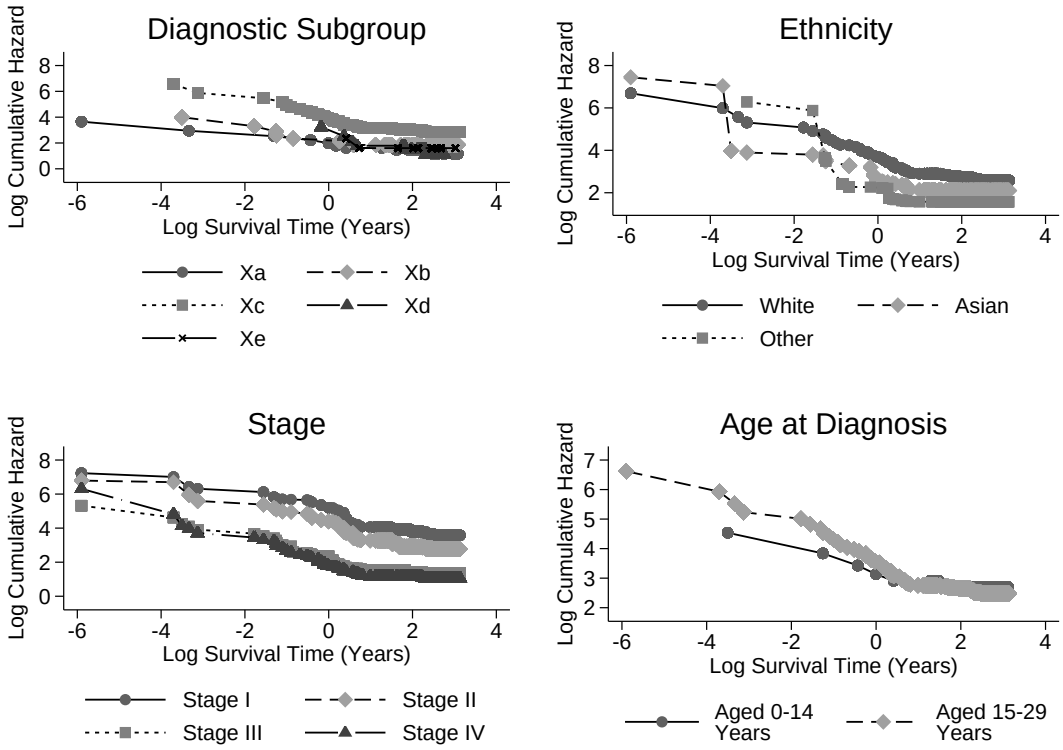


Figure 6.13: Log cumulative hazard plot for one imputation by diagnostic subgroup, ethnicity, stage and age group at diagnosis for germ cell tumours based on the final analysis model (see Table 6.18).
Diagnostic subgroups: Xa - Intracranial and Intraspinal GCTs; Xb - Malignant extracranial and extragonadal GCTs; Xc - Malignant Gonadal GCTs; Xd - Gonadal carcinomas; Xe - Other and unspecified malignant gonadal tumours

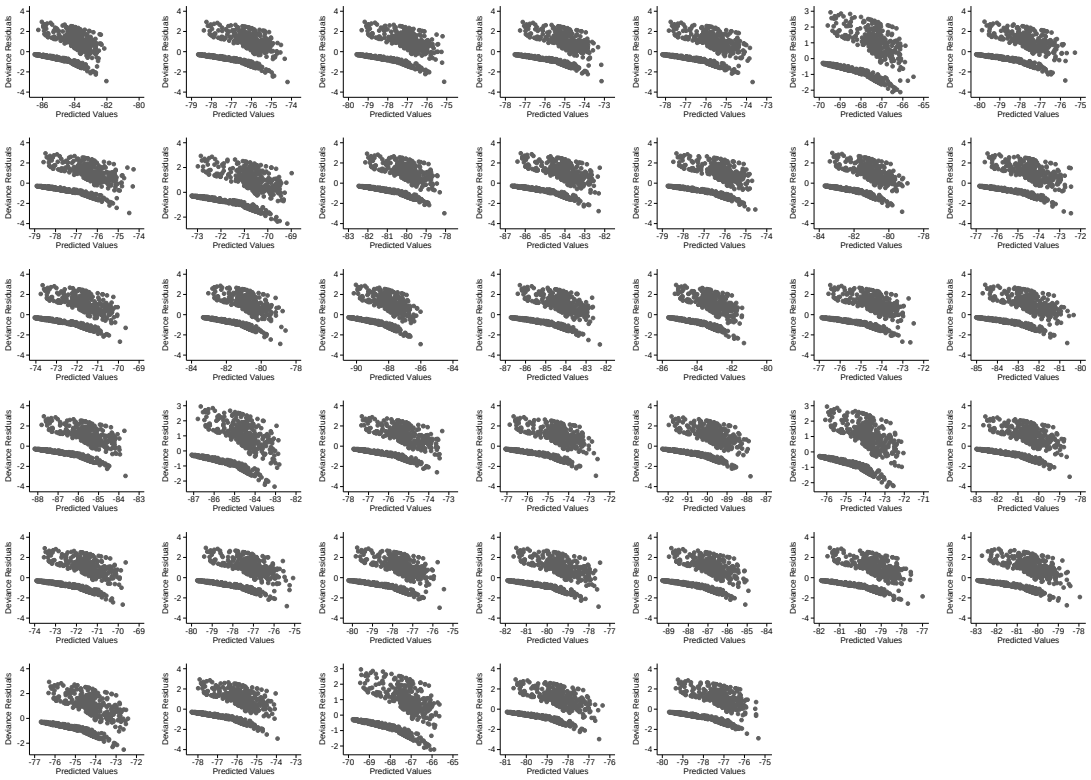


Figure 6.14: Deviance residual plots for each of the 40 Cox Proportional Hazards models based on individual imputations for central nervous system tumours

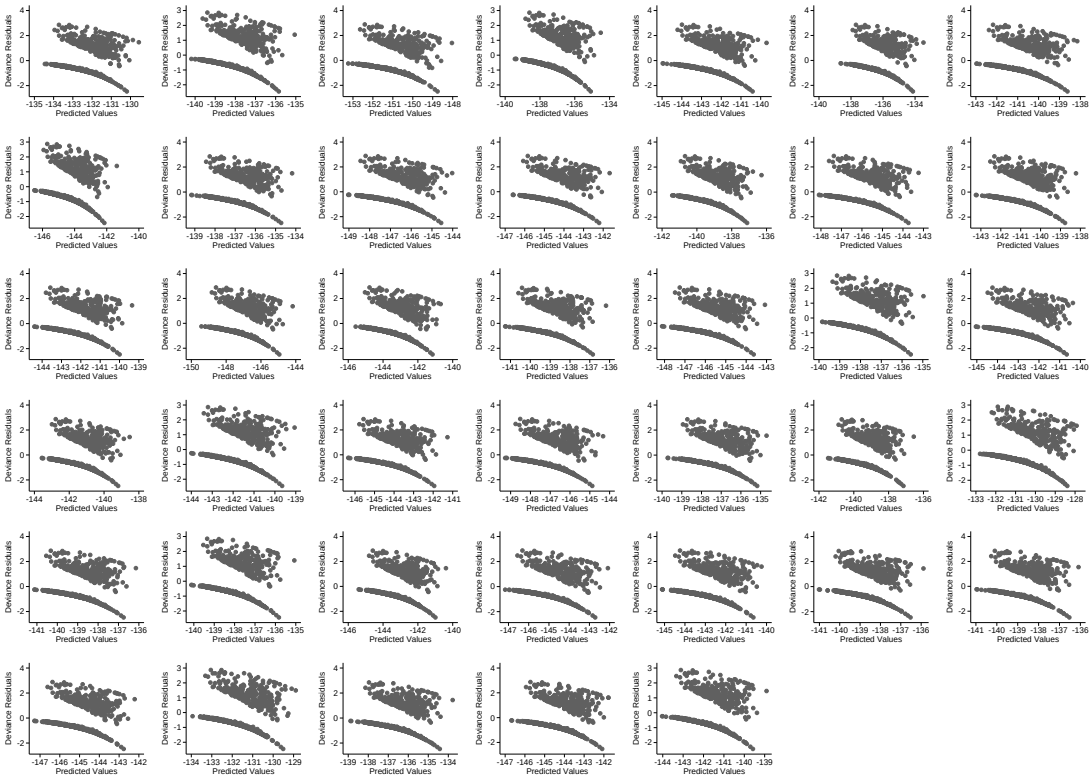


Figure 6.15: Deviance residual plots for each of the 40 Cox Proportional Hazards models based on individual imputations for leukaemia

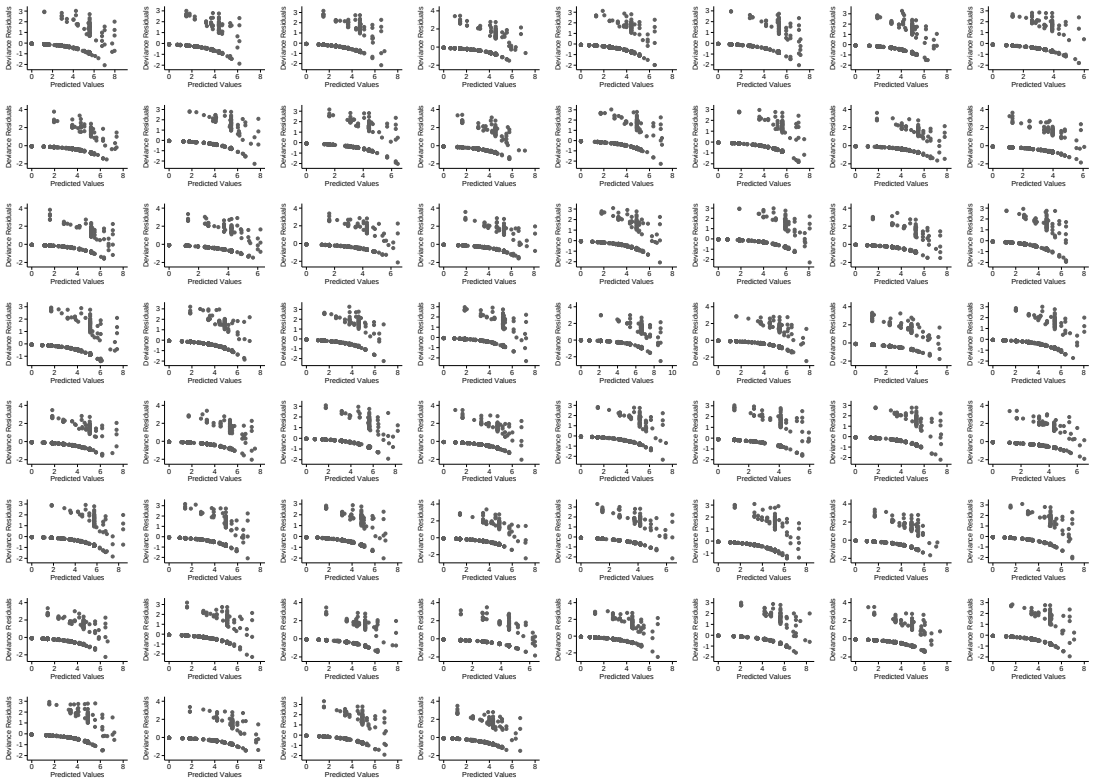


Figure 6.16: Deviance residual plots for each of the 60 Cox Proportional Hazards models based on individual imputations for germ cell tumours

6.6 Sensitivity Analysis

6.6.1 Deviations from MAR Assumption

Deviations from the MAR assumption were assessed by adjusting the probability of imputed values by a factor of θ=0.5 and θ=0.8 to assess sensitivity to the assumption that tumours with a higher grade, higher WBC measures or higher stage for CNS, leukaemia and GCT respectively, were associated with poorer survival. Figures 6.17, 6.18 and 6.19 show the HRs and standard errors obtained under the MAR assumption, compared to those obtained under an MNAR assumption by adjusting the imputed values by a factor of θ as indicated. The graph indicated that the multiple imputation process was robust to any deviations from the MAR assumption, as changes in the HRs were minimal and the 95% CIs for θ=0, 0.5 and 0.8 overlapped in all instances indicating that there were no significant differences to results when adjusting the probability of imputing disease severity. The greatest variation in HRs was seen amongst GCTs, with larger pooled HRs for stage IV tumours under the MNAR assumption, however, both CIs overlapped with the CI for the MAR assumption, showing that the assumption remained robust to deviations

from this assumption.

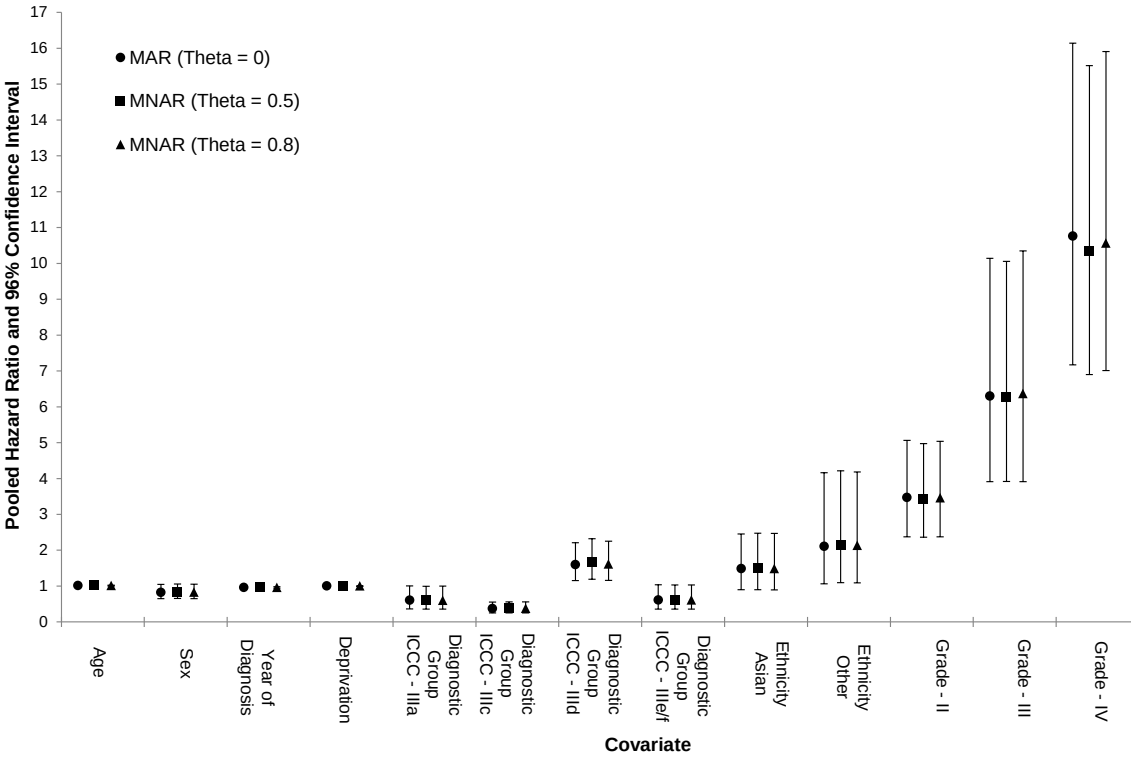


Figure 6.17: Hazard ratios and 95% confidence intervals obtained under the MAR assumption (θ=0) and the MNAR assumption θ=0.5 and θ=0.8, where θ was the adjustment factor of the probability of imputing higher grade central nervous system tumours

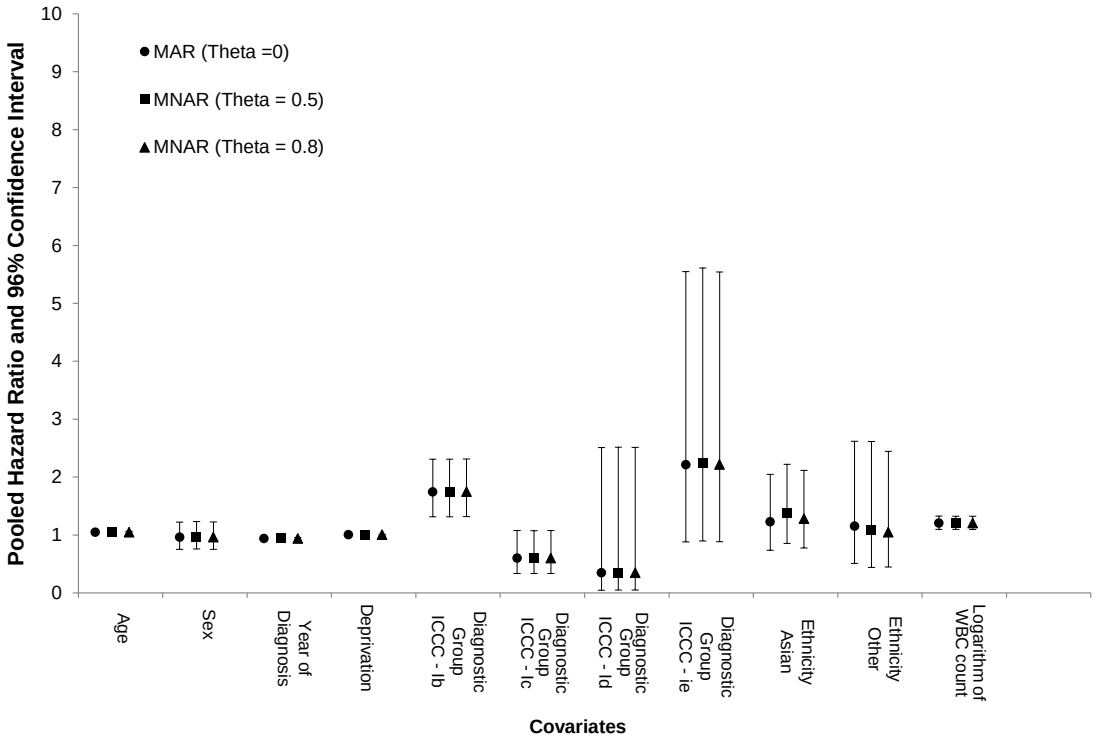


Figure 6.18: Hazard ratios and 95% confidence intervals obtained under the MAR assumption (θ=0) and the MNAR assumption θ=0.5 and θ=0.8, where θ was the adjustment factor of the probability of imputing higher WBC counts for leukaemia

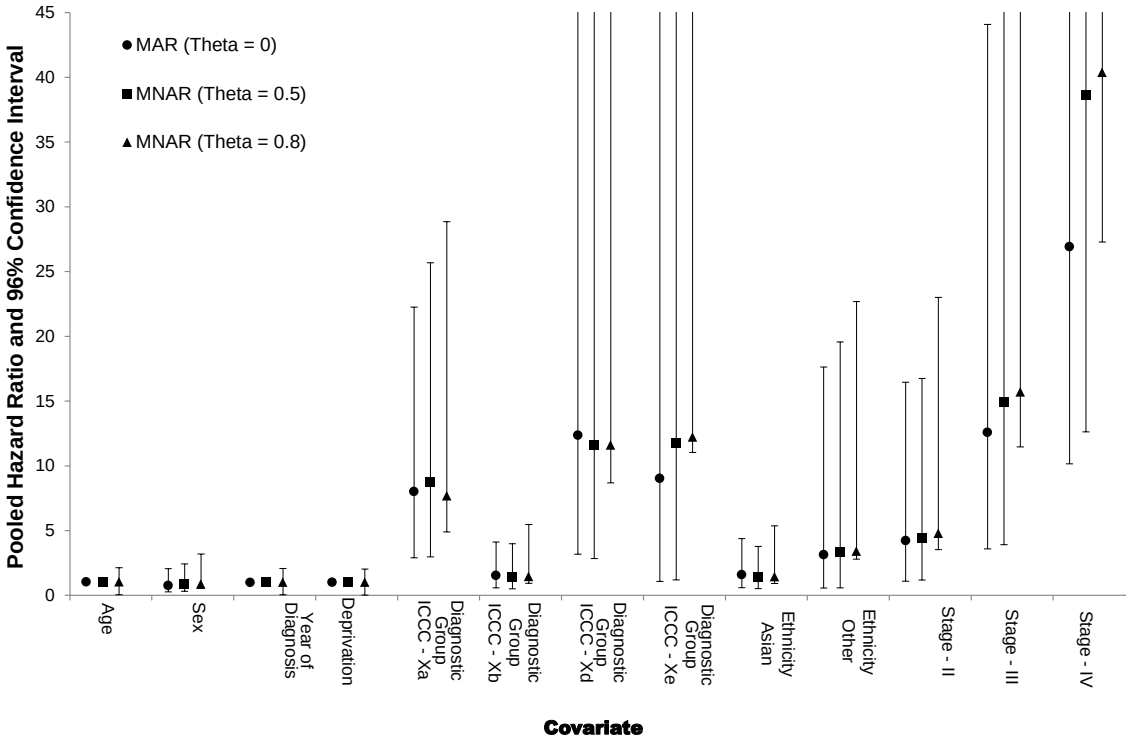


Figure 6.19: Hazard ratios and 95% confidence intervals obtained under the MAR assumption (θ=0) and the MNAR assumption θ=0.5 and θ=0.8, where θ was the adjustment factor of the probability of imputing higher stage germ cell tumours.
Upper confidence limits not shown on the plot were as follows: diagnostic group Xd - 68.1 (θ=0), 47.6 (θ=0.5) and 46.1 (θ=0.8); diagnostic group Xe - 117.1 (θ=0), 118.1 (θ=0.5) and 126.9 (θ=0.8); Stage III - 56.6 (θ=0.5) and 58.2 (θ=0.8); Stage IV - 130.2 (θ=0), 118.1 (θ=0.5) and 124.4 (θ=0.8)

6.7 Summary

The results presented in this chapter provide an important contribution to CYA cancer epidemiology, as well as the wider adult cancer epidemiology research field, showing important differences between the results from CCA and multiple imputation analyses and their effect upon conclusions drawn. Exploratory analyses revealed that the univariable survival curves for CNS tumours and leukaemia were significantly poorer for those cases which had missing data compared to those with complete data. Therefore, performing a CCA would result in overestimation of the true survival for CNS tumours and leukaemia. Furthermore, based on CCA, there were some variable levels for which HRs could not be estimated as there were instances in which some variable levels contained either no cases or events after listwise deletion. Specifically, there were no observed deaths for the myelodysplastic syndrome subgroup of leukaemia or the intracranial and intraspinal GCT subgroup and there were no cases of the other and unspecified subgroup of GCTs after listwise deletion. These subgroups were therefore excluded from the CCA, whereas the use of multiple imputation allowed these subgroups to be retained in the analysis.

Multiple imputation improved study power, as more cases were included in the analysis after imputation. There were several examples of increased study power to detect a significant difference, including effects of age, year of diagnosis, diagnostic subgroup and ethnicity for CNS tumours which were significant in the multiple imputation analysis but not in the CCA. For leukaemia, year of diagnosis and diagnostic subgroup were significant only after multiple imputation. These effects were not apparent amongst the GCT analysis, because power remained poor even after imputation due to the low number of deaths within this diagnostic group. Related to an improvement in the power of the analysis after multiple imputation, the precision of estimates was also improved and therefore smaller CIs were obtained, this occurred consistently across tumour groups.

SMC-FCS has been identified as a superior method to MICE when imputing for a non-linear model such as a Cox PH model, however, when results were compared to imputation under the MICE algorithm, any differences observed were negligible. Furthermore, the desired SMC-FCS analysis was not possible for the GCT tumour analysis, and a correctly specified imputation model using MICE was favoured compared to an incorrectly specified imputation model using SMC-FCS.

A summary of the survival analysis results per tumour group is provided below, with further in depth discussion of these results given in Chapter 9 as part of the main thesis discussion.

For CNS tumours, univariable analysis indicated very little difference in survival for children compared to TYAs at 1 and 3 years post diagnosis, however, a difference

began to emerge for longer term survival (5-years post diagnosis). The multivariable results confirmed a significant difference between children and TYAs, with 34% poorer survival for TYAs compared to children. There was a significant 2-fold increase in the risk of death for other ethnicity compared to white ethnicity, however, there was no difference in survival between Asian and white ethnic groups. The univariable analysis indicated poorest survival occurred for unspecified CNS tumours (29% 5-year survival), followed by other gliomas (51%) and embryonal tumours (61%). Improved survival was observed for other CNS tumours (90%), ependymomas (77%) and astrocytomas (70%). The multivariable analysis was consistent with the univariable analysis for other gliomas and unspecified CNS tumours which had a significant 50% and 3-fold increased risk of death compared to astrocytomas respectively. However, after adjustment for casemix, the embryonal tumours had a 60% improved chance of survival compared to astrocytomas which was not evident from the univariable analysis. Finally, the risk of death increased exponentially by 4, 6 and 10-fold for grade II, III and IV tumours respectively compared to grade I tumours. Importantly, survival improved by 4% per year on average over the study period, and there was no evidence against linearity of this effect. Neither deprivation or sex had significant effects on the survival for cases of CNS tumours.

For leukaemia, age was a strong predictor of survival, with those diagnosed under the age of 1 having very poor survival (29% 5-year survival). With the exception of those aged under 1, survival worsened with age which was observed in the univariable as well as multivariable analyses. Survival of leukaemia varied by diagnostic subgroup, with 5-year survival rates being highest for lymphoid leukaemia (80%) and myelodysplastic syndrome (89%) and poorer for AML (55%) and other and unspecified leukaemias (55%). This was confirmed in the univariable and multivariable analysis. The univariable results indicated slightly poorer 5-year survival of chronic myeloproliferative disease (72%), however, after adjustment for age, year, ethnicity and WBC count, those with chronic myeloproliferative disease were 50% less likely to die compared to those with lymhoid leukaemia. Survival became significantly poorer with the increase of WBC count, however, there was an improvement in survival overall of 7% per year on average during the study period. Ethnicity, deprivation and sex did not have an impact on the survival of leukaemia cases.

For GCTs, TYAs were at a 4-fold increased risk of death compared to children under the age of 15. In general, survival of GCTs was very high, ranging from 82% 5-year survival for intracranial and intraspinal GCTs and other and unspecified GCTs, to 96% 5-year survival for malignant gonadal GCTs. The multivariable model results indicated significant differences in the risk of death according to diagnostic subgroup, with a 20-fold, 10-fold and 15-fold increase in the risk of intracranial and intraspinal GCTs, gonadal carcinomas and other and unspecified GCTs compared to malignant gonadal

GCTs. However, these estimates were inflated and have very wide CIs due to the small number of deaths occurring across the whole diagnostic group, thus there was a large amount of uncertainty associated with the size of these estimates. The stage of tumour increased the risk of death exponentially, with a 5-fold, 16-fold and 32-fold increased risk of death for stage II, III and IV tumours respectively compared to grade I tumours. Again, the CIs surrounding these estimates were very wide. There was no significant effect of ethnicity on survival for GCTs, however, the non-significant increase in the risk of death was 70% and almost 3-fold for Asian and other ethnicity compared to white ethnicity. Although these effects were not significant, the size of the HRs suggest that there could be differences in survival according to ethnicity for GCTs which were not identified due to a lack of study power for this analysis. There were no effects of year of diagnosis, sex or deprivation on the survival of GCTs.

In conclusion, these results show the importance of multiple imputation through highlighting the serious errors in inferences which can occur from ignoring missing data. The novel application of this method to CYA cancer in Yorkshire has highlighted that results from this regional cancer registry could have been misrepresenting important patterns and changes in survival unless missing data was properly accounted for. Although the benefits of imputation were clear, the choice of imputation method was less obvious and this study showed there was limited benefit in using the recently developed SMC-FCS method compared to the more established MICE method. This study has, for the first time, produced survival estimates of CYA cancer which take into account disease severity as well as adequately handling missing data.

Survival for TYAs was poorer compared to that of children for CNS tumours, leukaemias (excluding survival of those aged under 1) and GCTs. There were large and significant decreases in survival according to disease severity for all three tumour groups as expected, highlighting the importance of this prognostic factor, which should not be excluded from survival analyses. There was a significant increase in the risk of death for other ethnicity compared to white ethnicity for CNS tumours, and despite effects not being significant, there was evidence of increased risk of death for Asian and other ethnic groups compared to white ethnicity amongst leukaemia and GCT cases. There were significant improvements in survival over the study period for CNS tumours and leukaemia, however, survival of GCTs did not change significantly over time. Finally, there were no significant differences between males and females or according to deprivation for any of the tumour groups studied.

Despite survival of CYA cancer being high in general, survival for certain subgroups of society remain poor even after adjustment for disease severity. Moreover, those with advanced disease at diagnosis were at a significant survival disadvantage. Ethnicity and

deprivation have both been shown to affect the severity of disease at presentation amongst adult cancers [95, 96, 98, 99, 100]. However, very little is known about the factors affecting disease severity at diagnosis for the CYA population and whether these are the same as seen amongst adults. Chapter 7 uses multiply imputed data arising from this chapter to determine, for the first time, whether age, ethnicity and deprivation play a role in the disease severity at diagnosis for CYAs in Yorkshire. Evidence of inequalities in disease severity at diagnosis could highlight potential inequalities in the quality of access to healthcare which could ultimately affect the survival of CYAs diagnosed with cancer. In addition, as survival has been shown to improve consistently for most tumour groups in Yorkshire, Chapter 8 focuses on one aspect of the quality of that survival, by quantifying the long term cardiovascular effects for a cohort of long term survivors of CYA cancer.

Chapter 7

Inequalities in Disease Severity at Diagnosis

7.1 Introduction

Advanced stage disease at diagnosis has clear survival implications for CYAs with cancer as evidenced by the analyses presented in Chapter 6. Furthermore, despite continued improvements in survival over time, outcomes remain significantly poorer for TYAs compared to children and for non-white ethnicity (Chapter 6).

For adults with cancer, ethnicity and deprivation have both been shown to be associated with disease severity at presentation as described in detail in Chapter 2 [95, 96, 98, 99, 100].

To date, no research in the UK has focused on the effects of ethnic group or deprivation on the stage or severity of cancer at diagnosis for CYAs. The results presented in this chapter aimed, for the first time, to determine whether inequalities in disease severity exist for the CYA cancer population in Yorkshire by assessing the impact of ethnicity (white, Asian, other) and deprivation (IMD quintile) upon disease severity at diagnosis. In addition, the impact of age group at diagnosis (children vs. TYAs) upon disease severity was also determined. The analysis presented here was based on multiply imputed data for missing values of ethnicity and disease severity for CNS tumours, leukaemia and GCTs as described in detail in Chapter 6. A detailed description of the methods used for this analysis is given in Chapter 4.

7.2 Disease Severity by Age, Ethnicity and Deprivation

Disease severity was missing in 9%, 22% and 46% of CNS tumour, leukaemia and GCT cases (Table 7.1). After imputation, the majority of CNS tumours, leukaemias and GCTs were low grade, standard risk or early stage respectively across the cohort (Table 7.1). The pattern mirrored that of the observed data prior to imputation.

For CNS tumours, there was little difference in the proportion of low grade tumours (grades I and II) between children and TYAs (Table 7.2). However, TYAs tended to have relatively more grade III tumours and fewer grade IV tumours compared to children. Similarly, there was little difference in the proportion of low grade tumours between ethnic groups, however, those of Asian ethnicity tended to be diagnosed with a higher proportion of grade IV tumours and fewer grade III tumours compared to white and other ethnic groups. There was little deviation from the average grade distribution by deprivation level for CNS tumours.

For leukaemia, there was no difference in the distribution of standard and high risk cases according to age group, with approximately 75% of cases being diagnosed with standard risk leukaemia (WBC of $< 50,000 \mu/L$) (Table 7.3). There was a smaller proportion of more advanced malignancies (high risk WBC count) amongst the 'other' ethnic group (13%) compared to white and Asian ethnic groups (24% and 21% respectively). There was little variation in disease severity according to deprivation level, although there was a slight increase in the proportion of high risk leukaemias in the least deprived group (quintile 1).

For GCTs, children were diagnosed with twice as many advanced stage tumours compared to TYAs (10% and 23% compared to 5% and 11% stage III and IV tumours for children and TYAs respectively) (Table 7.4). Furthermore, there was a higher than average proportion of advanced tumours amongst the Asian ethnic group (9% and 25% vs. 6% and 12% for stage III and IV tumours respectively). Similarly, there was a higher than average proportion of stage III tumours amongst 'other' ethnicity (13% vs. 6%), and only a small increase in stage IV tumours (14% vs. 12%). According to deprivation level, there appeared to be a decreasing proportion of stage III tumours with increasing deprivation, and a similar pattern (although not as clear) was observed amongst stage IV tumours.

Table 7.1: Number and percentage of observed and imputed cases of disease severity at diagnosis by tumour group

Disease Severity	**Observed**		**Imputed**[1]	
Central nervous system tumours				
Grade	**N**	**(%)**	**N**	**(%)**
I	319	(40.1)	354.0	(44.5)
II	163	(20.5)	178.9	(22.5)
III	46	(5.8)	53.8	(6.8)
IV	196	(24.7)	208.3	(26.2)
Missing	71	(8.9)	-	
Total	795		795	
Leukaemia				
WBC Count				
$< 50,000\mu/L$	543	(59.5)	695.3	(76.2)
$\geq 50,000\mu/L$	171	(18.8)	216.7	(23.8)
Missing	198	(21.7)	-	
Total	912		912	
Germ cell tumours				
Stage				
I	296	(35.0)	539.4	(63.8)
II	91	(10.8)	152.9	(18.1)
III	22	(2.6)	48.2	(5.7)
IV	51	(6.0)	105.6	(12.5)
Missing	386	(45.6)	-	
Total	846		846	

[1]Average number of cases over all imputations (m=40 for central nervous system tumours and leukaemia, m=60 for germ cell tumours)

Table 7.2: The distribution of WHO grade for central nervous system (CNS) tumours by age group at diagnosis, ethnicity and deprivation after multiple imputation by chained equations

	WHO Grade								
	I		II		III		IV		Total
Age at Diagnosis	**N**[a]	**%**	**N**[a]	**(%)**	**N**[a]	**(%)**	**N**[a]	**(%)**	
0-14 years	191.4	43.9	83.4	19.1	28.2	6.5	133.1	30.5	436.0
15-29 years	162.6	45.3	95.6	26.6	25.7	7.1	75.2	20.9	359.0
Ethnicity									
White	318.4	44.3	165.4	23.0	48.7	6.8	186.1	25.9	718.6
Asian	17.7	36.5	11.3	23.3	1.6	3.3	17.8	36.9	48.3
Other	18.0	63.8	2.3	8.3	3.5	12.4	4.4	15.5	28.1
Deprivation									
1 (Least Deprived)	51.8	38.4	32.7	24.2	9.2	6.8	41.4	30.6	135.0
2	74.6	44.1	35.5	21.0	14.8	8.8	44.2	26.1	169.0
3	57.5	43.2	38.7	29.1	6.5	4.9	30.4	22.8	133.0
4	65.9	47.7	28.1	20.4	6.4	4.6	37.7	27.3	138.0
5 (Most Deprived)	104.3	47.4	44.1	20.0	17.0	7.7	54.7	24.9	220.0
Total	354	44.5	179	22.5	54	6.8	208	26.2	795

[a] Average number of cases over m=40 imputations for partially observed variables WHO grade and ethnicity.

Table 7.3: The distribution of low and high risk white blood cell (WBC) count for leukaemia by age group at diagnosis, ethnicity and deprivation after multiple imputation by chained equations

	White Blood Cell Count				
	Standard Risk $< 50,000\mu/L$		**High Risk** $\geq 50,000\mu/L$		**Total**
Age at Diagnosis	**N**[a]	**%**	**N**[a]	**(%)**	
0-14 years	465.3	76.4	143.7	23.6	609.0
15-29 years	230.0	75.9	73.0	24.1	303.0
Ethnicity					
White	615.2	75.7	197.5	24.3	812.7
Asian	60.5	78.9	16.1	21.1	76.6
Other	19.7	86.7	3.0	13.3	22.8
Deprivation					
1 (Least Deprived)	88.3	71.8	34.7	28.2	123.0
2	129.1	75.5	42.0	24.5	171.0
3	124.1	79.0	33.0	21.0	157.0
4	135.3	81.0	31.7	19.0	167.0
5 (Most Deprived)	218.6	74.3	75.4	25.7	294.0
Total	695.3	76.2	216.7	23.8	912

[a] Average number of cases over m=40 imputations for partially observed variables white blood cell count and ethnicity.

Table 7.4: The distribution of stage for germ cell tumours (GCTs) by age group at diagnosis, ethnicity and deprivation after multiple imputation by chained equations

	Stage								
	I		II		III		IV		Total
Age at Diagnosis	**N**[a]	**%**	**N**[a]	**(%)**	**N**[a]	**(%)**	**N**[a]	**(%)**	
0-14 years	46.9	49.9	15.9	16.9	9.4	10.0	21.7	23.1	94.0
15-29 years	492.5	65.5	137.0	18.2	38.7	5.1	83.8	11.1	752.0
Ethnicity									
White	499.3	64.7	143.2	18.5	40.7	5.3	88.9	11.5	772.0
Asian	30.3	53.5	6.9	12.2	5.2	9.2	14.2	25.1	56.6
Other	9.9	56.6	2.9	16.4	2.2	12.6	2.5	14.4	17.4
Deprivation									
1 (Least Deprived)	65.9	56.8	23.6	20.3	8.1	7.0	18.4	15.9	116.0
2	103.9	62.9	30.9	18.7	12.3	7.4	18.0	10.9	165.0
3	104.6	61.9	34.8	20.6	9.4	5.5	20.3	12.0	169.0
4	117.9	69.8	21.5	12.7	7.6	4.5	22.0	13.0	169.0
5 (Most Deprived)	147.2	64.8	42.1	18.6	10.9	4.8	26.9	11.8	227.0
Total	326.3	64.9	87.2	17.3	29.2	5.8	60.3	12.0	846

[a] Average number of cases over m=60 imputations for partially observed variables stage and ethnicity.

7.2.1 Predictors of Advanced Stage Disease for CYAs with Cancer

The independent models of age group, ethnicity and deprivation (adjusted for sex and year of diagnosis) showed that there was no significant effect of these three factors upon disease severity at presentation for CNS tumours, leukaemias or GCTs (Table 7.5). This finding remained consistent after mutual adjustment of age, ethnicity and deprivation in the fully adjusted model (Table 7.5). For GCTs, the proportion of late stage tumours amongst children was double that of TYAs (Table 7.4, §7.2). Furthermore, there was a non-significant 40% decrease in the likelihood of being diagnosed with a late stage tumour for TYAs compared to children in the age model (OR = 0.57, 95% CI 0.23-1.40, P=0.220) and the fully adjusted model (OR = 0.60, 95% CI 0.24-1.49, P=0.268) (Table 7.5). The non-significant effect may be a reflection of the age group distribution for GCTs, as just 11% of GCTs were diagnosed amongst children and the remaining 89% amongst TYAs.

There was evidence of a decrease in the number of later stage diagnoses of GCTs by 4% on average over the study period (OR = 0.96, 95% CI 0.92-1.00, P=0.033), but similar patterns for CNS tumours and leukaemias were not observed. Finally, there was borderline significant evidence to suggest that females were 22% less likely to be diagnosed with higher grade CNS tumours compared to males (OR = 0.78, 95% CI 0.59-1.02, P=0.069 in the fully adjusted model). Model diagnostic checks for the fully adjusted models were performed in the next section (§7.2.2).

Table 7.5: Pooled odd ratios (OR) and 95% confidence intervals (CI) for being diagnosed with higher WHO grade central nervous system (CNS) tumours, high risk white blood cell (WBC) count or later stage germ cell tumours (GCTs) by age at diagnosis, ethnicity and deprivation for children and young adults with cancer in Yorkshire, 1990-2009

Model		CNS tumours			Leukaemia			GCT		
		OR[a]	CI	P-value	OR[b]	CI	P-value	OR[a]	CI	P-value
Age[c]	Age at Diagnosis									
	0-14 years	1			1			1		
	15-29 years	0.81	(0.60-1.08)	0.153	1.00	(0.66-1.53)	0.984	0.57	(0.23-1.40)	0.220
Ethnicity[c]	Ethnicity									
	White	1			1			1		
	Asian	1.54	(0.84-2.82)	0.161	0.81	(0.42-1.56)	0.528	1.82	(0.76-4.32)	0.175
	Other	0.53	(0.22-1.28)	0.158	0.41	(0.07-2.28)	0.307	1.31	(0.24-7.19)	0.753
Deprivation[c]	Deprivation									
	1 (Least Deprived)	1			1			1		
	2	0.82	(0.53-1.25)	0.351	0.84	(0.48-1.48)	0.548	0.77	(0.43-1.38)	0.376
	3	0.76	(0.49-1.20)	0.238	0.67	(0.36-1.25)	0.210	0.78	(0.42-1.44)	0.428
	4	0.70	(0.44-1.12)	0.140	0.60	(0.33-1.11)	0.104	0.57	(0.30-1.08)	0.084
	5 (Most Deprived)	0.75	(0.50-1.12)	0.154	0.89	(0.54-1.48)	0.657	0.74	(0.41-1.34)	0.313
Fully Adjusted	Age at Diagnosis									
	0-14 years	1			1			1		
	15-29 years	0.79	(0.59-1.06)	0.122	1.02	(0.67-1.57)	0.917	0.60	(0.24-1.49)	0.268
	Ethnicity									
	White	1			1			1.00		
	Asian	1.69	(0.89-3.21)	0.109	0.76	(0.39-1.50)	0.431	1.93	(0.80-4.66)	0.141
	Other	0.56	(0.23-1.35)	0.195	0.43	(0.08-2.41)	0.335	1.37	(0.24-7.71)	0.721
	Deprivation									
	1 (Least Deprived)	1			1			1		
	2	0.82	(0.53-1.26)	0.362	0.84	(0.47-1.49)	0.550	0.77	(0.43-1.39)	0.387
	3	0.75	(0.48-1.18)	0.214	0.69	(0.37-1.28)	0.235	0.77	(0.42-1.43)	0.414
	4	0.69	(0.43-1.10)	0.120	0.62	(0.33-1.13)	0.119	0.56	(0.29-1.06)	0.074
	5 (Most Deprived)	0.70	(0.46-1.08)	0.105	0.93	(0.56-1.56)	0.790	0.68	(0.37-1.24)	0.205
	Sex									
	Male	1			1			1		
	Female	0.78	(0.59-1.02)	0.069	1.00	(0.70-1.43)	0.987	1.40	(0.77-2.55)	0.270
	Year of Diagnosis	0.99	(0.96-1.01)	0.234	1.03	(0.99-1.03)	0.113	0.96	(0.92-1.00)	0.033

[a]Pooled OR based on ordered logistic regression for grade/stage of tumour (I-IV) over 40 and 60 imputations for CNS tumours and GCTs respectively
[b]Pooled OR based on logistic regression for high vs. standard risk WBC risk stratification over 40 imputations.
[c]Individual models for age, ethnicity and deprivation were additionally adjusted for sex and year of diagnosis

7.2.2 Model Assessment

As described in §4.4.4.2, all fully adjusted models presented in Table 7.5 were checked for model fit using the HL-GOF test averaged over all imputations. For all three tumours groups, the HL-GOF test provided no evidence to suggest poor model fit (P >0.05) overall. There was some suggestion of poor model fit for some of the individual imputations, with 1 out of 40 models for CNS tumours (P=0.025) and 1 out of 40 models for leukaemias showing poor fit (P=0.007). For GCTs, out of 60 imputations, 3 models displayed poor fit (P=0.022, P=0.028 and P=0.029), however this was not an unexpected finding over a large number of imputations and therefore there was no concern over lack of overall model fit for any of the tumour groups.

The mean discriminative power over all imputations was tested for individual components of the ordered logistic regression models for CNS tumours and GCTs (see methods §4.4.4.2) and were approximately 60% and 40% for CNS tumours and GCTs respectively (Table 7.6). For leukaemia, the model was a logistic regression model with binary outcome, therefore the mean discriminative power was calculated directly for the model of interest, and was 57%. These results indicate poor discriminative power for all three models, highlighting that the variables studied (age, ethnicity, deprivation, sex and year of diagnosis) were not strongly predictive of the disease severity levels and that there were other, unknown or unmeasured variables which may have a greater impact on the disease severity at diagnosis for CYAs with CNS tumours, leukaemias and GCTs. This was consistent with the model results which showed that age, ethnicity, deprivation, sex and year of diagnosis were not, in general, predictive of disease severity.

Table 7.6: Discriminative power (displayed as a percentage) for a logistic regression model for leukaemia risk and for component parts of ordered logistic regression models for disease severity for CNS tumours and GCTs (see Table 7.5)

Model	**Discriminative Power (%)**	
	Mean	**95% CI**
CNS tumours		
Binary outcome: Grade I vs. Grade II-IV	59%	(55%-63%)
Binary outcome: Grade I-II vs. Grade III-IV	59%	(53%-63%)
Binary outcome: Grade I-III vs. Grade IV	61%	(56%-65%)
Leukaemia		
Binary outcome: standard risk vs. high risk	57%	(52%-62%)
GCT		
Binary outcome: Stage I vs. Stage II-IV	39%	(33%-44%)
Binary outcome: Stage I-II vs. Stage III-IV	40%	(34%-46%)
Binary outcome: Stage I-III vs. Stage IV	42%	(33%-51%)

7.3 Summary

The results presented here form the first study in the UK to assess inequalities in disease severity at diagnosis for CYAs diagnosed with cancer. The use of advanced missing data techniques applied to a comprehensive population based cancer register allowed for a detailed analyses of the effects of age, ethnicity and deprivation upon the severity of disease at diagnosis for CYAs diagnosed with cancer in Yorkshire between 1990 and 2009.

Overall, the proportion of early stage tumours was higher across all three tumour groups when compared to advanced stage tumours. There was some evidence in the univariable analysis of an inverse relationship between disease severity and deprivation for leukaemia and GCTs such that more advanced tumours tended to occur in the least deprived groups. In addition, the univariable analysis indicated that despite poorer survival for the 'other' ethnic group amongst CNS tumours, they tended to be diagnosed with less advanced tumours compared to those of white or Asian ethnicity. A similar pattern was observed for leukaemia.

After adjusting for sex and year of diagnosis there was no evidence of a difference in disease severity at diagnosis according to ethnic group or deprivation for CNS tumours, leukaemia or GCTs. Importantly, the results indicate that in contrast to adults with cancer, ethnicity and deprivation did not appear to influence inequalities in disease severity for CYAs with cancer. This is supportive of the findings by Lightfoot et al. [65], which focused on childhood leukaemia, in which differences in survival by deprivation for children were thought to be attributed to treatment adherence rather than to inequalities in disease severity.

Furthermore, the Yorkshire register provided the unique opportunity to assess, for the first time, whether there were any differences in disease severity at diagnosis between children and TYAs. For CNS tumours and leukaemia, there was no evidence to suggest that children and TYAs differed in terms of their disease severity at diagnosis with cancer. Despite a two-fold increase in the proportion of children with advanced stage GCTs compared to TYAs, there was no significant evidence of later stage diagnoses for children compared to TYAs in the multivariable analyses. This was likely to be due to the small proportion of children diagnosed with GCTs compared to TYAs, and further larger scale studies are required to provide conclusive evidence of inequalities in disease severity for children and TYAs who develop GCTs. Importantly, there was significant evidence of a reduction in the chance of later stage diagnoses by 4% on average between 1990 and 2009 for GCTs. However, similar advances were not observed for CNS tumours or leukaemias. Finally, there was borderline significant evidence that males tended to

be diagnosed with more advanced grade CNS tumours compared to females, however, further study is required to determine the strength of this evidence.

Chapter 8

Long Term Effects amongst Survivors of Cancer

The following publications have arisen from the analysis and results in this chapter:

1 van Laar, M., R.G. Feltbower, C.P. Gale, D.T. Bowen, S.E. Oliver, and A. Glaser. (2014). Cardiovascular sequelae in long-term survivors of young peoples' cancer: a linked cohort study. *British Journal of Cancer* 110(5) 2014:1338-1341.

2 Simms A.D., M. van Laar, R.J. Birch, R.G. Feltbower, C.P. Gale, D.T. Bowen, S.E. Oliver, A. Glaser. (2011). Cardiovascular sequelae in long term survivors of childhood and young adult cancer. *European Heart Journal* 32:544-544. Conference Abstract.

3 van Laar, M., R.J. Birch, C.P. Gale, A. Glaser, D.T. Bowen, S.E. Oliver, R.G. Feltbower. (2010). Cardiovascular sequelae in long term survivors of childhood and young adult cancer. *Pediatric Blood & Cancer* 55:825-825. Conference Abstract.

8.1 Introduction

As described in Chapter 2, survival rates for CYA cancer are high, thus there is an increasing focus on improving the quality of that survival and the possible long term effects of treatment. This chapter describes the risk of cardiovascular late effects (LEs)

amongst survivors of CYA cancer using linked cancer registry and HES data. Details of the data sources and methodology used are provided in Chapter 4. Preliminary descriptive analysis of the cohort are given in §5.2.2.1, Chapter 5. The results in this chapter begin with a description of cardiovascular LEs (§8.2), followed by two main results sections focusing on the risk of cardiovascular LEs amongst survivors of CYA cancer compared to the general population (§8.2.1) and predictors of cardiovascular LEs (§8.2.2).

8.2 Cardiovascular Hospital Admissions

There were a total of 3247 survivors of CYA cancer (n=1367; 42% and n=1880; 58% for 0-14 and 15-29 year olds respectively) diagnosed between 1991 and 2006 who successfully linked to at least one inpatient HES record (see §5.2.1 for linkage rate details). There were 119 (3.6%) individuals with at least one cardiovascular LE (n=40 and 79 for 0-14 and 15-29 year olds respectively). The cumulative incidence of cardiovascular LEs was 7.5% (95% CI 5.3%-10.3%) and 14.0% (95% CI 9.9%-18.8%) for 0-14 and 15-29 year olds respectively at 20 years from diagnosis (see Figure 8.1). The majority of cases experienced one distinct cardiovascular LE (64%), 18% experienced two and the remainder experienced between 3 and 12 distinct cardiovascular LEs (Figure 8.2). The median time to a cardiovascular LE was 10.2 years (IQR = 6.8 to 13.4 years) from cancer diagnosis. The time to cardiovascular LE did not vary by cardiovascular event (see Figure 8.3).

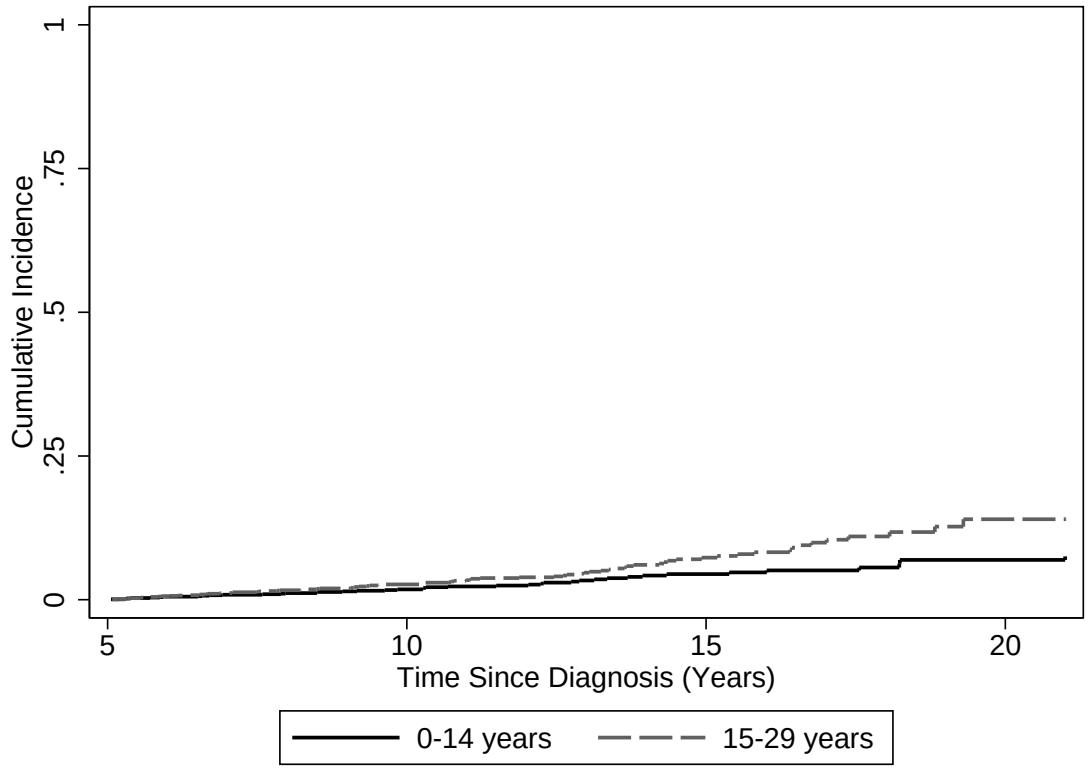


Figure 8.1: Cumulative incidence of cardiovascular late effects amongst cancer survivors by age group at diagnosis, for cancer diagnosis between 1991-2006 and hospital admissions between 1996-2011

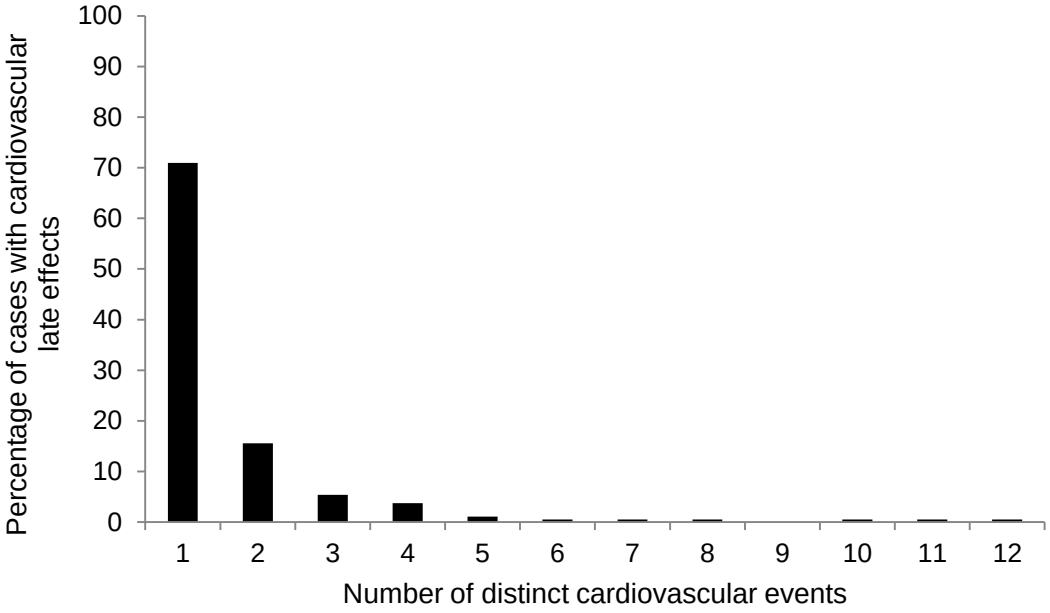


Figure 8.2: Bar graph showing the percentage of distinct cardiovascular late effects for 119 long term survivors of childhood and young adult cancer in Yorkshire, for cancer diagnosis between 1991-2006

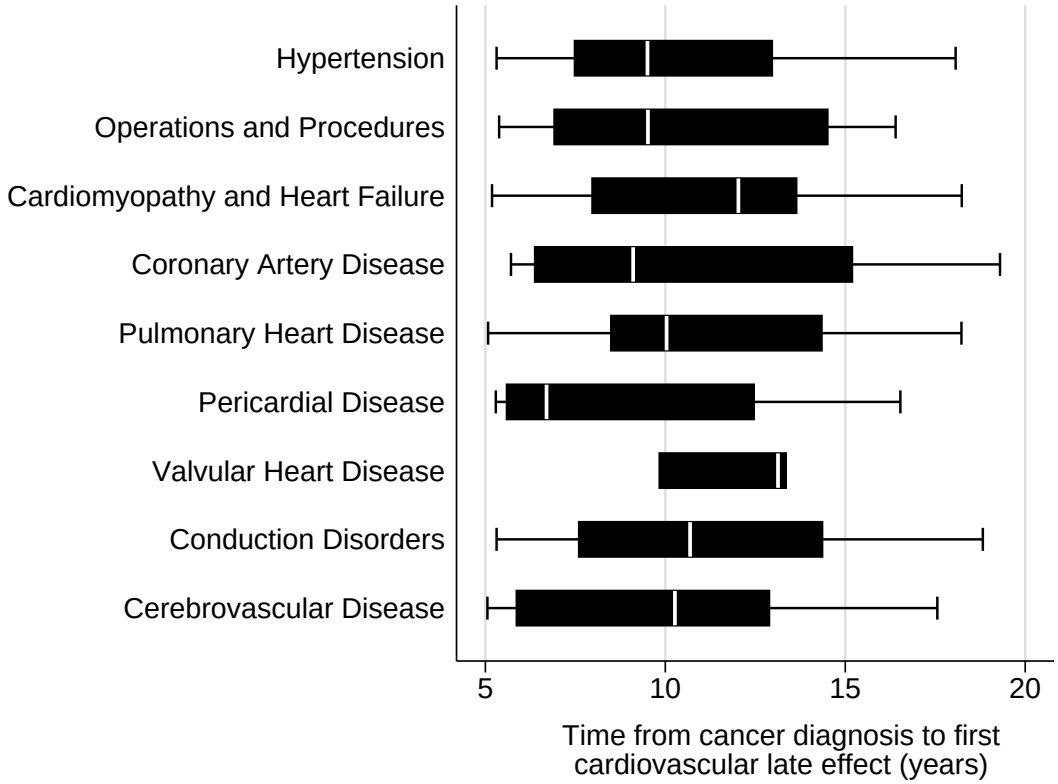


Figure 8.3: Box and whisker plot showing the time from diagnosis to first cardiovascular late effect by cardiovascular diagnosis amongst 119 long term survivors of childhood and young adult cancer in Yorkshire, for cancer diagnosis between 1991-2006

8.2.1 Comparison of Cancer Cohort to the General Population

Table 8.1 contains the number and crude incidence of cardiovascular LEs per 10,000 person-years (pys) amongst cancer survivors and the general population by age group. There were 181 cardiovascular LEs (counting each type of diagnosis once per person) amongst the cancer survivor cohort, compared to 112,118 events in the general population. Overall, the incidence was higher amongst cancer survivors than the general population (51.3 vs. 35.2 per 10,000 pys respectively). Hypertension was the most common type of cardiovascular event for the survivors cohort and the general population (13.3 and 9.9 per 10,000 pys respectively), except amongst the childhood survivor cohort, amongst which cardiomyopathy was most common (8.4 per 10,000 pys) and hypertension second most common (7.8 per 10,000 pys). Overall, the survivor cohort had notably higher incidence of hypertension (13.3 vs. 9.9 per 10,000 pys), cardiomyopathy and heart failure (7.4 vs. 1.8 per 10,000 pys), conduction disorders (6.5 vs. 5.0 per 10,000 pys), cerebrovscular disease (4.5 vs. 2.5 per 10,000 pys), pulmonary heart disease (4.3 vs. 1.8 per 10,000 pys) and pericardial disease (3.1 vs. 0.8 per 10,000 pys) compared to the general population. However, when assessing incidence according to age group, the incidence of hypertension and cerebrovascular disease was only increased compared to the general population for the childhood survivor cohort (7.6 vs. 2.8 per 10,000 pys for hypertension and 5.8 vs. 0.9 per 10,000 pys for cerebrovascular disease), but not the young adult survivor cohort (17.7 vs. 16.8 for hypertension and 3.5 vs. 3.2 for cerebrovascular disease). The incidence of cardiovascular operations and procedures (7.0 vs. 6.8 per 10,000 pys) and valvular heart disease (1.1 vs. 1.3 per 10,000 pys) was similar amongst the survivor cohort and general population, however, the incidence of coronary artery disease was lower amongst the survivor cohort compared to the general population (4.0 vs. 5.4 per 10,000 pys).

Figure 8.4a shows the excess risk of cardiovascular events amongst the cancer survivor cohort compared to the general population represented as hospitalisation rate ratios (HRRs). The rate of cardiovascular hospitalisations was higher for the survivor cohort compared to the general population overall (HRR = 1.4, 95% CI 1.2-1.7). Furthermore, there was a significant increase in the risk of hospitalisation for hypertension (HRR = 2.1, 95% CI 1.6-2.8), operations and procedures (HRR = 1.6, 95% CI 1.1-2.3), cardiomyopathy and heart failure (HRR = 5.9, 95% CI 4.0-8.6), pulmonary heart disease (HRR = 3.4, 95% CI 2.0-5.7), pericardial disease (HRR = 5.1, 95% CI 2.8-9.3), conduction disorders (HRR = 1.9, 95% CI 1.3-2.9) and cerebrovascular disease (HRR = 2.8, 95% CI 1.2-1.7).

Results by age group, Figure 8.4b, show that the rate of cardiovascular hospitalisations was higher for the childhood cohort compared to the general population overall (HRR = 2.6, 95% CI 1.9-3.6), but not for the young adult cohort (HRR = 1.2, 95% CI 0.91.5).

Amongst the younger cohort, there was a significant increased risk of cardiomyopathy and heart failure (HRR=12.7, 95% CI 7.4-21.9), cerebrovascular disease (HRR=7.9, 95% CI 4.1-15.2), pericardial disease (HRR=7.9, 95% CI 3.3-19.0), hypertension (HRR=4.0 95% CI 2.3-7.1), valvular heart disease (HRR=3.2, 95% CI 1.0-10.0) and operations and procedures (HRR=2.2, 95% CI 1.0-4.5). Despite no significant increased hospitalisation rate for young adults overall, there was a significant increase in the hospitalisation rate of pericardial disease (HRR=4.0, 95% CI 1.8-8.8), cardiomyopathy and heart failure (HRR=3.8, 95% CI 2.2-6.6), pulmonary heart disease (HRR=3.5, 95% CI 2.0-6.4), conduction disorders (HRR=2.0, 95% CI 1.2-3.2) and hypertension (HRR=1.8, 95% CI 1.3-2.5) in this age group.

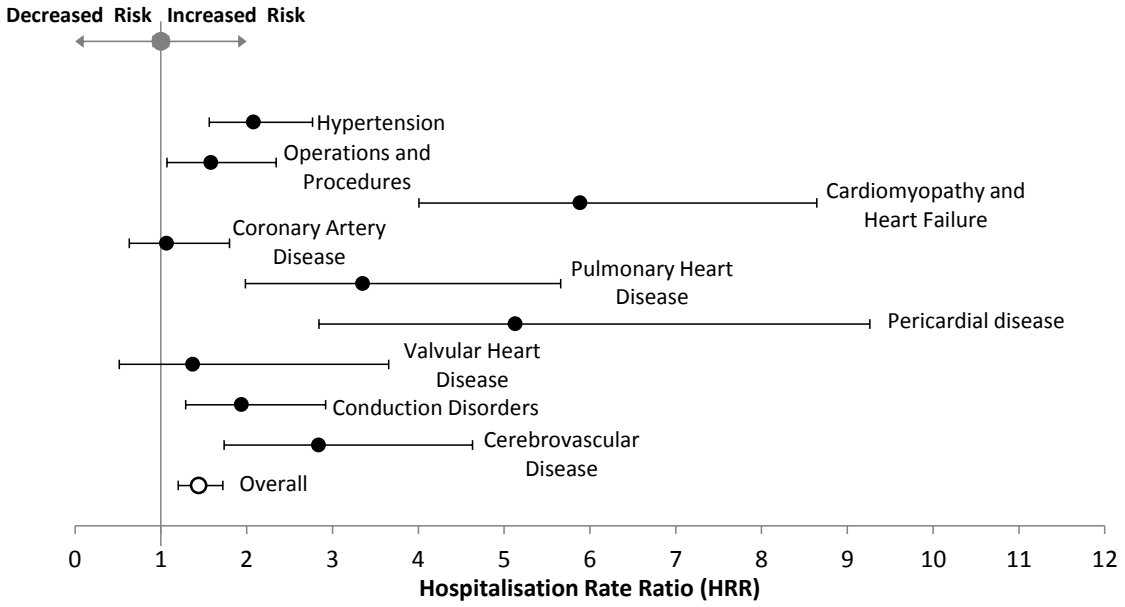
Table 8.1: Number of cardiovascular late effects[a] and crude incidence per 10,000 person-years (pys) by cardiovascular category and age group at cancer diagnosis for hospital admissions between 1996 and 2011

	Age at Diagnosis[b]					
Cardiovascular Category	**0-14 years**		**15-29 years**		**Total**	
	Cancer Survivors	**General Population**	**Cancer Survivors**	**General Population**	**Cancer Survivors**	**General Population**
	N (incidence/10,000 pys)		**N (incidence/10,000 pys)**		**N (incidence/10,000 pys)**	
Hypertension	12 (7.75)	4009 (2.57)	35 (17.68)	27293 (16.80)	47 (13.32)	31580 (9.91)
Cardiomyopathy & Heart failure	13 (8.39)	1343 (0.86)	13 (6.57)	3681 (2.27)	26 (7.37)	5786 (1.82)
Operations & Procedures	7 (4.52)	3413 (2.19)	18 (9.09)	11689 (7.19)	25 (7.08)	21497 (6.75)
Conduction Disorders	6 (3.87)	4740 (3.04)	17 (8.59)	10664 (6.56)	23 (6.52)	15790 (4.96)
Cerebrovascular Disease	9 (5.81)	1479 (0.95)	7 (3.54)	5252 (3.23)	16 (4.53)	7816 (2.45)
Pulmonary Heart Disease	3 (1.94)	1606 (1.03)	12 (6.06)	4599 (2.83)	15 (4.25)	5622 (1.76)
Coronary Artery Disease	2 (1.29)	847 (0.54)	12 (6.06)	10080 (6.20)	14 (3.97)	17328 (5.44)
Pericardial Disease	5 (3.23)	1386 (0.89)	6 (3.03)	2592 (1.60)	11 (3.12)	2661 (0.84)
Valvular Heart Disease	3 (1.94)	1032 (0.66)	1 (0.51)	1986 (1.22)	4 (1.13)	4038 (1.27)
Total[c]	60 (38.73)	19855 (12.71)	121 (61.12)	77836 (47.90)	181 (51.29)	112118 (35.19)
Total Person-Years	15492.43	15613297	19796.53	16246855	35288.96	31860152

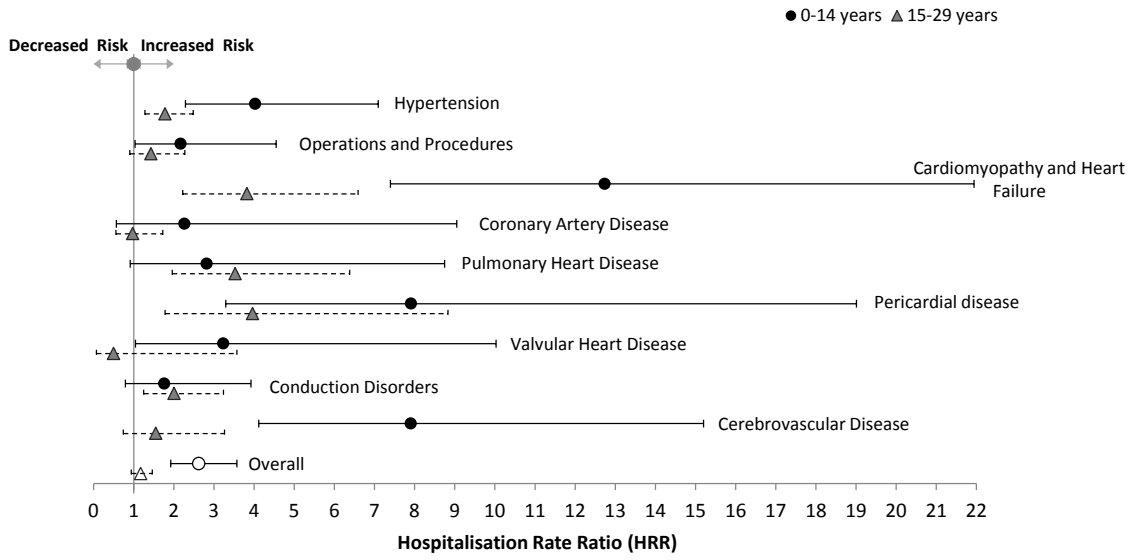
[a]For each case, multiple occurrences of the same diagnosis were not counted.

[b]Age at admission for general population corresponds to age of survivor cohort at admission dependant on their age at diagnosis.

[c]Total number of events does not equal the total number of cases as 30% of cases experienced multiple cardiovascular diagnoses.



(a)



(b)

Figure 8.4: Hospitalisation rate ratios and 95% confidence intervals comparing cardiovascular late effects amongst cancer survivors to the general population overall (a) and by age group (b), 1996-2011

8.2.2 Predictors of Cardiovascular LEs

There was significant evidence of an increased risk of cardiovascular LEs for those diagnosed aged 15-29 years compared to those diagnosed aged 0-14 years in the unadjusted analysis (HR=1.69, 95% CI 1.16-2.48 P-value=0.007) and a significant

reduced risk of cardiovascular LEs by 4% on average over the study period (HR=0.94, 95% CI 0.88-0.99, P-value=0.020); no further differences in demographic and clinical variables were observed (Table 8.2).

Table 8.3 gives the excess hazard ratios (EHR) for the risk of cardiovascular LEs adjusting for the rate of cardiovascular events in the background population. The EHR of age at diagnosis was 1.19 (95% CI 0.89-1.60), however, this effect was not significant. Furthermore, the EHR for year of diagnosis was very close to 1 and not significant (EHR=1.02, 95% CI 0.96-1.08), implying that although rates of cardiovascular LEs were decreasing for the cancer survivor cohort, this was mirrored by a decrease in cardiovascular events for the background population. The risk of cardiovascular LEs for those diagnosed with leukaemia (EHR=1.37, 95% CI 0.92-2.04), lymphoma (EHR=1.23, 95% CI 0.81-1.87) and CNS tumours (EHR=1.28, 95% CI 0.81-2.02) were all higher when compared to other solid tumours, however, non of the effects were statistically significant. In addition, despite large EHR for sex (EHR=1.19, 95% CI 0.89-1.60) there was no significant difference in cardiovascular LEs according to sex. Subgroup analysis model (i) including only survivors who received chemotherapy to assess the impact of the number of anthracyline drugs administered is given in Table 8.4. There was no significant difference in the risk of cardiovascular LEs according to the number of anthracycline drugs administered (EHR=0.77, 95% CI 0.53-1.12, P-value=0.171). Furthermore, there were no significant effects for age at diagnosis or year of diagnosis as in the main analysis (Table 8.3). However, there was borderline significant evidence at the 10% level of females having a higher risk of cardiovascular LEs compared to males (EHR=1.37, 95% CI 0.94-1.99). Subgroup analysis model (ii) including only survivors who received radiotherapy is given in Table 8.5. There was borderline significant evidence of an increased the risk of cardiovascular LEs (excluding cerebrovascular disease) for those who received radiotherapy to the chest (EHR=7.36, 95% CI 0.97-55.7, P-value=0.053). Furthermore, there was borderline significant evidence at the 10% level of a large increased risk in cardiovascular LEs for females compared to males (EHR=12.21, 95% CI 0.80-187.19, P-value=0.072) as in the chemotherapy subgroup analysis. The wide CIs surrounding both of these estimates were a reflection of the small number of cases in this subgroup analysis, which included 812 cases of whom 31 had a cardiovascular LE. Of those with a cardiovascular LE, there were 7 cases who received radiotherapy to the chest and 20 were female.

Table 8.2: Number of cases with and without cardiovascular late effects (LEs) and unadjusted hazard ratios (HRs) obtained from univariable Cox proportional hazards models for the time to cardiovascular LEs amongst survivors of childhood and young adult cancer, diagnosed between 1991 and 2006

	N			HR (95% CI)	P-value
	Late Effect	**No Late Effect**	**Total**		
Sex					
Male	69	1857	1926	1	
Female	50	1271	1321	1.02 (0.71-1.46)	0.933
Age at cancer diagnosis					
0-14 years	40	1327	1367	1	
15-29 years	79	1801	1880	1.69 (1.16-2.48)	0.007
Diagnostic Group[a]					
I: Leukaemia	23	564	587	1	
II: Lymphoma	32	681	713	1.10 (0.65-1.89)	0.719
III: CNS tumours	16	474	490	0.83 (0.44-1.57)	0.567
IV-XII: Other solid tumours	48	1409	1457	0.87 (0.53-1.43)	0.583
Year of Diagnosis	119	3128	3247	0.94 (0.88-0.99)	0.020
Treatment					
Chemotherapy (± Surgery)	41	1201	1242	1	
Radiotherapy (± Surgery)	9	312	321	0.70 (0.34-1.45)	0.340
Chemo & Radiotherapy (± Surgery)	27	464	491	0.86 (0.52-1.41)	0.543
Surgery only or no treatment	42	1151	1193	0.87 (0.57-1.34)	0.532
Number of Anthracyclines					
Zero	88	2089	2177	1	
One	21	835	856	0.70 (0.43-1.12)	0.137
Two	10	199	209	1.52 (0.79-2.93)	0.209
Three	0	5	5	-	-
Radiation to the Chest					
No	112	3026	3138	1	
Yes	7	102	109	1.64 (0.76-3.52)	0.204
Cranial Radiation					
No	108	2905	3013	1	
Yes	11	223	234	0.98 (0.53-1.83)	0.956
Radiation to the Neck					
No	111	2951	3062	1	
Yes	8	177	185	1.00 (0.49-2.06)	0.994
Deprivation Fifths[b]					
Most Deprived (5)	45	1062	1104	1	
4	19	653	672	0.72 (0.42-1.24)	0.230
3	24	587	611	1.02 (0.62-1.69)	0.484
2	25	500	525	1.19 (0.73-1.96)	0.927
Least Deprived (1)	9	326	335	0.64 (0.31-1.33)	0.235
Total	119	3128	3247		

[a]International classification of childhood cancer; [b]Index of Multiple Deprivation 2007

Table 8.3: Excess hazard ratios (EHR) and 95% confidence intervals (CI) obtained from a Royston-Parmar relative survival model (probit scale and 1 degree of freedom), modelling the risk of a cardiovascular late effects for long term survivors of cancer diagnosed between 1991 and 2006 aged 0-14 and 15-29 years inclusive

Variable	**EHR**	**95% CI**		P**-value**
		lower	**upper**	
Sex				
Male	1			
Female	1.19	0.89	1.60	0.234
Age at cancer diagnosis				
0-14 years	1			
15-29 years	1.09	0.79	1.51	0.610
Year of Diagnosis	1.02	0.96	1.08	0.508
Diagnostic Group[a]				
IV-XII: Other solid tumours	1			
I: Leukaemia	1.37	0.92	2.04	0.122
II: Lymphoma	1.23	0.81	1.87	0.337
III: CNS tumours	1.28	0.81	2.02	0.284
Treatment				
Surgery only or no treatment	1			
Chemotherapy ($\pm$ Surgery)	1.10	0.78	1.54	0.592
Radiotherapy ($\pm$ Surgery)	0.73	0.30	1.72	0.463
Chemo & Radiotherapy ($\pm$ Surgery)	0.93	0.57	1.50	0.763
Deprivation[b]	0.99	0.99	1.00	0.577

[a]International classification of childhood cancer; [b]Index of Multiple Deprivation 2007 included as a continuous score

Table 8.4: Excess hazard ratios (EHR) and 95% confidence intervals (CI) obtained from a Royston-Parmar relative survival model (probit scale and 1 degree of freedom), modelling the risk of a cardiovascular late effects for long term survivors of cancer who received chemotherapy treatment and were diagnosed between 1991 and 2006

Variable	**EHR**	**95% CI**		P**-value**
		lower	**upper**	
Sex				
Male	1			
Female	1.37	0.94	1.99	0.096
Age at cancer diagnosis				
0-14 years	1			
15-29 years	0.93	0.62	1.39	0.731
Year of Diagnosis	1.02	0.97	1.08	0.834
Number of Anthracyclines[a]	0.77	0.53	1.12	0.221

[a]Number of distinct anthracycline drugs administered modelled as a continuous variable

Table 8.5: Excess hazard ratios (EHR) and 95% confidence intervals (CI) obtained from a Royston-Parmar relative survival model, modelling the risk of a cardiovascular late effects for long term survivors of cancer who received radiotherapy treatment and were diagnosed between 1991 and 2006

Variable	**EHR**	**95% CI**		P**-value**
		lower	**upper**	
Sex				
Male	1			
Female	12.21	0.80	187.19	0.072
Age at cancer diagnosis				
0-14 years	1			
15-29 years	0.49	0.10	2.32	0.368
Year of Diagnosis	0.98	0.83	1.17	0.834
Chest Radiation				
No	1			
Yes	7.36	0.97	55.70	0.053

8.3 Model Assessments

Methods for model assessment are described in §4.4.5. As the flexible parametric RP survival models rely on cubic splines to estimate the baseline hazard function, the scale and the complexity of the splines need to be determined. Models were compared using the AIC and BIC for the proportional hazards (PH) model, proportional odds (PO) model and probit model between 1 and 5 degrees of freedom (df) (Table 8.6). For the PH model and the PO model, the AIC is lowest using 2 df and the BIC is lowest model using 1 df. For the probit model, the AIC and BIC are lowest for the simplest model with 1 df. Overall, the probit model had the lowest AIC and BIC, although the difference in AIC and BIC for PH and PO models compared to the probit model were very small. The probit model was chosen as the most appropriate model, further confirmed by a violation of the PH assumption §8.3.1. In general, BIC values were more sensitive to changes in df when compared to AIC, as differences between AIC values at each degree of freedom were much smaller compared differences in BIC. This is because as the df increase, there is increased model complexity, and BIC has a stronger penalty for model complexity when compared to AIC.

Table 8.6: Choice of baseline complexity based on the number of degrees of freedom (df) for the proportional hazards (PH) model, the proportional odds (PO) model and probit model. The optimal (lowest) AIC and BIC values are underlined for each model.

df	PH Model		PO Model		probit Model	
	AIC	BIC	AIC	BIC	AIC	BIC
1	1330.35	<u>1403.37</u>	1330.15	<u>1403.17</u>	<u>1328.23</u>	<u>1401.25</u>
2	<u>1329.74</u>	1408.85	<u>1329.72</u>	1408.83	1329.66	1408.76
3	1331.37	1416.56	1331.34	1416.53	1331.40	1416.60
4	1332.51	1423.78	1332.51	1423.78	1332.62	1423.90
5	1332.63	1430.00	1332.62	1429.99	1332.54	1429.90

8.3.1 Proportional hazards assumption

The PH assumption of the PH model with 1 df was checked using log cumulative hazard plots for each variable in the model (Figure 8.5). Overall, there did not appear to be a serious violation of the PH assumption except for the variable sex, in which the log cumulative hazard plot crossed at the mid-point on the time axis and thereby violating the PH assumption. This supports the choice of probit model based on the AIC and BIC which does not require a PH assumption (§8.3).

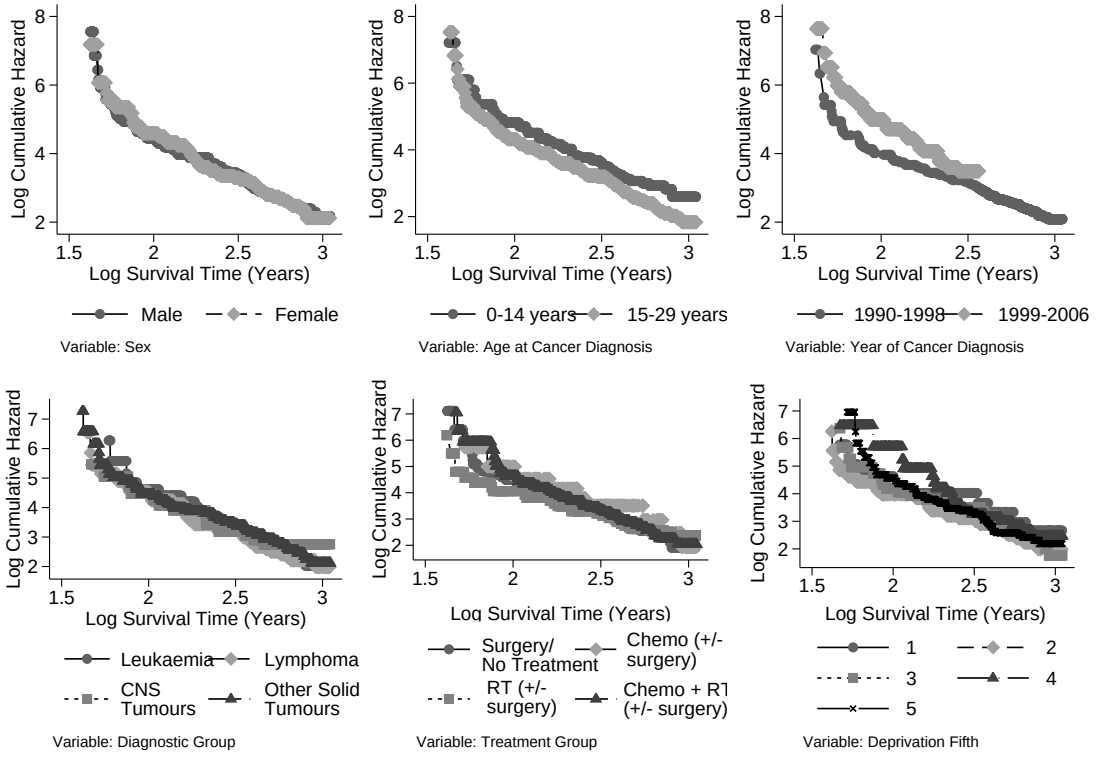


Figure 8.5: Log cumulative hazard plot for the proportional hazards Royston-Parmar survival model by sex, age group, year of diagnosis, diagnostic group, treatment group and Index of Multiple Deprivation

8.3.2 Sensitivity to Model Complexity and Scale

Differences in AIC and BIC shown in §8.3 were minimal, and the log cumulative hazard plots to assess the PH assumption are subjective. Sensitivity to the choice of model was assessed. Figure 8.6 shows there was little to no difference in the relative survival curves between using a 1 or 2 df model for the PH, PO or probit models with all confidence intervals overlapping. Figure 8.7 highlights there was very little difference in relative survival using 1 df between the PH, PO or probit models. Therefore, the models were robust to the choice of scale and df.

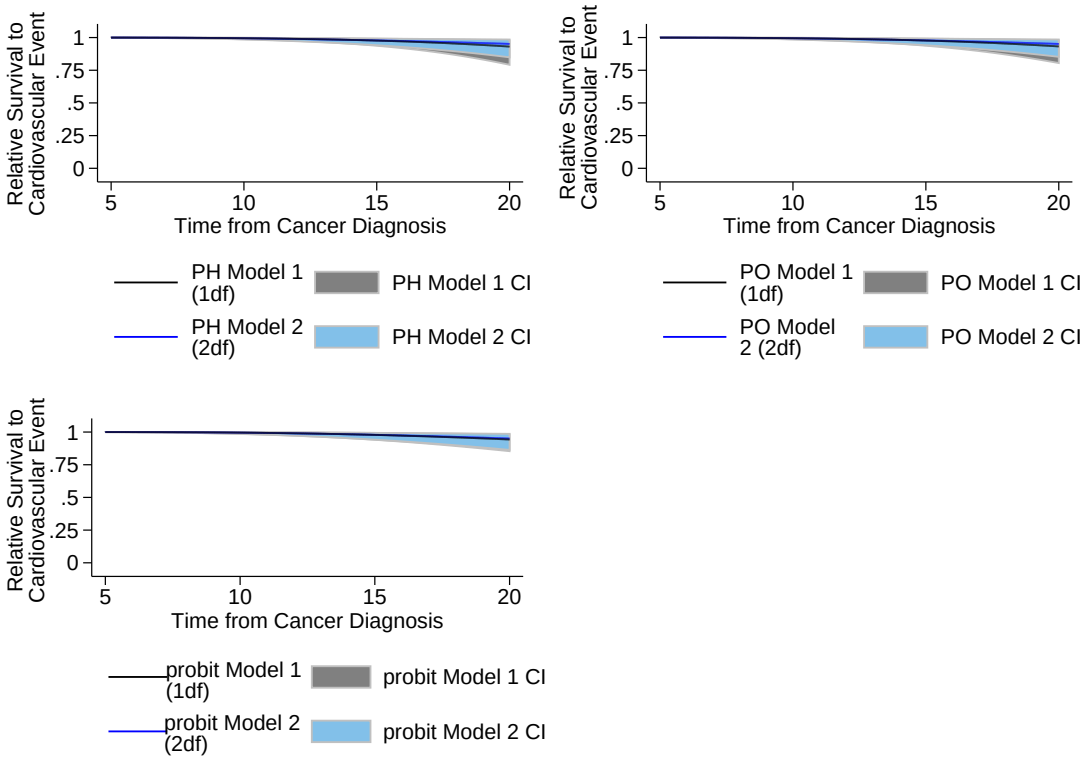


Figure 8.6: Relative survival curves for time to cardiovascular event obtained from multivariate Royston-Parmar relative survival models comparing models with 1 and 2 degrees of freedom (df) using the proportional hazards (PH), proportional odds (PO) and probit scales

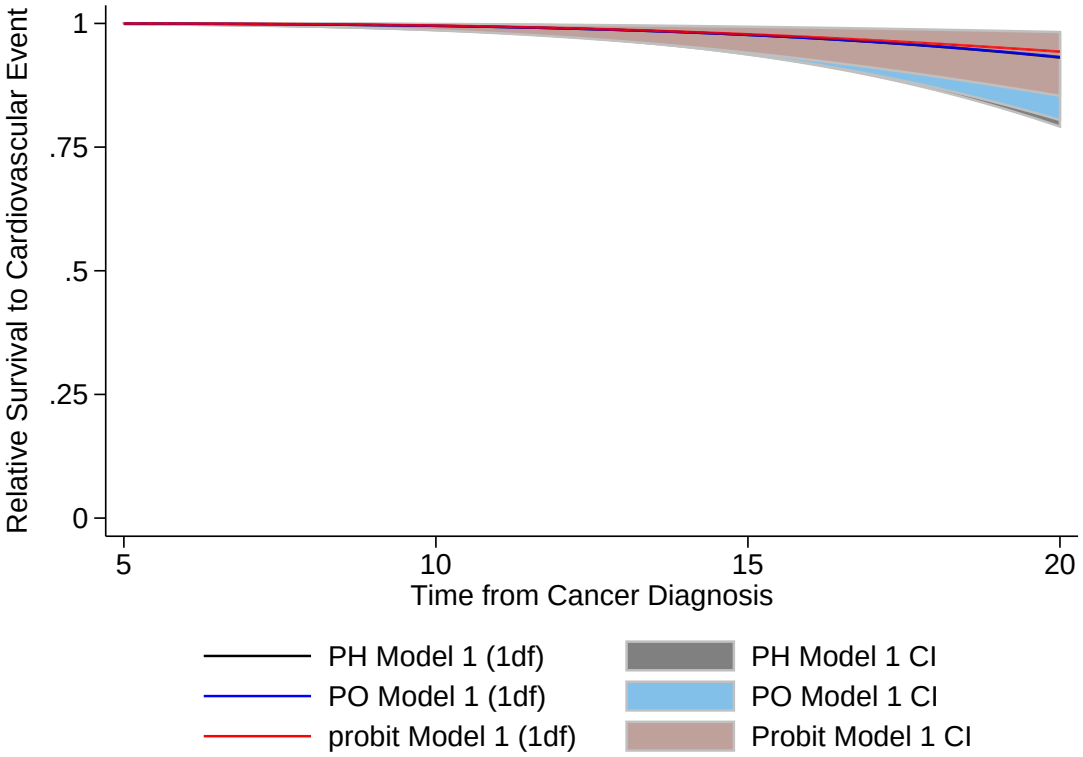


Figure 8.7: Relative survival curves for time to cardiovascular event obtained from multivariate Royston-Parmar relative survival models comparing the proportional hazards (PH), proportional odds (PO) and probit scales with 1 degree of freedom (df)

8.4 Conclusion

The results presented in this chapter provide the first population-based evidence of increased cardiovascular morbidity in a cohort of survivors of CYAs with cancer in Yorkshire through the use of cancer registry data linked to routinely collected inpatient HES data.

There was a significant increase in cardiovascular morbidity in survivors of childhood cancer compared with the general population. For TYAs, this increased burden on survivors was limited to pericardial disease, cardiomyopathy and heart failure, pulmonary heart disease, conduction disorders and hypertension compared to the general population. There was a significant increased risk of cardiovascular LEs for those who received chest radiation, however increases in the risk of cardiovascular LEs according to the use and number of anthracyclines were not observed. The rate of cardiovascular events decreased by 6% on average per year for survivors of CYA cancer, however, this was mirrored by a decreasing trend amongst the general population.

Although survival rates are improving (Chapter 6), the results presented here highlight an ongoing need for survivors of CYA cancer to receive continued monitoring. Awareness of the potential long term effects of cancer need to be raised not only with the survivors themselves, but also to potential future health carers.

Chapter 9 concludes the thesis with a detailed discussion of all the work presented in this thesis, alongside the novel contributions of the work, the study implications, limitations and recommended areas for future research.

Chapter 9

Discussion

9.1 Introduction

Cancer is the leading cause of death amongst CYAs in the UK, and improving outcomes from cancer remains a key national priority [22, 23]. Despite this, there is a paucity of research on CYA outcomes which take into account vital data on the severity of disease at diagnosis. Furthermore, missing data are often ignored, resulting in few conclusive results.

This study set out to improve upon the quality of research addressing variation in cancer outcomes amongst CYAs, and has identified key survival inequalities for a population of CYAs with cancer in Yorkshire. The novel application of multiple imputation to CYA cancer research enabled the first comprehensive population-based study of CYA cancer survival trends over time and in relation to ethnic group whilst adjusting for disease severity. The study has also sought to determine whether inequalities in disease severity according to age, ethnicity and deprivation exist for CYAs with cancer for the first time in the UK. Furthermore, despite a growing population of long term survivors of childhood and TYA cancer, the literature on late effects of cancer has largely overlooked the TYA population of survivors and has relied on self reported outcomes in retrospective cohort studies. This study therefore aimed to objectively quantify the burden of long term health effects, focusing on cardiovascular disease, for long term survivors of cancer for both children and TYAs in Yorkshire based on a unique data linkage approach.

The empirical and novel findings which have arisen from this thesis are summarised below. A detailed discussion and synthesis of these results is given in §9.2, §9.3, §9.4 and §9.5. The chapter concludes with the implications of the study, the strengths and limitations as well as recommendations for future research and an overall conclusion.

The empirical and novel findings arising from this thesis are as follows:

Missing Data and Multiple Imputation

For the first time, the high proportion of missing staging and ethnicity data has been quantified, and trends over time described, for CYAs with cancer in Yorkshire (Chapter 6). Complete case analysis resulted in biased estimates which lacked precision and power when compared to multiple imputation to correct for missingness. Improvements in survival of cancer over time were only evident after the use of multiple imputation.

Variation in Cancer Survival

CYA cancer survival has been described by age group, ethnic group and disease severity for the first time in a comprehensive analysis using appropriate methods to account for missingness. Survival was significantly poorer for TYAs compared to children, for those with advanced stage disease and for other compared to white ethnicity. There was significant improvement in survival over time for CNS tumours and leukaemia, but not GCTs.

Inequalities in Disease Severity at Diagnosis

Differences in disease severity at diagnosis were described for the first time for CYAs with cancer. Despite poorer survival of TYAs compared to children, there was no significant difference in the disease severity at diagnosis between these two age groups. Furthermore, ethnicity and deprivation did not influence the disease severity at diagnosis. Despite no significant improvements in survival over time for GCTs, the number of later stage diagnoses of GCTs decreased by 4% on average per year between 1990 and 2009.

Cardiovascular Late Effects amongst Long Term Survivors of CYA Cancer

In the first study assessing late effects amongst survivors of childhood and TYA cancer, a significant increase in cardiovascular morbidity was observed for childhood and subgroups of the TYA population compared with the general population. The rate of cardiovascular events decreased by 6% on average per year for survivors of CYA cancer, however, this was mirrored by a decreasing trend amongst the general population.

9.2 Missing Data and Multiple Imputation

The Yorkshire register's database system included only a single field for recording stage, and as such, a large number of recorded stage values were invalid when compared to existing staging mechanisms for CYA cancers, which contributed to an already low level of complete and accurate stage data. Further missing stage data could have arisen from doctors failing to record stage within medical notes or due to poor understanding of the information recorded within medical notes by data extractors. Importantly, this study showed that there was no improvement in the completeness of stage data within the

Yorkshire register over the 20 year study period for all cancers. This was in spite of the Department of Health's recognition in 2007, albeit towards the end of the study period, that stage data was a key gap in national cancer intelligence and was recommended as an important area for improvement [1, 22, 33].

This study identified that, in contrast to missing stage data, there was a significant improvement in the level of completeness of ethnicity data obtained from linked HES data records. This trend was primarily a consequence of the availability of HES data from 1996 onwards, with increased missing data at the beginning of the study period being due to a higher likelihood of non-linkage for these cases. However, despite this feature of the study design, data completeness of ethnic group continued to improve to the end of the study period.

In spite of an increase in the number of publications relating to advanced multiple imputation techniques in recent years, many applied health researchers fail to realise the potential problems of adopting inadequate ad-hoc techniques or simply ignoring missing data. There is a wealth of research focusing on survival of cancer across all ages worldwide, however, between 1997 and 2014, just 60 studies implemented MICE. In particular, of the 18 UK studies of childhood and TYA cancer survival identified from a systematic review, none adopted adequate methods for handling missing data, with the majority either performing a CCA or simply neglecting to mention any missing data. The novel application of two advanced multiple imputation techniques, MICE and SMC-FCS, highlighted for the first time in CYA cancer outcomes research the serious implications of ignoring missing data through comparisons with a CCA.

An analysis based on complete cases resulted in biased estimates, reduced power to identify significant effects and reduced precision. In particular, the study identified important survival inequalities according to age at diagnosis, year of diagnosis, diagnostic subgroup and ethnicity which only became apparent after the use of imputation techniques. Furthermore, multiple imputation avoided discarding non-missing stage data, and allowed for the inclusion of this prognostic factor within the analysis. This was in contrast to many previous studies of survival for CYA cancers that neglected to use any measure of disease severity, despite its importance in determining the prognosis of a patient. Imputation and inclusion of disease severity measures showed that key inequalities in survival remained evident for certain subgroups of society, even after adjustment for disease severity. These findings are discussed in detail in §9.3.

In summary, the comparison of CCA and imputation methods presented in this thesis show that failure to apply adequate missing data techniques can reduce statistical power and thereby the ability to identify potentially important factors related to CYA cancer outcomes. Furthermore, it can result in invalid and imprecise estimates and reinforces

the message that in most instances, complete case analysis is not an acceptable method in the presence of missing data. The use of CCA can have serious consequences should the results be subsequently used to determine clinical practice. These conclusions are not limited to CYA cancer outcomes research, but have wider applicability to other diseases.

9.3 Variation in Cancer Survival

Chapter 2 identified a key gap in the knowledge of the variation of survival amongst children and TYAs, in particular, there was a paucity of survival analysis studies which adjusted for disease severity. There was just one study in which the survival of TYAs was compared to survival rates amongst children [42], and the results, although covering the whole of England, did not adjust for year of diagnosis, ethnicity or stage which have all been identified as having significant effects on survival. Furthermore, although the authors acknowledged stage as an important factor, they excluded it from the analysis because of missing data. In addition, none of the identified studies on CYA cancer survival in the UK used adequate methods to handle missing data. The findings in this thesis relate to a novel population-based study examining survival trends over time and in relation to ethnic group encompassing the childhood and TYA age range whilst adjusting for disease severity. The key original findings according to ethnicity, deprivation, age group, sex, year of diagnosis and disease severity are discussed in detail below.

9.3.1 Ethnicity

Survival was poorer for other ethnicity compared to white ethnicity amongst those with CNS tumours, and marginally, but not significantly, poorer for Asian and other ethnicity compared to white ethnicity for those with leukaemia and GCTs.

Four previous studies have assessed ethnic differences for childhood leukaemia in the UK [52, 55, 56, 57], two of which also looked at ethnic differences for childhood CNS tumours [55, 56]. To date, no study has looked at ethnic differences amongst CYAs with GCTs or TYAs with any cancer in the UK.

Studies by McKinney et al. [55] and Stiller et al. [56] showed increased HRs of death for Asian compared to non-Asian children with leukaemia after adjusting for WBC count, however, the former studies effect was not significant, and the latter national study showed only a borderline significant effect (P=0.057). Stiller et al. [56] also observed that survival differences became more apparent after 5-years post diagnosis. The results in this thesis were consistent with these previous studies as no significant effects of Asian

or other ethnicity was observed compared to white ethnicity for those with leukaemia for diagnoses between 1990 and 2009, despite large effect sizes indicating poorer survival. It is possible that due to the rarity of the disease, the overall high level of survival and the small number of Asian and other ethnic minorities this study and previous research by McKinney et al. [55] and Stiller et al. [56] lack power to identify significant differences, despite consistent evidence of increased HRs for these ethnic groups. This is in contrast to studies by Powell et al. [57] and Oakhill and Mann [52] in which poorer survival for south Asians compared to non-south Asians with leukaemia was observed after adjusting for WBC count. However, differences were restricted to 1980s when overall survival of leukaemia was poorer, which may now be obsolete and could have led to increased study power due to a larger number of observed deaths. Furthermore, missing data was not accounted for in these studies.

This study was the first to identify a significant difference between other ethnicity and white ethnicity for CNS tumours in Yorkshire. The non-significant effect of Asian ethnicity for CNS tumours was consistent with previous studies [55, 56] for the CNS tumour group as whole. This study has identified that despite no difference for Asian ethnicity in terms of CNS tumour survival, the minority other ethnic group had poorer prognosis compared to those of white ethnicity. Although Stiller et al. [56] identified evidence of poorer survival for non-white children compared to white children for CNS tumours, this finding was restricted to the astrocytoma subgroup of CNS tumours, the early 1980s, did not include TYAs and did not adjust for grade of tumour. The novel contribution of this study, therefore, was the identification of a two-fold increase risk of death from CNS tumours for those of non-white and non-Asian ethnicity despite adjustment for disease severity at diagnosis through use of a more detailed breakdown of ethnicity based on routinely collected data rather than relying on name analysis programs as used in previous studies.

In addition, this study was the first to identify possible survival disadvantages from GCTs for non-white ethnic groups, although these results were not significant, they warrant further investigation.

9.3.2 Deprivation

There was no evidence of an effect on survival of deprivation for CNS tumours, leukaemia or GCTs. For CNS tumours, this was consistent with a study by Tseng et al. [60] in 2006 after adjustment for WHO grade, although the authors performed a CCA and did not provide details of the amount of missing data. Evidence from an earlier study in Yorkshire showed that children from middle affluent areas with CNS tumours were at an

increased risk of death compared to those from affluent areas, however, the results were not adjusted for WHO grade [55]. Therefore, it was not possible to ascertain whether these effects were truly due to deprivation, or whether children from middle affluent areas were more likely to be diagnosed with advanced grade tumours.

For leukaemia, the study by Lightfoot et al. [65] in 2012 showed that children from more deprived areas had poorer survival compared to children from more affluent areas, however, there was no adjustment for disease severity in this study, nor was missing data taken into account. Therefore, the results may be unreliable as deprivation effects could be confounded by disease severity and missing data could have led to biased estimates.

For GCTs, the study by Birch et al. [44] focusing on TYAs with cancer showed no effect of deprivation on survival. This is consistent with the results presented in this thesis, which have now extended this observation to childhood GCTs.

The results presented in this thesis provide evidence that deprivation does not affect survival of CNS tumours, leukaemia or GCTs for children and TYAs after adjustment for disease severity, in contrast to earlier studies which did not adjust for disease severity.

9.3.3 Age Group

The results show, in the first comprehensive study of childhood and TYA cancer survival adjusting for disease severity, that survival was significantly poorer for TYAs compared to children for CNS tumours, GCTs and leukaemia.

For CNS tumours, survival was 34% poorer for TYAs compared to children. One previous study comparing children and TYAs with CNS tumours showed a significant improvement in survival for TYAs compared to 0-14 year olds in contrast to our study [42], however, TYAs were defined as 15-24 year olds, rather than the definition of 15-29 year olds used here. O'Hara et al. [42] also provided a comparison with 25-49 year olds, and showed significantly poorer survival for this group compared to children under the age of 15. It is possible therefore that the effect of the 25-29 year olds included in the definition of TYAs for this thesis outweighed the positive effect of 15-24 year olds observed by O'Hara et al. [42]. Further differences in the results were likely to arise due to adjustment for WHO grade, year of diagnosis, ethnicity and diagnostic subgroup, which were not taken into account by O'Hara et al. [42]. Arguably, these adjustments were likely to have a much greater impact on the results than the choice of upper age boundary. There was some evidence from the TYA literature to indicate that older age was associated with poorer survival for CNS tumours, which was consistent with our results, however, the evidence was limited to the astrocytoma subgroup of CNS tumours, results were not compared to

children with CNS tumours, disease severity was not taken into account and a CCA was performed [44].

For leukaemia, the effects of age were consistent when compared to separate studies of childhood [4] and TYA [44] cancer. The increased risk of death for TYAs compared to children with ALL could be due to differences in disease biology, for example, the t(12,21) translocation seen in 25% of children with B-cell ALL is only seen in 3% of adult cases [76], and TYAs have a higher ratio of T-cell compared to B-cell ALL compared to children which is associated with poorer prognosis [77, 78]. The results in this thesis also confirm earlier childhood specific studies highlighting the poor survival rates of children diagnosed under the age of 1 [53, 64, 65]. Although it has been suggested that differences in presenting white cell count between TYAs and children could have resulted in poorer TYA survival compared to children with ALL [44], the results in this thesis show that the difference in survival between children and TYAs remained significant even after adjustment for presenting white cell count. Moreover, the results showed that there were no significant differences in presenting white cell count between children and TYAs.

For GCTs, there was a 4-fold increase in the risk of death for TYAs compared to children, which was in contrast to a TYA study by Birch et al. [44] indicating higher survival of GCTs for those aged 17-20 years (87%) and 21-24 years (90%) compared to 13-16 year olds (80%). However, the study by Birch et al. [44] did not include the full childhood age range in addition to the TYA age range, only included diagnoses up to 2001, did not adjust for stage and used a CCA. Furthermore, the 5-year survival rate for TYAs with GCTs within this thesis was 94%, which was higher than the 5-year survival shown for TYAs by Birch et al. [44].

Poorer survival of TYAs with cancer compared to children could be a result of differences in the biological or molecular characteristics of cancers occurring amongst TYAs, such as those described for leukaemia, or due to differences in access to clinical trials between children and TYAs. In 2006/2007, 19% of TYA patients were reported to be enrolled onto clinical trials, compared to 51% of children with cancer for the same time period. Specifically, enrolment to clinical trials for TYAs with CNS tumours was particularly low [87]. Similar evidence of clinical trial enrolment in the US highlights that only a small proportion of those aged 15-25 years were enrolled onto clinical trials compared to the majority of children diagnosed with cancer [293]. Detailed data on clinical trials for all children and TYAs in this study were not available to study the effects of trial enrolment on survival, however, information was available for the leukaemia subgroup, ALL, and a separate analysis of survival for ALL according to clinical trial era has been conducted using Yorkshire register data (van Laar et al, 2015 in press). In addition to clinical trial enrolment, it is possible that TYAs have poorer survival compared to children due to

the use of adult treatment protocols rather than more effective paediatric protocols, and the resulting differences in treatment toxicity could have impacted on survival. Data on treatment protocols and compliance of treatment for children and TYAs diagnosed with cancer were not available and could therefore not be studied here. Previous work has provided evidence to suggest that teenagers with CNS tumours, leukaemia and GCTs were amongst the diagnostic groups with the lowest referral rates to paediatric oncology centres [294]. Efforts are underway in the UK to ensure all TYAs have the opportunity to enter trials where appropriate and are treated within principal treatment centres so that individuals can be treated and followed-up within an environment with appropriate age and/or site-specific expertise [43].

9.3.4 Sex

For children and TYAs with CNS tumours, leukaemia and GCTs, sex was not a significant predictor of survival after adjustment for year of diagnosis, age group, ethnicity and disease severity. For CNS tumours, there was some evidence of poorer survival in the univariable analysis, however, this effect was no longer present after adjustment of covariates. For leukaemia, there were no univariable or multivariable effects of sex on survival. For GCTs, there was borderline significant evidence of poorer survival amongst females compared to males, however, this effect did not remain after adjustment for other covariates. This is in contrast to an earlier study of TYA cancers which has shown females have improved survival for most types of cancer [44], however, the authors conclude that a possible explanation could include earlier presentation of cancer for females compared to males although they were unable to explicitly assess this in their study. The analysis in this thesis covers a more recent study period (1990 to 2009 compared to 1979-2003 for Birch et al. [44]) and includes a more comprehensive analysis covering the childhood and TYA age range, in addition to adjustments for disease severity and ethnic group and the application of advanced missing data methods.

9.3.5 Year of Diagnosis

Importantly, survival rates showed a consistent improvement over the study period for CNS tumours and leukaemia. However, such an improvement was not observed for GCTs. Previous work focusing on TYAs in England [44] and the North of England [59] did show improvements in survival for GCTs, however, both studies covered earlier periods, with improvements seen from 80% to 94% between 1979 and 2001 [44] and 37% to 86% between 1968 and 1997 [59]. The analysis within this study included diagnoses between

1990 and 2009, and survival was already high (91%) at the start of this period, therefore, there was less opportunity for observing further improvements in this study. Further work should therefore focus on minimising the complications of treatment for GCTs to improve the quality of survival.

9.3.6 Disease Severity at Diagnosis

There were large and significant decreases in survival for advanced stage CNS tumours, leukaemia and GCTs. This was not unexpected due to the definition of disease severity relating strongly to survival, however, the results highlight the importance of including this prognostic variable within survival analysis. Previously reported differences in deprivation and gender were not evident within the analysis in this thesis, however, despite adjustment of disease severity, key survival disadvantages for ethnic minorities and TYAs were evident. The analyses adjusting for disease severity provided in this thesis covering the childhood and TYA age range provide an important benchmark for future larger studies of childhood and TYA cancer, in particular, focus should be directed towards other tumour groups, such as lymphoma, once data on stage and disease severity improves in completeness as well as quality.

9.4 Inequalities in Disease Severity at Diagnosis

In order to improve survival for all CYAs with cancer and to further determine the reasons for survival inequalities, it is important to understand whether potential inequalities in disease severity at diagnosis exist, and where possible, eliminate these through raising awareness of the importance of a timely diagnosis of cancer. To date, there were no previously reported studies assessing the effects of age at diagnosis, ethnicity and deprivation on the stage of cancer at diagnosis for CYAs, despite evidence of such effects amongst the adult cancer population.

Importantly, this study provides the first reported evidence showing that, despite significant survival disadvantages for TYAs compared to children, there were no differences in the likelihood of being diagnosed with advanced stage disease between children and TYAs for CNS tumours and leukaemia. Furthermore, the majority of CNS tumours and leukaemias were low grade and standard risk respectively at diagnosis. For GCTs, despite a two-fold increase in the proportion of late stage diagnoses amongst children compared to TYAs, the effect of age group on the severity of disease was not significant within the multivariable analysis. The increased proportion of children

with late stage tumours could simply be an artifact of this diagnostic group, in which the majority of cases are TYAs. However, further larger scale studies are required to ascertain whether children with GCTs are more likely to be diagnosed with late stage disease compared to TYAs. Importantly, despite no improvements in survival over time for GCTs, 5-year survival rates were around 75% to 80% even for late stage tumours, and the proportion of late stage diagnoses was shown to decrease over time by 4% on average over the study period. Furthermore, despite indications that children were more likely to be diagnosed with advanced stage disease, survival for TYAs was almost four-fold poorer compared to children. The reasons for continued inequalities in survival for TYAs compared to children do not appear to be due to inequalities in disease severity. Other reasons, as described in more detail in §9.3.3, could be differences in the biological characteristics of cancers between children and TYAs, differences in enrollment onto clinical trials or differences in the treatment protocols available. Further research is warranted in these areas to ascertain why TYAs have poorer survival compared to children diagnosed with the same disease of the same severity. Furthermore, the study showed that there were no differences in the likelihood of being diagnosed with more advanced stage tumours for different ethnic groups or levels of deprivation. In spite of this, CYAs of non-white and non-Asian ethnicity had a significant survival disadvantage from CNS tumours compared to those of white ethnicity. The results presented here indicate that ethnicity or deprivation were unlikely to have created barriers to receiving a timely diagnoses of cancer for CYAs in contrast to that observed for adults with cancer. Recent work by Lightfoot et al. [65] in 2012, focusing specifically on children with leukaemia, concluded that the observed deprivation gap in survival from leukaemia was more likely a result of differences in treatment adherence according to deprivation rather than a difference in access to healthcare. The lack of inequality in disease severity for leukaemia observed in this study is consistent with the conclusion by Lightfoot et al. [65], which was based on the premise that the deprivation gap arose when chemotherapy treatment moved from in-hospital to at home administration. However, these results were not adjusted for the severity of disease, and therefore it was unclear whether the deprivation gap observed in that study was partly or wholly accountable for by the severity of disease at diagnosis.

9.5 Cardiovascular Late Effects amongst Long Term Survivors of CYA Cancer

The systematic literature review in Chapter 2 revealed a paucity of timely research relating to objective population-based research on the long term effects of childhood cancer. Furthermore, research on the same topic was lacking even more so for the TYA

cancer survivor population. The work provided in this thesis sought to fill this gap in the knowledge by presenting the first population-based study of cardiovascular LEs in survivors of childhood and TYA cancer using linked Yorkshire register and routinely collected inpatient HES data.

There was evidence of a significant increase in cardiovascular LEs in survivors of childhood cancer compared with the general population, whilst for TYAs the increased risk of cardiovascular LEs were not observed overall, but did occur for pericardial disease, cardiomyopathy and heart failure, pulmonary heart disease, conduction disorders and hypertension compared to the general population. Subgroup analysis for those who received chest radiation as part of their cancer treatment were at a significant increased risk of developing one or more cardiovascular LE compared to children and TYAs who did not receive chest radiation which was consistent with previous studies [112, 113]. However, the use and number of anthracyclines was not associated with a significant increase in the risk of cardiovascular LEs in the analysis presented here.

9.6 Implications of the Study

Despite the use of advanced statistical techniques, the large amount of missing data had the important implication that survival patterns could not be studied for all CYA cancers, with imputations being limited to cancers with less that 60% of missing data. Therefore the improvement in data completeness and quality of staging for CYA cancers is of key importance. As part of this study, a comprehensive guide of CYA cancer staging mechanisms has been developed for all ICCC diagnostic groups that can be used to implement a tailored data entry mechanism for stage and disease severity within the Yorkshire registers database; the implementation of this is currently underway. The new system will encourage more accurate capturing of staging data due to automatic validation procedures to ensure increased validity of recorded stage data in future. In addition, data on the staging mechanism which was used will also be captured. It is anticipated that completeness will also be improved after implementation of this system as it may be easier to identify stage or disease severity within the medical notes with the availability of the names of all relevant staging mechanisms for CYA cancers on the registers database system. Continued training and awareness of the importance of complete and accurate staging data is required to ensure the improvement of future research into the outcomes of cancer as well as research into factors which affect late diagnosis.

Previously reported inequalities in survival from cancer amongst CYAs should be interpreted and acted upon with care due to their lack of adjustments for disease severity and lack of appropriate missing data methods. This study has provided contrasting

evidence of some previously identified survival inequalities which were not apparent in the comprehensive analysis within this thesis. For example, Birch et al. [44] previously reported poorer survival for females compared to males for most cancers amongst TYAs, however, sex did not significantly affect survival for any tumour group within this thesis. In addition, the national study by O'Hara et al. [42] showed poorer survival for children compared to TYAs with GCTs, which was the opposite effect observed within this thesis. Although the work presented in this thesis only focused on survival variation for CNS tumours, leukaemia and GCTs diagnosed within the Yorkshire region, the analysis was more comprehensive in terms of adjustment for disease severity as well as the use of multiple imputation, lack of which has been shown to produce biased estimates. Previous work has shown that Yorkshire is representative of the UK in terms of its demographic profile, therefore, results from this study are generalisable outside of the Yorkshire region [12]. Although large scale national studies are useful in providing more study power and confidence of UK wide applicability, more in depth and comprehensive research using appropriate and advanced statistical techniques are arguably more valuable to unpick real inequalities in survival.

Despite improvements in survival rates, survival for TYAs with cancer remains poor compared to children. This observation is in spite of evidence showing that the disease severity at presentation is the same for TYAs as children. Continued efforts should therefore be made to ensure TYAs have equal access to clinical trials in addition to improved treatment protocols to drive further improvements in survival for TYAs diagnosed with cancer. Furthermore, the two-fold increased risk of death for non-white and non-Asian CYAs with CNS tumours needs to be addressed, although the reasons for this survival inequality remain unclear. Importantly, the inequality cannot be explained by presentation with higher grade tumours at diagnosis compared to white children.

Finally, there is an ongoing need for survivors of both childhood and TYA cancer to receive continued monitoring for potential long term effects of cancer. This study quantified a 3-fold increased risk of cardiovascular LEs for survivors of childhood cancer, and although effects were not seen overall for TYAs, they remained at increased risk of many cardiovascular disease subgroups. It is important that those at risk of developing cardiovascular LEs are supported with strategies to maximise cardiovascular health and given access to appropriate health surveillance. Therefore, awareness of the potential long term effects of cancer need to be raised not only with the survivors themselves, but also to potential future health carers. Furthermore, surveillance of long term survivors should not only be targeted to survivors of childhood cancer, but should also include survivors of TYA cancer.

9.7 Strengths and Limitations of the Study

This study has provided comprehensive and original findings for a population of CYAs with cancer in Yorkshire relating to variation in cancer survival, inequalities in disease severity at diagnosis and the burden of cardiovascular disease for survivors of cancer. In addition to the strengths of this study in comparison to previous work, there were a number of limitations which need to be considered.

The initial survival analysis for this thesis was restricted to data between 1990 and 2005, and implemented multiple imputation of stage for all cancers combined (published in van Laar et al. [263] in 2012). There were several limitations of this preliminary work, including the large overall amount of missing stage data (two-thirds missing stage data), the lack of consideration of different staging mechanisms for specific cancers and the use of name analysis to identify ethnicity which limited the analysis to only two broad ethnic groups being studied (south Asian and non-south Asian). The work in this thesis built upon the preliminary published work by extending the time period to 2009, using linked HES data to identify more detailed ethnicity groupings (white, Asian and other) and analysing missing data more rigorously taking into account the validity of recorded values for specific diagnostic groups prior to analysis. This enhancement resulted in a more detailed analysis of novel and population-based findings of survival trends for both the childhood and TYA cohort. However, on the basis of the current multiple imputation literature, it was decided that missing data occurring in more than 60% of cases should be avoided due to an increased chance of errors within the imputation process leading to increased inaccuracies of the results. This meant that detailed survival analysis was restricted to CNS tumours, leukaemia and GCTs. Therefore, the survival patterns according to age group, ethnic group and deprivation after adjustment for disease severity remain unclear for other CYA cancers.

Furthermore, the preliminary work only focused on MICE, whereas a recent and more advanced imputation technique, SMC-FCS, was implemented for the final results assessing trends in survival for CYAs with cancer. Although Bartlett et al. [149] suggest that SMC-FCS offers a substantial advantage compared to MICE when imputing non-linear models including Cox PH models, the results in this thesis showed that differences in the analysis results between MICE and SMC-FCS methods were negligible. Simulations for n=100 and n=1000 subjects by Bartlett et al. [149] showed that FCS resulted in biased estimates, with a larger bias for continuous variables compared to binary variables, which was not observed for SMC-FCS. The simulations were restricted to data which were MCAR, and the true effect sizes, β, were small (β=1). Within the thesis, data were assumed to be MAR and estimates for disease severity were large, although the

true effect sizes were not known. The fact that there were no differences between MICE and SMC-FCS indicated that the imputation model was likely to be specified correctly, as the stated advantage of SMC-FCS over MICE was to avoid mis-specification of the imputation model.

This study was based on a detailed population-based register of CYA cancers in Yorkshire. Although previous work has shown that Yorkshire is representative of the UK in terms of its demographic profile, thereby making the study results generalisable outside of the Yorkshire region [12], the main disadvantage of focusing only on the Yorkshire region was a lack of study power for some analyses. For example, in depth analyses of interaction terms, to explore how the effects disease severity on survival varied between diagnostic subgroups was not possible. In addition, it was not possible to study the cardiovascular LEs amongst survivors of CYA cancer according to individual tumour group, therefore, we were unable to determine whether the overall risk was the same for all diagnostic groups. Data on cause specific mortality or relapse deaths were not available in this study and therefore only overall survival was assessed. Cause-specific death data is available from the Office of National Statistics, however, such data has been shown to be unreliable for cancer, and the cost of such data far outweigh its benefit [295, 296]. An alternative method to overall survival which does not require cause-specific mortality data, is relative survival [297], which accounts for the mortality in an age and sex matched background population. This method is important for elderly populations in which the underlying mortality rates are high, however, the benefits of relative survival compared with overall survival for childhood and TYA studies are limited. Furthermore, despite availability of a detailed ethnicity breakdown within HES records, there were too few non-white and non-Asian cases to assess further individual ethnic categories, which may have been avoided if the study had been performed on a national level. Nonetheless, linkage with HES data allowed for an additional 'other' ethnic group to be included compared to previous studies which were limited to south Asian and non-south Asian on the basis of name analysis.

Despite detailed data on ethnicity, there were some limitations to the accuracy of ethnicity data recorded in HES and multiple ethnicity codes were associated with one person. This is a known problem within HES data, and could lead to misclassification of ethnicity [298]. The most commonly recorded value of ethnicity was used to classify patients into ethnic categories, which will have resulted in fewer misclassifications of ethnicity compared to a simpler approach of using the latest recorded ethnicity value. Furthermore, individual review of 8 cases was required to determine their ethnicity as there was no unique ethnicity code which occurred most commonly. For these cases in particular, the likelihood of misclassification of ethnicity was greater, which ultimately could have led to some bias in the estimates.

In addition to data on ethnicity, the linkage of Yorkshire register data with HES data allowed the long term cardiovascular burden for survivors of CYAs to be quantified in an objective manor, rather than previous studies which relied on self-reported long term effects [135]. Despite this improvement upon previous studies, some methodological limitations remained. Firstly, the study was based in Yorkshire and the overall number of cardiovascular admissions identified for survivors of CYA cancer was n=119 (3.6%). As the number of survivors with cardiovascular LEs was small, it was not possible to perform subgroup analysis and determine whether the risk of LEs varied between diagnostic groups. In addition, the study did not provide evidence to show that the use and number of anthracyclines were significant predictors of cardiovascular LEs. Whereas other studies, which focused on childhood survivors, show a clear dose-dependent effect [109, 111, 116, 118]. It was not possible to explore a dose-dependent effect within this study due to the lack of data on cumulative anthracycline doses within the Yorkshire register's database. It was likely that the number of different anthracycline agents administered (the only information available for this thesis) was a poor surrogate for cumulative dose effects.

The overall linkage rate for the study was 98%, which although high, meant that 2% of the cohort of survivors of CYA cancer did not link to any hospital record. It is not possible to determine whether these cases did not link due to data errors or whether or not 2% of cases simply did not have any hospital admissions. The latter is possible, although unlikely for a cohort of CYAs diagnosed with cancer. In addition to non-linkage of 2% of cases, there could also be a certain degree of misclassification caused by errors in the identifiers used to perform the linkage. This would have to effect of matching hospital records to an incorrect diagnosis of cancer which could bias results. Nonetheless, 98% of cases matched on the combination of exact NHS number, gender and date of birth, thus mismatching would be unlikely for the majority of cases. For the remaining 2% of cases, matches were made on exact NHS number, gender, postcode and partial date of birth or on exact date of birth, postcode and gender but not NHS number. There could have been a small number of mismatches in these 2% of cases, which could have resulted in some biases in the estimates.

Finally, the overall burden of cardiovascular disease for long terms survivors of CYAs with cancer was likely to be underestimated. The data provided in the present study include only those cardiovascular LEs which were severe enough to warrant a hospital admission, or which were discovered at the same time of an admission for a non-cardiovascular related reason. Although the Yorkshire register was successfully linked to outpatient HES data in addition to inpatient data, the former dataset could not be used for identification of cardiovascular LEs due to the poor quality of this dataset. In addition, further cardiovascular events could have been recorded within primary care records which

were not available for the current study. Nonetheless, the results presented in this thesis show a significant increased burden of cardiovascular disease amongst survivors of CYA cancer compared to an age-sex matched general population in Yorkshire by considering only part of the patient pathway post cancer diagnosis. The true burden of cardiovascular disease amongst survivors could therefore be even greater.

One of the main strengths of this thesis is the use of multiple imputation techniques rather than ignoring missing data or using CCA. However, ultimately, availability of accurate and complete data is the gold standard, and imputation methods are only as good as the partially observed data it is based on. Without comparison to the gold standard of complete and accurate data, it is not possible to determine whether, even after imputation, some bias has remained. Nevertheless, the use of multiple imputation has minimised the level of bias compared to a method which simply ignores missing data, and has made the best use of all the available recorded data. In addition, there are likely to be further unknown or unmeasured confounders which were excluded from the analysis which may result in further bias. However, a comprehensive set of possible confounders given the available data were included in the analysis based on a systematic literature review in addition to clinical input.

9.8 Future Research Recommendations

This study has shown that despite clear disadvantages of using inadequate missing data techniques, the use of multiple imputation within cancer epidemiology remains rare. Future studies should ensure the use of detailed multiple imputation techniques, as described in this thesis, are applied as standard in order to minimise the effects of missing data. More importantly, the recording of disease severity in both the medical records and cancer registries needs to be improved.

Increased severity of disease at diagnosis results in poorer prognosis for patients, however, the factors affecting disease severity at diagnosis remain unclear. This study has shown that sociodemographic factors, including ethnicity and deprivation, which are thought to affect disease severity at diagnosis through inequalities in access to healthcare for adults with cancer, do not determine disease severity for CYAs with cancer. In order to improve survival outcomes for all CYAs with cancer, further research into possible predictors of late stage diagnoses is required for the CYA population.

The burden of cardiovascular disease amongst survivors of CYA cancer which warrant submission to hospital has been quantified for the first time in the UK using data linkage methods. The work by Woodward et al. [133], in addition to a paucity of data on LEs

amongst survivors of TYA cancer, also identified a need to characterise survivors into risk groups to ensure appropriate screening programs could be developed. Further work should therefore focus on expanding the current study to a larger geographical area to determine whether the risk of cardiovascular LEs differs between diagnostic subgroups. However, despite additional study power offered by a national project, there may be a compromise in the level of detailed case information when using data from the national cancer registration service compared to the specialist register. Nonetheless, ongoing changes to national cancer registration system may make this a viable option for future research. In addition, cardiovascular LEs form only a small part of the overall burden of disease in survivors of CYAs and the methods presented within this study should be applied to quantify the burden of other long term effects amongst survivors of cancer including secondary tumours and respiratory late effects. In addition, further linkages to other electronic health records, such as primary care datasets should be sought to evaluate the burden of late effects on survivors of CYA cancer which occur outside of inpatient hospital admissions.

9.9 Future planned publications

In addition to the already published papers arising from this thesis (van Laar et al. [263] and van Laar et al. [299]), the results from Chapter 7 will be submitted for a publication entitled: 'Poorer TYA survival for GCTs is not due to increased stage at diagnosis in a population based sample' to the British Journal of Cancer.

9.10 Conclusion

Ignoring missing data can have serious consequences on the conclusions drawn from applied research. In spite of this, many researchers continue to ignore missing data. This study has identified contrasting findings to earlier studies which reported inequalities in survival, which due to the lack of adjustment for disease severity at diagnosis and inappropriate methods of handling missing data, could be incorrect.

Survival rates continue to improve over time for CNS tumours and leukaemia, and the number of advanced stage GCTs at diagnoses has decreased significantly between 1990 and 2009. However, those under the age of 1 year and diagnosed with leukaemia, TYAs compared to children with CNS tumours, leukaemia and GCTs and those of non-white and non-Asian ethnicity diagnosed with CNS tumours remain at a significant survival disadvantage. Continued efforts should be made to ensure that TYAs have equal access to

clinical trials and improved treatment protocols to drive further improvements for TYAs diagnosed with cancer. In addition, the long term cardiovascular effects of cancer have been shown, in the first population based study, to exist not only for children, but also for TYAs surviving their cancer. Continued efforts must therefore be made to monitor both children and TYAs for early signs of cardiovascular disease in order to maximise cardiovascular health in this growing population of survivors.

Appendices

A International Classification of Childhood Cancer

Table 1: The International Classification of Childhood Cancer, Third Edition [7].

Diagnostic Group	**ICD-O-3 code(s)**[1]	
	Morphology	**Topography**
I. Leukemias, myeloproliferative diseases, and myelodysplastic diseases		
a. Lymphoid leukemias	9820, 9823, 9826, 9827, 9831-9837, 9940, 9948	
b. Acute myeloid leukemias	9840, 9861, 9866, 9867, 9870-9874, 9891, 9895-9897, 9910, 9920, 9931	
c. Chronic myeloproliferative diseases	9863, 9875, 9876, 9950, 9960-9964	
d. Myelodysplastic syndrome and other myeloproliferative diseases	9945, 9946, 9975, 9980, 9982-9987, 9989	
e. Unspecified and other specified leukemias	9800, 9801, 9805, 9860, 9930	
II. Lymphomas and reticuloendothelial neoplasms		
a. Hodgkin lymphomas	9650-9655, 9659, 9661-9665, 9667	

[1] International Classification of Diseases for Oncology, Third Edition [30]

b. Non-Hodgkin lymphomas (except Burkitt lymphoma)	9591, 9670, 9671, 9673, 9675, 9678-9680, 9684, 9689-9691, 9695, 9698-9702, 9705, 9708, 9709, 9714, 9716-9719, 9727-9729, 9731-9734, 9760-9762, 9764-9769, 9970	
c. Burkitt lymphoma	9687	
d. Miscellaneous lymphoreticular neoplasms	9740-9742, 9750, 9754-9758	
e. Unspecified lymphomas	9590, 9596	
III. CNS[2] and miscellaneous intracranial and intraspinal neoplasms		
a. Ependymomas and choroid plexus tumor	9383, 9390-9394[c]	
b. Astrocytomas	9380[3] 9384, 9400-9411, 9420, 9421-9424, 9440-9442[c]	C72.3
c. Intracranial and intraspinal embryonal tumors	9470-9474, 9480, 9508[c] 9501-9504[c]	C70.0-C72.9
d. Other gliomas	9380[c] 9381, 9382, 9430, 9444, 9450, 9451, 9460[c]	C70.0-C72.2, C72.4-C72.9, C75.1-C75.3
e. Other specified intracranial and intraspinal neoplasms	8270-8281, 8300, 9350-9352, 9360-9362, 9412, 9413, 9492, 9493, 9505-9507, 9530-9539, 9582[c]	
f. Unspecified intracranial and intraspinal neoplasms	8000-8005[c]	C70.0-C72.9, C75.1-C75.3
IV. Neuroblastoma and other peripheral nervous cell tumors		
a. Neuroblastoma and ganglioneuroblastoma	9490, 9500	
b. Other peripheral nervous cell tumors	8680-8683, 8690-8693, 8700, 9520-9523, 9501-9504	

[2]Central Nervous System
[3]Tumors with non-malignant behaviour are included for all morphology codes on the line

	9501-9504	C00.0-C69.9, C73.9-C76.8, C80.9
V. Retinblastoma	9510-9514	
VI. Renal Tumours		
a. Nephroblastoma and other nonepithelial renal tumors	8959, 8960, 8964-8967 8963, 9364	C64.9
b. Renal carcinomas	8010-8041, 8050-8075, 8082, 8120-8122, 8130-8141, 8143, 8155, 8190-8201, 8210, 8211, 8221-8231, 8240, 8241, 8244-8246, 8260-8263, 8290, 8310, 8320, 8323, 8401, 8430, 8440, 8480-8490, 8504, 8510, 8550, 8560-8576, 8311, 8312, 8316-8319, 8361	C64.9
c. Unspecified malignant renal tumors	8000-8005	C64.9
VII. Hepatic tumors		
a. Hepatoblastoma	8970	
b. Hepatic carcinomas	8010-8041, 8050-8075, 8082, 8120-8122, 8140, 8141, 8143, 8155, 8190-8201, 8210, 8211, 8230, 8231, 8240, 8241, 8244-8246, 8260-8264, 8310, 8320, 8323, 8401, 8430, 8440, 8480-8490, 8504, 8510, 8550, 8560-8576, 8160-8180	C22.0, C22.1
c. Unspecified malignant hepatic tumors	8000 - 8005	C22.0, C22.1
VIII. Malignant Bone Tumours		
a. Osteosarcoma	9180-9187, 9191-9195, 9200	C40.0-C41.9, C76.0-C76.8, C80.9

b. Chondrosarcoma	9210, 9220, 9240	C40.0-C41.9, C76.0-C76.8, C80.9
	9221, 9230, 9241-9243	
c. Ewing tumor and related sarcomas of bone	9260 9363-9365	C40.0-C41.9, C76.0-C76.8, C80.9 C40.0-C41.9
d. Other specified malignant bone tumors	8810, 8811, 8823, 8830, 8812, 9250, 9261, 9262, 9270-9275, 9280-9282, 9290, 9300-9302, 9310-9312, 9320-9322, 9330, 9340-9342, 9370-9372	C40.0-C41.9
e. Unspecified malignant bone tumors	8000-8005, 8800, 8801, 8803-8805	C40.0-C41.9
IX. Soft tissue and other extraosseous sarcomas		
a. Rhabdomyosarcomas	8900-8905, 8910, 8912, 8920, 8991	
b. Fibrosarcomas, peripheral nerve sheath tumors, and other fibrous neoplasms	8810, 8811, 8813-8815, 8821, 8823, 8834-8835, 8820, 8822, 8824-8827, 9150, 9160, 9491, 9540-9571, 9580	C00.0-C39.9, C44.0-C76.8, C80.9
c. Kaposi sarcoma	9140	
d. Other specified soft tissue sarcomas	8587, 8710-8713, 8806, 8831-8833, 8836, 8840-8842, 8850-8858, 8860-8862, 8870, 8880, 8881, 8890, 9040-9044, 9120-9125, 9130-9133, 9135, 9136, 9141, 9142, 9161, 9170-9175, 9231, 9251, 9252, 9373, 9581	
	8830	C00.0-C39.9, C44.0-C76.8, C80.9
	8963	C00.0-C39.9, C65.5-C69.9, C73.9-C76.8, C80.9
	9180, 9210, 9220, 9240	C49.0-C49.9
	9260	C00.0-C39.9, C47.0-C75.9

	9364 9365	C00.0-C39.9, C47.0-C75.9 C00.0-C39.9, C47.0-C63.9, C65.9-C69.9, C73.9-C76.8, C80.9 C00.0-C39.9, C47.0-C63.9,
e. Unspecified soft tissue sarcomas	8800-8805	C00.0-C39.9, C44.0-C76.8
X. Germ cell tumors, trophoblastic tumors, and neoplasms of gonads		
a. Intracranial and intraspinal germ cell tumors	9060-9065, 9070-9072, 9080-9085, 9100, 9101[c]	C70.0-C72.9, C75.1-C75.3
b. Malignant extracranial and extragonadal germ cell tumors	9060-9065, 9070-9072, 9080-9085, 9100-9105	C00.0-C55.9, C57.0-C61.9, C63.0-C69.9, C73.9-C75.0, C75.4-C76.8, C80.9
c. Malignant gonadal germ cell tumors	9060-9065, 9070-9073, 9080-9085, 9090, 9091, 9100, 9101	C56.9, C62.0-C62.9
d. Gonadal carcinomas	8010-8041, 8050-8075, 8082, 8120-8122, 8130-8141, 8143, 8190-8201, 8210, 8211, 8221-8241, 8244-8246, 8260-8263, 8290, 8310, 8313, 8320, 8323, 8380,8384, 8430, 8440, 8480-8490, 8504, 8510, 8550, 8560-8573, 9000, 9014, 9015, 8441-8444, 8451, 8460-8473	C56.9, C62.0-C62.9
e. Other and unspecified malignant gonadal tumors	8590-8671 8000-8005	C56.9, C62.0-C62.9
XI. Other malignant epithelial neoplasms and malignant melanomas		
a. Adrenocortical carcinomas	8370-8375	

b. Thyroid carcinomas	8010-8041, 8050-8075, 8082, 8120-8122, 8130-8141, 8190, 8200, 8201, 8211, 8230, 8231, 8244-8246, 8260-8263, 8290, 8310, 8320, 8323, 8430, 8440, 8480, 8481, 8510, 8560-8573 8330-8337, 8340-8347, 8350	C73.9
c. Nasopharyngeal carcinomas	8010-8041, 8050-8075, 8082, 8120-8122, 8130-8141, 8190, 8200, 8201, 8211, 8230, 8231, 8244-8246, 8260-8263, 8290, 8310, 8320, 8323, 8430, 8440, 8480, 8481, 8200-8576	C11.0-C11.9
d. Malignant melanomas	8720-8780, 8790	
e. Skin carcinomas	8010-8041, 8050-8075, 8078, 8082, 8090-8110, 8140, 8143, 8147,8190, 8200, 8240, 8246, 8247, 8260, 8310, 8320, 8323, 8390-8420, 8430, 8480, 8542, 8560, 8570-8573, 8940, 8941	C44.0-C44.9
f. Other and unspecified carcinomas	8010-8084, 8120-8157, 8190-8264, 8290, 8310, 8313-8315, 8320-8325, 8360, 8380-8384, 8430-8440, 8452-8454, 8480-8586, 8588-8589, 8940, 8940, 8941, 8983, 9000, 9010-9016, 9020, 9030	C00.0-C10.9, C12.9-C21.8, C23.9-C29.9, C48.8, C50.0-C55.9, C57.0-C61.9, C63.0-C63.9, C65.9-C72.9, C75.0-C76.8, C80.9
XII. Other and unspecified malignant neoplasms		
a. Other specified malignant tumors	8930-8936, 8950, 8951, 8971-8981, 9050-9055, 9110 9363	C00.0-C39.9, C47.0-C75.9

b. Other unspecified malignant tumors	8000-8005	C00.0-C21.8, C23.9-C39.9, C42.0-C55.9, C57.0-C61.9, C63.0-C63.9, C65.9-C69.9, C73.9-C75.0, C75.4-C80.9

B Classification Scheme for Cancers in 15-24 year olds

Table 2: Classification Scheme for Cancers in 15-24 year olds

GROUP 1 - Leukaemias	
1.1.	Acute lymphoid leukaemia (ALL)
1.2.	Acute myeloid leukaemia (AML)
1.3.	Chronic myeloid leukaemia (CML)
1.4.	Other and unspecified leukaemia (Other Leuk)
1.4.1.	*Other and unspecified lymphoid leukaemias*
1.4.2.	*Other and unspecified myeloid leukaemias*
1.4.3.	*Other specified leukaemias, NEC*
1.4.4.	*Unspecified leukaemia*
GROUP 2 - Lymphomas	
2.1.	Non-Hodgkin lymphoma (NHL)
2.1.1.	*Non-Hodgkin lymphoma, specified subtype*
2.1.2.	*Non-Hodgkin lymphoma, subtype not specified*
2.2.	Hodgkin lymphoma (HL)
2.2.1.	*Hodgkin lymphoma, specified subtype*
2.2.2.	*Hodgkin lymphoma, subtype not specified*
GROUP 3 - Central Nervous System & other Intracranial & Intraspinal Neoplasms (CNS tumours)	
3.1.	Astrocytoma
3.1.1.	*Pilocytic astrocytoma*
3.1.2.	*Other low grade astrocytoma*
3.1.3.	*Glioblastoma and anaplastic astrocytoma*
3.1.4.	*Astrocytoma not otherwise specified*
3.2.	Other gliomas
3.2.1.	*Oligodendroglioma*
3.2.2.	*Other specified glioma*
3.2.3	*Glioma NOS*
3.3.	Ependymoma
3.4	Medulloblastoma and other primitive neuroectodermal tumours
3.4.1	*Medulloblastoma*
3.4.2	*Supratentorial PNET*
3.5.	Other specified intracranial and intraspinal neoplasms (Other CNS)
3.5.1	*Craniopharyngioma*
3.5.2	*Pituitary tumours*

3.5.3	*Pineal tumours*
3.5.4	*Choroid plexus tumours*
3.5.5	*Meningioma*
3.5.6	*Nerves sheath tumour of the brain*
3.5.7	*Other specified tumours*
3.6	Unspecified intracranial and intraspinal neoplasms tumours
3.6.1.	*Unspecified malignant intracranial and intraspinal neoplasms*
3.6.2.	*Unspecified non-malignant intracranial and intraspinal neoplasms*
GROUP 4 - Osseous and Chondromatous Neoplasms, Ewing tumour and other Neoplasms of Bone (Bone Tumours)	
4.1.	Osteosarcoma
4.2.	Chondrosarcoma
4.3.	Ewing sarcoma
4.3.1.	*Ewing sarcoma of bone*
4.3.2.	*Extraskeletal Ewing sarcoma*
4.3.3.	*Ewing sarcoma of unknown site*
4.4.	Other specified and unspecified bone tumours (Other bone tumours)
4.4.1.	*Other specified bone tumours*
4.4.2.	*Unspecified bone tumours*
GROUP 5 - Soft Tissue Sarcomas (STS)	
5.1.	Fibromatous neoplasms (Fibrosarcoma)
5.1.1.	*Fibrosarcoma*
5.1.2.	*Malignant fibrous histiocytoma*
5.1.3.	*Dermatofibrosarcoma*
5.2.	Rhabdomyosarcoma
5.3.	Other specified soft tissue sarcomas
5.3.1.	*Liposarcoma*
5.3.2.	*Leiomyosarcoma*
5.3.3.	*Synovial sarcoma*
5.3.4.	*Clear cell sarcoma*
5.3.5	*Blood vessel tumours*
5.3.6	*Nerve sheath tumours*
5.3.7	*Alveolar soft part sarcoma*
5.3.8	*Miscellaneous specified soft tissue sarcoma*
5.4	Unspecified soft tissue sarcomas
GROUP 6 - Germ Cell & Trophoblastic Neoplasms (Germ cell tumours)	
6.1	Gonadal germ cell & trophoblastic neoplasms
6.2	Germ cell & trophoblastic neoplasms of non-gonadal sites

6.2.1.	*Intracranial germ cell and trophoblastic tumours*
6.2.2.	*Other non-gonadal germ cell and trophoblastic tumours*
GROUP 7 - Melanoma and Skin Carcinoma	
7.1.	Melanoma
7.2.	Skin carcinoma
GROUP 8 - Carcinomas (except of skin)	
8.1.	Carcinoma of thyroid
8.2.	Other carcinoma of head and neck
8.2.1.	*Nasopharyngeal carcinoma*
8.2.2.	*Carcinoma of other sites in lip oral cavity and pharynx*
8.2.3.	*Carcinoma of nasal cavity, middle ear, sinuses, larynx and other ill-defined sites in head and neck*
8.3.	Carcinoma of trachea, bronchus, lung and pleura
8.4.	Carcinoma of breast
8.5.	Carcinoma of genito-urinary (GU) tract
8.5.1.	*Carcinoma of kidney*
8.5.2.	*Carcinoma of bladder*
8.5.3.	*Carcinoma of ovary*
8.5.4.	*Carcinoma of cervix*
8.5.5.	*Carcinoma of other and ill-defined sites in GU*
8.6.	Carcinoma of gastro-intestinal (GI) tract
8.6.1.	*Carcinoma of colon and rectum*
8.6.2.	*Carcinoma of stomach*
8.6.3.	*Carcinoma of liver and intrahepatic bile ducts*
8.6.4.	*Carcinoma of pancreas*
8.6.5.	*Carcinoma of other and ill-defined sites in GI tract*
8.7.	Carcinomas of other & ill-defined sites not elsewhere classified (NEC)
8.7.1.	*Adrenocortical carcinoma*
8.7.2.	*Other carcinomas NEC*
GROUP 9 - Miscellaneous Specified Neoplasms NEC	
9.1.	Embryonal tumours NEC
9.1.1.	*Wilms tumour*
9.1.2.	*Neuroblastoma*
9.1.3.	*Other embryonal tumours NEC*
9.2	Other rare miscellaneous specified neoplasms
9.2.1.	*Paraganglioma and glomus tumours*
9.2.2.	*Other specified gonadal tumours NEC*

9.2.3.	*Myeloma, mast cell tumours and miscellaneous reticuloendothelial neoplasms NEC*
9.2.4.	*Other specified neoplasms NEC*
GROUP 10 - Unspecified Malignant Neoplasms NEC	

C Literature Review on Survival of Cancer Amongst Children and Young Adults in the UK

Table 3: Medline and Web of Science Search Strategy, 1980-2014, English Language Articles only

Index	Search Terms	Results (N)
1	survival OR survival analys* OR survival rate* OR prognos*	1,446,307
2	child* OR pediatric* OR paediatric* OR teenage* or young adult* OR adolescen* or TYA or AYA or CYA or CTYA	3,055,665
3	neoplasm* OR cancer* OR cancer regist* OR tumor* OR tumour*	4,776,978
4	England OR English OR UK OR United Kingdom OR GB OR Great Britain	46,125
5	1 and 2 and 3 and 4	1276
6	Title and Abstract Review	18

D WHO Grading of Tumours of the Central Nervous System

	I	II	III	IV
Astrocytic tumours				
Subependymal giant cell astrocytoma	•			
Pilocytic astrocytoma	•			
Pilomyxoid astrocytoma		•		
Diffuse astrocytoma		•		
Pleomorphic xanthoastrocytoma		•		
Anaplastic astrocytoma			•	
Glioblastoma				•
Giant cell glioblastoma				•
Gliosarcoma				•
Oligodendroglial tumours				
Oligodendroglioma		•		
Anaplastic oligodendroglioma			•	
Oligoastrocytic tumours				
Oligoastrocytoma		•		
Anaplastic oligoastrocytoma			•	
Ependymal tumours				
Subependymoma	•			
Myxopapillary ependymoma	•			
Ependymoma		•		
Anaplastic ependymoma			•	
Choroid plexus tumours				
Choroid plexus papilloma	•			
Atypical choroid plexus papilloma		•		
Choroid plexus carcinoma			•	
Other neuroepithelial tumours				
Angiocentric glioma	•			
Chordoid glioma of the third ventricle		•		
Neuronal and mixed neuronal-glial tumours				
Gangliocytoma	•			
Ganglioglioma	•			
Anaplastic ganglioglioma			•	
Desmoplastic infantile astrocytoma and ganglioglioma	•			
Dysembryoplastic neuroepithelial tumour	•			

	I	II	III	IV
Central neurocytoma		•		
Extraventricular neurocytoma		•		
Cerebellar liponeurocytoma		•		
Paraganglioma of the spinal cord	•			
Papillary glioneuronal tumour	•			
Rosette-forming glioneuronal tumour of the fourth ventricle	•			
Pineal tumours				
Pineocytoma	•			
Pineal parenchymal tumour of intermediate differentiation		•	•	
Pineoblastoma				•
Papillary tumour of the pineal region		•	•	
Embryonal tumours				
Medulloblastoma				•
CNS primitive neuroectodermal tumour (PNET)				•
Atypical teratoid / rhabdoid tumour				•
Tumours of the cranial and paraspinal nerves				
Schwannoma	•			
Neurofibroma	•			
Perineurioma	•	•	•	
Malignant peripheral nerve sheath tumour (MPNST)		•	•	•
Meningeal tumours				
Meningioma	•			
Atypical meningioma		•		
Anaplastic / malignant meningioma			•	
Haemangiopericytoma		•		
Anaplastic haemangiopericytoma			•	
Haemangioblastoma	•			
Tumours of the sellar region				
Craniopharyngioma	•			
Granular cell tumour of the neurohypophysis	•			
Pituicytoma	•			
Spindle cell oncocytoma of the adenohypophysis	•			

Figure D.1: WHO Grading of Tumours of the Central Nervous System. Reprinted from Louis et al. [6]

E The Expectation-Maximization (EM) Algorithm within the Medical Literature

Table 4: Medline and Web of Science Search Strategy, 1925-present, English Language Articles only

Index	Search Terms	Results (N)
1	EM Algorithm	953
2	Expectation-Maximization Algorithm	480
3	Expectation-Maximisation Algorithm	17
4	maximum likelihood estimation	1042
5	MLE	604
6	1 or 2 or 3 or 4 or 5	2894
7	missing data	3494
8	missing values	751
9	missing cases	61
10	incomplete data	938
11	incomplete cases	35
12	incomplete values	0
13	missing completely at random	90
14	MCAR	94
15	missing at random	238
16	MAR	2658
17	(missing not at random)	32201
18	MNAR	59
19	non response	1341
20	7 or 8 or 9 or 10 or 11 or 12 or 13 or 14 or 15 or 16 or 17 or 18 or 19	36451
21	6 and 20	179
22	Remove duplicates from 21	169
23	Remove non-relevant articles	98

F Inverse Probability Weighting within the Medical Literature

Table 5: Medline and Web of Science Search Strategy, 1946-present, English Language Articles only

Index	Searches	Results
1	(inverse adj probability adj weighting)	76
2	(IPW not Prader–Willi)	39
3	(IPW not inferior-posterior)	36
4	(missing adj values)	749
5	(missing adj data)	3431
6	(missing adj cases)	65
7	(incomplete adj values)	0
8	(incomplete adj cases)	34
9	(incomplete adj data)	955
10	MCAR	84
11	(missing adj completely adj at adj random)	87
12	MAR	2684
13	(missing adj at adj random)	241
14	MNAR	57
15	(non adj response)	1347
16	1 or (2 and 3)	76
17	4 or 5 or 6 or 7 or 8 or 9 or 10 or 11 or 12 or 13 or 14 or 15	9036
18	16 and 17	8

G Multiple Imputation and Cancer Survival within the Medical Literature

Table 6: Medline and Web of Science Search Strategy, 1978-present, English Language Articles only

Index	Search Terms	Results (N)
1	(multiple adj2 imputation)	588
2	(multiple adj2 imputa*)	618
3	exp Neoplasms/	2097627
4	neoplasms	1767329
5	cancer	804104
6	(cancer adj regist*)	10073
7	Survival Analysis/ or Survival/ or Survival Rate/	190152
8	Treatment Outcome/	497985
9	(survival adj analysis)	91658
10	survival	635694
11	(survival adj rate)	140627
12	1 or 2	622
13	3 or 4 or 5 or 6	2384201
14	7 or 8 or 9 or 10 or 11	1051097
15	12 and 13 and 14	40

H Kaplan-Meier curves for multiply imputed variables in Stata

The following Stata code was used to create K-M plots for ethnic group by imputation (Figure 6.8):

```
set more off
#delimit ;
sts graph if _mj!=0, by(eth_grp _mj) scheme(sol)
addplot(line survivor_eth _t if eth_grp==1, sort c(J) lcolor(black) ||
        line survivor_eth _t if eth_grp==2, sort c(J) lcolor(red) ||
        line survivor_eth _t if eth_grp==3, sort c(J) lcolor(blue)
                legend(order(121 "White" 122 "Asian" 123 "Other"))) ;
#delimit cr
```

In the graph editor, select "record" and then manually change 1 line to a different colour. Save the grec file. Locate the grec file and open it as a text file, then edit the file by adding in line colour changes for all lines required. For a graph with 3 groups (i.e. ethnicity: white, asian, other), imputed 40 times, you need to colour 3 times 40 lines as follows:

```
//plot1 color
plotregion1.plot1.style.editstyle line(color(gray)) editcopy
// plot2 color
plotregion1.plot2.style.editstyle line(color(gray)) editcopy
//repeat until plot number 40

// plot41 color
plotregion1.plot41.style.editstyle line(color(erose)) editcopy
// plot42 color
plotregion1.plot42.style.editstyle line(color(erose)) editcopy
//repeat until plot number 90

// plot81 color
plotregion1.plot81.style.editstyle line(color(eltblue)) editcopy
// plot82 color
plotregion1.plot82.style.editstyle line(color(eltblue)) editcopy
//repeat until plot number 120
```

Once the grec file has been amended with the above, save the edited text file with extension .grec. Return to graph editor, with your original graph, and "play" the .grec file to colour lines accordingly.

I Staging of Childhood and Young Adult Cancer

Figure I.2 contains details of all prognostic staging mechanisms recommended for implementation into the cancer registry database for the Yorkshire Specialist Register of Cancer in Children and Young People based on a detailed description of missing staging data given in Chapter 5.

Figure I.2: Prognostic disease severity and staging mechanisms for childhood and young adult cancers by ICCC diagnostic group

Diagnostic group	ICCC	Data name	Description	Data value	Value description
Leukaemia	Ia, Ib, Ic, Id, Ie	White Blood Cell Count (WBC)	Highest white blood cell count pre-treatment (x10^9g per litre)	Range from 0.0 to 999.9	
		French-American-British (FAB)	French-American-British (FAB) Calssification. Bennett et al, 1976 [252]	M0	Undifferentiated acute myeloblastic leukemia
				M1	Acute myeloblastic leukemia with minimal maturation
				M2	Acute myeloblastic leukemia with maturation
				M3	Acute promyelocytic leukemia
				M4	Acute myelomonocytic leukemia
				M5	Acute monocytic leukemia
				M6	Acute erythroid leukemia
				M7	Acute megakaryoblastic leukemia
				L1	Small, monomorphic
				L2	Large, heterogenous
				L3	Burkitt-cell type
Lymphoma	IIa,IIb, IIc, IId, IIe	Ann Arbor Stage	Staging System Based on Location of Detected Disease	1	l = One region of lymph nodes, or spleen or thymus or Waldeyer's ring enlarged
				2	II = 2 regions of lymph nodes enlarged, on same side of diaphragm
				3	III = lymph nodes enlarged on both sides of diaphragm
				4	IV = disease outside lymph nodes e.g. liver, bone marrow excluding E
		Ann Arbor Symptoms	Additional stage designation based on presence or absence of specific symptoms	A	No Symptoms
				B	Presence of any of the following: unexplained persistent or recurrent fever (greater than 38C/ 101.5F, drenching night sweats, unexplained weight loss of 10% or more within the last 6 months)
		Ann Arbor Extranodality	Additional stage designation based on extranodal involvement	E	Code `E' if there is involvement of a single extranodal site that directly adjoins or is next to the known nodal group.

CNS tumours	IIIa, IIIb, IIIc, IIId, IIIe, IIIf	WHO Grade	World Health Organization Grading of Tumours of the Central Nervous System. Edge and Compton, 2010 [30]	1	Grade I
				2	Grade II
				3	Grade III
				4	Grade IV
	IIIc provided morphology code is M9470/3-M9474/3	CHANG Stage	CHANG Staging System for Medulloblastoma	M0	No evidence of metastasis
				M1	Microscopic Tumour Cells found in CSF
				M2	Gross nodular seeding in cerebellum, cerebral subarachnoid space, or in the third or fourth ventricles
				M3	Gross nodular seeing in spinal subarachnoid space
Neuroblastoma	IVa, Ivb	INSS	International Neuroblastoma Staging System. Brodeur et al, 1993 [256]	1	Localised tumour with complete gross excision, with or without microscopic residual disease; representative ipsilateral lymph nodes negative for tumour microscopically (nodes attached to and removed with the primary tumour may be positive).
				2A	Localised tumour with incomplete gross excision; representative ipsilateral nonadherent lymph nodes negative for tumour microscopically.
				2B	Localised tumour with or without complete gross excision, with ipsilateral nonadherent lymph nodes positive for tumour. Enlarged contralateral lymph nodes must be negative microscopically.
				3	Unresectable unilateral tumour infiltrating across the midline, with or without regional lymph node involvement; or localised unilateral tumour with contralateral regional lymph node involvement; or midline tumour with bilateral extension by infiltration (unresectable) or by lymph node involvement. The midline is defined as the vertebral column. Tumours originating on 1 side and crossing the midline must infiltrate to or beyond the opposite side of the vertebral column.
				4	Any primary tumour with dissemination to distant lymph nodes, bone, bone marrow, liver, skin, and/or other Organs (except as defined for stage 4S).

Neuroblastoma (Continued)			International Neuroblastoma Staging System. Brodeur et al, 1993 [256]	4S	Localised primary tumour (as defined for stage 1, 2A, or 2B), with dissemination limited to skin, liver, and/or bone marrow (limited to infants younger than 1 year). Marrow involvement should be minimal (<10% of total nucleated cells identified as malignant by bone biopsy or by bone marrow aspirate). More extensive bone marrow involvement would be considered to be stage 4 disease. The results of the MIBG scan (if performed) should be negative for disease in the bone marrow.
		INRG Staging System	International Neuroblastoma Risk Group Staging System.Cohn et al, 2009 [257]	L1 L2 M MS	
	IVa, Ivb	INRG Risk Group	International Neuroblastoma Risk Group - pre-tretament risk group	VL	Very low
				L	Low
				I	Intermediate
				H	High
Retinoblastoma[a]	Va				
Renal tumours	Via	Wilms Tumour Staging		1	Stage I - tumour is limited to the kidney and completely resected.
				2	Stage II - tumour is completely resected, and there is no evidence of tumour at or beyond the margins of resection but the tumour extends beyond the kidney (penetration of capsule, invasion of blood vessels outside renal parenchyma).
				3	Stage III - there is residual tumour following surgery that is confined to the abdomen.
				4	Stage IV - there are distant metastases (lung, liver, bone, brain), or lymph node metastases outside the abdominopelvic region.
				5	Stage V - involvement of both kidneys is present at diagnosis
	VIa, VIb Vic	TNM Staging (Renal)		1	Stage I
				2	Stage II
				3	Stage III
				4	Stage IV

Hepatic tumours	VIIa, VIIb, VIIc	PRETEXT Staging System	PRETEXT Staging System for Hepatablastoma and Hepaocellular Carcinomas	1	PRETEXT 1
				2	PRETEXT 2
				3	PRETEXT 3
				4	PRETEXT 4
		PRETEXT Staging Outside Liver	Additional Staging Information For Tumours Outside the Liver	V	`extension' into the vena cava and/or all three hepatic veins
				P	`extension' into the main and/or both left and right branches of the portal vein
				E	extra-hepatic disease
				M	presence of distant metastases
Malignant bone tumours	VIIIa, VIIIb, VIIIc, IIId, IIIe	TNM Staging (Bone)	TNM Staging System for Bone Tumours. Edge and Compton, 2010 [30]	1A	Stage IA
				1B	Stage IB
				2A	Stage IIA
				2B	Stage IIB
				3	Stage III
				4A	Stage IVA
				4B	Stage IVB
Soft tissue and other extraosseuous sarcomas	IXa, IXb, IXc, IXd	TNM Staging (STS)	TNM Staging System for Soft Tissue Sarcomas. Edge and Compton, 2010 [30]	1A	Stage IA
				1B	Stage IB
				2A	Stage IIA
				2B	Stage IIB
				3	Stage III
				4	Stage IV
Germ Cell Tumours	Xa	Beta Human Chorionic Gonadotropin	Maximum serum level of HCG at diagnosis in IU/l	Format max n3	
	Xc, Xd, Xe if topography code is C569	TNM Staging (Ovary)	TNM Staging System for Ovarian tumours. Edge and Compton, 2010 [30]	1	
				1A	
				1B	
				1C	

Germ Cell Tumours (continued)	Xc, Xd, Xe if topography code is C569	TNM Staging (Ovary)	TNM Staging System for Ovarian tumours. Edge and Compton, 2010 [30]	2 2A 2B 2C 3 3A 3B 3C 4
	Xc, Xd, Xe if topography code is C569	FIGO Staging System		1A 1B 1C 2A 2B 2C 3A 3B 3C 4
	Xc, Xd, Xe if topography code is C620-C629	TNM Staging (Testis)		0 1 1A 1B 1S 2 2A 2B 2C 3 3A 3B 3C

Germ Cell Tumours (continued)	Xc, Xd, Xe if topography code is C620-C629	Royal Marsden Staging System		1	No evidence of metastasis
				1M	Rising concentrations of serum markers with no other evidence of metastasis
				2A	Abdominal node metastasis <=2cm in diameter
				2B	Abdominal node metastasis 2-5cm in diameter
				2C	Abdominal node metastasis >=5cm in diameter
				3M	Supradiaphragmatic nodal metastasis - mediastinal
				3N	Supradiaphragmatic nodal metastasis - Supraclavicular, cervical, or axillary
				3O	Supradiaphragmatic nodal metastasis - No abdominal node metastasis
				3A	Supradiaphragmatic nodal metastasis - node <=2cm in diameter
				3B	Supradiaphragmatic nodal metastasis - node 2-5cm in diameter
				3C	Supradiaphragmatic nodal metastasis - node >=5cm in diameter
				4L1	Extralymphatic metastasis, Lung <3 metastases
				4L2	Extralymphatic metastasis, Lung >3 metastases, all <=2cm in diameter
				4L3	Extralymphatic metastasis, Lung >=3 metastases, one or more of which are >2cm in diameter
				4H+	Extralymphatic metastasis, Liver
				4Br+	Extralymphatic metastasis, Brain
				4Bo+	Extralymphatic metastasis, Bone
	Xb	IGCCCG Classification	The International Germ Cell Cancer Collaborative Group Classification	1	Stage I
				2	Stage II
				3	Stage III
				4	Stage IV
Other tumours or unspecified staging mechanisms[b]	XIa, XIb, XIc, XId, XIe, XIf, XIIa, XIIb	Stage		Any value	To be accompanied by a description of staging mechanism if known

[a]There is no relevant prognostic staging mechanism for retinoblastoma

[b]To be used for ICCC XI or XII, or when the staging mechanism is unknown.

References

[1] A Morse. *Delivering the Cancer Reform Strategy*. Department of Health, 2010.

[2] National Registry of Childhood Tumours. Childhood Cancer Research Group, 2012. URL http://www.ccrg.ox.ac.uk/home/about.shtml.

[3] NWCIS. Cancer incidence in teenagers and young adults in England, 2012. URL http://www.nwcis.nhs.uk/tya/tya-incidence.aspx. [Date Accessed: 26/09/2014].

[4] Cancer Research UK. Childhood Cancer - Cancer Statistics Key Facts., June 2014. URL http://publications.cancerresearchuk.org/downloads/Product/CS_KF_CHILDHOOD.pdf. [Accessed: 25/09/2014].

[5] Cancer Research UK. Teenage and Young Adult Cancer - Cancer Statistics Key Facts., March 2013. URL http://www.cancerresearchuk.org/cancer-info/cancerstats/teenage-and-young-adult-cancer/survival/. [Accessed: 25/09/2014].

[6] DN Louis, H Ohgaki, OD Wiestler, and WK Cavenee. WHO Classification of tumours of the central nervous system. *IARC, Lyon*, 2007.

[7] E Steliarova-Foucher, C Stiller, B Lacour, and P Kaatsch. International Classification of Childhood Cancer, Third Edition. *Cancer*, 103:1457–67, 2005.

[8] PD Sasieni, J Shelton, N Ormiston-Smith, CS Thomson, and PB Silcocks. What is the lifetime risk of developing cancer?: the effect of adjusting for multiple primaries. *British Journal of Cancer*, 105(3):460–465, 2011.

[9] J Little. *Epidemiology of Childhood Cancer*. International Agency for Research on Cancer, Lyon, 1999.

[10] Cancer Research UK. Childhood Cancer Incidence Summary, Great Britain, 1996-2005, February 2012. URL http://publications.cancerresearchuk.org/downloads/Product/CS_DT_CHILDHOOD.pdf. [Accessed: 25/09/2014].

[11] LA Fern, C Campbell, TOB Eden, R Grant, I Lewis, U Macleod, D Weller, and J Whelan. How frequently do young people with potential cancer symptoms present in primary care? *British Journal of General Practice*, 61(586):e223–e230, 2011.

[12] RG Feltbower, IJ Lewis, S Picton, M Richards, AW Glaser, SE Kinsey, and PA McKinney. Diagnosing childhood cancer in primary care: a realistic expectation? *British Journal of Cancer*, 90(10):1882–1884, 2004.

[13] RM Dommett, MT Redaniel, MCG Stevens, W Hamilton, and RM Martin. Features of childhood cancer in primary care: a population-based nested case–control study. *British journal of cancer*, 106(5):982–987, 2012.

[14] CA Stiller. Childhood Cancer in Britain: Incidence, Survival, Mortality. *Oxford University Press*, 2007.

[15] Cancer Research UK. All cancers combined - Key Facts., 2011. URL `http://www.cancerresearchuk.org/cancer-info/cancerstats/keyfacts/Allcancerscombined/`. [Accessed: 25/09/2014].

[16] KC Oeffinger, AC Mertens, CA Sklar, T Kawashima, MM Hudson, AT Meadows, DL Friedman, N Marina, W Hobbie, and NS Kadan-Lottick. Chronic health conditions in adult survivors of childhood cancer. *New England Journal of Medicine*, 355(15):1572, 2006.

[17] G Gatta, G Zigon, R Capocaccia, JW Coebergh, E Desandes, P Kaatsch, G Pastore, R Peris-Bonet, and CA Stiller. Survival of European children and young adults with cancer diagnosed 1995-2002. *European Journal of Cancer*, 45(6):992–1005, 2009.

[18] RD Neal. Do diagnostic delays in cancer matter? *British Journal of Cancer*, 101: S9–S12, 2009.

[19] MP Coleman, G Gatta, A Verdecchia, J Esteve, M Sant, H Storm, C Allemani, L Ciccolallo, M Santaquilani, and F Berrino. EUROCARE-3 summary: cancer survival in Europe at the end of the 20th century. *Annals of Oncology*, 14(suppl 5): v128–v149, 2003.

[20] 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. *British journal of cancer*, 107(8):1220–1226, 2012.

[21] Department of Health. The NHS Outcomes Framework 2014/15. ©Crown copyright 2010, November 2013. URL `https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/256456/NHS_outcomes.pdf`. [Accessed: 26/09/2014].

[22] Department of Health. Improving Outcomes: A Strategy for Cancer. ©Crown copyright 2010, January 2011. URL `https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/213785/dh_123394.pdf`. [Accessed: 26/09/2014].

[23] I Lewis and C Lenehan. Report of the Children and Young People's Health Outcomes Forum. *www.gov.uk/dh*, 2012.

[24] NS Kadan-Lottick, KK Ness, S Bhatia, and JG Gurney. Survival variability by race and ethnicity in childhood Acute Lymphoblastic Leukemia. *The Journal of the American Medical Association*, 290(15):2008–2014, 2003.

[25] AM Linabery and JA Ross. Childhood and Adolescent Cancer Survival in the US by Race and Ethnicity for the Diagnostic Period 1975-1999. *Cancer*, 113(9): 2575–2596, 2008.

[26] Office for National Statistics. Registrations of cancer diagnosed in 2009, England. *Cancer Statistics Registrations*, MB1 40, 2011.

[27] BH Pollock and JM Birch. Registration and classification of adolescent and young adult cancer cases. *Pediatric Blood & Cancer*, 50(S5):1090–1093, 2008.

[28] Cancer Research UK. What is Cancer?, June 2014. URL `http://www.cancerresearchuk.org/about-cancer/cancers-in-general/what-is-cancer/`. [Accessed: 25/09/2014].

[29] National Cancer Institute. Cancer Topics, June 2014. URL `http://www.cancer.gov/cancertopics`. [Accessed: 25/09/2014].

[30] AG Fritz. *International Classification of Diseases for Oncology (ICD-O)*. World Health Organisation, Geneva, 3rd edition, 2000.

[31] JM Birch, RD Alston, AM Kelsey, MJ Quinn, P Babb, and RJQ McNally. Classification and incidence of cancers in adolescents and young adults in England 1979-1997. *British Journal of Cancer*, 87(11):1267, 2002.

[32] SB Edge and CC Compton. The American Joint Committee on Cancer: the 7th Edition of the AJCC Cancer Staging Manual and the Future of TNM. *Ann Surg Oncol*, 17:1471–1474, 2010.

[33] Department of Health. The Cancer Reform Strategy, 2007. URL `http://webarchive.nationalarchives.gov.uk/20130107105354/http://www.dh.gov.uk/prod_consum_dh/groups/dh_digitalassets/documents/digitalasset/dh_081007.pdf`. [Accessed: 25/09/2014].

[34] NCIN. *Developing a UK wide training specification for non medical staff involved in collecting and using data for improving cancer clinical outcomes: final report*. National Cancer Intelligence Network, 2009.

[35] NCIN. *Cancer Incidence and Survival By Major Ethnic Group, England, 2002-2006*. National Cancer Intelligence Network, June 2009 2009.

[36] NR Kressin, B Chang, A Hendricks, and LE Kazis. Agreement Between Administrative Data and Patients: Self-Reports of Race/Ethnicity. *American Journal of Public Health*, 93(10):1734–1739, 2003.

[37] PJ Aspinall and B Jacobson. Why poor quality of ethnicity data should not preclude its use for identifying disparities in health and healthcare. *BMJ Quality and Safety*, 16:176–180, 2007.

[38] World Health Organisation. Cancer Fact Sheet No 297, February 2014. URL http://www.who.int/mediacentre/factsheets/fs297/en/.

[39] Office for National Statistics. Cancer survival in England: one-year and five-year survival for 21 common cancers, by sex and age Patients diagnosed 2003-2007 and followed up to 2008. *Statistical Bulletin*, 2010.

[40] C Magnani, G Pastore, JW Coebergh, S Viscomi, C Spix, and E Steliarova-Foucher. Trends in survival after childhood cancer in Europe, 1978–1997: report from the Automated Childhood Cancer Information System project (ACCIS). *European Journal of Cancer*, 42(13):1981–2005, 2006.

[41] G Bahadur and P Hindmarsh. Age definitions, childhood and adolescent cancers in relation to reproductive issues. *Human Reproduction*, 15(1):227–227, 2000.

[42] C O'Hara, A Moran, and TYA National Cancer Intelligence Advisory Group. Survival of Teenagers and Young Adults (TYA) with Cancer in the UK. National Cancer Intelligence Network, August 2012. URL http://www.ncin.org.uk/publications/reports/survival_in_teenagers_and_young_adults_with_cancer_in_the_uk. [Accessed: 26/09/2014].

[43] National Collaborating Centre for Cancer. *Improving Outcomes in Children and Young People with Cancer; The Manual*. National Institute for Health and Clinical Excellence, 2005.

[44] JM Birch, D Pang, RD Alston, S Rowan, M Geraci, A Moran, and TOB Eden. Survival from cancer in teenagers and young adults in England, 1979-2003. *British Journal of Cancer*, 99(5):830, 2008.

[45] M Geraci, TOB Eden, RD Alston, A Moran, RS Arora, and JM Birch. Geographical and temporal distribution of cancer survival in teenagers and young adults in England. *British Journal of Cancer*, 101(11):1939–1945, 2009.

[46] M Geraci, JM Birch, RD Alston, A Moran, and TOB Eden. Cancer mortality in 13 to 29-year-olds in England and Wales, 1981–2005. *British Journal of Cancer*, 97 (11):1588–1594, 2007.

[47] A Bleyer, A Viny, and R Barr. Cancer in 15 to 29 year-olds by primary site. *Oncologist*, 11(6):590, 2006.

[48] A Bleyer, R Barr, B Hayes-Lattin, D Thomas, C Ellis, and B Anderson. The distinctive biology of cancer in adolescents and young adults. *Nature Reviews Cancer*, 8(4):288–298, 2008.

[49] EE Kent, LS Sender, JA Largent, and H Anton-Culver. Leukemia survival in children, adolescents, and young adults: influence of socioeconomic status and other demographic factors. *Cancer Causes and Control*, 20:1409–1420, 2009.

[50] International Agency for Research on Cancer. *The Automated Cancer Information System*. World Health Organisation http://accissiarcfr Accessed: 22/02/2012, 2012.

[51] RD Alston, S Rowan, TOB Eden, A Moran, and JM Birch. Cancer incidence patterns by region and socioeconomic deprivation in teenagers and young adults in England. *British Journal of Cancer*, 96:1760–1766, 2007.

[52] A Oakhill and JR Mann. Poor prognosis of acute lymphoblastic leukaemia in Asian children living in the United Kingdom. *BMJ*, 286(6368):839, 1983.

[53] P Badrinath, NE Day, and D Stockton. Population-based survival trends for leukaemia in East Anglia, United Kingdom. *Journal of Public Health Medicine*, 19(4):403–407, 1997.

[54] JA Schillinger, PC Grosclaude, S Honjo, MJ Quinn, A Sloggett, and MP Coleman. Survival after acute lymphocytic leukaemia: effects of socioeconomic status and geographic region. *Archives of Diseases in Childhood*, 80:311–317, 1999.

[55] PA McKinney, RG Feltbower, RC Parslow, IJ Lewis, S Picton, SE Kinsey, and CC Bailey. Survival from childhood cancer in Yorkshire, UK: effect of ethnicity and socio-economic status. *European Journal of Cancer*, 35(13):1816–1823, 1999.

[56] CA Stiller, KJ Bunch, and IJ Lewis. Ethnic group and survival from childhood cancer: report from the UK Children's Cancer Study Group. *British Journal of Cancer*, 82(7):1339–1343, 2000.

[57] JE Powell, E Mendez, SE Parkes, and JR Mann. Factors affecting survival in White and Asian children with acute lymphoblastic leukaemia. *British Journal of Cancer*, 82(9):1568, 2000.

[58] D Joshi, JR Anderson, C Paidas, J Breneman, DM Parham, and W Crist. Age is an independent prognostic factor in rhabdomyosarcoma: a report from the soft tissue sarcoma committee of the Childrens Oncology Group. *Paediatric Blood and Cancer*, 42:64–73, 2004.

[59] MS Pearce, L Parker, KP Windebank, SJ Cotterill, and AW Craft. Cancer in adolescents and young adults aged 15–24 years: a report from the North of England young person's malignant disease registry, UK. *Pediatric Blood & Cancer*, 45(5): 687–693, 2005.

[60] MY Tseng, JH Tseng, and E Merchant. Comparison of effects of socioeconomic and geographic variations on survival for adults and children with glioma. *Journal of Neurosurgery*, 105:297–305, 2006.

[61] C Croucher, JS Whelan, H Møller, and EA Davies. Trends in the incidence and survival of cancer in teenagers and young adults: regional analysis for South East England 1960–2002. *Clinical Oncology*, 21(5):417–424, 2009.

[62] R Eyre, RG Feltbower, PW James, K Blakey, E Mubwandarikwa, D Forman, PA McKinney, MS Pearce, and RJQ McNally. The epidemiology of bone cancer in 0-39 year olds in northern England, 1981-2002. *BMC cancer*, 10(1):357, 2009.

[63] WT Johnston, TJ Lightfoot, J Simpson, and E Roman. Childhood cancer survival: A report from the United Kingdom Childhood Cancer Study. *Cancer Epidemiology*, 34:659–666, 2010.

[64] NO Basta, PW James, B Gomez-Pozo, AW Craft, and RJQ McNally. Survival from childhood cancer in northern England, 19682005. *British Journal of Cancer*, 105:1402–1408, 2011.

[65] TJ Lightfoot, WT Johnston, J Simpson, AG Smith, and P Ansell. Survival from childhood acute lymphoblastic leukaemia: the impact of social inequality in the United Kingdom. *European Journal of Cancer*, 48:263–269, 2012.

[66] NO Basta, PW James, B Gomez-Pozo, AW Craft, P Norman, and RJQ McNally. Survival from teenage and young adult cancer in Northern England, 1968–2008. *Pediatric Blood & Cancer*, 61(5):901–906, 2014.

[67] EJ Estlin, RJ Gilbertson, and RF Wynn. *Pediatric Hematology and Oncology - Scientific Principles and Clinical Practice*. Wiley & Blackwell, 2010.

[68] G Gatta, R Luksch, MP Coleman, I Corazziari, and the EUROCARE Working Group. Survival from acute non-lymphocytic leukaemia (ANLL) and chronic myeloid leukaemia (CML) in European children since 1978: a population-based study. *European Journal of Cancer*, 37:695–702, 2001.

[69] EJ Estlin, RJ Gilbertson, and RF Wynn. *Pediatric Hematology and Oncology - Scientific Principles and Clinical Practice*, chapter 12: Neuroblastoma, Vaidya, SJ and Pearson, ADJ. Wiley & Blackwell, 2010.

[70] KK Matthay, C Perez, RC Seeger, GM Brodeur, H Shimada, JB Atkinson, CT Black, R Gerbing, GM Haase, DO Stram, P Swift, and JN Lukens. Successful treatment of stage III neuroblastoma based on prospective biologic staging: a Children's Cancer Group study. *Journal of Clinical Oncology*, 16(4):1256–1264, 1998.

[71] SN Zafar, SQ Ahmad, and N Zafar. Retinoblastoma in an adult. *BMC research notes*, 6(1):1–3, 2013.

[72] EJ Estlin, RJ Gilbertson, and RF Wynn. *Pediatric Hematology and Oncology - Scientific Principles and Clinical Practice*, chapter 13: Renal Tumors, Estlin, EJ and Garf, N. Wiley & Blackwell, 2010.

[73] EJ Estlin, RJ Gilbertson, and RF Wynn. *Pediatric Hematology and Oncology - Scientific Principles and Clinical Practice*, chapter 10: Bone Tumors, Gorlick, R and Perisoglou, M and Whelan, J. Wiley & Blackwell, 2010.

[74] EJ Estlin, RJ Gilbertson, and RF Wynn. *Pediatric Hematology and Oncology - Scientific Principles and Clinical Practice*, chapter 14: Soft Tissue Sarcoma, Bisogno, G and Anderson, J. Wiley & Blackwell, 2010.

[75] MJ Murray and JC Nicholson. Germ cell tumours in children and adolescents. *Paediatrics and Child Health*, 20(3):109–116, 2010.

[76] RCT Aguiar, J Sohal, F Van Rhee, M Carapeti, IM Franklin, AH Goldstone, JM Goldman, and NCP Cross. TEL-AML1 fusion in Acute Lymphoblastic Leukaemia of Adults. *British Journal of Haematology*, 95(4):673–677, 1996.

[77] DI Marks, EM Paietta, AV Moorman, SM Richards, G Buck, G DeWald, A Ferrando, AK Fielding, AH Goldstone, and RP Ketterling. T-cell Acute Lymphoblastic Leukemia in adults: clinical features, immunophenotype, cytogenetics, and outcome from the large randomized prospective trial (UKALL XII/ECOG 2993). *Blood*, 114(25):5136–5145, 2009.

[78] MF Greaves, G Janossy, J Peto, and H Kay. Immunologically defined subclasses of Acute Lymphoblastic Leukaemia in children: their relationship to presentation features and prognosis. *British Journal of Haematology*, 48(2):179–197, 1981.

[79] BES Gibson, K Wheatley, IM Hann, RF Stevens, D Webb, RK Hills, SSN De Graaf, and CJ Harrison. Treatment strategy and long-term results in paediatric patients treated in consecutive UK AML trials. *Leukemia*, 19(12):2130–2138, 2005.

[80] N Wollner, JH Burchenal, P H Lieberman, P Exelby, G D'Angio, and M L Murphy. Non–Hodgkin's lymphoma in children A comparative study of two modalities of therapy. *Cancer*, 37(1):123–134, 1976.

[81] EJ Estlin, RJ Gilbertson, and RF Wynn. *Pediatric Hematology and Oncology - Scientific Principles and Clinical Practice*, chapter 16: Hepatic Tumours, Roebuck, DJ and Plaschkens, J. Wiley & Blackwell, 2010.

[82] S Gururangan, A O'Meara, C Macmahon, EJ Guiney, B O'Donnell, RJ Fitzgerald, and F Breatnach. Primary hepatic tumours in children: A 26–year review. *J Surg Oncol*, 50:30–36, 1992.

[83] MD Stringer, S Hennayake, ER Howard, L Spitz, EA Shafford, G Mieli-Vergani, R Saxena, M Malone, C Dicks-Mureaux, J Karani, AP Mowat, and J Pritchard. Improved outcome for children with hepatoblastoma. *British Journal of Surgery*, 82:386–391, 1995.

[84] E Copson, T Maishman, S Gerty, B Eccles, L Stanton, RI Cutress, DG Altman, L Durcan, P Simmonds, and L Jones. Ethnicity and outcome of young breast cancer patients in the United Kingdom: the POSH study. *British Journal of Cancer*, 110 (1):230–241, 2014.

[85] B Rachet, LM Woods, E Mitry, M Riga, N Cooper, MJ Quinn, J Steward, H Brenner, J Esteve, R Sullivan, and MP Coleman. Cancer survival in England and Wales at the end of the 20th century. *British Journal of Cancer*, 99:S2–S10, 2008.

[86] RG Feltbower, SE Kinsey, M Richards, G Shenton, MP Michelagnoli, and PA McKinney. Survival following relapse in childhood haematological malignancies diagnosed in 1974– 2003 in Yorkshire, UK. *British Journal of Cancer*, 96:1147–1152, 2007.

[87] L Fern, S Davies, T Eden, R Feltbower, R Grant, M Hawkins, I Lewis, E Loucaides, C Rowntree, S Stenning, and J Whelan. Rates of inclusion of teenagers and young adults in England into National Cancer Research Network clinical trials: Report from the National Cancer Research Institute (NCRI) Teenage and Young Adult

Clinical Studies Development Group. *British Journal of Cancer*, 99(12):1967–1974, 2008.

[88] LM Woods, B Rachet, and MP Coleman. Origins of socio-economic inequalities in cancer survival: a review. *Annals of Oncology*, 17(1):5, 2006.

[89] KM Gorey, EJ Holowaty, G Fehringer, E Laukkanen, NL Richter, and CM Meyer. An international comparison of cancer survival: metropolitan Toronto, Ontario, and Honolulu, Hawaii. *American journal of public health*, 90(12):1866–1872, 2000.

[90] AL Potosky, RM Merrill, R Ballard-Barbash, GF Riley, SH Taplin, W Barlow, and BH Fireman. Breast cancer survival and treatment in health maintenance organization and fee-for-service settings. *Journal of the National Cancer Institute*, 89(22):1683–1691, 1997.

[91] K Robin Yabroff and Leon Gordis. Does stage at diagnosis influence the observed relationship between socioeconomic status and breast cancer incidence, case-fatality, and mortality? *Social Science & Medicine*, 57(12):2265–2279, 2003.

[92] DK Whynes, EJ Frew, CM Manghan, JH Scholefield, and JD Hardcastle. Colorectal cancer, screening and survival: the influence of socio-economic deprivation. *Public Health*, 117(6):389–395, 2003.

[93] H Wrigley, P Roderick, S George, J Smith, M Mullee, and J Goddard. Inequalities in survival from colorectal cancer: a comparison of the impact of deprivation, treatment, and host factors on observed and cause specific survival. *Journal of Epidemiology and Community Health*, 57(4):301–309, 2003.

[94] GK Singh, BA Miller, BF Hankey, and BK Edwards. Persistent area socioeconomic disparities in US incidence of cervical cancer, mortality, stage, and survival, 1975–2000. *Cancer*, 101(5):1051–1057, 2004.

[95] GR Prout, MN Wesley, PG McCarron, VW Chen, RS Greenberg, RM Mayberry, and BK Edwards. Survival experience of black patients and white patients with bladder carcinoma. *Cancer*, 100(3):621–630, 2004.

[96] CD O'Malley, GM Le, SL Glaser, SJ Shema, and DW West. Socioeconomic status and breast carcinoma survival in four racial/ethnic groups. *Cancer*, 97(5):1303–1311, 2003.

[97] L Meng, G Maskarinec, and L Wilkens. Ethnic differences and factors related to breast cancer survival in Hawaii. *International Journal of Epidemiology*, 26(6): 1151–1158, 1997.

[98] F Kaffashian, S Godward, T Davies, L Solomon, J McCann, and SW Duffy. Socioeconomic effects on breast cancer survival: proportion attributable to stage and morphology. *British Journal of Cancer*, 89(9):1693–1696, 2003.

[99] A Auvinen and S Karjalainen. Possible explanations for social class differences in cancer patient survival. *IARC scientific publications*, 138:377–397, 1997.

[100] RG Roetzheim, N Pal, EC Gonzalez, JM Ferrante, DJ van Durme, and JP Krischer. Effects of health insurance and race on colorectal cancer treatments and outcomes. *American Journal of Public Health*, 90(11):1746, 2000.

[101] G Velikova, L Booth, C Johnston, D Forman, and P Selby. Breast cancer outcomes in South Asian population of West Yorkshire. *British Journal of Cancer*, 90(10): 1926–1932, 2004.

[102] CI Li, KE Malone, and JR Daling. Differences in breast cancer stage, treatment, and survival by race and ethnicity. *Archives of Internal Medicine*, 163(1):49, 2003.

[103] KL Schwartz, H Crossley-May, FD Vigneau, K Brown, and M Banerjee. Race, socioeconomic status and stage at diagnosis for five common malignancies. *Cancer Causes & Control*, 14(8):761–766, 2003.

[104] MT Halpern, EM Ward, AL Pavluck, NM Schrag, J Bian, and AY Chen. Association of insurance status and ethnicity with cancer stage at diagnosis for 12 cancer sites: a retrospective analysis. *The Lancet Oncology*, 9(3):222–231, 2008.

[105] F Sassi, HS Luft, and E Guadagnoli. Reducing racial/ethnic disparities in female breast cancer: screening rates and stage at diagnosis. *American Journal of Public Health*, 96(12):2165–2172, 2006.

[106] U Macleod, S Ross, C Gillis, A McConnachie, C Twelves, and GCM Watt. Socio-economic deprivation and stage of disease at presentation in women with breast cancer. *Annals of Oncology*, 11(1):105–107, 2000.

[107] F Islami, AR Kahn, NA Bickell, MJ Schymura, and P Boffetta. Disentangling the effects of race/ethnicity and socioeconomic status of neighborhood in cancer stage distribution in New York City. *Cancer Causes & Control*, pages 1–10, 2013.

[108] J Maddams, D Brewster, A Gavin, J Steward, J Elliott, M Utley, and H Møller. Cancer prevalence in the United Kingdom: estimates for 2008. *British Journal of Cancer*, 101(3):541–547, 2009.

[109] JP Neglia, DL Friedman, Y Yasui, AC Mertens, S Hammond, M Stovall, SS Donaldson, AT Meadows, and LL Robison. Second malignant neoplasms in five-year survivors of childhood cancer: childhood cancer survivor study. *Journal of National Cancer Institute*, 93(8):618, 2001.

[110] BJ Zebrack, JG Gurney, K Oeffinger, J Whitton, RJ Packer, A Mertens, N Turk, R Castleberry, Z Dreyer, and LL Robison. Psychological outcomes in long-term survivors of childhood brain cancer: a report from the childhood cancer survivor study. *Journal of Clinical Oncology*, 22(6):999–1006, 2004.

[111] RK Mulhern, TE Merchant, A Gajjar, WE Reddick, and LE Kun. Late neurocognitive sequelae in survivors of brain tumours in childhood. *The Lancet Oncology*, 5(7):399–408, 2004.

[112] DA Mulrooney, MW Yeazel, T Kawashima, AC Mertens, P Mitby, M Stovall, SS Donaldson, DM Green, CA Sklar, and LL Robison. Cardiac outcomes in a

cohort of adult survivors of childhood and adolescent cancer: retrospective analysis of the Childhood Cancer Survivor Study cohort. *BMJ*, 339:b4606, 2009.

[113] M Tukenova, C Guibout, O Oberlin, F Doyon, A Mousannif, N Haddy, S Guerin, H Pacquement, A Aouba, and M Hawkins. Role of Cancer Treatment in Long-Term Overall and Cardiovascular Mortality After Childhood Cancer. *Journal of Clinical Oncology*, 28(8):1308, 2010.

[114] RC Reulen, DL Winter, C Frobisher, ER Lancashire, CA Stiller, ME Jenney, R Skinner, MC Stevens, and MM Hawkins. Long-term Cause-Specific Mortality Among Survivors of Childhood Cancer. *Journal of the American Medical Association*, 304(2):172, 2010.

[115] RC Reulen, C Frobisher, DL Winter, J Kelly, ER Lancashire, CA Stiller, K Pritchard-Jones, HC Jenkinson, MM Hawkins, and British Childhood Cancer Survivor Study Steering Group. Long-term risks of subsequent primary neoplasms among survivors of childhood cancer. *JAMA*, 305(22):2311–2319, 2011.

[116] JG Blanco, C-L Sun, W Landier, L Chen, D Esparza-Duran, W Leisenring, A Mays, DL Friedman, JP Ginsberg, and MM Hudson. Anthracycline-related cardiomyopathy after childhood cancer: Role of polymorphisms in carbonyl reductase genesA report from the Children's Oncology Group. *Journal of Clinical Oncology*, 30(13):1415–1421, 2012.

[117] LB Travis, AK Ng, JM Allan, C-H Pui, AR Kennedy, XG Xu, JA Purdy, K Applegate, J Yahalom, and LS Constine. Second malignant neoplasms and cardiovascular disease following radiotherapy. *Journal of the National Cancer Institute*, 2012.

[118] HJ van der Pal, EC van Dalen, E van Delden, IW van Dijk, WE Kok, RB Geskus, E Sieswerda, F Oldenburger, CC Koning, and FE van Leeuwen. High risk of symptomatic cardiac events in childhood cancer survivors. *Journal of Clinical Oncology*, 30(13):1429–1437, 2012.

[119] HC Jenkinson, MM Hawkins, CA Stiller, DL Winter, HB Marsden, and MCG Stevens. Long-term population-based risks of second malignant neoplasms after childhood cancer in Britain. *British Journal of Cancer*, 91(11):1905–1910, 2004.

[120] C-H Pui, C Cheng, W Leung, SN Rai, GK Rivera, JT Sandlund, RC Ribeiro, MV Relling, LE Kun, and WE Evans. Extended follow-up of long-term survivors of childhood acute lymphoblastic leukemia. *New England Journal of Medicine*, 349(7):640–649, 2003.

[121] JG Gurney, KK Ness, SD Sibley, M O'Leary, DR Dengel, JM Lee, NM Youngren, SP Glasser, and KS Baker. Metabolic syndrome and growth hormone deficiency in adult survivors of childhood acute lymphoblastic leukemia. *Cancer*, 107(6): 1303–1312, 2006.

[122] R Mody, S Li, DC Dover, S Sallan, W Leisenring, KC Oeffinger, Y Yasui, LL Robison, and JP Neglia. Twenty-five year follow-up among survivors of childhood acute lymphoblastic leukemia: a report from the Childhood Cancer Survivor Study. *Blood*, 111(12):5515–5523, 2008.

[123] SE Lipshultz, SD Colan, RD Gelber, AR Perez-Atayde, SE Sallan, and SP Sanders. Late cardiac effects of doxorubicin therapy for acute lymphoblastic leukemia in childhood. *New England Journal of Medicine*, 324:808–815, 1991.

[124] L Guldnera, N Haddy, F Pein, A Diallo, A Shamsaldin, M Dahan, J Lebidois, P Merlet, E Villain, D Sidi, O Sakiroglu, O Hartmann, D Leftakopoulos, and F de Vathaire. Radiation dose and long term risk of cardiac pathology following radiotherapy and anthracyclin for a childhood cancer. *Radiotherapy and Oncology*, 81(1):47–56, 2006.

[125] SS Donaldson and HS Kaplan. Complications of treatment of Hodgkin's Disease in children. *Cancer Treatment Reports*, 66(4):977–989, 1982.

[126] SL Hancock, SS Donaldson, and RT Hoppe. Cardiac Disease Following Treatment of Hodgkin's Disease in Children and Adolescents. *Journal of Clinical Oncology*, 11:1208–1215, 1993.

[127] JG Gurney, NS Kadan-Lottick, RJ Packer, JP Neglia, CA Sklar, JA Punyko, M Stovall, Y Yasui, HS Nicholson, and S Wolden. Endocrine and cardiovascular late effects among adult survivors of childhood brain tumors. *Cancer*, 97(3):663–673, 2003.

[128] A MacCarthy, AM Bayne, GJ Draper, EM Eatock, ME Kroll, CA Stiller, TJ Vincent, MM Hawkins, HC Jenkinson, JE Kingston, R Neale, and Murphy MFG. Non-ocular tumours following retinoblastoma in Great Britain 1951 to 2004. *British Journal of Ophthalmology*, 93:1159–1162, 2009.

[129] O Fletcher, D Easton, K Anderson, C Gilham, M Jay, and J Peto. Lifetime Risks of Common Cancers Among Retinoblastoma Survivors. *J Natl Cancer Inst*, 96(5): 357–363, 2004.

[130] L Aung, RG Gorlick, W Shi, H Thaler, NA Shorter, J Healey, AG Huvos, and PA Meyers. Second malignant neoplasms in long term survivors of osteosarcoma: Memorial Sloan–Kettering Cancer Center experience. *Cancer*, 95:1728–1732, 2002.

[131] MC Stevens. Treatment for childhood rhabdomyosarcoma: the cost of cure. *Lancet Oncology*, 6:77–84, 2005.

[132] L Sung, JR Anderson, SS Donaldson, SL Spunt, WM Crist, and AS Pappo. Late events occurring five years or more after successful therapy for childhood rhabdomyosarcoma: a report from the Soft Tissue Sarcoma Committee of the Children's Oncology Group. *European Journal of Cancer*, 40(12):1878–1885, 2004.

[133] E Woodward, M Jessop, A Glaser, and D Stark. Late effects in survivors of teenage and young adult cancer: does age matter? *Annals of Oncology*, 22(12):2561–2568, 2011.

[134] R Skinner, WHB Wallace, and GA Levitt. Long-term follow-up of people who have survived cancer during childhood. *The Lancet Oncology*, 7(6):489–498, 2006.

[135] MM Hawkins. Survivorship outcomes research based on record linkage. *Pediatric Blood & Cancer*, 55(2):224–225, 2010.

[136] Y Zhang, MF Lorenzi, K Goddard, JJ Spinelli, C Gotay, and ML McBride. Late morbidity leading to hospitalization among 5-year survivors of young adult cancer: A report of the childhood, adolescent and young adult cancer survivors research program. *International Journal of Cancer*, 134(5):1174–1182, 2014.

[137] O Harel and XH Zhou. Multiple imputation: review of theory, implementation and software. *Statistics in medicine*, 26(16):3057–3077, 2007.

[138] SR Lipsitz, JG Ibrahim, MH Chen, and H Peterson. Non ignorable missing covariates in generalized linear models. *Statistics in Medicine*, 18(17):2435–2448, 1999.

[139] NJ Horton and NM Laird. Maximum likelihood analysis of logistic regression models with incomplete covariate data and auxiliary information. *Biometrics*, 57 (1):34–42, 2001.

[140] P Royston. Multiple imputation of missing values. *The Stata Journal*, 4(3):227–241, 2004.

[141] TD Pigott. A review of methods for missing data. *Educational Research and Evaluation*, 7(4):353–383, 2001.

[142] JAC Sterne, IR White, JB Carlin, M Spratt, P Royston, MG Kenward, AM Wood, and JR Carpenter. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ*, 338, 2009.

[143] S van Buuren. Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, 16:219–242, 2007.

[144] CK Enders. *Applied missing data analysis*. The Guilford Press, 2010.

[145] JK Vermunt, JR van Ginkel, LA van der Ark, and K Sijtsma. Multiple imputation of incomplete categorical data using latent class analysis. *Sociological Methodology*, 38(1):369–397, 2008.

[146] Mulugeta Gebregziabher and Stacia M DeSantis. Latent class based multiple imputation approach for missing categorical data. *Journal of Statistical Planning and Inference*, 140(11):3252–3262, 2010.

[147] Yajuan Si and Jerome P Reiter. Nonparametric Bayesian multiple imputation for incomplete categorical variables in large-scale assessment surveys. *Journal of Educational and Behavioral Statistics*, 38(5):499–521, 2013.

[148] MJ Daniels and JW Hogan. *Missing data in longitudinal studies: Strategies for Bayesian modeling and sensitivity analysis*, volume 109. Chapman & Hall, 2008.

[149] JW Bartlett, SR Seaman, IR White, and JR Carpenter. Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model. *Statistical Methods in Medical Research*, pages 1–26, 2014.

[150] ED De Leeuw. Reducing missing data in surveys: an overview of methods. *Quality & Quantity*, 35(2):147–160, 2001.

[151] DB Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

[152] BA Cattle, PD Baxter, DC Greenwood, CP Gale, and RM West. Multiple imputation for completion of a national clinical audit dataset. *Statistics in Medicine*, 30(22):2736–2753, 2011.

[153] JL Schafer. *Analysis of incomplete multivariate data*, volume 72. Chapman & Hall/CRC, 1997.

[154] LM Collins, JL Schafer, and CM Kam. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, 6(4):330, 2001.

[155] H Demirtas and JL Schafer. On the performance of random–coefficient pattern–mixture models for non–ignorable drop–out. *Statistics in Medicine*, 22(16):2553–2575, 2003.

[156] DB Rubin, HS Stern, and V Vehovar. Handling "don't know" survey responses: the case of the Slovenian plebiscite. *Journal of the American Statistical Association*, pages 822–828, 1995.

[157] M Jamshidian and S Jalal. Tests of homoscedasticity, normality, and missing completely at random for incomplete multivariate data. *Psychometrika*, 75(4):649–674, 2010.

[158] M Jamshidian, SJ Jalal, and C Jansen. MissMech: an R package for testing homoscedasticity, multivariate normality, and missing completely at random (MCAR). *Journal of Statistical Software*, 56(6), 2014.

[159] Douglas M Hawkins. A new test for multivariate normality and homoscedasticity. *Technometrics*, 23(1):105–110, 1981.

[160] S van Buuren, HC Boshuizen, and DL Knook. Multiple Imputation of Missing Blood Pressure Covariates in Survival Analysis. *Statistics in Medicine*, 18:681–694, 1999.

[161] RJA Little and DB Rubin. *Statistical analysis with missing data*. Wiley New York, second edition, 2002.

[162] CJ Pannucci and EG Wilkins. Identifying and avoiding bias in research. *Plastic and Reconstructive Surgery*, 126(2):619, 2010.

[163] F Godlee. Milestones on the long road to knowledge. *BMJ*, 334(suppl 1):s2–s3, 2007.

[164] KJ Thomas, H MacPherson, L Thorpe, J Brazier, M Fitter, MJ Campbell, M Roman, SJ Walters, and J Nicholl. Randomised controlled trial of a short course of traditional acupuncture compared with usual care for persistent non-specific low back pain. *BMJ*, 333(7569):623, 2006.

[165] Y Ginsberg and N Lindefors. Methylphenidate treatment of adult male prison inmates with attention-deficit hyperactivity disorder: randomised double-blind placebo-controlled trial with open-label extension. *The British Journal of Psychiatry*, 200(1):68–73, 2011.

[166] PN Tariot, MR Farlow, GT Grossberg, SM Graham, S McDonald, and I Gergel. Memantine treatment in patients with moderate to severe Alzheimer disease already receiving donepezil. *Journal of the American Medical Association*, 291 (3):317–324, 2004.

[167] RJA Little and DB Rubin. *Statistical analysis with missing data*. Wiley New York, 1987.

[168] SS Wilks. Moments and distributions of estimates of population parameters from fragmentary samples. *The Annals of Mathematical Statistics*, 3(3):163–195, 1932.

[169] S Ju, D Zhang, and J Pacha. Employability Skills Valued by Employers as Important for Entry-Level Employees With and Without Disabilities. *Career Development for Exceptional Individuals*, 2011.

[170] N Frasure-Smith and F Lesperance. Depression and anxiety as predictors of 2-year cardiac events in patients with stable coronary artery disease. *Archives of General Psychiatry*, 65(1):62, 2008.

[171] EML Beale and RJA Little. Missing values in multi-variate analysis. *Journal of the Royal Statistical Society*, Series B(37):129–145, 1975.

[172] SF Buck. A method of estimation of missing values in multivariate data suitable for use with electronics computer. *Journal of the Royal Statistical Society*, Series B(22):302–306, 1960.

[173] JNK Rao and J Shao. Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, 79(4)(4):811–822, 1992.

[174] AP Dempster, NM Laird, and DB Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society Series B (Methodological)*, 39(1):1–38, 1977.

[175] XL Meng and DB Rubin. Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm. *Journal of the American Statistical Association*, pages 899–909, 1991.

[176] AG McKendrick. Applications of mathematics to medical problems. *Proceedings of the Edinburgh Mathematical Society*, 44(1):98–130, 1925.

[177] T Orchard and MA Woodbury. A missing information principle: Theory and applications. *Proceedings of the 6th Berkley Symposium on Mathematical Statistics and Probability*, 1:697–715, 1972.

[178] R Sundberg. Maximum likelihood theory for incomplete data from an exponential family. *Scandinavian Journal of Statistics*, pages 49–58, 1974.

[179] HO Hartley. Maximum likelihood estimation from incomplete data. *Biometrics*, 14(2):174–194, 1958.

[180] LE Baum, T Petrie, G Soules, and N Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The annals of mathematical statistics*, 41(1):164–171, 1970.

[181] XL Meng and DB Rubin. Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, 80(2):267, 1993.

[182] CK Enders. Using the Expectation Maximization Algorithm to Estimate Coefficient Alpha for Scales With Item-Level Missing Data. *Psychological Methods*, 8(3):322, 2003.

[183] G Claeskens and F Consentino. Variable selection with incomplete covariate data. *Biometrics*, 64(4):1062–1069, 2008.

[184] JG Ibrahim. Incomplete Data in Generalized Linear Models. *American Statistical Association*, 85(411):765–769, 1990.

[185] SR Lipsitz and JG Ibrahim. Using the EM-algorithm for survival data with incomplete categorical covariates. *Lifetime Data Analysis*, 2(1):5–14, 1996.

[186] SR Lipsitz and JG Ibrahim. Estimating equations with incomplete categorical covariates in the Cox model. *Biometrics*, pages 1002–1013, 1998.

[187] JG Ibrahim, SR Lipsitz, and MH Chen. Missing covariates in generalized linear models when the missing data mechanism is non-ignorable. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(1):173–190, 1999.

[188] C Jenkinson, C Heffernan, H Doll, and R Fitzpatrick. The Parkinsons Disease Questionnaire (PDQ-39): evidence for a method of imputing missing data. *Age and Ageing*, 35(5):497–502, 2006.

[189] C Jenkinson, R Harris, and R Fitzpatrick. The Amyotrophic Lateral Sclerosis Assessment Questionnaire (ALSAQ-40): Evidence for a method of imputing missing data. *Amyotrophic Lateral Sclerosis*, 8(2):90–95, 2007.

[190] DB Rubin. *Multiple imputations in sample surveys - a phenomenological Bayesian approach to nonresponse*. Proceedings of the Section on Survey Research Methods Section American Statistical Association, 1978.

[191] DB Rubin. *Multiple imputation for non-response in surveys*. New York: John Wiley & Sons, 1987.

[192] S van Buuren and K Groothuis-Oudshoorn. mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, 45(3), 2011.

[193] S van Buuren and K Oudshoorn. Flexible multivariate imputation by MICE. *Leiden, The Netherlands: TNO Prevention Center*, 1999.

[194] K-H Li. Imputation using Markov Chains. *Journal of Statistical Computing and Simulation*, 30:57–79, 1988.

[195] DB Rubin and JL Schafer. Efficiently creating multiple imputations for incomplete multivariate normal data. *Proceedings of the Statistical Computing Section*, pages 83–88, 1990.

[196] TE Raghunathan, JM Lepkowski, J van Hoewyk, and P Solenberger. A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, 27(1):85–96, 2001.

[197] P Royston. Multiple imputation of missing values: Further update of ice, with an emphasis on categorical variables. *The Stata Journal*, 9(3):466–477, 2009.

[198] F Li, Y Yu, and DB Rubin. Imputing missing data by fully conditional models: Some cautionary examples and guidelines. *Duke University Department of Statistical Science Discussion Paper*, 24(11), 2012.

[199] JL Schafer. Multiple imputation: a primer. *Statistics Methods in Medical Research*, 8:3–15, 1999.

[200] TE Bodner. What Improves with Increased Missing Data Imputations? *Structural Equation Modeling: A Multidisciplinary Journal*, 15(4):651–675, 2008.

[201] JW Graham, AE Olchowski, and TD Gilreath. How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science*, 8(3)(3):206–213, 2007.

[202] IR White, P Royston, and AM Wood. Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*, 30:377–399, 2011.

[203] KGM Moons, RART Donders, T Stijnen, and FE Harrell. Using the outcome for imputation of missing predictor values was preferred. *Journal of Clinical Epidemiology*, 59(10):1092–1101, 2006.

[204] J Hippisley-Cox, C Coupland, Y Vinogradova, J Robson, M May, and P Brindle. Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study. *BMJ*, 335(7611):136, 2007.

[205] IR White and P Royston. Imputing missing covariate values for the Cox model. *Statistics in Medicine*, 28(15):1982–1998, 2009.

[206] R Young and DR Johnson. *Proceedings of the AAPOR Conference Abstracts*, chapter Imputing the Missing Y's: Implications for Survey Producers and Survey Users. American Statistical Association, 2010.

[207] PD Allison. Missing data - Quantitative Applications in the Social Sciences (Vol 136). *Thousand Oaks Sage Publications*, 2002.

[208] PT von Hippel. Regression with missing Y's: An improved strategy for analyzing multiply imputed data. *Sociological Methodology*, 37(1):83–117, 2007.

[209] MA Klebanoff and SR Cole. Use of multiple imputation in the epidemiologic literature. *American Journal of Epidemiology*, 168(4):355, 2008.

[210] AM Wood, IR White, and SG Thompson. Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. *Clinical Trials*, 1:368–376, 2004.

[211] S Greenland and WD Finkle. A critical look at methods for handling missing covariates in epidemiologic regression analyses. *American Journal of Epidemiology*, 142(12):1255–1264, 1995.

[212] DB Rubin. Multiple imputation after 18+ years. *Journal of the American Statistical Association*, pages 473–489, 1996.

[213] TG Clark, ME Stewardt, DG Altman, H Gabra, and JF Smyth. A prognostic model for ovarian cancer. *British Journal of Cancer*, 85(7):944–952, 2001.

[214] TG Clark and DG Altman. Developing a prognostic model in the presence of missing data: an ovarian cancer case study. *Journal of Clinical Epidemiology*, 56: 28–37, 2003.

[215] MK Ferguson, J Siddique, and T Karrison. Modeling major lung resection outcomes using classification trees and multiple imputation techniques. *European Journal of Cardio-thoracic Surgery*, 34:1085–1089, 2008.

[216] U Nur, B Rachet, MKB Parmar, MR Sydes, N Cooper, C Lepage, JMA Northover, R James, and MP Coleman. No socioeconomic inequalities in colorectal cancer survival within a randomised clinical trial. *British Journal of Cancer*, 99:1923–1928, 2008.

[217] JS Mandleblatt, VB Sheppard, A Hurria, G Kimmick, C Isaacs, KL Taylor, AB Kornblith, AM Noone, G Luta, M Tallarico, WT Barry, L Hunegs, R Zon, M Naughton, E Winer, C Hudis, SB Edge, HJ Cohen, and H Muss. No socioeconomic inequalities in colorectal cancer survival within a randomised clinical trial. *Journal of Clinical Oncology*, 28(19):3146–3153, 2010.

[218] SR Seaman and IR White. Review of inverse probability weighting for dealing with missing data. *Statistical Methods in Medical Research*, 2011.

[219] M Höfler, H Pfister, R Lieb, and H Wittchen. The use of weights to account for non-response and drop-out. *Social Psychiatry and Psychiatric Epidemiology*, 40: 291–299, 2005.

[220] RC Kessler, RJA Little, and RM Groves. Advances in Strategies for Minimizing and Adjusting for Survey Nonresponse. *Epidemiologic Reviews; The Johns Hopkins University School of Hygiene and Public Health*, 17(1):192–204, 1995.

[221] JM Robins. Non-response models for the analysis of non-monotone non-ignorable missing data. *Statistics in Medicine*, 16(1):21–37, 1997.

[222] D Clayton, Spiegelhalter D, Dunn G, and Pickles A. Analysis of longitudinal binary data from multi-phase sampling (with discussion). *Journal of the Royal Statistical Society Series B (statistical methodology)*, pages 71–87, 1998.

[223] S van Buuren. Multiple imputation of multilevel data. *Handbook of advanced multilevel analysis*, pages 173–196, 2011.

[224] C Beunckens, C Sotto, and G Molenberghs. A simulation study comparing weighted estimating equations with multiple imputation based estimating equations for longitudinal binary data. *Computational Statistics & Data Analysis*, 52(3): 1533–1548, 2008.

[225] J Carpenter, M Kenward, and S Vansteelandt. A comparison of multiple imputation and inverse probability weighting for analyses with missing data. *Journal of the Royal Statistical Society, Series A*, 169(3)(3):571–84, 2006.

[226] H Kang. The prevention and handling of the missing data. *Korean Journal of Anesthesiology*, 64(5):402–406, 2013.

[227] RJ Little, R D'Agostino, ML Cohen, K Dickersin, SS Emerson, JT Farrar, C Frangakis, JW Hogan, G Molenberghs, and SA Murphy. The prevention and treatment of missing data in clinical trials. *New England Journal of Medicine*, 367 (14):1355–1360, 2012.

[228] MG Kenward and J Carpenter. Multiple imputation: current perspectives. *Statistical Methods in Medical Research*, 16(3):199–218, 2007.

[229] S van Buuren, JPL Brand, CGM Groothuis-Oudshoorn, and DB Rubin. Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, 76(12):1049–1064, 2006.

[230] Health & Social Care Information Centre. HES User Guide, 2010. URL `www.hesonline.nhs.uk`. [Accessed: 24/11/2011].

[231] A Lakhani, H Olearnik, and D Eayres. Compendium of clinical and health indicators - Annex 4. *London: Department of Health, National Centre For Health Outcomes Development*, 2011.

[232] NHS Connecting for Health. *OPCS Classification of Interventions and Procedures Version 45, Volume 1: Tabular List*. The Stationery Office, 2009.

[233] AM Pollock and N Vickers. Trends in colorectal cancer care in southern England, 1989-1993: using HES data to inform cancer services reviews. *Journal of Epidemiology, Community and Health*, 52(7):433–438, 1998.

[234] E Morris, P Quirke, JD Thomas, L Fairley, B Cottier, and D Forman. Unacceptable variation in abdominoperineal excision rates for rectal cancer: time to intervene? *Gut*, 57(12):1690–1697, 2008.

[235] J Maddams, M Utley, and H Møller. Levels of acute health service use among cancer survivors in the United Kingdom. *European Journal of Cancer*, 47(14): 2211–2220, 2011.

[236] StataCorp. Stata: Release 13 Statistical Software. *College Station, Texas: StataCorp LP*, 2013.

[237] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014. URL `http://wwwR-projectorg`.

[238] J Carpenter and M Kenwards. Brief comments on computational issues with multiple imputation, 2008. URL `http://missingdata.lshtm.ac.uk/downloads/mi_comp_issues.pdf`.

[239] RJA Little and N Schenker. Missing Data. In *Handbook for Statistical Modeling in the Social and Behavioral Sciences*, pages 39–75. Springer, 1995.

[240] N Eisemann, A Waldmann, and A Katalinic. Imputation of missing values of tumour stage in population-based cancer registration. *BMC Medical Research Methodology*, 11(129):1471–2288, 2011.

[241] F Barzi and M Woodward. Imputations of missing values in practice: results from imputations of serum cholesterol in 28 cohort studies. *American Journal of Epidemiology*, 160(1):34–45, 2004.

[242] P Royston. Multiple imputation of missing values: update of ice. *The Stata Journal*, 5(2):188, 2005.

[243] Stef van Buuren. *Flexible imputation of missing data*. Chapman & Hall, 2012.

[244] A Marshall, DG Altman, RL Holder, and P Royston. Combining estimates of interest in prognostic modelling studies after multiple imputation: current practice and guidelines. *BMC Medical Research Methodology*, 9(1):57, 2009.

[245] D Collett. *Modelling survival data in medical research*. CRC press, 2003.

[246] J Fox. *Applied regression analysis, linear models, and related methods*. Sage, 1997.

[247] Xiao-Li Meng and Donald B Rubin. Performing likelihood ratio tests with multiply-imputed data sets. *Biometrika*, 79(1):103–111, 1992.

[248] JR Wilkinson, RG Feltbower, IJ Lewis, RC Parslow, and PA McKinney. Survival from adolescent cancer in Yorkshire, UK. *European Journal of Cancer*, 37(7): 903–911, 2001.

[249] Communities and Local Government. The English Indices of Deprivation 2010. Neighbourhoods Statistical Release, March 2011. URL `https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/6871/1871208.pdf`. [Accessed: 25/09/2014].

[250] M Bajekal, S Scholes, M OFlaherty, R Raine, P Norman, and S Capewell. Unequal Trends in Coronary Heart Disease Mortality by Socioeconomic Circumstances, England 1982–2006: An Analytical Study. *PloS one*, 8(3):e59608, 2013.

[251] H Jordan, P Roderick, and D Martin. The Index of Multiple Deprivation 2000 and accessibility effects on health. *Journal of Epidemiology and Community Health*, 58(3):250–257, 2004.

[252] JP Archie Jr. Mathematic coupling of data: a common source of error. *Annals of Surgery*, 193(3):296, 1981.

[253] J Adams and M White. Removing the health domain from the Index of Multiple Deprivation 2004effect on measured inequalities in census measure of health. *Journal of Public Health*, 28(4):379–383, 2006.

[254] P Townsend, P Phillimore, and A Beattie. *Health and deprivation: inequality and the North*. Routledge Kegan & Paul, 1988.

[255] V Carstairs and R Morris. Deprivation: explaining differences in mortality between Scotland and England and Wales. *BMJ*, 299(6704):886, 1989.

[256] Caroline Bennette and Andrew Vickers. Against quantiles: categorization of continuous variables in epidemiologic research, and its discontents. *BMC medical research methodology*, 12(1):21, 2012.

[257] KJ Rothman, S Greenland, and Timothy L Lash. *Modern Epidemiology*. Lippincott Williams & Wilkins, 2008.

[258] S Greenland. Avoiding power loss associated with categorization and ordinal scores in dose-response and trend analysis. *Epidemiology*, 6(4):450–454, 1995.

[259] CR Weinberg. How bad is categorization? *Epidemiology*, pages 345–347, 1995.

[260] TJ Hastie and RJ Tibshirani. *Generalized additive models*, volume 43. CRC Press, 1990.

[261] S Greenland. Dose-response and trend analysis in epidemiology: alternatives to categorical analysis. *Epidemiology*, pages 356–365, 1995.

[262] P Royston. A strategy for modelling the effect of a continuous covariate in medicine and epidemiology. *Statistics in Medicine*, 19(14):1831–1847, 2000.

[263] M van Laar, PA McKinney, DP Stark, A Glaser, SE Kinsey, IJ Lewis, SV Picton, M Richards, PD Norman, and RG Feltbower. Survival trends of cancer amongst the south Asian and non-south Asian population under 30 years of age in Yorkshire, UK. *Cancer Epidemiology*, 36(1):e13–e18, 2012.

[264] P Peduzzi, J Concato, AR Feinstein, and TR Holford. Importance of events per independent variable in proportional hazards regression analysis II: Accuracy and precision of regression estimates. *Journal of Clinical Epidemiology*, 48(12):1503–1510, 1995.

[265] R Simon and DG Altman. Statistical aspects of prognostic factor studies in oncology. *British Journal of Cancer*, 69(6):979, 1994.

[266] DG Altman. *Practical statistics for medical research*. CRC Press, 1990.

[267] E Vittinghoff and CE McCulloch. Relaxing the rule of ten events per variable in logistic and Cox regression. *American Journal of Epidemiology*, 165(6):710–718, 2007.

[268] PT von Hippel. How to impute interactions, squares, and other transformed variables. *Sociological Methodology*, 39(1):265–291, 2009.

[269] TM Therneau, PM Grambsch, and TR Fleming. Martingale-based residuals for survival models. *Biometrika*, 77(1):147–160, 1990.

[270] MG Wilson. Assessing model adequacy in proportional hazards regression. *Statistics and Data Analysis*, 431:1–36, 2013.

[271] FE Harrell Jr, RM Califf, DB Pryor, KL Lee, and RA Rosati. Evaluating the yield of medical tests. *JAMA*, 247(18):2543, 1982.

[272] DW Hosmer and S Lemeshow. A goodness of fit test for the multiple logistic regression model. *Communication in Statistics*, A10:1043–1069, 1980.

[273] DW Hosmer and S Lemeshow. *Applied Logistic Regression*. Wiley series in probability and statistics. Wiley, 2nd edition, 2000.

[274] E Lesaffre. Logistic discriminant analysis with applications in electrocardiography. *PhD Thesis, Katholieke Universiteit Leuven*, 1986.

[275] CB Begg and R Gray. Calculation of polychotomous logistic regression parameters using individualized regressions. *Biometrika*, 71(1):11–18, 1984.

[276] S Juul. Chapter 14: Incidence, Mortality and Survival. In *An Introduction to Stata for Health Researchers*, pages 227–232. Stata Press, Texas, 2006.

[277] P Royston and PC Lambert. Flexible parametric survival analysis using Stata: beyond the Cox model. *Stata Press books*, 2011.

[278] JM Bennett, D Catovsky, M-Th Daniel, G Flandrin, DAG Galton, HR Gralnick, and C Sultan. Proposals for the Classification of the Acute Leukaemias French-American-British (FAB) Co-operative Group. *British Journal of Haematology*, 33 (4):451–458, 1976.

[279] H Dohner, S Stilgenbauer, A Benner, E Leupolt, A Krober, L Bullinger, K Dohner, M Bentz, and P Lichter. Genomic aberrations and survival in chronic lymphocytic leukemia. *New England Journal of Medicine-Unbound Volume*, 343(26):1910–1916, 2000.

[280] S Nguyen, T Leblanc, P Fenaux, F Witz, D Blaise, A Pigneux, X Thomas, F Rigal-Huguet, B Lioure, and A Auvrignon. A white blood cell index as the main prognostic factor in t (8; 21) acute myeloid leukemia (AML): a survey of 161 cases from the French AML Intergroup. *Blood*, 99(10):3517, 2002.

[281] BJ Lange, FO Smith, J Feusner, DR Barnard, P Dinndorf, S Feig, NA Heerema, C Arndt, RJ Arceci, and N Seibel. Outcomes in CCG-2961, a children's oncology group phase 3 trial for untreated pediatric acute myeloid leukemia: a report from the children's oncology group. *Blood*, 111(3):1044, 2008.

[282] GM Brodeur, J Pritchard, F Berthold, NL Carlsen, V Castel, RP Castelberry, B De Bernardi, AE Evans, M Favrot, and F Hedborg. Revisions of the international criteria for neuroblastoma diagnosis, staging, and response to treatment. *Journal of Clinical Oncology*, 11(8):1466–77, 1993.

[283] SL Cohn, ADJ Pearson, WB London, T Monclair, PF Ambros, GM Brodeur, A Faldum, B Hero, T Iehara, and D Machin. The International Neuroblastoma Risk Group (INRG) classification system: an INRG task force report. *Journal of Clinical Oncology*, 27(2):289–297, 2009.

[284] BH Kushner, MP LaQuaglia, K Kramer, and NKV Cheung. Radically different treatment recommendations for newly diagnosed neuroblastoma: pitfalls in assessment of risk. *Journal of Pediatric Hematology/Oncology*, 26(1):35–39, 2004.

[285] T Monclair, GM Brodeur, PF Ambros, HJ Brisse, G Cecchetto, K Holmes, M Kaneko, WB London, KK Matthay, JG Nuchtern, D von Schweinitz, T Simon, SL Cohn, AD Pearson, and The IRNG Task Force. The International Neuroblastoma Risk Group (INRG) staging system: an INRG Task Force report. *Journal of Clinical Oncology*, 27(2):298–303, 2009.

[286] AB Reese and RM Ellsworth. The evaluation and current concept of retinoblastoma therapy. *Transactions-American Academy of Ophthalmology and Otolaryngology*, 67:164, 1963.

[287] MA Linn. Intraocular retinoblastoma: the case for a new group classification. *Ophthalmology Clinics of North America*, 18(1):41, 2005.

[288] A Horwich. Testicular Cancer. In *Oncology a multidisciplinary textbook*, pages 485–498. Chapman and Hall, London, 1995.

[289] RJ Packer, BH Cohen, and K Cooney. Intracranial germ cell tumors. *The Oncologist*, 5(4):312–320, 2000.

[290] CH Chang, EM Housepian, and C Herbert. An operative staging system and a megavoltage radiotherapeutic technic for cerebellar medulloblastomas. *Radiology*, 93(6):1351–1359, 1969.

[291] JC Allen. Controversies in the management of intracranial germ cell tumors. *Neurologic clinics*, 9(2):441, 1991.

[292] PT von Hippel. Should a Normal Imputation Model be Modified to Impute Skewed Variables? *Sociological Methodology*, 42(1):105–138, 2012.

[293] WA Bleyer. Cancer in older adolescents and young adults: epidemiology, diagnosis, treatment, survival, and importance of clinical trials. *Medical and Pediatric Oncology*, 38:1–10, 2002.

[294] C Stiller. Epidemiology of cancer in adolescents. *Medical and Pediatric Oncology*, 39(3):149–155, 2002.

[295] C Percy, E Stanek 3rd, and L Gloeckler. Accuracy of cancer death certificates and its effect on cancer mortality statistics. *American Journal of Public Health*, 71(3): 242–250, 1981.

[296] CB Begg and D Schrag. Attribution of deaths following cancer treatment. *Journal of the National Cancer Institute*, 94(14):1044–1045, 2002.

[297] Fred Ederer, Lillian M Axtell, and Sidney Joshua Cutler. The relative survival rate: a statistical methodology. *National Cancer Institute Monograph*, 6:101–121, 1961.

[298] G Hagger-Johnson, K Harron, A Gonzalez-Izquierdo, M Cortina-Borja, N Dattani, B Muller-Pebody, R Parslow, R Gilbert, and H Goldstein. Identifying Possible False Matches in Anonymized Hospital Administrative Data without Patient Identifiers. *Health services research*, 2014.

[299] M van Laar, RG Feltbower, CP Gale, DT Bowen, SE Oliver, and A Glaser. Cardiovascular sequelae in long-term survivors of young peoples cancer: a linked cohort study. *British Journal of Cancer*, 110(5):1338–1341, 2014.