Estimating Prevalence of Haematological Malignancies Using Data from the Haematological Malignancy Research Network (HMRN)

Jinlei Li

PhD

University of York

Health Sciences

February 2014

Abstract

The prevalence of the haematological malignancies enumerates those currently living with past diagnosis of this class of diseases, and provides insights regarding survivor populations and their burden. However, there is a lack of accurate information regarding the prevalence of the haematological malignancies. This is partly because of changing disease classifications and the fact that the current methods available to estimate total prevalence have not always been appropriate due to the characteristics of the disease including age at diagnosis and the introduction of novel treatments that have altered outcomes.

Using data from the Haematological Malignancy Research Network (HMRN) a method was developed to estimate the prevalence of the haematological malignancies, according to current disease classification, in HMRN region and for the UK as a whole. The method used a mathematical model and flexible statistical methods to estimate the total prevalence on 31st, August, 2011.

Total prevalence estimates that about 19,700 cases in HMRN area are living with a prior diagnosis of haematological malignancy on the index date. Among them, about 9,600 living cases were diagnosed before the establishment of HMRN registry. Using observed prevalence, it was estimated that in the UK there are 165,841 cases of haematological malignancies; however, total prevalence estimates 327,818 cases. Subtypes showed different disease burdens due to their own characteristics.

This thesis is the first study to calculate the prevalence of haematological malignancies using current disease classification (ICD-O-3). It provides indicators of real burden of haematological malignancies for each of the subtypes in HMRN area; these can then be extrapolated to the UK as a whole.

List of contents

List of figures

List of tables

Acknowledgments

I am deeply indebted to my supervisors, Alex Smith and Simon Crouch, for their support and help in my PhD study and research. Their guidance helped me throughout the research and writing of this thesis. They provided me with the guidance, assistance, and expertise that I needed over the course of these three years. I could not have imagined having better supervisors for my PhD study. I would also like to thank Alex for providing me with the chance to join the EUMDS project. I got a lot of practice and experience from it.

Besides my supervisors, I would like to thank the other members of my Thesis Advisory Panel (TAP), Eve Roman and Steven Oliver, for their time and valuable feedback on my research. Their encouragement, insightful comments, and hard questions made my thesis better and better after every meeting.

I would like to express my sincere gratitude to my university, the University of York and the Department of Health Sciences, which offered me this chance to study for a PhD, and the Epidemiology & Cancer Statistics Group that offered me a scholarship to support my research for these three years.

I would like to say a heartfelt thank you to my parents. They gave me unlimited love and always believed in me, encouraging me to follow my dream. I would especially like to thank my best friend Chengcheng Hao, who shared her knowledge and always talked with me when I met problems. My deep appreciation goes out to all my family and friends in the UK and China.

Author’s declaration

I hereby certify that I am the author of this thesis. No part of this thesis has been published or submitted for publication. All research presented in this thesis was conducted by the author, with guidance from members of the University of York.

Chapter 1 Introduction

1.1 Background

1.1.1 Prevalence

The prevalence of a disease in a population is the proportion of people who have received a diagnosis of that disease in the past and that are alive on a specified date, which is called the index date. Unlike incidence, which can provide information on disease prevalence for diseases of short duration (the patient died or was cured), prevalence is more informative for diseases of relatively long duration. It helps to measure the burden of disease in a population and is an important measure for health and social care planning. Conceptually, it seems straightforward to obtain the prevalence of disease by counting the number of survivors of the disease alive at any point in time. There are, however, many methodological challenges associated with estimating disease prevalence and these will be discussed in the following chapter.

1.1.1.1 Cancer registration

This thesis is concerned with the prevalence of cancer. Obtaining accurate estimates of cancer prevalence requires either broad sampling by surveillance or estimates based on available registry data. For survey-based reporting, recall bias may lead to over or under reporting of diseases. The accuracy of self-reported data may vary by disease type, and the likelihood of misreporting is different

among subgroups of patients. Therefore, it is common and convenient to estimate cancer prevalence using data from cancer registries.

Here, the term "cancer registry" usually refers to a population-based cancer registry, which collects data on cancer occurring in a well-defined population. Information for these registries usually comes from treatment facilities, such as hospitals and private clinics, and diagnostic services, such as pathology departments, radiology departments and death certificates. Data can be actively obtained by personnel visiting different departments or passively by heath care workers notifying cancer registries. More recently, there is electronic capture of data, which may integrate cancer registration into the patient administrative systems (Gjerstorff, 2011). The data items collected are determined by the aims of the registry. These usually, but not only, include personal identification (such as name, sex, and date of birth), demographics (such as address and ethnicity), the cancer and its investigations (such as diagnosis, classification), treatment, and follow-up (Silva, 1999). Reporting of cancer cases to a registry may be voluntary, or compulsory by legislation or administrative order. Confidentiality should be taken into account to protect individual privacy (Jensen, et al., 1991).

From registry data, information regarding patients' gender, age at diagnosis, cancer sites, and status at last follow-up can be obtained relatively easily. In terms of cancer registration it is generally assumed that once a patient has been diagnosed with cancer, they remain a prevalent case until death (Silva 1999). Within a registry, for example, the prevalence of a disease diagnosed within a limited duration can be captured conveniently from the available data. However, although it is relatively simple to calculate, this measurement may potentially lead to an underestimate of actual prevalence for diseases which have longer survival periods than spans the time period a registry has been in existence.

When the registry has been in operation for many years, the prevalent cases may simply be enumerated from registry data. Therefore, within a cancer registry

observation period, such direct methods simply exploit incidence and life status to count the number of cancer patients living at a certain time in the population. However, this numerical direct method can only provide prevalence for *L* years (where *L* is the length of registry period). This is known as **observed prevalence**, which covers all patients diagnosed after the start of the registry. **n-year** prevalence measures the proportion of the population alive on the index date that have received a diagnosis of the disease in the period of n years before the index date. For example, 5-year prevalence is based on the most recent 5 years of available registry data. Both of these measure prevalence in limited durations, and may provide biased estimates of the total prevalence of the disease in the population. Developments of a method to estimate such a **total prevalence** (that is, the proportion of the population alive on the index date who have ever received a diagnosis of the disease), can correct for this bias and provide better guidance for the planning of health care services. (All definitions are described in section 1.2.)

The population-based cancer registry used in this thesis is the Haematological Malignancy Research Network (HMRN), and is further described in Chapter Three.

1.1.1.2 Motivation for this work

Despite many reports in the literature on the incidence, mortality and survival of cancer, relatively few studies exist describing prevalence (Merrill, et al., 2000; Capocaccia, et al., 2002; Verdecchia, et al., 2002; Forman, et al., 2003; Lutz, et al., 2003; Möller, et al., 2003). n-year prevalence, which is abstracted simply from registry data, is the usual way that cancer prevalence is reported in the literature; such as 1-year prevalence, 5-year prevalence, and 10-year prevalence. Observed prevalence is highly related to the length of registry. Therefore, measures of observed prevalence are not comparable among different registries. Estimates

based on total prevalence are rarely reported, possibly reflecting the challenges associated with calculating prevalence using this method.

The primary information gained from total prevalence is an understanding of the proportion of people in a given population on the index date who remain alive after having received a diagnosis of the certain diseases. As a vital indicator, cancer prevalence is a measure of the number of cancer patients who require health and social services resources, and can be used to adequately plan future allocation of such resources. It should be useful for government to make health care planning, and for doctors to know the cost for diseases (such as treatment cost and cost of monitoring activities). This study focuses on estimating prevalence for haematological malignancies based on registry data from HMRN. There is rare report about haematological malignancies due to the difficulties in classification and methodology. Additionally, the heterogeneity of the haematological malignancies and their treatments make it useful to estimate prevalence for subtypes. For example, some of the subtypes such as myeloma can be treated as chronic disease (Barlogie et al., 2004). So a patient may undergo treatment for their rest of their lives. This is in contrast to other cancers, or other haematological malignancies such as diffuse large B- cell lymphoma where patients may be cured after first line treatment Sehn et al., 2005). For subtypes such as monoclonal gammopathy of undetermined significance and monoclonal B-cell lymphocytosis, the patients are usually asymptomatic, and do not require treatment. However, the monitoring is needed for them for relatively long time (Marti et al., 2005; Shanafelt et al., 2010). In other words, the estimation of prevalence for different subtypes of haematological malignancy helps to make suggestions about the cost of disease management and health resources allocation.

1.1.2 Haematological malignancies

1.1.2.1 What are haematological malignancies?

Haematological malignancies are a group of cancers associated with the blood, bone marrow, and lymph systems. It is useful to recall some basic facts about the operation of the blood and lymph systems. Blood cells are divided from stem cells. There are two main stem lineages: myeloid and lymphoid. Myeloid stem cells produce red cells, platelets and some types of white cell. Lymphoid stem cells produce two types of white cell: T-cells and B-cells (Hoffbrand, et al., 2006; Howard and Hamilton, 2007) (see Figure 1-1).

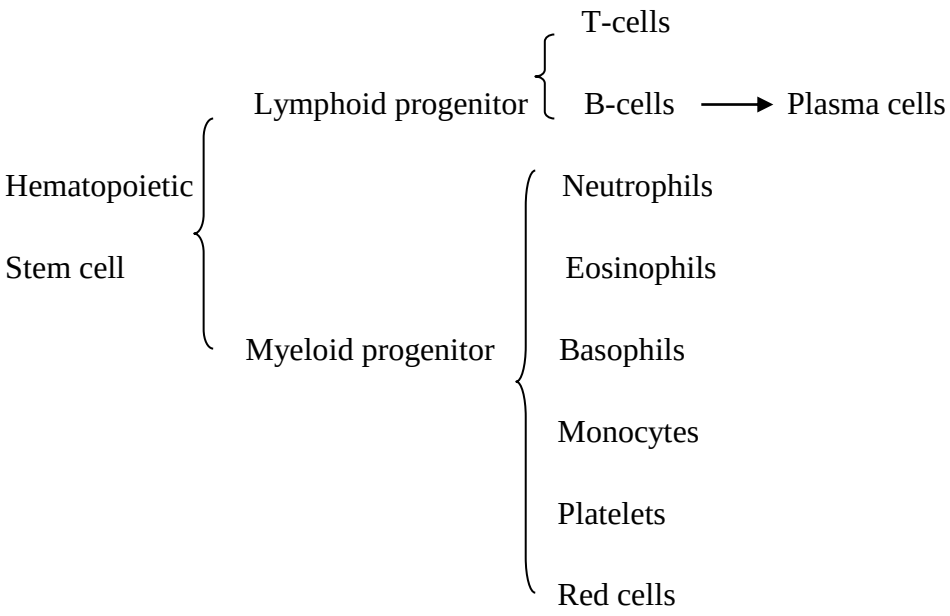


Figure 1- 1 Haematopoiesis map of blood cells

Haematopoiesis occurs in the bone marrow, and blood cells are released to the blood stream when they are mature enough. (Hoffbrand, et al., 2006; Howard and Hamilton, 2007). If something goes wrong in this process, especially during the various stages of differentiation, haematological diseases may occur. Generally,

from an anatomical perspective, if immature white blood cells fill up the bone marrow, preventing normal blood cells from being made, leukaemia occurs. Myeloma is associated with blood plasma, which is developed from B-lymphocytes. Abnormal plasma cells accumulating in the bone marrow will interfere with the production of normal blood cells, destroying normal bone tissue and causing pain. Lymphoma results when a lymphocyte (either a B or T lymphocyte) undergoes a malignant change and multiplies out of control. Eventually, healthy cells are crowded out and malignant lymphocytes amass in the lymph nodes, liver, spleen or other sites in the body. Unlike other haematological malignancies lymphoma usually present as a solid tumour of lymphoid cells. (Hoffbrand, et al., 2006; Howard and Hamilton, 2007). Classification of haematological malignancies not only depends on the place and stage that errors have occurred, but is also related to other clinical factors such as immunophenotype and genetic abnormalities (HMRN, 2011). This classification is explored in more detail in the next section.

1.1.2.2 Classification of haematological malignancy

The classification of haematological malignancies is complex, and has changed over time as knowledge about the disease has developed. For lymphomas, the Rappaport classification developed in the mid- 1950s was purely based on morphology (Rappaport, 1966). In 1982, the Working Formulation (Rosenberg, et al., 1982) based on morphology and clinical prognosis became the standard classification in the U.S. During the same time period, a different classification called Kiel was being used in Europe (Lennert, 1978), which was based on cell lineage and lymphocyte differentiation. This lack of consensus on lymphoma classification made effective comparison between the U.S. and Europe almost impossible. Thus, the Revised European- American Lymphoma (REAL) classification that was published in 1994 by the International Lymphoma Study Group (ILSG) rapidly became the standard in all countries of the world (Harris, et al., 2000a). For leukaemia, the classification made by the French, American, and

British Cooperative Group (FAB) in 1976 was a milestone. This was based on morphology, cytochemistry, and immunophenotype (Bennett, et al., 1976).

In 1995, the European Association of Pathologists and the Society for Haematopathology developed a new classification of haematological malignancies (Harris, et al., 2000b). In 2001, the World Health Organization (WHO) adopted the REAL classification for lymphomas and expanded the principle of REAL to the classification of myeloid malignancies, producing an international classification (known as the WHO classification) of haematological malignancies based on morphology, immunophenotype, genetic abnormalities and clinical features (Harris, et al., 2000a). This classification is incorporated into the International Classification of Diseases for Oncology (ICD-O-3) (Fritz, 2000), which codes tumours or cancers with site, morphology, behaviour, and grading of neoplasms.

Although the WHO classification of haematological malignancies has been widely used in clinical practice around the world (Smith, et al., 2010), many population-based cancer registries still report under the broader classification definitions of ICD-10 (WHO, 1994). This is because, compared with other cancers, the complex data required to classify using ICD-O-3 is difficult for registries to access systematically and it is difficult to bridge code between classifications (Roman and Smith, 2011). In the literature, data on haematological malignancies are traditionally presented using the conventional groupings of leukaemia, Hodgkin lymphoma, non-Hodgkin lymphoma, and myeloma (Ferlay, et al., 2010; NORDCAN, 2010; NCIN, 2012; SEER, 2012). However, there may be diversities within one traditional category, for example different prognoses and age distributions. Furthermore, one category may contain a mix of lineage. The broad category of leukaemia contains both myeloid and lymphoid leukaemias (Figure 1- 2). In addition, it includes both precursor and mature B-cell and T-cell subtypes which again are of considerable significance for the interpretation of epidemiological data. Myelodyplastic syndromes (MDS) and myeloproliferative neoplasms (MPN) are classified as D codes (classified as neoplasms of uncertain

behaviour) in ICD-10, but in fact, they have been clinically recognized as malignancies for at least a decade (Fritz, 2000). Presentation of haematological malignancies in this traditional broad way may be of little value for health resource allocation and for making comparisons of outcomes due to the high level of diversity among the subtypes contained within each of the traditional groupings (Smith, et al., 2010). For example, mantle cell lymphoma and follicular lymphoma appear to have little in common in incidence and survival, therefore there may be doubts about of the usefulness of epidemiological studies that do not distinguish among these disease categories.

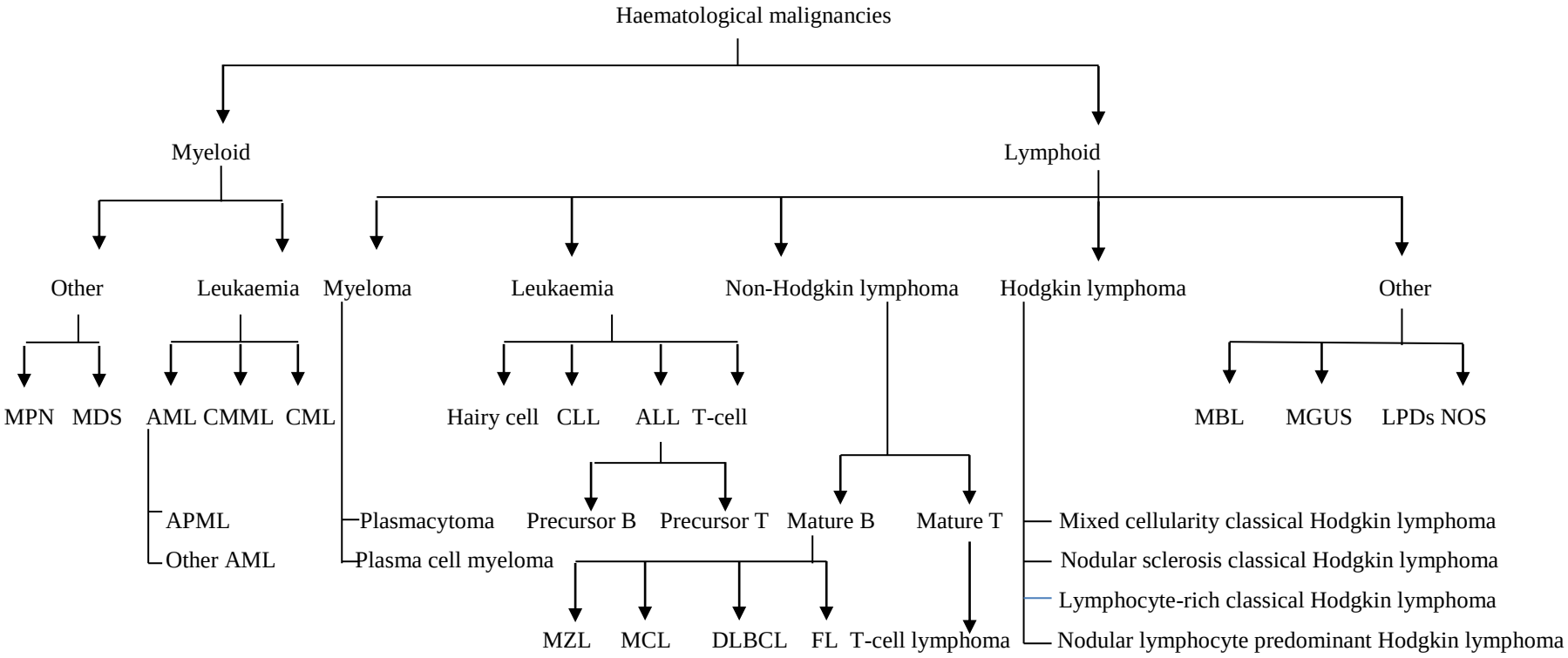


Figure 1- 2 The lineage of subtypes of haematological malignancies (Harris, et al., 2000b) (MPN: myeloproliferative neoplasms, MDS: myelodysplastic syndromes, CML: chronic myelogenous leukaemia, CMML: chronic myelomonocytic leukaemia, AML: acute myeloid leukaemia, ALL: acute lymphoblastic leukaemia, CLL: chronic lymphocytic leukaemia, MBL: monoclonal B-cell lymphocytosis, MGUS: monoclonal gammopathy of undetermined significance, LPDs NOS: lymphoproliferative disorder not otherwise specified, APML: acute promyelocytic myeloid leukaemia, MZL: marginal zone lymphoma, MCL: mantle cell lymphoma, DLBCL: diffuse large B-cell lymphoma, FL: follicular lymphoma[further classification and subtypes can be found in Section 3.1.2])

1.1.2.3 Transformation of haematological malignancies

Haematological malignancies have the ability to transform in to more aggressive subtypes. Within the myelodysplastic syndromes, for example, a general progression to more aggressive disease, acute myeloid leukaemia, is a relatively common pathway (Shi et al., 2004). Normally, immature cells known as "blasts" make up less than five per cent of all cells in the marrow. In myelodysplastic syndromes, blasts often constitute more than five per cent of the cells, whilst the more aggressive subtype--acute myeloid leukaemia, has more than 20 per cent blasts in the marrow (Hoffbrand, et al., 2006; Howard and Hamilton, 2007).

Lower grade subtypes may grow slowly, and remain stable for a long time (for example, follicular lymphoma) (Horning and Rosenberg, 1984). On the other hand, more aggressive subtypes have cancerous cells that multiply quickly (for example, diffuse large B-cell lymphoma) (Davies et al., 2007; Lossos et al., 2002). The designations "indolent" and "aggressive" are often applied to subtypes of non-Hodgkin lymphomas (Figure 1-3). The cells in indolent subtypes do not die off within their normal lifespan, and can sustain additional damage over time. This usually causes the cells to begin to grow rapidly, and makes the lower grade subtypes transform in to higher-grade subtypes. When a transformation occurs, there is a mix of indolent and aggressive cells, and the lower and higher grade diseases often coexist within the same patient. (Horning and Rosenberg, 1984; Kyle et al., 2010; Landgren et al., 2009; Lossos et al., 2002; Shanafelt et al., 2010).

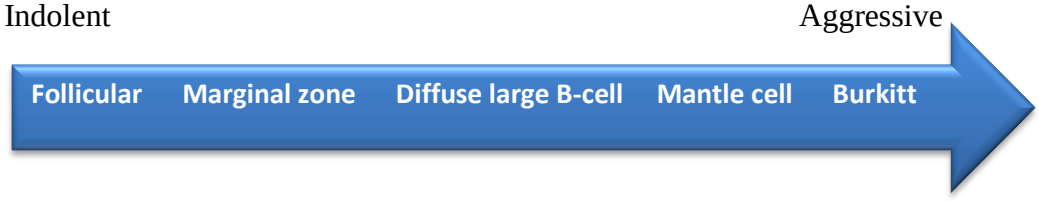


Figure 1- 3 Examples of indolent and aggressive B-cell lymphoma

Transformations tend to follow well-known pathways, for example, follicular lymphoma, which can transform into the more aggressive lymphoma called diffuse large B-cell lymphoma (Davies et al., 2007; Lossos et al., 2002). Myelodysplastic syndromes have a general progression to acute myeloid leukaemia (Shi et al., 2004). Monoclonal B-cell lymphocytosis can transform in to chronic lymphocytic leukaemia (Shanafelt et al., 2010), and monoclonal gammopathy of undetermined significance can transform in to myeloma (Kyle et al., 2010; Landgren et al., 2009). Monoclonal B-cell lymphocytosis is a precursor condition that resembles chronic lymphocytic leukaemia, but the total B-cell count is below the threshold for diagnosis of chronic lymphocytic leukaemia ($<5.0 \times 10^9$ cells/L) (Marti et al., 2005; Rawstron et al., 2008). Monoclonal gammopathy of undetermined significance is the precursor condition for myeloma and is similarly characterised by the presence of a paraprotein, but at a lower level, and the patient is usually asymptomatic, and does not require treatment (Smith et al. 2010).

1.1.2.4 Challenges in estimating prevalence of haematological malignancies

Estimating prevalence for some subtypes of hematological malignancies is hampered by both the difficulty in obtaining data and complexity in developing methods:

I. Difficulty in obtaining data of haematological malignancies under WHO classification

This difficulty in calculating prevalence of haematological malignancies lies in their diagnostic complexity and classification (Smith, et al., 2010). The diagnostic parameters of haematological malignancies include a combination of histology,

cytology, immunophenotype, cytogenetics, imaging and clinical data. This range of diagnostic criteria usually results in difficulty in making an integrated diagnosis of disease. The broad ICD-10 classification is still used to report national data, for example by the National Cancer Intelligence Network (NCIN, 2012) in the UK, the Surveillance Epidemiology and End Results Program (SEER, 2012) in the U.S., and the International Agency for Research on Cancer (IARC, 2012) (the reports are discussed in Chapter Two).

Furthermore, the WHO classification for haematological malignancies was only established in 2001. Therefore, for some datasets, cases diagnosed prior to the WHO classification may be classified according to various older schemes, however there is no standard for translating from these historical classifications (Morton, et al., 2007). Indeed, it is difficult to 'bridge' code diagnoses classified to ICD-10 to current WHO classification.

The above barriers make estimating the prevalence of haematological malignancies more challenging than for other cancers.

II. Haematological malignancies have characteristics that are different from other common cancers

Haematological malignancies can be diagnosed at any age, and show different survival patterns between childhood and adulthood (further discussion about this is given in Chapter Three). This is another reason why it is challenging to calculate the prevalence of haematological malignancies. The methods used in the literature to estimate total prevalence estimation may not be suitable for haematological malignancies (since most cancers tend to occur in later adulthood), and it is necessary to make some adaptations to the model, using more flexible statistical tools. Furthermore, some subtypes of haematological malignancy show uncommon age distributions, such as Hodgkin lymphoma

which has a bimodal incidence curve. The log linear model used in the literature will fail to fit data for them, so it is necessary to develop a flexible model to make descriptions and estimations.

The total prevalence of haematological malignancies is estimated by modeling the mathematical relationship among prevalence, incidence, survival, and general mortality. The model is easy to understand, since prevalent cases are actually the patients who were diagnosed with the particular disease in the past (incidence of disease), and who keep being patients (survival) without dying (mortality). To make the calculation simple and practical, some assumptions are made. (The methods of calculation and the assumptions made are discussed in Chapter Three.) Since total prevalence figures are unavailable until the registry is sufficiently mature to capture all patients ever diagnosed with a cancer, the aim of this study is to try and demonstrate the burden of haematological malignancies by estimating its prevalence from limited length registry data - HMRN.

1.2 Aims and objectives of this thesis

The aim of this thesis is to estimate the prevalence of haematological malignancies, under the WHO classification, using data from the Haematological Malignancy Research Network (HMRN). The main objectives are summarized as follows:

1. To demonstrate n-year prevalence using HMRN data for subtypes of haematological malignancies under the latest disease classification.
2. To develop a general method to estimate total prevalence.
3. To calculate total prevalence for all subtypes using the method in this study.
4. To find a suitable method to calculate total prevalence for subtypes where survival has changed significantly in the past.

5. To estimate prevalence in the UK and make suggestions for the burden of haematological malignancies.

This is the first time that the prevalence of haematological malignancies has been estimated under WHO classification. High quality data from HMRN that overcomes the difficulties in diagnosis and classification of haematological malignancies makes these estimates possible. Although HMRN has a limitation of a relatively short follow-up period for some subtypes with longer survival, the statistical models used help to achieve the final goal of estimating total prevalence.

Total prevalence estimates can be used to show the real burden of subtypes of haematological malignancies, and to suggest reasonable health resource allocation. Besides total prevalence, 1-year and 5-year prevalence estimates, as supplementary information, are calculated in this study by simply counting the number of prevalent cases on the index date. Furthermore, suggestions for making prevalence estimates for the diseases in which there have been significant survival changes due to new treatments regimes are also made.

1.3 Definitions

1.3.1 General notations in this thesis

The following notation will be used throughout this study.

Let t be the age at diagnosis (in years).

Let x be the current age on the index date (in years).

Let u be the age at death (in years).

Let Y be a calendar year.

Let d be duration of disease (in years).

Let D be the number of deaths in a period of time.

Let $V(d)$ be the number of patients who survive for d years from diagnosis.

Let U be the number of patients lost to follow-up within the registry period of time.

Let $I(t)$ denote the probability that an individual will be diagnosed with a specified cancer at the age of t (age t years means between t and $t+1$ years old).

Let $C(t)$ denote the number of new patients (or diagnosis) at age t.

Let $G(u)$ denote the probability of an individual in the background population dying at age u (age u years means between u and $u+1$ years old).

Let $S(t, d)$ denote the probability that an individual, who has a confirmed cancer diagnosis at the age of t, survives for time d after diagnosis (age t years means between t and $t+1$ years old) .

Let $N(x)$ denote the number of patients (or diagnoses) alive on the index date at age x who had a diagnosis of cancer in the past.

Let $P(x)$ denote the probability that an arbitrary person of age x in the population has received a diagnosis of cancer in the past (age x years means between x and $x+1$ years old).

For a patient, she or he is considered to be **disease-free** before age t (age at diagnosis), which means the patient does not have the disease of interest in the study. After age t, the patient becomes a **prevalent case** until age u (age at death), which means between age t and age u the patient survives with the disease. At any particular time, all diagnosed patients who have not previously died are prevalent cases at that time, and their age at that time is x. In other words, “cure” is not

considered in this study, and once the patient is diagnosed with cancer, he or she is considered a prevalent case for all the rest of his/her life (we discuss this assumption in Chapter Five).

The length of registry is L years (for example, HMRN has $L = 7$ from 2004 to 2011). The observations under study are cases with registry information. *P(x)* is the prevalence rate at age *x*, and *N(x)* is the number of prevalent cases at age *x*.

1.3.2 Prevalence, incidence, and survival

In epidemiology, the prevalence of a disease in a population can be given as a count or as a proportion. It is defined either as the total number of cases in the population at a given time, or the total number of cases in the population, divided by the number of individuals in the population at that time. In this thesis, proportion is used as the default meaning for prevalence and it will be specified explicitly when it means count.

$$prevalence\ rate = \frac{total\ number\ of\ cases}{population} \qquad (1.\ 1)$$

Point prevalence is a measure of the proportion of people in a population who have a disease at a particular time, such as on a particular date. This date is called the **index date**. Point prevalence is like a snap shot of the disease at a particular time.

$$point\ prevalence = \frac{Number\ of\ existing\ cases\ on\ a\ specific\ date}{population\ on\ the\ specific\ date} \qquad (1.\ 2)$$

Prevalence at a certain age is **age-specific prevalence**, and it can be understood as:

$$\boldsymbol{age-specific\ prevalence} = \frac{\boldsymbol{number\ of\ cases\ who\ are\ at\ age\ x}}{\boldsymbol{number\ of\ population\ who\ are\ at\ age\ x}} \quad (1.3)$$

Age-specific prevalence focuses on current age *x* (*x* years means between *x* and *x+1* years old.) rather than diagnosed age *t*. It counts patients of a certain age on the index date, no matter when they were diagnosed.

These measures of prevalence can be estimated by counting the number of people found to have the disease in question and by comparing this with the total number of people studied. In order to estimate the number of observed patients in a registry and unobserved patients diagnosed before the start of the registry, prevalence considered as a proportion can also be considered as a probability. At this point, prevalence can be estimated as the probability of being found at a particular time, having had present a previous diagnosis of the disease in question.

Prevalence, working as a proportion, summarizes the observations. Within the registry, it shows the real phenomenon — number of live patients who can be observed in the data. Conversely, probability is a measure of the expectation of people being found at a certain time, as having had present or past diagnosis for the disease. Here, practically, proportion of an event can be considered as its probability, with observable proportions being used as the expected probabilities in the calculation.

Prevalence can tell us how widespread a disease is in a given population. It depends on both the frequency of cancer and its survival characteristics. In other words, it is related to incidence and survival duration. For example, for a disease with good survival characteristics and high incidence prior to and in year Y_1 but

with low incidence by and after year Y_2, we will find both high incidence and prevalence in year Y_1, but after year Y_2 the incidence will decrease, while the decrease in prevalence will show a significant time lag due to the long survival. Conversely, a disease that has a short duration might spread widely during Y_1 but is likely to have a low prevalence in Y_2 (due to its short duration).

Incidence is a measure of the risk of developing a disease within a specified period of time. It is the number of newly diagnosed cases during a specific time period. When expressed as a rate, it is the number of new cases per standard unit of population during the time period. It is often expressed as, for example, a number per 100, 000 per year or number per 100, 000 per age group. It is calculated as:

$$\boldsymbol{incidence\ rate} = \frac{\boldsymbol{the\ number\ of\ new\ cases\ wihin\ a\ specified\ time\ period}}{\boldsymbol{person{-}time\ of\ the\ at\ risk\ population}} \quad (1.4)$$

In equation 1.4, the "at risk population" is the population minus the number of people who already have a certain disease at the beginning of that period of time. Since the number of patients with haematological malignancies is small compared to the population, the size of population in the study area is considered to be equal to the size of the population initially at risk.

It is assumed that incidence is constant for the registry period. This assumption is extended to "the incidence is constant over the period of interest" (details are shown in Chapter Three) in the total prevalence estimates.

Incidence rates can also be calculated based on a number of factors, such as age or sex, for example:

$$\boldsymbol{age-specific\ incidence\ rate} = \frac{\boldsymbol{the\ number\ of\ new\ cases\ diagnosed\ at\ a\ certain\ age}}{\boldsymbol{population\ at\ a\ certain\ age}} \quad (1.5)$$

Prevalence is a measure taken at a certain point in time and is cross-sectional, whilst incidence is longitudinal, looking at the occurrence of the disease of interest. Therefore, unlike age-specific prevalence which focuses on a patient's current age *x* on the index date, age-specific incidence refers to the age at diagnosis *t*.

Incidence rates can be used as expected probabilities. If *t* is the age at diagnosis of a patient, then the incidence rate at age *t* can be interpreted as being the probability that an arbitrary person in the population will be diagnosed at age *t*. For example, if the incidence of acute myeloid leukaemia (AML) in age 0-4 years is 1.3 per 100,000, it can be also considered as a probability of receiving a diagnosis of acute myeloid leukaemia in the age rage 0-4 years of $1.3 * 10^{-5}$ in the area of study. The notation $I(t)$ indicates the incidence at age *t* in the following calculations.

Similar to the definitions of incidence as rates or probabilities, the survival rate indicates the percentage of people in a study who are alive for a given period of time *d* after diagnosis. For an individual it is defined as the probability, $S(d)$, that an individual survives longer than *d* (Cleves, et al., 2010).

$$\boldsymbol{S(d) = Pr(an\ individual\ survives\ longer\ than\ d)} \quad (1.6)$$

$$\boldsymbol{S(d) = 1 - Pr(an\ individual\ dies\ before\ d)} \quad (1.7)$$

In medical research, survival may also be considered as a function of the age at diagnosis or of other explanatory variables. $S(t, d)$ is the proportion of people diagnosed at age *t* who survive for *d* years after diagnosis. It can be also

considered as the probability that a patient who is diagnosed at age *t* is still alive after d years. The probability of death during a very small time interval is an instantaneous death rate, called the hazard function (Cleves, et al., 2010).

However, this definition of survival refers to overall survival. As a measure, it does not take into account what the subject actually dies from. Other causes of death can be understood as "die of another cause but with the disease present" (Ederer, Axtell, and Cuter, , 1961). Relative survival, S_r captures how survival is affected by the disease (net survival rate):

$$\boldsymbol{relative\ survival\ rate} = \frac{\boldsymbol{overall\ survival\ rate}}{\boldsymbol{the\ expected\ survival\ rate\ in\ the\ population}} \quad (1.8)$$

When relative survival is less than 1(100 percent), then mortality in the patients in the study exceeds that of disease- free persons in the population. When it reaches 1(100 percent), it indicates that the death rate of patients is equal to that of the general population.

In equation 1.8, the expected survival rate in the population is actually the survival rate of those who do not have the specific disease under consideration. This group of people can be considered as a control group and their survival characteristics can be used to adjust the overall survival characteristics of the patient group. It is often the case that the mortality from a specific cancer constitutes a negligible contribution to total mortality (Ederer, Axtell, and Cuter, , 1961). In this situation, the survival rates of the general population provide satisfactory estimates for expected survival rates when the relative survival of patients with the cancer under consideration is analyzed.

Relative survival is widely used in prevalence calculation in the literature (see Chapter Two). However, as described above, relative survival is easy to define,

but is not easy to estimate, especially when parametric models are involved (details are explored in Chapters Two). Furthermore, relative survival is a reasonable indicator of the survival experience of patients in a population, but is less informative if used to predict the prognosis of an individual (Parker, et al., 1996).

Confounding by age may occur when we predict the prevalence rate in other populations (Pop), since disease incidence and survival varies across age groups; age usually has a powerful influence on the risk of cancer and on survival. **Age standardisation** (age adjustment), is used to control for differences between the age structures of different populations (Leon, 2008). It is accomplished first by multiplying the age specific rates (r) of disease in an age group (in area 1) by the population size in the corresponding age group in the target area (area 2). Next, the sum of those products is divided by the total population size in the target area (area 2).

$$\boldsymbol{age-standardised\ rate} = \frac{\sum \boldsymbol{Pop(i)} * \boldsymbol{r(i)}}{\sum \boldsymbol{Pop(i)}} \qquad (1.\ 9)$$

$Pop(i)$ indicates the population sizes in the relevant age groups *(i)* in the target area (area 2), and $r(i)$ are the age-specific rates in age groups *i* in the local area (area 1).

1.3.3 Different types of prevalence in this thesis

After the establishment of a cancer registry, new cases are registered every year. If the status (dead or alive) of a case on the index date is available, you therefore know the number of live cases on that day within the registry. One can even calculate the prevalence for a special group of people, such as those diagnosed in the most recent years. However, registry data does not include the patients who

were diagnosed before the start of the registry and are still alive. Therefore, in the calculations, different types of prevalence are defined as follow (Figure 1- 4):

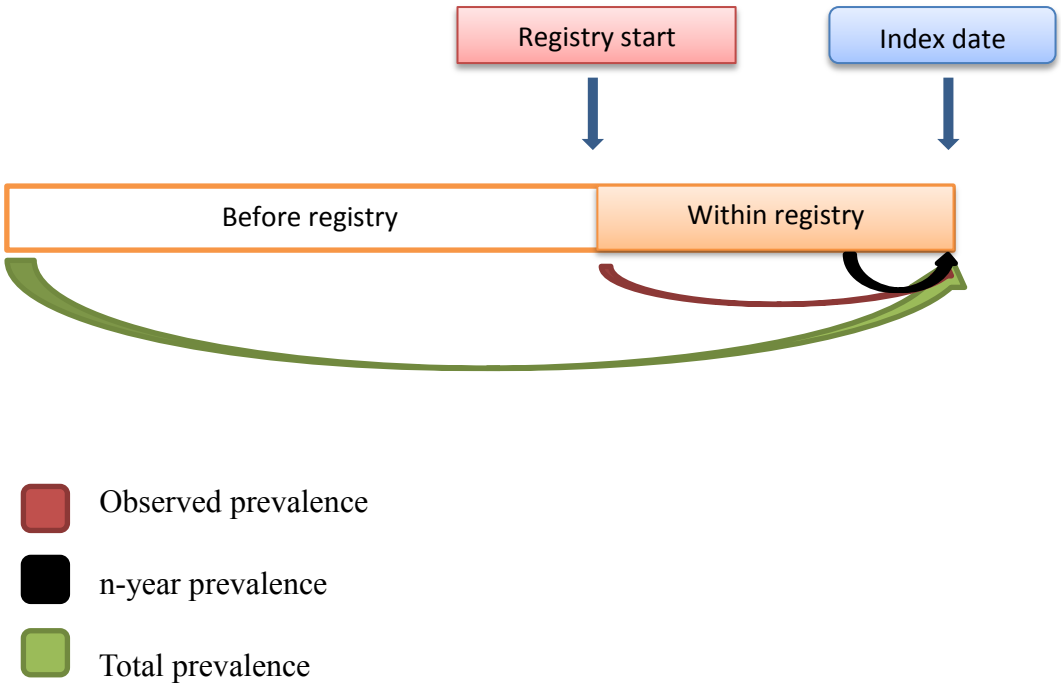


Figure 1- 4 different types of prevalence

The prevalence calculated for the patients who were diagnosed within the registry period is called the **observed prevalence** (can also be called limited duration prevalence).

n- year prevalence includes all persons who were diagnosed with the disease in question within n years of the index date. When n equals the length of the registry, n-year prevalence is observed prevalence. Therefore, when $n \leq the\ length\ of\ regisry$, n-year prevalence can be obtained directly. It is a convenient and commonly used method, calculated based on available data in the registry, for example, 5-year prevalence based on 5 years of available registry data (Figure 1- 5).

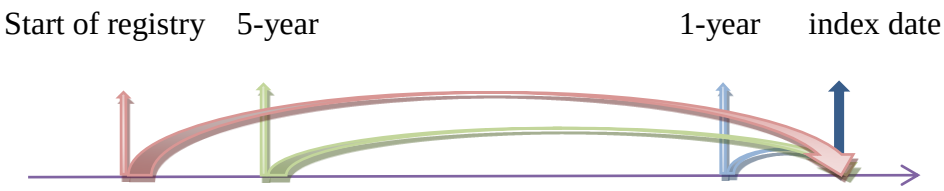


Figure 1- 5 n-year (1-year and 5-year) prevalence, and observed prevalence

Total prevalence refers to all persons in a given population diagnosed in the past with the disease under consideration and who are still alive on a specified index date. It is also called unlimited duration prevalence in the literature. Total prevalence calculated in this work is an estimate (equation 1.4), and cannot be calculated directly by the definition of prevalence. The method used to estimate total prevalence is described in Chapter Three.

Total prevalence is estimated from observed data, and in fact, it is a measure of the expectation that people being found on the index date, having had present or past diagnosis of the disease. The real number of patients who are alive on a certain date is unavailable until the length of the registry is long enough to cover all living patients. The prevalence that includes all live patients in the real world is called true **complete prevalence**. Total prevalence is an estimate, and it is the expected complete prevalence.

Prevalence can be calculated on person basis or on a diagnosis basis. **Person prevalence** only considers the first malignancy diagnosed in each person, and is a measure of the number of people actually surviving having received a previous diagnosis. On the other hand, **diagnosis prevalence** refers to diagnosis and considers all malignancies in a patient. Although most prevalence studies (Capocaccia, et al., 2002; Micheli, et al., 2002a; Verdecchia, et al., 2002; Forman, et al., 2003; Lutz, et al., 2003; Møller, et al., 2003) only consider person prevalence, diagnosis prevalence is more appropriate for haematological

malignancies due to its characteristic ability to transform. Therefore the prevalence estimates of haematological malignancies include all diagnoses, regardless of whether they were first or subsequent cancer.

Beyond those main definitions, the definition of **prevalence range** is introduced**,** which is used to specify a reasonable range of possible total prevalence for diseases where survival characteristics have changed dramatically in the past due to the introduction of new treatments. It contains an upper limit of total prevalence and a lower limit of a certain, n-year prevalence, (details are given in Chapter Three Section 3.5).

1.4 Structure and outline of this thesis

To meet the objectives of this study, tasks are carried out in several phases; here these are shown in sequential order. The first phase is a review of the literature. The second phase is to introduce and to describe the data from HMRN to be used in this study. With data from HMRN, simple calculations are conducted in phase three. n-year prevalence is obtained by counting the number of alive patients within the registry for $n=1$ and $n=5$. In the next, more difficult phase four, a method was developed to estimate total prevalence from the limited prevalence data that are available. The final phase estimates prevalence ranges for the haematological malignancy subtypes where survival characteristics have changed significantly due to the introduction in the past of new treatments.

The working phases are summarized in chapters to keep the whole structure clear. The structure of the thesis is shown in Figure1- 6.

There are three main sections in Chapter One. Firstly, it discusses background information about haematological malignancies and cancer registration. Secondly,

it demonstrates the motivation for the study, and identifies the aims for the work. Lastly, key definitions used throughout this work concerning prevalence are made.

In the second chapter, a literature review is presented of both the methods of calculation of cancer prevalence and for reported prevalence figures for haematological malignancies in previous studies. The methods are introduced and compared to each other in order to find the appropriate methodology to estimate the prevalence of haematological malignancies based on the data from HMRN. Furthermore, the summary of reported prevalence figures from previous studies forms an overview of the prevalence of haematological malignancies, and also demonstrates the limitations in these studies.

Chapter Three introduces the methods used in this study. It includes data and materials used in calculating and estimating throughout the study, including the direct method used to calculate n-year prevalence, how the model for total prevalence estimation is built, as well as the statistics involved in the methods. There is a further study at the end of this chapter to find a method to show prevalence for some diseases that have had great survival improvements in the past due to the introduction of new treatments.

Chapter Four describes the results. Firstly, the demographic characteristics of hematological malignancies are described. Secondly, n-year prevalence was calculated for all diagnostic subtypes. Thirdly, total prevalence is calculated and to demonstrate the processes acute myeloid leukaemia and Hodgkin lymphoma are used as examples.

In the final chapter, the main findings and contributions of this work are described, followed by a discussion of the methodology and the results obtained. The advantages and the limitations of this study are explained in this chapter, as

well as the comparison with other reports in the world. There are also suggestions about future study at the end of this chapter.

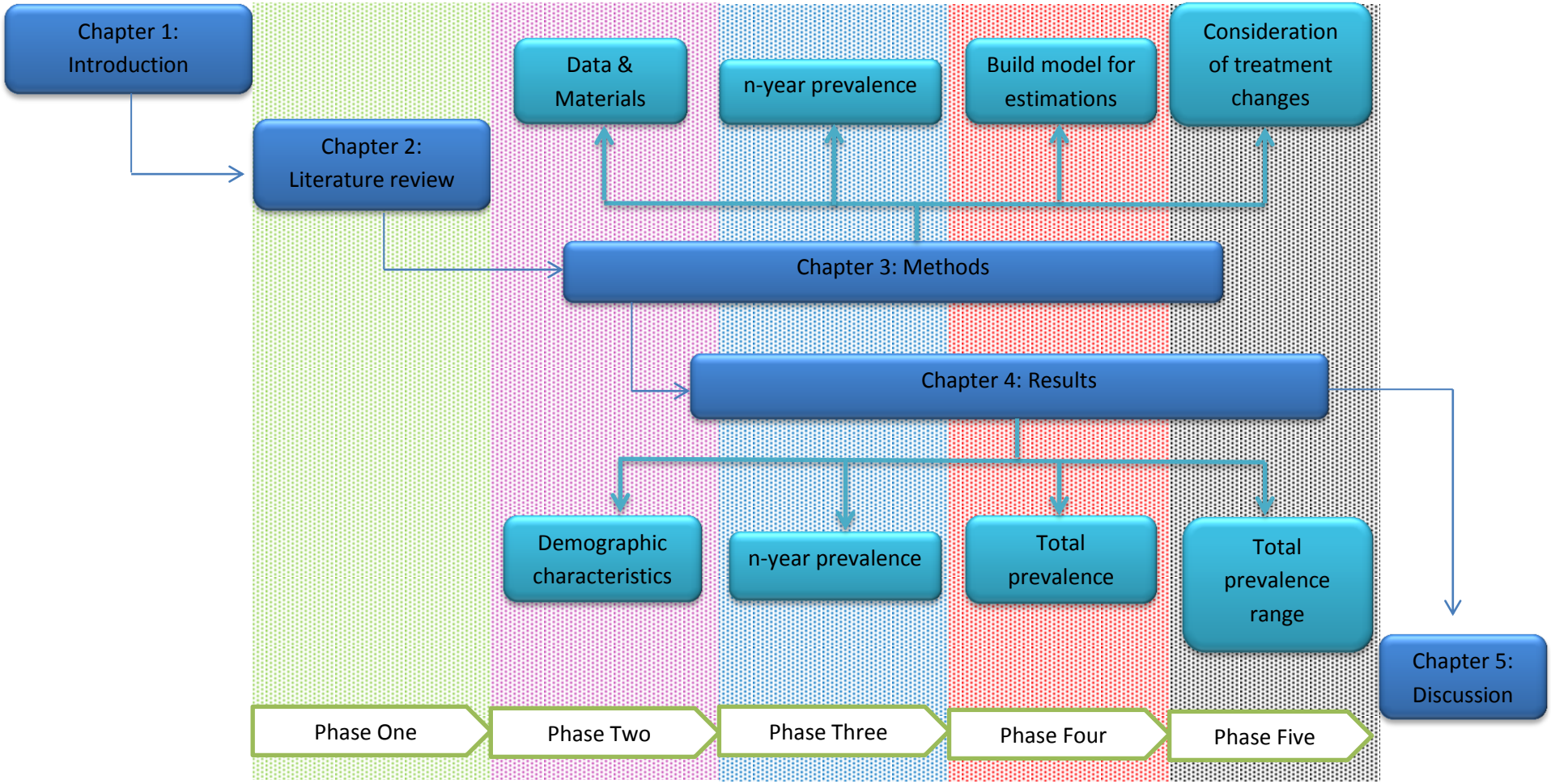


Figure 1- 6 Structure of thesis

1.5 Summary

This chapter has described the different concepts of prevalence and the challenges faced in trying to calculate the burden of disease for haematological malignancies. This study attempts to surmount the difficulties in calculating prevalence estimates for haematological malignancies based on current disease classification. General notations have been made for further estimations and the whole thesis follows the structure shown in this chapter.

Chapter 2 Literature review

Before estimating the prevalence of haematological malignancies based on data from HMRN, it is necessary to consider work that has already been published. The first aim of this chapter is to compare the methods used to calculate/estimate prevalence in previous studies. It also assesses whether the methods mentioned in the literature can be used to estimate the prevalence of haematological malignancies using data from HMRN. The second aim is to summarize the prevalence of haematological malignancies among countries or areas in the world. The temporal and geographic variability of haematological malignancy prevalence are described at the end of this chapter.

The terms "cancer prevalence", "cancer registry", "prevalence of leukemia", "prevalence of Hodgkin lymphoma", "prevalence of non-Hodgkin lymphoma", "prevalence of myeloma", as well as prevalence of acute myeloid leukemia (AML), chronic myelogenous leukemia (CML), acute lymphoblastic leukemia (ALL), and chronic lymphocytic leukemia (CLL) were used to search online database (last search, July 2013). The search is not restricted by date. The earliest paper used in this section is in 1975, and the most recent one is in 2013. Besides Medline, Google scholar was also used to search papers. Web based reports on-line searches were also conducted, and only articles in English were reviewed. Compared with incidence, mortality, and survival, the information on cancer prevalence is limited, and the systematic impact of haematological malignancies on health systems has not been fully described. The literature review identifies some methods, which will be discussed in the following section.

2.1 The methodology for estimation of cancer prevalence

Broadly speaking, methods for calculating cancer prevalence can be divided into two categories: direct calculation and indirect estimation. Direct calculation

computes the prevalence observed from data, whilst indirect estimation provides estimates of unobserved prevalence based on the observations in the registry. Different methods in the two categories can be found in the literature. These are summarized in Figure 2-1.

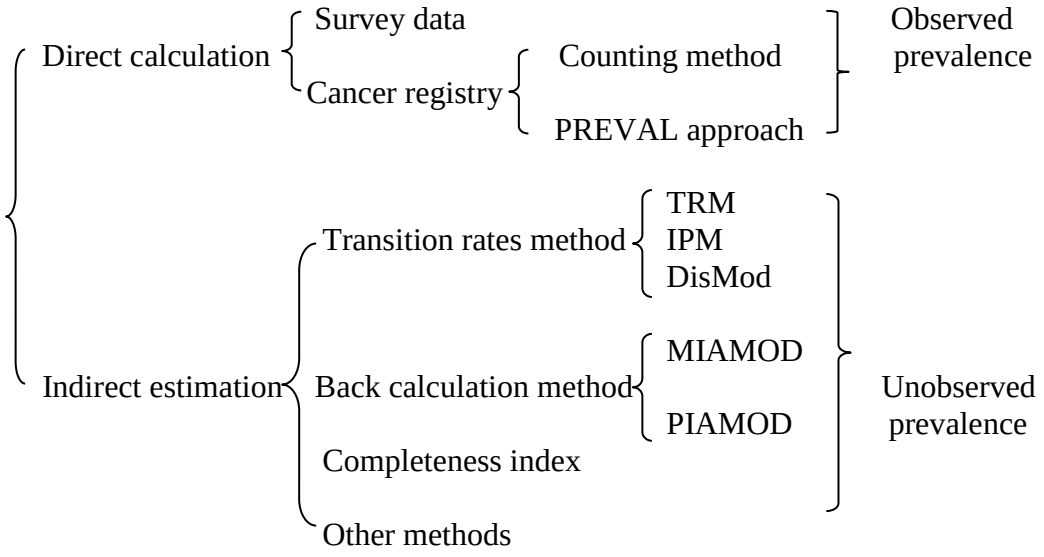


Figure 2- 1 The main methods for calculating prevalence (TRM: transition rates method. IPM: incidence-Prevalence-Mortality Model, DisMod: Disease Model. MIAMOD: Mortality Incidence Analysis MODel; PIAMOD: Prevalence Incidence Analysis MODel)

Direct calculation refers to the methods that involve the number of observed cases and population only. Cross sectional research (such as surveys) is the most straightforward way to assess prevalence, and prevalence can cover all prevalent cases in a defined time. However, longitudinal studies collect data for a period and only provide prevalence within the registry. Indirect estimation usually uses incidence, mortality and survival probabilities abstracted from cancer registry data to estimate prevalence. In the literature, this group of methods is used to either predict prevalence in the future or to estimate total prevalence that covers all patients diagnosed before and after the start of the registry.

The different methods in the literature pertaining to prevalence calculation and estimation are shown one by one in the following sections. The methods are

briefly introduced, as well as examples given using certain methods. Each method has its advantages and disadvantages. Appropriateness for this study is described after making comparisons between the methods.

2.1.1 Cross-sectional population based surveys

To obtain the total prevalence, a conceptually straightforward method is to conduct a cross-sectional population based survey. In 1987, the prevalence of cancer was estimated based on a sample in the U.S. through the National Health Interview Survey (NHIS) (Byrne, Kessler and Devesa, 1992). Weighting procedures were developed by the Bureau of the Census and National Center for Health Statistics to reflect the civilian population of the U.S. in 1987. In 1991, cancer prevalence was calculated using this method in the Netherlands, and the results were compared with cancer registry records (Schrijvers, et al., 1994). However, this method is conceptually easy but in practice hard to apply to a large population. Moreover, when using this method one must consider the problem of self-reporting, such as underreporting and misclassification.

2.1.2 The Counting method

Using cancer registry data, prevalent cases observed for a period can be counted directly. On the desired index date, the cases still alive are simply counted, whilst adjustments are made to estimate the proportion of cases lost to follow-up who could have made it to the prevalence date. The survival probability of each lost case is estimated from the subset of followed patients belonging to the same sex, age, and period of diagnosis. In 1999, Gail developed the Counting method (Gail, et al., 1999). Recalling the definition and notations made in Chapter One, for a patient, t is the age at diagnosis, x is the current age, $S(t, d)$ is the probability that an individual who developed cancer at age t will survive beyond duration d ($d = x - t$) after cancer incidence, and the registry is L years long. Further notation is given specifically for this method:

Let Y_{index} be the calendar time of the index date,
Let Y_I be the calendar time of cancer incidence for a typical patient,
Let Y_D be the calendar time of death,
Let Y_U be the calendar time of loss from follow- up,
Let $Pop(x)$ be the total population at age *x*.

Suppose F is an indicator function equalling one when the argument is true and zero otherwise. The number of cases that can be observed to survive to age x is the summation over all members in the registry:

$$N_1(x) = \sum F(x, Y_{index} - L \leq Y_I < Y_{index},\ Y_D \geq Y_{index}, Y_U \geq Y_{index}) \quad (2.\ 1)$$

(1) (2) (3)

This includes patients at age *x* t: (1) diagnosed in a certain year within the registry period, (2) not deceased before the index date, (3) not lost to follow-up before the index date.

For the patient lost to follow-up, the probability of being alive on the index date is estimated from the appropriate survival function of the cohort, conditional on the time of loss- to- follow-up. Each case lost to follow-up has conditional survival (Gail, et al., 1999). It is the probability that a patient will survive at least until the index date, given that patient was diagnosed at Y_I, and lost to follow-up at Y_U. The number of cases of age *x* alive on the index date among those of the same age who were lost from follow- up before age *x* is:

$$N_2(x) = \sum \{F(x, Y_{index} - L \leq Y_I < Y_{index}, Y_D > Y_U, Y_U < Y_{index})\}$$

$$* \frac{S(Y_{index} - Y_i, t)}{S(Y_u - Y_i, t)} \quad (2.\ 2)$$

The individual estimate of survival ($S(d,t)$) is obtained using the life table (or known as the actuarial) method (Cutler and Ederer, 1958; Gail, et al., 1999).

The probability *P(x)*, of a person being alive on the index date at age *x* is calculated by:

$$P(x) = \frac{N_1(x)+N_2(x)}{Pop(x)} \qquad (2.\ 3)$$

The calculation can be implemented using SEER* Stat software (National Cancer Institute, 2010). The prevalence calculated using this method is also called "limited duration prevalence" (National Cancer Institute, 2010). This method was also used to calculate the prevalence of cancer in Quebec, Canada (Louchini, et al. 2006). An advantage of this method is that it is easy to understand. It calculates observed prevalence adjusted for losses in the follow-up. When the cancer registries for those papers are long, and most patients diagnosed before the start of the registry die before the index date, this method can provide relatively unbiased prevalence estimates for the corresponding population.

2.1.3 The PREVAL approach

Observed prevalence based on cancer registry data can also be calculated using the PREVAL approach (Krogh and Micheli, 1996). This method uses the same idea as the counting method; the number of cases alive on the index date is the sum of the number of observed cases and the number of cases lost to follow-up and still alive. However, it estimates the prevalence according to the time *d* since diagnosis. Following the general notations of this study, *D* is the number of deaths; *N* is the number of surviving cases; *U* is the number of patients lost to follow-up. Unlike the counting method, which calculates age-specific prevalence, duration is the determinant in the PREVAL approach. Besides the notation of *d* years from diagnosis to index date, it also sets *s* to be the number of years from diagnosis to death, and *m* to be the number of years from diagnosis to loss to follow-up. In a

cohort, N_d is the number of patients who have survived for *d* years until the index date, and N_s is the number of cases surviving for s years until death, D_s is the number of cases that died after s years since diagnosis, U_m is the number of patients lost to follow- up *m* years from diagnosis. The formula to calculate the expected number of prevalent cases at the index date who have survived for *d* years is:

$$E_d = N_d + \sum_{m=0}^{d} U_m \prod_{s=m}^{d} \frac{N_s}{N_s + D_s} \qquad (2.4)$$

The formula, $\frac{N_s}{N_s+D_s}$ is used for making adjustments to account for lost- to- follow-up. E_d is calculated using strata of age, gender, and race. It is the expected number of patients who survive for a certain number of years until the index date, taking into account the survival of those lost to follow-up.

This method was also used to calculate the observed prevalence in 1992 in Connecticut, Iowa, and Utah (Micheli, et al. 2002b). The correction in this method assumes that the lost-to-follow-up cases have the same survival characteristics as those not lost to follow-up. If lost-to-follow-up cases differ from other cases by some factor that influences their survival, the assumption of the same survival rate will be flawed.

Both the counting method and the PREVAL approach calculate observed prevalence based on registry data. Some observed prevalence data pertaining to the U.S. and some European countries that have been covered by cancer registration for many decades have been published (Hakama, et al., 1975; Feldman, et al., 1986; Adami, et al., 1989; Polednak, 1997; Gail, et al., 1999). However, the prevalence estimated using these methods is the observed prevalence within a limited period of follow-up time. This ignores patients diagnosed before the establishment of the registry who are still alive. When the registry is young and cannot offer enough of a follow-up period, the bias may be large.

2.1.4 The transition rate method

All of the above methods are direct calculations that provide prevalence estimates of number of prevalent cases in the population. In this section, it is introduce indirect estimate methods. These methods estimate prevalence based on probabilities and proportions (recall the definitions in Chapter One, equation 1.4). A group of methods that are based on stochastic process modelling are named transition rate methods:

2.1.4.1 The Transition Rate Method (TRM)

These methods are based on the assumption that individuals are in different states at different times in their life history. They move from one state to another according to some state transition probabilities. In the transition rate method (TRM), these probabilities are called transition rates. The transition rate method (Gras, Daurès and Tretarre, 2006) estimates cancer prevalence using a stochastic process with three states as follows:

1. alive and cancer free, state H
2. alive with cancer, state I
3. dead, state D

At a point in time on the calendar, the healthy state H may transit to the disease state I with transition rate $r_1(x)$ which depends on age x. Alternatively, the individual may die directly from state H with transition rate $r_2(x)$. A subject in state I is at risk of death with transition rate $r_3(x, d)$ which depends on the duration of disease d as well as age x.

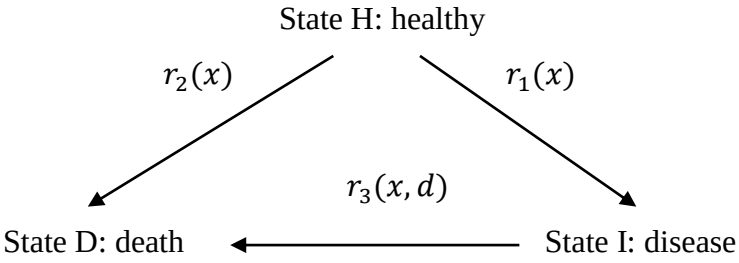


Figure 2- 2 The three states stochastic model

After estimating the transition rates between the states, the model is then allowed to run to estimate cancer prevalence under a set of specified conditions (Gail, et al., 1999).

It is assumed that no cancer case is ever "cured", which means the patients are considered as prevalent cases until they have died once they have been diagnosed with cancer. Consider a single birth cohort; the probability of being alive with disease in state I at age *x* is:

$$P_I(x,L) = \int_{x-L}^{x} \exp\left(-\int_0^t (r_1 + r_2)(u)du\right) r_1(t) \exp\left(-\int_t^x r_3(u, u-t)du\right) dt \quad (2.5)$$

1 2 3

1 represents the probability of surviving cancer-free up to age *t*,

2 represents the probability of cancer onset at age *t*,

3 represents the probability of surviving to age x given that the individual is diagnosed with cancer at age *t*

It is assumed that the overall survival of the population at age x is $S_{overal}(x)$. Next, the L- year prevalence of cancer is:

$$P(x) = \frac{P_I(x,L)}{S_{overall}(x)} \qquad (2.\ 6)$$

Followed by,

$$P(x) = \frac{\int_{x-L}^{x} exp\left(-\int_0^t (r_1+r_2)(u)du\right) r_1(t)\, exp\left(-\int_y^x r_3(u,u-t)du\right)dt}{S_{overall}(x)} \qquad (2.\ 7)$$

The probabilities of being in various states of the process are required to estimate prevalence. There is an assumption that the transition rates are constant over time (Gail, et al., 1999). This method is relatively flexible in prevalence estimates since it divided the life history by status. However, as equation 2.7 shows, it is not an easy calculation, and includes many probabilities: three transition rates, and overall survival. To estimate prevalence using this method, the transition rates should be abstracted from data according to age and gender.

2.1.4.2 The Incidence- Prevalence- Mortality Model (IPM)

Using a similar theory to the TRM, the prevalence of cancer in the Netherlands was calculated in 2000 (Hoogenveen and Gijsen, 2000), using a model called the incidence- prevalence- mortality (IPM) model. Similar to the TRM, it is a two-state transition model. For a given cancer, in addition to the state "Death", the two states are distinguished as "disease- free" and "with the disease" (prevalent). The difference to the TRM is that it differentiates between causes of death. There are three transition rates: disease incidence rates (disease- free to prevalent), disease-related excess mortality rates (prevalent to dead), and mortality rates for all other causes (disease- free to dead, and prevalent to dead). The model structure can be expressed as in Figure 2-3. It is also assumed that there is no remission, which

means once the individual is diagnosed as a cancer patient, he or she will be a prevalent case for the rest of their life.

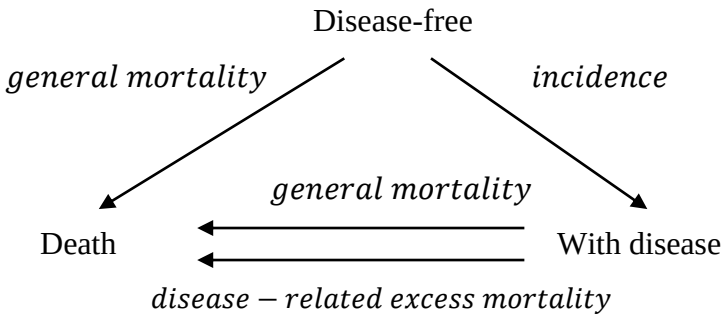


Figure 2- 3 Incidence- prevalence- mortality model structure. (Hoogenveen and Gijsen, 2000)

In this model, the mortality rate of the prevalent cases (patients) is considered to be the sum of the general mortality rate for people without the disease and the disease- related excess mortality. In other words, the mortality rates for all other causes are assumed to be the same for persons with and without the disease (Hoogenveen and Gijsen, 2000). There is another important assumption: there are no trends in the incidence and mortalities in the model, which means that the transition rates between states are consistant with calendar years.

2.1.4.3 The Disease Model (DisMod)

Based on the same theory as IPM, in the Disease Model (DisMod) the population is described as being in different states, whilst transition rates determine how people move from one state to another. The model structure of the DisMod can be expressed by Figure 2- 4. Being different to the IPM, it also includes remission as a fourth transition rate, however it can be set to zero when cure is not taken into account in the registered cancer prevalence (Kruijshaar, Barendregt and Hoeymans, 2002). Unlike the epidemiological terminology (disease-related excess mortality) in the IPM, the DisMod uses fatality rates to describe "the excess of mortality rate due to the disease, as well as the increased susceptibility

to the force of general mortality" (Kruijshaar, Barendregt and Hoeymans, 2002).

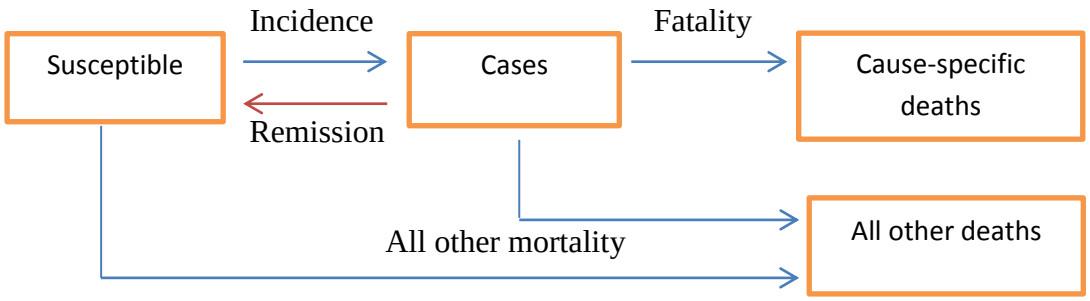


Figure 2- 4 Schematic representation of the DisMod for cancers

The prevalence of cases at an age *x* should be calculated as:

$$Prevalence(x) = \int_0^x Incidence(t) * \big(1 - Remission(x-t)\big) * \exp\left(-\int_t^x Fatality(t,u)du\right)dt \quad (2.8)$$

Where *t* is the age at diagnosis, and $Remission(x-t)$ is the proportion of cases, among the survivors at time $(x-t)$ since diagnosis that have been cured and have consequently been removed from the prevalence. The part in equation 2.8 $\exp\left(-\int_t^x Fatality(t,u)du\right)$ can be understood as the net survival function of patients in absence of mortality from other causes.

Briefly, the TRM, the IPM, and the DisMod share the same theory in building their models, and same assumptions about time-constant cancer incidence, mortality and survival probabilities. In this steady-state situation, the prevalence estimate for a birth cohort coincides with that of the current population. The differences between them are different epidemiological terms used for life status and transition rate descriptions for example, the disease-related excess mortality in the IPM is called fatality in the DisMod.

2.1.5 Back calculation methods

Back calculation methods form a broad family of methods. However, they are put to special uses in the literature, such as to provide prevalence estimate projections for the future. Therefore these methods are considered as a single group. Back calculation methods produce statistical solutions to estimate prevalence, using the parameter trends estimated from observed data to make estimates for the unobserved part. There are two methods in this group:

2.1.5.1 The MIAMOD method

The MIAMOD (Mortality Incidence Analysis Model) (Verdecchia, et al., 1989) can be used to estimate cancer prevalence using mortality data. It considers prevalent cases as the result of several phenomena acting together on the population throughout a period. These include contracting the disease, not dying from the disease, or from other causes. An important assumption in the MIAMOD is that the disease process is considered irreversible.

Figure 2- 5 shows a compartmental representation of the model with two live states (disease-free and with disease) and two death states (from specific disease and from all other causes). This model is similar to transition rate methods, where the states transition according to different rates. However, there are more kinds of death hazard rates. From demographic sources, death hazard rates from the specific cause $r_a(x)$, and from all causes together $r_b(x)$ at age *x* are usually known. From registry data, all-cause death hazards at age *x* $r_c(x,t)$ for people who became ill at age *t*, and the corresponding specific cause death hazard $r_d(x,t)$ can be estimated.

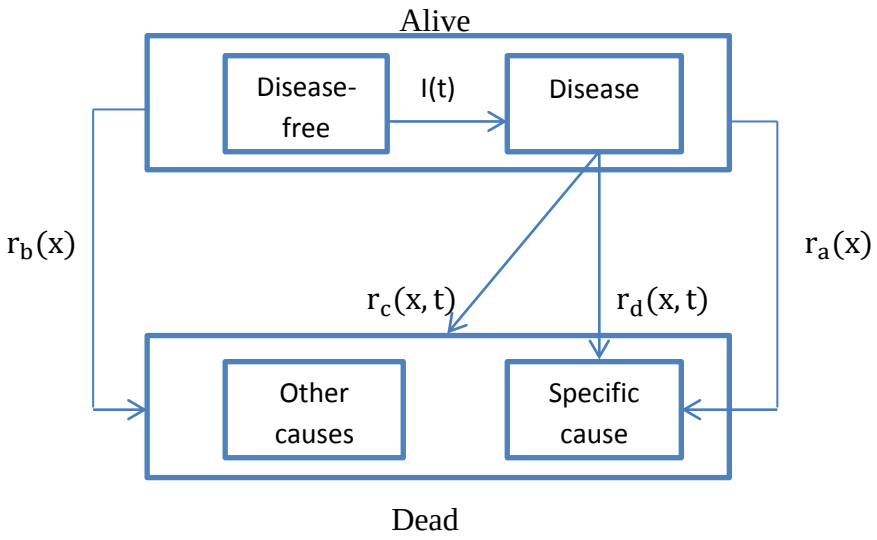


Figure 2- 5 A compartmental representation of irreversible disease-death processes

After an estimate of the incidence *I(t)* in Figure 2-5 (the disease hazard for disease- free people of age *t*) is obtained, the probability of being in the disease state for people of age *x* in the cohort can be expressed as:

$$P(x) = \int_0^x (1 - P(t)I(t)exp\{-\int_t^x [r_c(u,t) - r_b(u)]du\}dt \quad (2.9)$$

Therefore, the probability of being in the disease state for people at age *x* is the integral over all younger ages *t* of the probability of becoming diseased from disease-free at each age *t*, times the probability of surviving the extra death risk between age t and *x*. $r_a(x)$ and $r_d(x,t)$ do not appear in the prevalence equation 2.9. In fact, they are used to perform other related calculations in the MIAMOD (Verdecchia, et al., 1989).

For the MIAMOD, software is freely available using this method to calculate prevalence (EUROCARE, 2011). It can be used to estimate current prevalence and to calculate future prevalence projections. One of the features of this method is that there are many parameters involved in the estimation, which means that the results heavily depend on the estimates of those parameters.

2.1.5.2 The PIAMOD method

Unlike the MIAMOD method that focuses on mortality, the PIAMOD (Prevalence Incidence Analysis Model) (Verdecchia, et al., 2002) estimates prevalence from incidence and survival by fitting a parametric incidence model to incidence data. Following the notation in Chapter One, *t* is patient age at diagnosis and *x* is the current age on the index date. The theory of this method can be expressed as the function:

$$P(x) = \sum_{t=0}^{x-1}\big(1 - P(t)\big)I(t)S_r(t,x) \qquad (2.\ 10)$$

Where $P(x)$ is the probability of being a cancer patient, who is still alive at age x, $I(t)$ is the probability of being ill between age *t* and *t+1*, $S_r(t,x)$ is the probability of surviving the extra death hazard specifically due to the disease under consideration, and $(1 - P(t))$ represents the proportion of disease-free people at age *t* (Verdecchia, et al., 2002 (a)). The cohort-specific prevalence at age *x* is the summation over all ages up to *x* of the probabilities. This theory is similar to the transition rate method, if it is assumed that incidence, survival and population are constant with calendar years.

The PIAMOD can be used to calculate future prevalence projections. It is assumed that the projection of modelled incidence to future years is the same as during the observation period. Age-specific incidence in every year of observed period can be obtained directly. For survival, it is usually assumed that survival improvements will no longer be observed in the future. Therefore the hypothesis is that cancer patients' survival remains stable for future years. Or, in a more optimistic scenario, cancer patients' survival is assumed as continuing to improve at the same rate as observed in recent past years. Lastly, population evolution: The numbers of new born and age-specific general mortality in the population are assumed to keep constant during the projection period (Verdecchia, De Angelis, and Capocaccia, 2002).

Unlike the transition rate methods in the previous sections, the PIAMOD method is formulated as a discrete time model. The advantage of this is that it is easy to obtain probabilities in discrete time, because practical applications usually deal with discrete data (for example, incidence for a single age in every calendar year). The disadvantage is that more attention has to be paid to building models. Verdecchia, De Angelis, and Capocaccia (2002) assumed that, "events (that is, diagnosis, death) can only occur at the midpoint between two consecutive birthdays". However, if patients are diagnosed and die within the same calendar year, this assumption results in zero survival time. Equation 2.10 shows that the prevalence at age *x* does not include the patients who are diagnosed at age *x* and who are still alive.

The MIAMOD and the PIAMOD are widely used to estimate prevalence and to make future projections. Prevalence of cancer in Italy was estimated and projected to the year 2000 using the MIAMOD method (Mariotto, et al., 1999). In 2007, it was used to calculate prevalence in Italian regions (Verdecchia, et al., 2007), and the prevalence of cancer in 2010 was derived with the MIAMOD method (De Angelis, et al., 2007). The PIAMOD method was used to make long-term projections of cancer prevalence up to 2030 in the U.S., based on the data from 1973 to 1993 (Verdecchia, De Angelis, and Capocaccia, 2002). The PIAMOD was also used to estimate the number of patients with colorectal carcinoma by phases of care in the U.S. from 2000 to 2020 (Mariotto, et al., 2006), and cancer survivors in Switzerland in 2020 (Herrmann, et al., 2013).

2.1.6 Completeness index

The completeness index is a statistical model that estimates total prevalence from limited duration prevalence data (Capocaccia and De Angelis, 1997; Merrill, et al., 2000).

Limited duration prevalence (observed prevalence) represents the proportion of people alive on an index date that had a diagnosis of cancer within the period of

registry. It can be obtained using the counting method. Parametric incidence and survival models are used to estimate the proportion of modelled prevalence that is observed; the proportion is called the completeness index. This in turn is used to inflate the limited duration prevalence (Gigli, et al., 2006). Therefore, together with incidence and follow-up data from the registry, the total (unlimited) prevalence may be estimated using the completeness index method (Capocaccia and De Angelis 1997).

Total prevalence can be estimated by modeling a mathematical relationship between prevalence, incidence and survival. This is done in a single cohort, observed for a time period of *L* years. If the disease is not reversible, the relationship between prevalence, incidence and relative survival can be expressed as:

$$N(x) = \int_0^x I(t) S_r(x-t, t) dt \qquad (2.\ 11)$$

Where $N(x)$ is the proportion in the population of individuals with cancer at age *x*, $I(t)$ is the incidence rate at age *t* and $S_r(x-t, t)$ indicates the probability that a single patient diagnosed with a certain cancer at age *t* is still alive at age *x* (*t*: diagnosed age, *x*: current age).

Both incidence and relative survival in equation 2.11 are estimated using parametric functions. Usually the model assumes that there is an exponential relationship between incidence and age, adjusting for birth cohort.

$$I(t) = at^b \qquad (2.\ 12)$$

Where *a* is a scale parameter which is dependent on the birth cohort, *b* is the age slope parameter, and *t* is the incidence age. If we take the logarithm of both sides of this equation we find a linear relationship between log (incidence) and log (age):

$$logI(t) = loga + blogt \qquad (2.\ 13)$$

Because age and cohort parameters are additive on the logarithmic scale, this model has the advantage of mathematical simplicity. However, there is disadvantage in this incidence model, which is the underlying assumption of lack of interaction between age and cohort. Nevertheless, this model does provide a good fit for the incidence data of some cancers (details are in Chapter Four).

For the survival model, it considered that a proportion of the patients are cured. It is assumed that they are exposed to the same mortality rates as the general population. Under the assumption that only a proportion of the patients have an excess mortality, whilst the remainder, share the same death rate as the general population, a mixture model is used for relative survival. If *A* is the proportion of individuals with cancer who will die of the cancer with a relative survival function following the Weibull distribution, whilst the remaining proportion *(1-A)* have the same mortality rate as the general population, then:

$$S_r(x-t,t) = \left[(1-A) + Aexp\left(-\lambda(x-t)^{\beta}\right)\right]^{exp(\Upsilon(t-t_0))} \quad (2.\ 14)$$

Where *t* is the age at diagnosis and *x* is the current age, λ and β are the scale and shape parameters of Weibull distribution. In this model, $(1-A)$ represents the proportion of patients who are not exposed to excess risk of death. The parameters $\boldsymbol{\Upsilon}$ are the log relative risk of being diagnosed one year older; the constant t_0 is reference age. Usually the median age of diagnosis is used as the reference age in the calculation.

The computation of prevalence is particularly simple if incidence and relative survival are known parametric functions. However, the estimates depend on the goodness of fit of the chosen models. Therefore, Capocaccia and De Angelis (1997) continue to produce the completeness index method. It uses the incidence and relative survival parametric models to estimate the proportion of modelled prevalence that is observed, which in turn is used to inflate the limited duration prevalence.

The total prevalence could be divided into two parts: $N_o(x, L)$ which is the observed prevalent cases, and unobserved prevalent cases $N_u(x, L)$, which were diagnosed before the start of the registry and who are still alive:

$$N(x) = N_o(x, L) + N_u(x, L) \quad (2.15)$$

It can be expressed as:

$$N(x) = \int_{x-L}^{x} I(t)S_r(x-t, t)dt + \int_0^{x-L} I(t)S_r(x-t, t)dt \quad (2.16)$$

The proportion of observed prevalence of the total prevalence is given by the ratio R, and called the completeness index:

$$R = \frac{N_o(x,L)}{N(x)} = 1 - \frac{N_u(x,L)}{N(x)} \quad (2.17)$$

The completeness index R, is in turn used to inflate the observed prevalence to total prevalence:

$$N(x) = \frac{N_o(x,L)}{R} \quad (2.18)$$

This method has been widely used in European countries and in the U.S. In the U.S., cancer prevalence was estimated using this method based on tumour registry data from the Surveillance Epidemiology and End Results (SEER) program (Merrill, et al., 2000). The completeness index method was also used to estimate cancer prevalence in Europe in the EUROPREVAL program (Micheli, et al. 2002a; Capocaccia, et al. 2002). Cancer prevalence in France, Italy and Spain, Northern Europe and Central Europe, and in the UK was calculated using this method separately (Verdecchia, et al., 2002; Forman, et al., 2003; Lutz, et al., 2003; Møller et al., 2003). The total prevalence of leukaemia in Australia based on this method in1997 was reported in 2002 (Brameld, et al., 2002); cancer

prevalence was estimated in Queensland in 2002 using this method (Youlden, Health, and Baade, 2005). In 2010, the completeness index method was applied to calculate total prevalence based on the North Carolina Central Cancer Registry (NCCCR) data (Wobker, Yeh and Carpenter, 2010). This method was also used to estimate the complete childhood prevalence of acute lymphocytic leukaemia and all cancer combined based on SEER cancer registry data (Simonetti, et al., 2008).

The completeness index method has an obvious advantage. It estimates the proportion of observed prevalence and uses it to calculate the total prevalence. Therefore, compared with the methods outlined in previous sections, the completeness index method comes closer to the observed data (Capocaccia and De Angelis, 1997; Merrill, et al., 2000; Gigli, et al., 2006).

However, in this study, some subtypes of haematological malignancies show different incidence curves (see Appendix A5) and cannot be modelled using equation 2.12. Furthermore, there is no evidence to show that there is a linear relationship between "the log relative risk" and diagnosed age for some subtypes. Indeed, for some haematological malignancies, the mortality goes down with age in the young groups and increases in the elderly (details are shown in Chapter Three and Four).

2.1.7 Additional Methods

Other methods used to calculate cancer prevalence appear in the literature, and these are discussed in this section. Since they are used less in literature, they are discussed together and only briefly outlined.

2.1.7.1 The relationship between incidence, mortality and prevalence

This method uses the relationship that exists between the risk of getting cancer, the net risk of dying of a given cancer, and the age-specific prevalence of cancer (Estève, Benhamou and Raymond, 1994).

$$\frac{p(x)}{1-p(x)} = \frac{R_i(x)-R_d(x)}{1-R_i(x)} \quad (2.\ 19)$$

Where $p(x)$ is the prevalence within people of age *x*, $R_i(x)$ is the risk of getting cancer, and $R_d(x)$ is the net risk of dying of a certain cancer for the same individuals. A birth cohort is used to calculate the total prevalence. It is used to calculate prevalence, assuming the population size and structure are stationary (Estève, Benhamou and Raymond, 1994). In 2000, the total prevalence of cancer in France was estimated using this method (Colonna, et al., 2000). The advantage of this method is its simple formula (equation 2.19), whilst the disadvantage is that the net risk of dying of the specific cancer is not usually directly available from registry data.

2.1.7.2 Age specific n-year prevalent cases

This method can provide age-specific prevalence. The International Agency for Research on Cancer (IARC) reported the world-wide prevalence of cancer by estimating 1-, 2-3, and 4-5 year prevalence in 1990 (Pisani, Bray and Parkin, 2002). Prevalent cases of a given age were estimated from incidence rates and year-specific survival probabilities according to the following formula:

$$P(n-year\ cases) = \sum C * S(i-0.5) \quad (2.\ 20)$$

Where C is the annual number of new cases in age *x*, $S(d)$ represents the proportion of cancers diagnosed at age *x* and alive at time *d* after diagnosis, and *n* is the number of years as cases. Age-specific n-year prevalence includes all patients at a certain age that were diagnosed within n years before the index date, and who are still alive. For example, 5-year prevalent cases of age 45 in year 1990 are those diagnosed at age 41 in 1986 and who survive 4.5 years (from mid time of a year), plus those diagnosed at age 42 in 1987 and who survive 3.5 years, and so on until those diagnosed in 1990 at age 45 and who survive 0.5 year (see Figure 2-6).

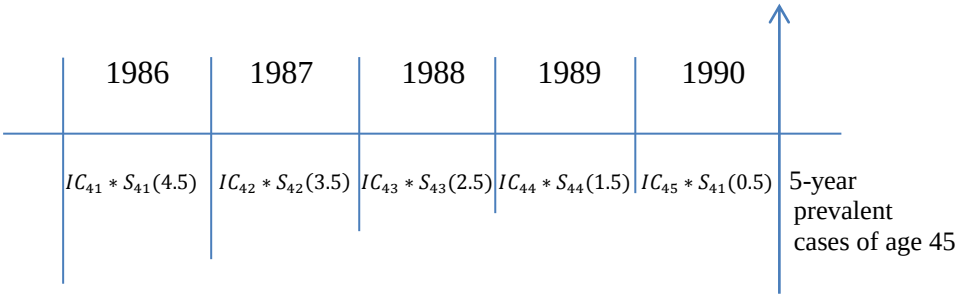


Figure 2- 6 Explanation of the method using the example of 5-year prevalent cancers at the age of 45 (Pisani, Bray and Parkin, 2002).

In the example, "5-year prevalent cases of age 45" are the patients at age 45 in 1990 who were diagnosed between 1986 and 1990 and who were still alive in 1990. This can be expressed as:

$$P_{45}(5-year) = C_{41} * S_{41}(4.5) + C_{42} * S_{42}(3.5) + C_{43} * S_{43}(2.5) + C_{44} * S_{44}(1.5) + C_{45} * S_{45}(0.5) \quad (2.\ 21)$$

At IARC, information on incidence and survival rates was directly obtained from the different countries and areas (Pisani, Bray and Parkin, 2002). In 2008, in Japan this method was used to estimate future prevalence for the year 2020 (Tabata, et al., 2008). Colorectal cancer and gastric cancer prevalence was calculated using this method in 2009 and in 2010, according to incidence and survival data in Iran (Mehrabian, et al., 2010; Esna-Ashari, et al., 2012). 5-year prevalence in Germany in 2004, based on this method, was published in 2010 (Haberland, et al., 2010). Recently, the n-year prevalence has been updated to 2008 world-wide using this method, in the GLOBOCAN project (GLOBOCAN is a project to provide contemporary estimates of the incidence of, mortality and prevalence from major types of cancer, at national level, for 184 countries of the world) (Bray, et al., 2013).

2.1.7.3 Future prevalence based on trends

Fiorentino, et al. (2011) showed a method to estimate future trends in prevalence, taking data concerning the current prevalence and externally generated trends in cancer incidence and survival as input. In addition to the general notation given in Chapter One, the following are also to be noted:

Let Y_{index} denote the most recent year for which data concerning cancer prevalence is available.

Let Y_{future} denote the year for which we want to forecast cancer prevalence.

Let Y_I denote the year of incidence, with $Y_I = Y_0$ the earliest year of diagnosis.

Let $C(t, Y_I)$ denote the number of cancer diagnoses confirmed during year $Y_I > Y_{index}$ for patients at age t.

Let A denote the contribution to cancer prevalence at Y_{future} from those diagnosed at a time $\leq Y_{index}$.

Let B denote the contribution to cancer prevalence at $Y_{future,}$ from those diagnosed at a time $> Y_{index}$.

Let *Tot* denote the total cancer prevalence at $Y_{future.}$

The conditional probability of an individual, diagnosed at $Y_0 \leq Y_I \leq Y_{index}$, of age *t,* surviving at least a time of $Y_{future} - Y_I$ after diagnosis, given that they have survived a period of $Y_{index} - Y_I$, is

$$\frac{S(t,Y_{future}-Y_I,Y_I)}{S(t,Y_{index}-Y_I,Y_I)} \qquad (2.\ 22)$$

The number of patients expected to be alive at Y_{future}, given that they have survived to Y_{index} is:

$$A = \sum_{Y_I=Y_0}^{Y_{index}} \sum_{t=0}^{99} N(t, Y_I) * \frac{S(t, Y_{future} - Y_I, Y_I)}{S(t, Y_{index} - Y_I, Y_I)} \quad (2.\ 23)$$

Next, the prevalence attributable to those cancer diagnoses occurring after Y_{index} needs to be added:

$$B = \sum_{Y_I=Y_{index}+1}^{Y_{future}} \sum_{t=0}^{99} C(t, Y_I) * S(t, Y_{future} - Y_I, Y_I) \quad (2.\ 24)$$

The total prevalence in some year Y_{future} in the future is:

$$Tot = A + B \quad (2.\ 25)$$

This method is used to estimate the prevalence up to 20 years after index an date in the UK (Fiorentino, et al., 2011). It uses data from 1985 to 2004, which means that in the calculation, $Y_{index} = 2004$ and $Y_0 = 1985$. It was also applied to data from the NCIN (NCIN is the National Cancer Intelligence Network that works to drive improvement in standards of cancer care and clinical outcomes using information collected about cancer patients for research and analysis in the UK. [NCIN, 2012]) to estimate the prevalence in 2040 in the UK, with $Y_{index} = 2008$ and $Y_0 = 1971$ (Maddams, Utley, and Møller, 2012). The significant advantage of this method is that the future prevalence can be estimated using a simple analytical model. Incidence and survival in the future are assumed to follow the observed trend in data. However, this method requires a relatively long registry period to avoid the bias from surviving patients who were diagnosed before the start of the registry.

2.1.8 Summary of the methods

References describing the major methodologies used to estimate prevalence are summarised in Table 2-1. These methods calculate or estimate various different types of prevalence. In the earlier methods, prevalence was calculated directly without estimation or adjustment. This means that there were frequent underestimates of prevalence at that time (Hakama, et al., 1975; Feldman, et al., 1986; Adami, et al., 1989). After the 1990s, most estimates were made from cancer registries based on mathematical models. Such estimation-based methods were used to calculate total prevalence in Europe (Capocaccia, et al., 2002; Micheli, et al., 2002a), and provided country- specific estimations of total prevalence in the UK, France, Italy, Sweden, Norway, Denmark, and so on (Verdecchia, et al., 2002; Forman, et al., 2003; Lutz, et al., 2003; Möller, et al., 2003). Total prevalence in the U.S. was also estimated using SEER (SEER is the Surveillance, Epidemiology, and End Results Program in the National Cancer Institute. It works to provide information on cancer statistics with the aim to reduce the burden of cancer among the U.S. population. [SEER, 2012]) data from the mathematical model (Merrill et al. 2000). There are also some studies that have focussed on examining the trends and projecting forward prevalence estimates in the U.S. (Mariotto, et al., 2006), UK (Fiorentino, et al., 2011; Maddams, Utley, and Møller., 2012), and Switzerland (Herrmann, et al., 2013).

Most of the prevalence estimates were based on patients, however some estimated tumour prevalence based on diagnoses. For example, patients with multiple malignant primaries were included in the computation of total prevalence in Italy in 2006 (Guzzinati, et al., 2012). The reports in Canada focused on tumour-based prevalence instead of person- based prevalence (Ellison and Wilkins, 2009).

Table 2- 1 Summary of methods used to calculate prevalence

Author	Published date	Region	Prevalence	Method
Hakama, et al.	1975	Finland	Observed prevalence	Direct method (count numbers)
Feldman, et al.	1986	US	Observed prevalence	Direct method (count numbers)
Adami, et al.	1989	Sweden	Observed prevalence	Direct method (count numbers)
Byrne, Kessler and	1992	US	Observed prevalence	Cross-sectional population-based surveys
Devesa Schrijvers, et al.	1994	Netherland	Observed prevalence	Cross-sectional population-based surveys
De Angelis, et al.	1994	Italy	Future prevalence	MIAMOD
Polednak	1997	US	Observed prevalence	Direct method (count numbers)
Capocaccia and De Angelis	1997	NA.	Total prevalence	Completeness index
Gail, et al.	1999	US	Observed prevalence Total prevalence	Counting method Transition rate
Mariotto, et al.	1999	Italy	Future prevalence	MIAMOD
Zanetti, et al.	1999	Nordic countries, EU, Connecticut, Italy	Observed prevalence Total prevalence	Counting method Completeness index
Hoogenveen, et al.	2000	Netherland	Total prevalence	IPM
Colonna, et al	2000	France	Total prevalence	Other
Merrill, et al.	2000	US	Total prevalence	Completeness index
Merrill	2001	US	Observed prevalence	Counting method
Parkin, et al.	2001	Worldwide	n-year prevalence	Other

Table 2-1 continued

Author	Published date	Region	Prevalence	Method
Kruijshaar, Barendregt and Hoeymans	2002	Netherland	Total prevalence	DisMod
Verdecchia, et al.	2002	Europe and US	Future prevalence	PIAMOD
Pisani, Bray and Parkin	2002	Worldwide	n-year prevalence	Other
Micheli, et al.	2002b	US	Observed prevalence	PREVAL
Micheli,, et al.	2002a	Europe	Total prevalence	Completeness index
Capocaccia, et al	2002	Europe	Total prevalence	Completeness index
Verdecchia, et al.	2002	France, Italy, Spain	Total prevalence	Completeness index
Brameld, et al.	2002	Western Australia	Total prevalence	Completeness index
Forman, et al.	2003	UK	Total prevalence	Completeness index
Möller, et al.	2003	Northern Europe	Total prevalence	Completeness index
Lutz, et al.	2003	Central Europe	Total prevalence	Completeness index
Youlden, Health, and Baade	2005	Queensland	Total prevalence	Completeness index
Louchini, et al.	2006	Quebec (Canada)	Observed prevalence	Counting method
Gigli, et al.	2006	US	Total prevalence	Completeness index
Mariotto, et al.	2006	US	Future prevalence	PIAMOD
De Angelis, et al.	2007	Italy	Future prevalence	MIAMOD
Verdecchia, et al.	2007	Italy	Total prevalence	MIAMOD
Tabata, et al.	2008	Japan	n-year prevalence	Other
Simonetti, et al.	2008	US	Total prevalence	Completeness index

Table 2-1 continued

Author	Published date	Region	Prevalence	Method
Ellison and Wilkins	2009	Canada	n-year prevalence	Counting method
Mehrabian, et al.	2010	Iran	n-year prevalence	Other
Haberland, et al.	2010	Germany	n-year prevalence	Other
Wobker, Yeh and Carpenter	2010	US	Total prevalence	Completeness index
Fiorentino, et al.	2011	UK	Future prevalence	Other
Esna-Ashari, et al.	2012	Iran	n-year prevalence	Completeness index
Guzzinati, et al.	2012	Italy	Total prevalence	Completeness index
Maddams, Utley and Møller	2012	UK	Future prevalence	Other
Visser, et al.	2012	Europe	Total prevalence	Completeness index
Bray, et al.	2013	Worldwide	n-year prevalence	Other
Herrmann, et al.	2013	Switzerland	Future prevalence	PIAMOD

MIAMOD: Mortality Incidence Analysis Model; PIAMOD: Prevalence Incidence Analysis Model; IPM: incidence, prevalence, and mortality; DisMod: disease model; SEER: Surveillance Epidemiology and End Results; IARC: International Agency for Research on Cancer; EU: European Union. NA: Not Available.

2.2 Comparison of the methods

2.2.1 The differences between the methods

Amongst the main methods mentioned in the previous section, direct calculations (the counting method and the PREVAL approach) and the transition rate methods are non-parametric methods to calculate prevalence, whilst back calculation and completeness index methods estimate prevalence using parametric models. Direct calculations provide prevalence for patients diagnosed within the registry period, and transition rate methods are more flexible and can provide n-year prevalence or total prevalence. Back calculation methods project future prevalence, as well as estimate total prevalence, whilst the completeness index is a method purely designed to estimate total prevalence.

The relationships between incidence, survival, mortality, and prevalence are used in models to make estimates, and the models are built based on assumptions that make the estimation process feasible. The assumptions used in prevalence estimation are not in order to adopt the best hypothesis, but to keep the model easy to be understood and calculate, as well as to provide plausible results (Verdecchia, De Angelis, and Capocaccia, 2002). The most convenient assumption is to keep probabilities (such as incidence and transition rates) constant over time in the calculation. However for the projection of prevalence in the future, the incidence is modelled to follow the observed trend into future years. Usually cancer is considered as an irreversible disease, and all incident cases are counted as prevalent cases up to death, even if treatment is effective (Estève, Benhamou and Raymond, 1994). However, there are some methods in the literature that involve remission rates and calculate prevalence for curable disease (Capocaccia and De Angelis, 1997; Hoogenveen and Gijsen 2000). Sometimes the advantages and disadvantages are not so obvious, and it is difficult to say which one is the best method. Table 2-2 summarizes the characteristics of the different methods of prevalence estimation.

Table 2- 2 Comparison of methods in literature

Method		Estimating model	Assumption about probabilities trends	Cured	Strength	Weakness
Direct calculation	Counting method	Survival	Observed and lost in follow-up patients have same survival	-	Age-specific prevalence	Only calculate for patients diagnosed within the registry period
	PREVAL approach	Mortality	Lost cases have the same mortality as not lost to follow-up	-	Prevalence for n years	
Transition rate method	TRM	Incidence &Mortality	Constant over time	No	Non-parametric, Easy to use and robust	Requires many probabilities
	IPM	Incidence &Mortality	Constant over time	No		
	DisMod	Incidence &Mortality	Constant over time	Yes		
		Remission rate	-			
Back calculation method	MIAMOD	Incidence & Mortality	Change follow the observed trends	No	Project prevalence in future	Strongly depends on parametric assumptions on incidence and survival models
	PIAMOD	Incidence	Change follow the observed trends	No	Project prevalence in future	
		Survival	Change follow the observed trends OR Constant over time		Formulate in discrete time	
Completeness Index	Completeness Index	Incidence	Cohort parameter in incidence can be omitted in calculation	Yes	More closed to observed data	Parametric models do not suitable for the data in this study
		Survival	Constant over time			
Other	Other (Estève, et al., 1994)	Incidence & Mortality	Constant over time	No	Use easily available information	Net risk of dying is not usually available from registry data
	Other (Pisani et al. 2002)	Incidence & Survival	-	No	Age-specific prevalence	Restriction of limited duration prevalence to 5 years
	Other (Fiorentino, et al., 2011)	Incidence & Survival	Constant over time	No	Project prevalence in future	Simplicity of model used to project cancer prevalence

In this study, the aim was to estimate the number of past cases that were not registered. Since it was not concerned with the extrapolation of prevalence estimates for the future, back calculation methods were not appropriate for this work. Apart from this, transition rate methods and completeness index methods can be used to estimate total prevalence based on the observed data. Usually the completeness index method uses parametric functions to estimate incidence and relative survival. The transition rates are obtained from actuarial estimates.

At first sight, those methods have different models for estimation. However, there is a relationship between them. This is because there are theoretical relationships which link incidence, mortality, and survival (Estève, Benhamou and Raymond, 1994). The feasibility of estimating prevalence can be assessed from this.

2.2.2 The relationship between the transition rate method and the completeness index method

This section gives an insight into the relationship between two prevalence models. In this calculation, the relationships between the transition rate method and completeness index method can be found. Although the models appear totally different to each other, a certain amount of algebraic manipulation will show them to be similar. These manipulations can be found in the literature (Verdecchia, et al., 1989; Gras, Daurès andTretarre,, 2006); the relationship between transition rate models and the completeness index model is shown as follows:

Following the definitions given for transition rate methods, let $r_2^*(x)$ indicate the general mortality rates at age *x*. $r_1(x)$ is the incidence rate at age *x*. $r_3(x, x-t)$ is the death rate at age *x* for patients who had a cancer diagnosed at age *t*. If $S_r(x, x-t)$ is the relative survival, then it can be shown that:

$$S_r(x, x-t) = exp(-(\int_y^x r_3(u, u-t)du - \int_y^x r_2^*(u)du)) \quad (2.\ 26)$$

It is assumed that:

Firstly, the disease is rare and the incidence is low: $r_1 \ll 1$. So,

$$exp(-\int_0^u r_1(t)dt) \approx 1 \qquad (2.27)$$

Furthermore, since the model is used to estimate prevalence in a population, *u* represents age and will not be very large (usually less than 100). In addition, the mortality rate of non-cases r_2 is approximated using the mortality rate of the general population $r_2^*(x)$. Therefore,

$$r_2(x) \approx r_2^*(x) \qquad (2.28)$$

The age-specific L- year partial prevalence using the method of transition rates is:

$$P = \frac{P_I(x,L)}{S(x)} = \frac{\int_{x-L}^{x} exp\left(-\int_0^y (r_1+r_2)(u)du\right) r_1(t)\, exp\left(-\int_t^x r_3(u,u-t)du\right)dt}{S_{overall}(x)} \qquad (2.29)$$

Because of the first assumption of $r_1 \ll 1$ and equation (2.26), equation (2.28) can be reformulated as follows:

$$P = \frac{\int_{x-L}^{x} exp\left(-\int_0^t r_2(u)du\right) r_1(t)\, exp\left(-\int_t^x r_3(u,u-t)du\right)dt}{S_{overall}(x)} \qquad (2.30)$$

Because of the second assumption $r_2(x) \approx r_2^*(x)$, equation (2.29) can be continued to be reformulated as follow:

$$P = \frac{\int_{x-L}^{x} exp\left(-\int_0^t r_2^*(u)du\right) r_1(t) exp\left(-\int_t^x r_3(u,u-t)du\right)dt}{S_{overall}(x)} \qquad (2.31)$$

This leads to:

$$P = \int_{x-L}^{x} r_1(t) * \frac{exp\left(-\int_0^t r_2^*(u)du\right)}{exp\left(-\int_0^x r_2^*(u)du\right)} * exp\left(-\int_t^x r_3(u, u-t)du\right)dt$$

$$= \int_{x-L}^{x} r_1(t) exp\left(\int_t^x r_2^*(u)du\right) exp\left(-\int_t^x r_3(u, u-t)du\right)dt \quad (2.\ 32)$$

Because of the equation (2.26), this can be expressed as follows:

$$P = \int_{x-L}^{x} I(t) S_r(x, x-t)dt \quad (2.\ 33)$$

Where *I(t)* is incidence at age t, $S_r(x, x-t)$ is the relative survival at age *x* surviving for *x-t* years. This is the basic model of the completeness index method. In other words, under certain conditions, the completeness index method is approximately equal to the transition rates method in calculating the L- year prevalence.

2.3 Reported prevalence figures of haematological malignancies in the literature

2.3.1 Prevalence reports from main cancer registries in the world

Some well-established cancer registries that provide prevalence figures are summarized in Table 2-3. These cancer registries can provide total prevalence or n-year prevalence for the UK, Italy, Canada, Australia, the U.Ss, and the Nordic European countries, with latest reports from 2006 to 2011. N-year cancer prevalence in the UK in 2006 is obtained from the NCIN (National Cancer Intelligence Network (NCIN), 2010). For Italy, data is derived from the network of cancer registries, AIRTUM (Guzzinati, et al., 2012) to estimate total prevalence. For Canada, the latest n-year prevalence in 2009 was derived from the Canadian Cancer Registry (CCR) maintained by Statistics Canada (Ellison and Wilkins, 2009; CCR, 2012). For the U.S., total prevalence can be obtained from the Surveillance, Epidemiology and End Results (SEER) Program (SEER, 2012), in which information is from specific geographic areas representing 28 percent of the U.S. population. Longer period cancer prevalence can be obtained directly from registries that operated for longer times. For Australia, data is from the

Australian Institute of Health and Welfare (AIHW) (AIHW, 2012), which provides 26- year prevalence, and for the Nordic European countries, data is from NORDCAN, which provides up to 48- year prevalence in Denmark (Engholm, et al., 2013).

Apart from the reports from these main cancer registries, international n-year prevalence was estimated using GLOBOCAN (GLOBOCAN, 2008) which covers 184 countries. Some prevalence figures in this chapter are derived from a specific publication (Pisani, Bray and Parkin, 2002), whilst for countries in the world, the prevalence has been updated to 2008 on its web site however are not shown here (GLOBOCAN, 2008; Bray, et al., 2013). The estimated prevalence in GLOBOCAN relies on incidence and survival at country level, therefore it is greatly limited in terms of some of its data sources. For example, compared to developed countries, many low-income countries rarely have survival estimates and proxies of their survival are created under assumptions. The European Cancer Registry-based study on survival and Care (EUROCARE) is a large cooperative registry based study which covers 23 countries and 89 cancer registries in Europe (Sant, et al., 2009). The EUROPREVAL and HAEMACARE are projects that were set up to estimate total prevalence of cancer, and to improve the standardization of data on haematological malignancies respectively, archived by EUROCARE. The EUROPREVAL project presented total prevalence in European countries in 1992. However, in some countries, such as France, the cancer registries providing the data only covered small fractions of the populations (Verdecchia, et al., 2002; Crocetti, et al., 2013). The HAEMACARE project provided incidence and survival for haematological malignancies under WHO classification in 48 cancer registries, however the reports concerning prevalence were not available (Sant, et al., 2010; Marcos-Gragera, et al., 2011; Maynadié, et al., 2013). Apart from these projects, total prevalence of myeloid malignancies and other rare cancers were provided by RARECARE, which extracted data from 64 cancer registries (excluding cancer registries which did not report cancers according to ICD- O- 3) (Gatta, et al., 2011; Visser, et al., 2012).

Table 2- 3 Main cancer registries in this section, projects for prevalence, percentage of the population of the country, and latest year for prevalence reports

Country	Cancer registry	Project	Population covered by cancer registry	Year of prevalence	Web-site
United Kingdom	Cancer Networks	NCIN	100%	2006	http://www.ncin.org.uk/
Italy	AIRTUM	ITACAN	40%	2006	http://www.registri-tumori.it/cms/en
Canada	CCR	-	100%	2009	http://www.statcan.gc.ca/start-debut-eng.html
Australia	AACR	AIHW	100%	2007	http://www.aihw.gov.au/
United States	SEER	SEER	28%	2010	http://seer.cancer.gov/
Nordic European Countries*	ANCR	NORDCAN	100%	2010(2011)	http://www.ancr.nu/nordcan.asp

*Norway, Sweden, Denmark, Finland, Iceland (Norway and Sweden provided prevalence in 2010; Denmark, Finland, Iceland provided prevalence in 2011).

Abbreviations:
NCIN: National Cancer Intelligence Network. AIRTUM: The Italian Association of Cancer Registries. CCR: Canadian Cancer Registry. AACR: Australasian Association of Cancer Registries. AIHW: Australian Institute of Health and Welfare. SEER: Surveillance, Epidemiology and End Results. ANCR: The Association of the Nordic Cancer Registries.

2.3.2 Lack of systematic reports about haematological malignancies

Although there are many methods in the literature that are used to calculate the prevalence of cancer, few publications have focused on prevalence studies for haematological malignancies, therefore their prevalence is not routinely available. For example, Gail (1999) used breast and brain cancers only as the example to show the Transition Rates Method and Counting method (Gail, et al., 1999). In 2006, Gigli (2006) calculated the prevalence of colon cancer from an Italian registry. The PIAMOD was used to calculate and project cancer prevalence in the future, but only the prevalence of breast cancer was reported in that study (Verdecchia, De Angelis, and Capocaccia 2002). Breast cancer, prostate cancer, lung cancer, and colon cancer were routinely reported in the studies, however not all the papers about prevalence demonstrated the prevalence of haematological malignancies.

Furthermore, haematological malignancies were considered as a whole group for prevalence estimates in France (Colonna, et al., 2000) and Norway (Skjelbakken, et al., 2002) in 2000 and 1996 respectively. In the EUROPREVAL project, only the prevalence of leukemia and Hodgkin lymphoma were reported (Verdecchia, et al., 2002; Micheli, et al., 2002a; Capocaccia, et al., 2002; Forman, et al., 2003; Lutz, et al., 2003). The SEER project only reported the number of prevalent cases without the prevalence rates (Merrill, et al., 2000). The reported prevalence of haematological malignancies is summarized in the Table 2-4 and Table 2-5. The prevalence of haematological malignancies is abstracted from the results in the literature, shown in order of increasing index date. Not all of them are total prevalence, and the types of prevalence they calculated are indicated in the column "Note". Among those reports, GLOBOCAN estimated prevalence for adults only (over 15 years old) (GLOBOCAN, 2008), whilst others included all age groups.

Although there are some reports of prevalence in the literature, most of the studies were published before the new classification of haematological malignancies by the WHO (WHO, 2008). Therefore prevalence is only reported by the four broad

categories—leukaemia, Hodgkin lymphoma, non-Hodgkin lymphoma, and myeloma. It is very interesting that some of the studies exclude non-Hodgkin lymphoma, as this is the biggest group, yet the reasons for this are not given (Verdecchia, et al., 2002; Micheli, et al., 2002a; Capocaccia, et al., 2002; Forman, et al., 2003; Lutz, et al., 2003). It is worth noting that the prevalence indicated of Hodgkin lymphoma in some countries (for example the UK [Forman, et al., 2003]) is not total prevalence but 15-year prevalence due to the marked changes in treatment (Capocaccia, et al., 2002). For some recent web based measures, the prevalence of some subtypes were provided: NORDCAN presented prevalence of acute leukaemia and other leukaemia separately (Engholm, et al., 2013); SEER provided prevalence of acute myeloid leukaemia, chronic myelogenous leukaemia, acute lymphocytic leukaemia, and chronic lymphocytic leukaemia in the broad leukaemia group, as well as of myelodysplastic syndromes and chronic myelomonocytic leukaemia (SEER, 2012); RARECARE showed the prevalence of myeloid malignancies based on ICD- O- 3 classification (Visser, et al., 2012).

Table 2- 4 Prevalence of haematological malignancies per 100, 000 for males

Country/ Area	Project	Leukaemia		HL		NHL		Myeloma		Index date	Note
		Counts	Rates*	Counts	Rates*	Counts	Rates*	Counts	Rates*		
Canada			1.4							1990	5-year prevalence
SSA	GLOBOCAN		1.4		1.7		4.4		0.5	1990	1 year prevalence
MENA			2.1		1.3		2.9		0.2	1990	
LAC			2.8		1.6		3.7		0.6	1990	
North America			9.4		3.2		15.8		4.5	1990	
China and OEA			1.7		0.3		0.9		0.1	1990	
Japan			5.4		0.5		7.5		1.8	1990	
South- Eastern Asia			2.0		0.6		2.7		0.4	1990	
South- Central Asia			2.0		1.1		2.1		0.3	1990	
Eastern Europe			6.6		3.7		4.7		1.7	1990	
EU and EEA			9.0		3.5		12.0		3.9	1990	
Oceania			8.8		2.1		12.7		3.6	1990	
Developed			8.1		3.1		10.6		3.3	1990	
Developing			1.9		0.9		2.2		0.3	1990	
World			3.4		1.4		4.3		1.0	1990	

Table 2-4 continued

Country/ Area	Project	Leukaemia		HL		NHL		Myeloma		Index date	Note
		Counts	Rates*	Counts	Rates*	Counts	Rates*	Counts	Rates*		
SSA	GLOBOCAN		1.7		2.7		6.4		0.7	1990	2-3 years prevalence
MENA			2.7		2.1		4.2		0.3	1990	
LAC			3.5		2.6		5.1		0.7	1990	
North America			14.4		6.0		26.2		6.2	1990	
China and OEA			2.3		0.5		1.3		0.1	1990	
Japan			8.6		0.9		12.5		2.4	1990	
South- Eastern Asia			2.5		1.0		3.8		0.4	1990	
South- Central Asia			2.4		1.9		3.3		0.4	1990	
Eastern Europe			8.7		6.3		7.0		2.2	1990	
EU and EEA			12.3		6.5		18.7		5.2	1990	
Oceania			12.0		3.8		19.4		4.8	1990	
Developed			11.5		5.6		16.9		4.3	1990	
Developing			2.4		1.5		3.1		0.4	1990	
World			4.6		2.5		6.5		1.3	1990	
SSA	GLOBOCAN		1.1		2.2		4.9		0.5	1990	4-5 years prevalence
MENA			1.9		1.8		3.2		0.2	1990	
LAC			2.4		2.1		3.8		0.5	1990	
North America			11.3		5.6		22.2		3.7	1990	

Table 2-4 continued

Country/ Area	Project	Leukaemia		HL		NHL		Myeloma		Index date	Note
		Counts	Rates*	Counts	Rates*	Counts	Rates*	Counts	Rates*		
China and OEA	GLOBOCAN		1.7		0.4		1.0		0.1	1990	4-5 years prevalence
Japan			6.9		0.8		10.6		1.5	1990	
South- Eastern Asia			1.8		0.8		2.9		0.3	1990	
South- Central Asia			1.4		1.6		2.2		0.3	1990	
Eastern Europe			6.0		5.3		5.3		1.3	1990	
EU and EEA			8.7		5.8		14.7		3.2	1990	
Oceania			8.5		3.3		15.2		3.0	1990	
Developed			8.5		5.0		13.7		2.7	1990	
Developing			1.7		1.2		2.4		0.2	1990	
World			3.3		2.1		5.1		0.8	1990	
UK	EUROPREVAL		38.0		28.0					1992	Total prevalence
France			61.6		25.4					1992	
Italy			42.6		32.2					1992	
Spain			39.1		29.2					1992	
Denmark			47.2		28.4					1992	
Finland			37.4		24.3					1992	
Iceland			30.2		27.5					1992	
Estonia			37.4		17.9					1992	
Sweden			52.6		20.6					1992	

Table 2-4 continued

Country/ Area	Project	Leukaemia		HL		NHL		Myeloma		Index date	Note
		Counts	Rates*	Counts	Rates*	Counts	Rates*	Counts	Rates*		
Netherland	EUROPREVAL		31.6		20.2					1992	Total prevalence
Germany			46.2		35.0					1992	
Austria			49.9		30.1					1992	
Switzerland			64.6		25.6					1992	
Slovenia			30.4		22.2					1992	
Slovakia			42.3		19.7					1992	
Poland			15.2		18.7					1992	
All European countries			39.5		26.3					1992	
Worldwide	GLOBOCAN	233,100		113,500		381,400		74,300		2000	5-year prevalence
Canada		3,426	21.5	883	5.5	4,930	31	1,362	8.6	2005	2-year prevalence
		6,720	42.2	2,079	13.1	10,015	92.9	2,428	15.2	2005	5-year prevalence
		10,170	63.9	3,806	23.9	15,316	96.2	3,126	19.6	2005	10-year prevalence
England	NCIN	2,192	7.9	673	2.6	3,440	12.1	1,294	4.4	2006	1-year prevalence
Scotland		244	8.9	65	2.6	322	11.1	148	4.9	2006	
Wales		172	9.8	34	2.2	228	12.9	97	5.1	2006	
Northern Ireland		60	6.8	20	2.3	104	12.1	56	6.3	2006	
United Kingdom		2,668	8.0	792	2.6	4,094	12.1	1,595	4.5	2006	

Table 2-4 continued

Country/ Area	Project	Leukaemia		HL		NHL		Myeloma		Index date	Note
		Counts	Rates*	Counts	Rates*	Counts	Rates*	Counts	Rates*		
England	NCIN	8,225	29.8	2,939	11.5	12,898	45.6	4,277	14.5	2006	5-year prevalence
Scotland		1,008	36.2	313	12.5	1,284	44.8	472	15.8	2006	
Wales		891	33.5	161	10.9	794	45.4	325	17.3	2006	
Northern Ireland		229	26.1	84	9.6	402	46.2	173	19.7	2006	
United Kingdom		10,053	30.5	3,497	11.5	15,378	45.6	5,247	14.9	2006	
England	NCIN	12,876	46.7	5,521	21.6	24,204	71.5	5,659	19.2	2006	10-year prevalence
Scotland		1,619	57.9	550	21.8	2,020	70.3	606	20.2	2006	
Wales		897	51.2	303	20.6	1,146	65.7	427	22.7	2006	
Northern Ireland		346	39.3	162	18.9	634	72.9	229	26.0	2006	
United Kingdom		15,738	47.7	6,536	21.5	24,004	71.1	6,921	19.7	2006	
England	NCIN	17,392	63.2	9,498	38.1	27,144	108.9	6,327	21.4	2006	20-year prevalence
Scotland		2,145	77.3	990	38.6	2,734	95.8	660	22.0	2006	
Italy	ITACAN	7,608	97.0	6,418	82.0	14,102	180.0	3,194	40.0	2006	Total prevalence
Texas	TCR	5,740		2,406		9,389		1,854		2006	10-year prevalence
Australia	AIHW	-	-	3,877	36.7	16,547	156.7	3,030	28.7	2007	1982-2007
Worldwide	GLOBOCAN	278,754	11.4	114,537	4.7	427,038	17.4	112,421	4.6	2008	5-year prevalence

Table 2-4 continued

Country/ Area	Project	Leukaemia		HL		NHL		Myeloma		Index date	Note
		Counts	Rates*	Counts	Rates*	Counts	Rates*	Counts	Rates*		
Canada		4,182		900		5,896		1,562		2009	2-year prevalence
		8,502		2,099		12,440		3,111		2009	5-year prevalence
		13,044		3,890		19,140		4,103		2009	10-year prevalence
United States	SEER	162,651		93,890		266,487		42,185		2010	Total prevalence
Nordic countries	NORDCAN	12,920	101.7	6,438	50.7	19,355	152.4	3,896	30.7	2010	1980-2010
Norway		2,400	97.5	1,297	52.7	3,662	148.8	859	34.9	2010	1973-2010
Sweden		5,099	108.7	2,080	44.4	6,734	143.6	1,549	33.0	2010	1980-2010
Denmark		2,988	108.0	1,489	53.8	4,463	161.4	807	29.2	2011	1963-2011
Finland		2,355	88.8	1,523	57.4	4,757	179.3	692	26.1	2011	1973-2011
Iceland		166	103.5	102	63.6	225	140.2	52	32.4	2011	1975-2011

*per 100, 000

Abbreviations in the table:
HL: Hodgkin lymphoma, NHL: non-Hodgkin lymphoma, SEER: the Surveillance, Epidemiology, and Ends Results, IARC: International Agency for Research on Cancer, SSA: sub-Saharan Africa, MENA: Middle East and Northern Africa, EU: European Union, EEA: European Economic Area, LAC: Latin America and Caribbean, OEA: Korea, Mongolia, and Hong Kong, NCIN: National Cancer Intelligence Network, TCR: Texas Cancer Registry, AIHW: Australian Institute of Health and Welfare, ASP: age standardized proportion, NA: not available

Table 2- 5 Prevalence of haematological malignancies per 100, 000 for females

Country/ Area	Project	Leukaemia		HL		NHL		Myeloma		Index date	Note
		Counts	Rates*	Counts	Rates*	Counts	Rates*	Counts	Rates*		
Canada			1.4							1990	5-year prevalence
SSA	IARC		1.1		0.8		3.0		0.4	1990	1 year prevalence
MENA			1.3		0.8		2.4		0.3	1990	
LAC			2.4		1.0		2.6		0.6	1990	
North America			6.8		2.5		12.4		4.0	1990	
China and OEA			1.3		0.3		0.7		0.1	1990	
Japan			3.7		0.2		5.0		1.9	1990	
South- Eastern Asia			1.7		0.3		2.0		0.3	1990	
South- Central Asia			1.4		0.5		1.3		0.3	1990	
Eastern Europe			5.2		2.3		3.3		1.9	1990	
EU and EEA			6.8		2.1		9.4		3.6	1990	
Oceania			5.8		1.6		9.9		3.2	1990	
Developed			6.0		2.1		8.0		3.1	1990	
Developing			1.5		0.5		1.6		0.3	1990	
World			2.7		0.9		3.3		1.0	1990	

Table 2-5 continued

Country/ Area	Project	Leukaemia		HL		NHL		Myeloma		Index date	Note
		Counts	Rates*	Counts	Rates*	Counts	Rates*	Counts	Rates*		
SSA	IARC		1.4		1.3		4.4		0.5	1990	2-3 years prevalence
MENA			1.6		1.3		3.4		0.5	1990	
LAC			2.9		1.6		3.5		0.7	1990	
North America			10.4		4.7		20.1		5.4	1990	
China and OEA			1.8		0.5		1.0		0.1	1990	
Japan			5.9		0.4		8.2		2.6	1990	
South- Eastern Asia			2.1		0.5		2.9		0.3	1990	
South- Central Asia			1.7		0.7		1.8		0.4	1990	
Eastern Europe			6.9		4.1		4.9		2.6	1990	
EU and EEA			9.7		3.9		14.7		4.8	1990	
Oceania			8.3		3.0		15.3		4.2	1990	
Developed			8.7		3.8		12.7		4.1	1990	
Developing			1.9		0.8		2.2		0.3	1990	
World			3.7		1.6		5.0		1.3	1990	
SSA	IARC		0.9		1.0		3.4		0.3	1990	4-5 years prevalence
MENA			1.1		1.1		2.6		0.3	1990	
LAC			2.0		1.3		2.6		0.5	1990	
North America			8.2		4.4		16.8		3.2	1990	

Table 2-5 continued

Country/ Area	Project	Leukaemia		HL		NHL		Myeloma		Index date	Note
		Counts	Rates*	Counts	Rates*	Counts	Rates*	Counts	Rates*		
China and OEA	IARC		1.4		0.4		0.8		0.1	1990	4-5 years prevalence
Japan			4.8		0.4		6.9		1.6	1990	
South- Eastern Asia			1.5		0.4		2.2		0.2	1990	
South- Central Asia			1.0		0.6		1.3		0.2	1990	
Eastern Europe			4.9		3.7		3.7		1.6	1990	
EU and EEA			7.2		3.6		11.6		2.9	1990	
Oceania			6.1		2.7		12.1		2.5	1990	
Developed			6.5		3.4		10.2		2.5	1990	
Developing			1.3		0.7		1.7		0.2	1990	
World			2.7		1.4		3.9		0.8	1990	
UK	EUROPREVAL		31.0		19.0					1992	Total prevalence
France			50.6		19.7					1992	
Italy			34.5		29.2					1992	
Spain			25.6		18.4					1992	
Denmark			36.9		17.8					1992	
Finland			33.4		17.9					1992	
Iceland			22.8		11.5					1992	
Estonia			35.7		13.4					1992	
Sweden			35.7		17.1					1992	

Table 2-5 continued

Country/ Area	Project	Leukaemia		HL		NHL		Myeloma		Index date	Note
		Counts	Rates*	Counts	Rates*	Counts	Rates*	Counts	Rates*		
Netherland	EUROPREVAL		26.6		17.0					1992	Total prevalence
Germany			26.1		19.2					1992	
Austria			31.8		23.4					1992	
Switzerland			38.3		17.4					1992	
Slovenia			31.5		15.1					1992	
Slovakia			32.9		18.3					1992	
Poland			13.7		15.9					1992	
All European countries			31.9		19.3					1992	
Worldwide	GLOBOCAN	187,500		83,300		291,200		69,300		2000	5-year prevalence
Canada		2,368	14.6	735	4.5	4,323	26.6	1,175	7.2	2005	2-year prevalence
		4,791	29.5	1,672	10.3	8,976	55.3	2,136	13.2	2005	5-year prevalence
		7,514	46.3	3,100	19.1	14,303	88.1	2,776	17.1	2005	10-year prevalence
England	NCIN	1,533	4.8	541	2.0	2,890	8.8	1,050	2.9	2006	1-year prevalence
Scotland		138	4.4	60	2.3	317	9.0	108	2.6	2006	
Wales		123	5.8	27	1.9	203	9.4	78	3.3	2006	
Northern Ireland		53	5.4	17	1.9	87	8.5	58	4.8	2006	
United Kingdom		1,847	4.9	645	2.0	3,497	8.8	1,294	3.0	2006	

Table 2-5 continued

Country/ Area	Project	Leukaemia		HL		NHL		Myeloma		Index date	Note
		Counts	Rates*	Counts	Rates*	Counts	Rates*	Counts	Rates*		
England	NCIN	5,792	18.7	2,236	8.5	11,309	34.4	3,404	9.6	2006	5-year prevalence
Scotland		659	20.2	265	9.9	1,248	35.5	383	9.4	2006	
Wales		432	21.0	114	9.2	722	34.9	242	10.0	2006	
Northern Ireland		175	17.6	78	8.6	395	37.6	146	13.0	2006	
United Kingdom		7,058	18.9	2,693	8.6	13,674	34.6	4,175	9.7	2006	
England	NCIN	9,353	30.2	4,133	15.7	18,023	54.7	4,521	12.8	2006	10-year prevalence
Scotland		1,143	34.3	479	17.9	2,006	56.4	506	12.5	2006	
Wales		670	33.3	218	14.4	1,078	52.0	324	13.4	2006	
Northern Ireland		264	26.4	129	14.5	661	63.1	193	17.3	2006	
United Kingdom		11,430	30.6	4,959	16.1	21,768	54.9	5,544	12.9	2006	
England	NCIN	13,072	42.6	7,127	26.8	24,010	92.9	5,065	14.2	2006	20-year prevalence
Scotland		1,566	47.7	786	28.8	2,668	74.7	564	13.9	2006	
Italy	ITACAN	6,479	78.0	5,305	64.0	14,360	173.0	3,162	38.0	2006	Total prevalence
Texas	TCR	4,235		2,147		8,623		1,546		2006	10-year prevalence
Australia	AIHW	-	-	3,291	30.8	14,099	132.0	2,415	22.6	2007	1982-2007
Worldwide	GLOBOCAN	221,120	9.0	81,808	3.3	344,983	14.0	98,276	4.0	2008	5-year prevalence

Table 2-5 continued

Country/ Area	Project	Leukaemia		HL		NHL		Myeloma		Index date	Note
		Counts	Rates*	Counts	Rates*	Counts	Rates*	Counts	Rates*		
Canada		2,971		783		4,864		1,324		2009	2-year prevalence
		6,118		1,805		10,704		2,506		2009	5-year prevalence
		9,470		3,271		17,082		3,358		2009	10-year prevalence
United States	SEER	125,312		88,038		242,587		35,432		2010	Total prevalence
Nordic countries	NORDCAN	10,568	82.0	5,050	39.2	17,323	134.5	3,258	25.3	2010	1980-2010
Norway		1,894	77.0	916	37.2	3,358	136.5	697	28.3	2010	1973-2010
Sweden		4,193	88.7	1,670	35.3	5,714	120.9	1,288	27.3	2010	1980-2010
Denmark		2,419	86.0	1,149	40.9	3,874	137.8	647	23.0	2011	1963-2011
Finland		2,093	76.1	1,299	47.3	4,571	166.3	606	22.0	2011	1973-2011
Iceland		109	68.5	67	42.1	174	109.3	40	25.1	2011	1975-2011

*per 100, 0000

Abbreviations in the table:
HL: Hodgkin lymphoma, NHL: non-Hodgkin lymphoma, SEER: the Surveillance, Epidemiology, and Ends Results, IARC: International Agency for Research on Cancer, SSA: sub-Saharan Africa, MENA: Middle East and Northern Africa, EU: European Union, EEA: European Economic Area, LAC: Latin America and Caribbean, OEA: Korea, Mongolia, and Hong Kong, NCIN: National Cancer Intelligence Network, TCR: Texas Cancer Registry, AIHW: Australian Institute of Health and Welfare, ASP: age standardized proportion, NA: not available.

2.3.3The reported prevalence figures in the literature vary according to geography, time, and method of calculation

2.3.3.1 Geographic variability

Developed countries show higher prevalence (Table 2-4 and Table 2-5) than developing countries (Pisani, Bray and Parkin, 2002). For example, for males, according to GLOBOCAN reports in 1990, the 1 year prevalence of leukemia in developed countries was more than four times that of developing countries; the highest prevalence figure which appeared in America was 6.2 times greater than that in sub- Saharan Africa (SSA), with a prevalence of leukemia of 9.4 per 100, 000 and 1.4 per 100, 000 respectively. This may be because people in developed countries generally enjoy a higher standard of living and the life expectancy is higher when compared with developing countries (Lutz, et al., 2003). The reason may also lie in the relatively poor registration in developing countries (Parkin, 2006). Figure 2-8 indicates the percentage of population covered by cancer registries; 83% in North America and 32% in Europe, compared with only 6% in Central and South America, 4% in Asia and 1% in Africa (IARC, 2013a). Furthermore, not all of these cancer registries can produce data of a sufficiently high quality to provide accurate and unbiased estimates. Although there are large cancer problems in low and middle-income countries, Asia, the Middle East, North and Sub-Saharan Africa, and Central and South America, there still remains a lack of high-quality population-based cancer registries (Curado, et al., 2007; IARC, 2013a).

In addition, geographic heterogeneity of cancer prevalence may be influenced by different age structures in populations. Age standardized 5-year prevalence in some developed countries are calculated to make comparisons (Crocetti, et al., 2013). After age adjustment, Italy showed the highest 5-year prevalence of Hodgkin lymphoma and myeloma out of the countries of U.S., Italy, Australia, France, and the Nordic European countries, whilst the U.S. showed the highest 5-year prevalence of non-Hodgkin lymphoma and leukaemia. The population was

younger in the U.S. compared with the Nordic European countries, whilst Italy had an older population (Crocetti, et al., 2013). This may explain why higher total prevalence of myeloma (which is primarily a disease of later adulthood; see Appendix A5) emerged in Italy in comparison with the Nordic European countries (Guzzinati, et al., 2012; Engholm, et al., 2013).

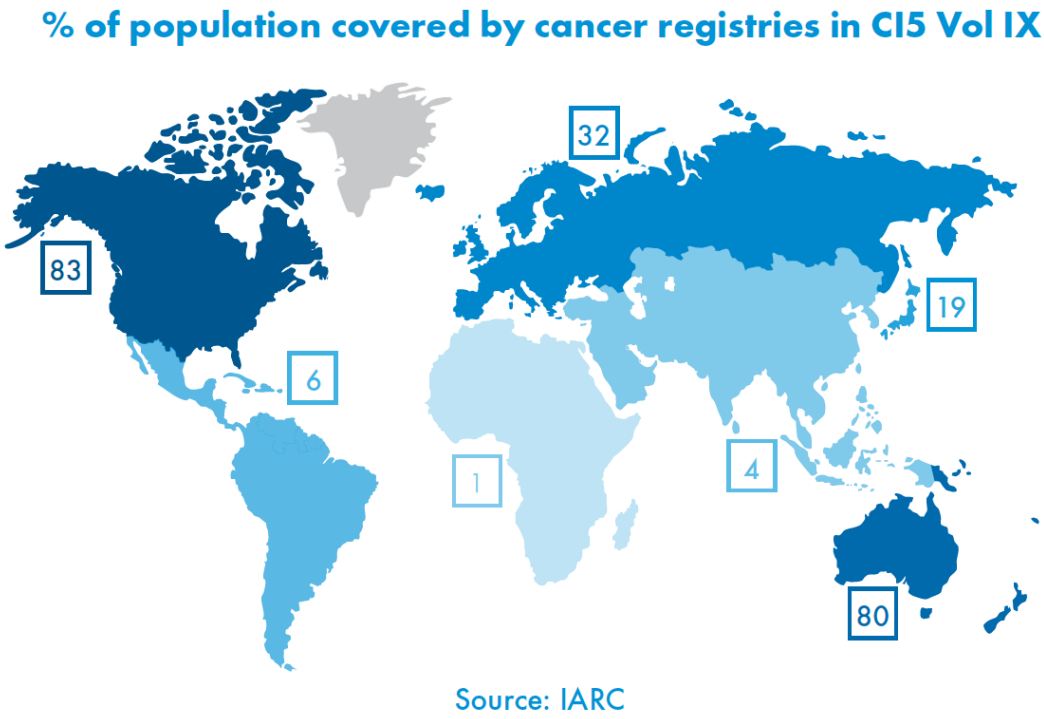


Figure 2- 8 Cancer registry coverage; the geographic coverage (per cent of total population) of cancer registries by region. (IARC, 2013a).

2.3.3.2 Increasing prevalence with calendar years

In general, the prevalence of haematological malignancies increases with increasing calendar year. For example, the prevalence of male leukemia in Denmark calculated in the EUROPREVAL project in 1992 was 47.2 per 100, 000, whilst by 2011, it had increased to 108.0 per 100, 000 as reported by the NORDCAN project (Møller, et al., 2003; Engholm, et al., 2013). Similarly, the total prevalence of male leukaemia in Italy increased from 42.6 per 100,000 in 1992 to 97.0 per 100,000 in 2006 (Verdecchia, et al., 2002; Guzzinati, et al., 2012). Several explanations exist for the marked increase in prevalence estimates

between the previous and more recent studies. Firstly, the survival prognosis tends to become better with increasing calendar year of diagnosis (Capocaccia and De Angelis, 1997). This could be linked to early diagnosis and the improvement of therapies. Secondly, improvements in data collection and reporting may result in more cases being recorded in cancer registries. Increasing prevalence may also be the result of other multiple factors: Population aging can augment the number of prevalent cases even with stable or decreasing incidence trends (Guzzinati, et al., 2012). Furthermore, increasing life expectancy or other reasons may result in increasing prevalence (see the discussion in Chapter Five).

2.4 Summary

After reviewing the literature, two problems appear: although some of the available methods in the literature can be used to calculate cancer prevalence, it is necessary to make more suitable method for haematological malignancies; in addition compared with other common cancers, there are fewer reports about prevalence of haematological malignancies because of its complexity in classification and difficulty in getting high quality data, as discussed in Chapter One. This study calculates prevalence of haematological malignancies based on the data from HMRN.

Chapter 3 Methodology

3.1 Database and materials

3.1.1 The Haematological Malignancy Research Network (HMRN)

The data for this study comes from the Haematological Malignancy Research Network (HMRN) (HMRN, 2011; Smith, et al., 2010). This section describes the area of study, data collection, and study period as follow:

3.1.1.1 Area of study

At the time of the inception of the study, cancer care in the UK was co-ordinated through a series of 34 area-based cancer networks: 28 cancer networks in England, three cancer networks in Wales, and three cancer networks in Scotland (see Appendix A1) (NHS, 2011). HMRN covers two adjacent UK Cancer Networks: Yorkshire, and Humber and Yorkshire Coast (Smith, et al., 2010), and a population of 3.6 million (Figure 3- 1).

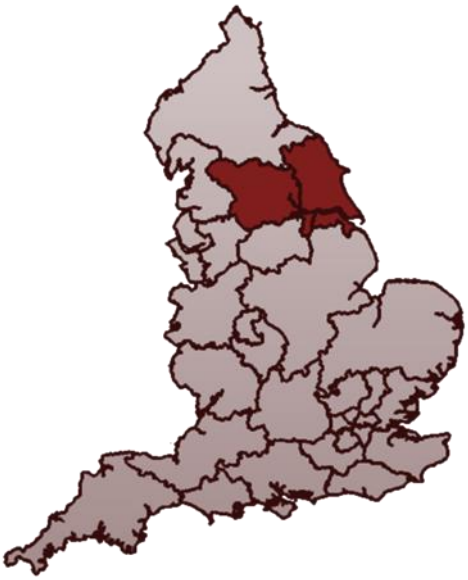


Figure 3- 1 Map of Cancer Networks in England and the HMRM region (shaded dark red) (HMRN, 2011)

In these two cancer networks, 14 hospitals provide clinical care to patients diagnosed with a haematological malignancy (Figure 3-2); each year around 2,000 patients are newly diagnosed (Smith, et al., 2010).

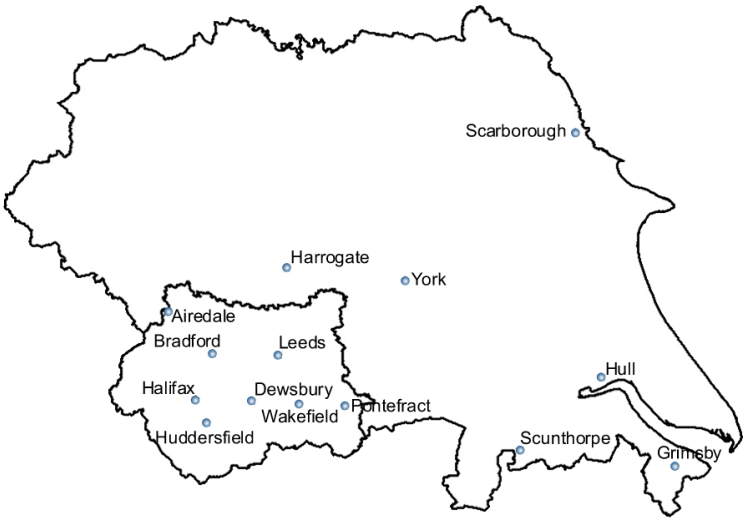


Figure 3- 2 14 hospitals in the Haematological Malignancy Research Network (HMRN) (HMRN, 2011)

3.1.1.2 Data collection

HMRN is a population-based registry (HMRN, 2011), and a collaboration between the Clinical Network, the specialist integrated diagnostic laboratory (Haematological Malignancy Diagnostic Service [HMDS] [HMDS, 2011]), and the Epidemiology and Cancer Statistics Group (ECSG), based at the University of York (HMRN, 2011; Smith, et al., 2010). The processes of case ascertainment and data collection are summarised in Figure 3-3, and are discussed in more detail in the next sections.

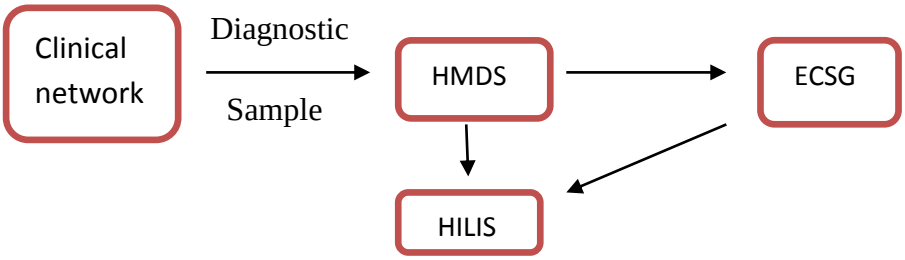


Figure 3- 3 Case ascertainment and data collection in the Haematological Malignancy Research Network (HMDS: Haematological Malignancy Diagnostic Service; ECSG: Epidemiology & Cancer Statistics Group; HILIS: HMDS Integrated Laboratory Information System)

The network provides the clinical care for patients diagnosed with a haematological malignancy. Patients' samples are sent to the centralized diagnostic laboratory HMDS and information is logged onto a bespoke web-based laboratory information system- HMDS Integrated Laboratory Information System (HILIS), which provides a tracking system for each patient (HMDS, 2011). In HMDS, diagnoses are made in a single department that contains all relevant expertise and technologies to provide an integrated diagnostic service including histology, cytology, immunophenotyping and molecular cytogenetics. All diagnoses are coded to current WHO classification (WHO, 2008).

The ECSG are responsible for collecting the detailed information of patients newly diagnosed with a haematological malignancy in the Network. In the ECSG, a list of newly diagnosed patients is downloaded on a weekly basis, and a group of trained nurses abstract clinical data from patients' medical records. They collect relevant information that includes demographic details, prognostic factors, and treatment and response to treatment for all patients. These data extracted by the ECSG are input into HILIS linking patients' diagnostic information with their clinical data (HMRN, 2011; Smith et al., 2010).

All HMRN patients are registered at the NHS Central Register and the date of death and the cause of death are updated monthly. This data along with gender, diagnosis, age at diagnosis, and date of diagnosis were downloaded from HILIS in order to estimate the prevalence of all haematological malignancies.

3.1.1.3 Study period

HMRN was established and began to collect information on newly diagnosed haematological malignancies patients on 1st, September 2004 (Smith, et al., 2010). Subjects diagnosed up to the 31st August 2011 had been flagged for death certification, so it is chosen as the index date. Therefore, all patients newly diagnosed between 1st, September 2004 and 31st, August 2011 were included in the estimation of prevalence.

3.1.2 Diagnostic subtypes

In HMRN, all diagnoses are coded to International Classification of Diseases for Oncology, 3rd Edition (ICD-O-3) (WHO, 2008). There are more than 60 ICD-O-3 codes in HMRN data from 2004 to 2011. Table 3-1 summarizes the diagnoses

with ICD-O-3, ICD-10, and lineage of diseases. It is shown on the basis of HMRN bridge-coded data. It is worth noting that not all ICD-O-3 codes have clear ICD -10 counterparts. To interpret the findings in this study, the bridge coding here may provide a reasonable approximation for conditions such as Hodgkin lymphoma, whilst for others, it may not (for example, T- cell leukaemia, hairy cell leukaemia, and chronic myelomonocytic leukaemia could be coded as leukaemia or other). Furthermore, conditions such as myeloproliferative neoplasms and myelodysplastic syndromes that are classified as in situ neoplasms in the ICD-10 are recognized as malignancies in the ICD-O-3 (Fritz, 2000).

It is not possible to analyse separately each subtype defined by the ICD- O- 3 separately, since there are too many entities and many of them are too rare to enable a robust estimation of prevalence. Therefore for estimation purposes, 21 main subtypes were used to estimate total prevalence: chronic myelogenous leukaemia, chronic myelomonocytic leukaemia, acute myeloid leukaemia, acute lymphoblastic leukaemia, chronic lymphocytic leukaemia, hairy cell leukaemia, T-cell leukaemia, marginal zone lymphoma, follicular lymphoma, mantle cell lymphoma, diffuse large B-cell lymphoma, Burkitt lymphoma, T-cell lymphoma, Hodgkin lymphoma, plasma cell myeloma, plasmacytoma, myelodysplastic syndromes, myeloproliferative neoplasms, monoclonal B-cell Lymphocytosis, monoclonal gammopathy of undetermined significance, and lymphoproliferative disorder not otherwise specified. The third column in Table 3-1 lists the subtypes used in this study. Although there may be diversities within one main subtype, it seemed the most reasonable way since sample size is an important factor in making estimations.

Table 3- 1 HMRN diagnoses with ICD-O-3, ICD-10, and lineage from 2004 to 2011

Broad Category	ICD-10 group	Main WHO groups	Diagnosis	ICD-O-3	Lineage
Leukaemia	Myeloid leukaemia (C92-C94)	Chronic myelogenous leukaemia	Chronic myelogenous leukaemia	9875/3	Myeloid
			Atypical chronic myeloid leukaemia	9876/3	Myeloid
		Chronic myelomonocytic leukaemia	Chronic myelomonocytic leukaemia	9945/3	Myeloid
			Juvenile chronic myelomonocytic leukaemia	9946/3	Myeloid
		Acute myeloid leukaemia	AML with inv(16)(p13;q22) or t(16;16)	9871/3	Myeloid
			AML NOS	9861/3	Myeloid
			AML - probable therapy related	9861/3	Myeloid
			AML NOS	9895/3	Myeloid
			APML t(15;17)(q22;q11-12)	9866/3	Myeloid
			AML t(8;21)(q22;q22)	9896/3	Myeloid
			AML with NPM mutation as sole abnormality	9861/3	Myeloid
			AML - probable therapy related	9920/3	Myeloid
			AML with MLL (11q23) rearrangement	9897/3	Myeloid
			Blastic plasmacytoid dendritic cell neoplasm	9727/3	Myeloid
	Lymphoid leukaemia (C91)	Acute lymphoblastic leukaemia	B-lymphoblastic leukaemia NOS	9811/3	Lymphoid
			B-lymphoblastic leukaemia with hyperdiploidy	9815/3	Lymphoid
			B-lymphoblastic leukaemia with t(12;21)	9814/3	Lymphoid
			B-lymphoblastic leukaemia with t(9;22)	9812/3	Lymphoid
			B-lymphoblastic leukaemia with MLL rearrangement	9813/3	Lymphoid
			B-lymphoblastic leukaemia with hypodiploidy	9816/3	Lymphoid
			Precursor T-lymphoblastic leukaemia	9837/3	Lymphoid
		Chronic lymphocytic leukaemia	B-cell chronic lymphocytic leukaemia	9823/3	Lymphoid
		Hairy cell leukaemia	Hairy cell leukaemia	9940/3	Lymphoid
		T-cell leukaemia	T-cell or NK cell large granular lymphocytosis	9831/3	Lymphoid
			T-cell prolymphocytic leukaemia	9834/3	Lymphoid

Table 3-1 Continued

Broad Category	ICD-10 group	Main WHO groups	Diagnosis	ICD-O-3	Lineage
Non-Hodgkin lymphoma	Non-Hodgkin lymphoma (C82–C85)	Marginal zone lymphoma	Systemic marginal zone lymphoma	9689/3	Lymphoid
			Extranodal marginal zone lymphoma	9699/3	Lymphoid
		Follicular lymphoma	Follicular lymphoma	9690/3	Lymphoid
			Follicular lymphoma with large cell transformation	9698/3	Lymphoid
		Mantle cell lymphoma	Mantle cell lymphoma	9673/3	Lymphoid
		Diffuse large B-cell lymphoma	Diffuse large B-cell lymphoma	9680/3	Lymphoid
			Plasmablastic large B-cell lymphoma	9735/3	Lymphoid
			T-cell/histiocyte-rich large B-cell lymphoma	9688/3	Lymphoid
			Mediastinal large B-cell lymphoma	9679/3	Lymphoid
			Diffuse large B-cell lymphoma	9596/3	Lymphoid
			Intravascular large B-cell lymphoma	9712/3	Lymphoid
		Burkitt lymphoma	Burkitt lymphoma	9687/3	Lymphoid
		T-cell lymphoma	Anaplastic large cell lymphoma of T/null type ALK+	9714/3	Lymphoid
			Mycosis fungoides	9700/3	Lymphoid
			Extranodal NK/T-cell lymphoma, nasal type	9719/3	Lymphoid
			Anaplastic large cell lymphoma of T/null type ALK-	9702/3	Lymphoid
			Peripheral T-cell lymphoma - common; unspecified	9702/3	Lymphoid
			Enteropathy-type T-cell lymphoma	9717/3	Lymphoid
			Angioimmunoblastic T-cell lymphoma	9705/3	Lymphoid
			Primary cutaneous CD30 positive T-cell lymphoproliferative disorder	9718/3	Lymphoid
			Sezary syndrome	9701/3	Lymphoid
			Anaplastic large cell lymphoma of T/null type	9714/3	Lymphoid
			Adult T-cell lymphoma/leukaemia (HTLV-1 positive)	9827/3	Lymphoid
Hodgkin lymphoma	Hodgkin's disease (C81)	Hodgkin Lymphoma	Mixed cellularity classical Hodgkin lymphoma	9652/3	Lymphoid
			Nodular sclerosis classical Hodgkin lymphoma	9663/3	Lymphoid
			Lymphocyte-rich classical Hodgkin lymphoma	9651/3	Lymphoid
			Nodular lymphocyte predominant Hodgkin lymphoma	9659/3	Lymphoid

Table 3-1 Continued

Broad Category	ICD-10 group	Main WHO group	Diagnosis	ICD-O-3	Lineage
Myeloma	Myeloma (C90)	Plasma cell myeloma	Plasma cell myeloma	9732/3	Lymphoid
		Plasmacytoma	Extraosseous plasmacytoma	9734/3	Lymphoid
			Solitary plasmacytoma of bone	9731/3	Lymphoid
Myelodysplastic syndromes	Myelodysplastic syndromes (D46)	Myelodysplastic syndromes	Refractory cytopenia with multilineage dysplasia	9985/3	Myeloid
			Refractory anaemia with ring sideroblasts	9982/3	Myeloid
			Refractory anaemia with excess blasts	9983/3	Myeloid
			Myelodysplastic syndrome (5q-)	9986/3	Myeloid
Other	Other neoplams of uncertain or unknown behaviour (D47)	Myeloproliferative neoplasms	Myeloproliferative neoplasm, unclassifiable	9960/3	Myeloid
			Chronic eosinophilic leukaemia	9964/3	Myeloid
			Systemic mastocytosis	9741/3	Myeloid
			Chronic myeloproliferative neoplasm with myelofibrosis	9961/3	Myeloid
			Myelodysplastic / Myeloproliferative neoplasms unclassifiable	9975/3	Myeloid
		Monoclonal B-cell Lymphocytosis	Monoclonal B-cell lymphocytosis (CLL phenotype)	9823/3	Lymphoid
		Monoclonal gammopathy of undetermined significance	Monoclonal gammopathy of undetermined significance	9765/1	Lymphoid
			Monoclonal gammopathy of undetermined significance	9769/1	Lymphoid
		Lymphoproliferative disorder NOS	Lymphoproliferative disorder NOS	9591/3	Lymphoid
			Lymphoproliferative disorder NOS	9823/3	Lymphoid

NOS: Not Otherwise Specified

3.1.3 Population in the study area

For the purpose of calculating prevalence, the population in the defined area is needed.

3.1.3.1 Population in the UK and HMRN

Population data were obtained from the 2001 UK census (Office for National Statistics, 2001), which was the most recent available when the study began. Census Area Statistics on the Web (CASWEB) (Office for National Statistics, 2001) provides online access to UK census aggregate data. It was developed by the Census Dissemination Unit (CDU), based within Mimas at the University of Manchester (Census Dissemination Unit, 2001).

Figure 3-4 is a diagram depicting the geographical structure of England; there are similar structures for Wales, Scotland and Northern Ireland. From this figure, it can be seen that the statistics are available from country level to Output Areas (OA). OAs are the base unit for census data releases, and allow for a finer resolution of data analysis due to their small size. OAs are based on postcodes, and were designed to have similar population sizes and be as socially homogenous as possible (Office for National Statistics, 2008). Therefore if we know the output area codes of HMRN area, the defined population can be abstracted from the census data.

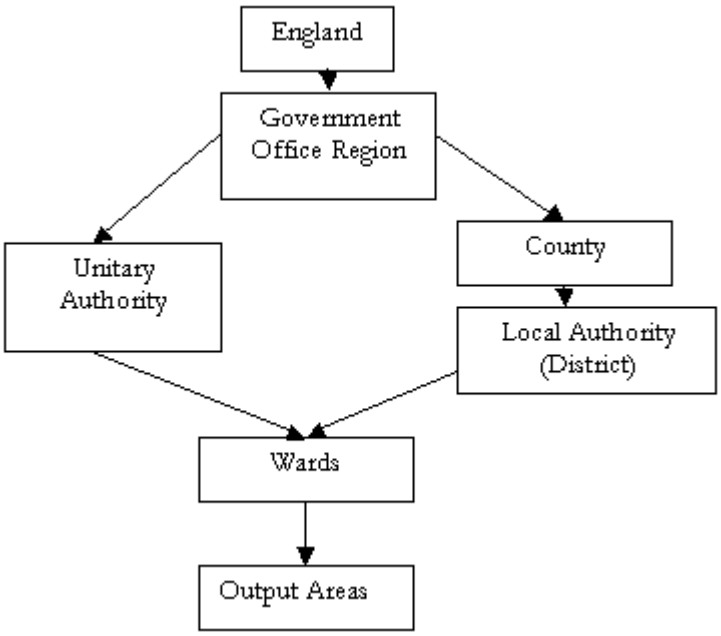


Figure 3- 4 The hierarchy of administrative areas in England for the 2001 Census. (There are 56 unitary authorities in England, and 27 shire counties split into 201 districts. Counties, districts and unitary authorities are subdivided into electoral wards) (Census Dissemination Unit, 2001)

HMRN covers two cancer networks, and the NHS postcode directory was used to identify which output area codes were in the two cancer networks (see Figure 3-5).

Population data were downloaded for England and then restricted to the two cancer networks. The detailed steps were:

(1) Downloaded cancer network codes (CANNET) and the corresponding Output area codes (OACODE) from the NHS postcode directory;

(2) Merged that information with the data of the 2001 census matching with Census Output Area Codes (population in every age group by cancer networks is then obtained);

(3) Only kept the population data with the cancer network codes N06 (Yorkshire Cancer Network) and N07 (Humber and Yorkshire Coast Cancer Network).

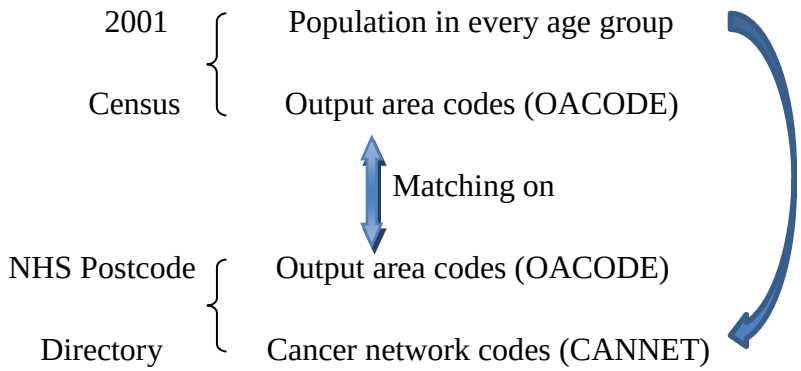


Figure 3- 5 Process identifying HMRN population

3.1.3.2 Comparing the population in HMRN area and in the UK

Populations that were obtained from the census are shown in Table 3-2. The peak of the population is in the age group 35- 39; males and females have broadly a similar age distribution. A slightly higher population of males are in the 0-4 category compared to females, however more females than males survive to reach old age.

Table 3- 2 Population in the UK and HMRN (from the 2001 census)

Age group	UK						HMRN					
(Years)	Male	%	Female	%	Total	%	Male	%	Female	%	Total	%
0-4	1784418	6.2	1703181	5.6	3487599	5.9	107160	6.2	104373	5.7	211533	5.9
5-9	1915964	6.7	1823598	6.0	3739562	6.4	119103	6.9	113668	6.2	232771	6.5
10-14	1987442	7.0	1891124	6.3	3878566	6.6	124558	7.2	120175	6.5	244733	6.9
15-19	1870485	6.5	1794571	5.9	3665056	6.2	117089	6.8	114423	6.2	231512	6.5
20-24	1766041	6.2	1780229	5.9	3546270	6.0	107301	6.2	110671	6.0	217972	6.1
25-29	1895216	6.6	1971412	6.5	3866628	6.6	108539	6.3	114524	6.2	223063	6.2
30-34	2199746	7.7	2293926	7.6	4493672	7.6	128167	7.4	134554	7.3	262721	7.4
35-39	2277756	8.0	2348442	7.8	4626198	7.9	133384	7.7	138410	7.5	271794	7.6
40-44	2056382	7.2	2095058	6.9	4151440	7.1	123962	7.2	125696	6.8	249658	7.0
45-49	1851535	6.5	1884582	6.2	3736117	6.4	113034	6.5	113814	6.2	226848	6.4
50-54	2003276	7.0	2037455	6.7	4040731	6.9	124644	7.2	125792	6.8	250436	7.0
55-59	1651372	5.8	1687710	5.6	3339082	5.7	99325	5.7	99606	5.4	198931	5.6
60-64	1409740	4.9	1470273	4.9	2880013	4.9	86445	5.0	90097	4.9	176542	4.9
65-69	1241343	4.3	1355789	4.5	2597132	4.4	75680	4.4	83945	4.6	159625	4.5
70-74	1058882	3.7	1280770	4.2	2339652	4.0	63721	3.7	79433	4.3	143154	4.0
75-79	817783	2.9	1149010	3.8	1966793	3.3	50210	2.9	70475	3.8	120685	3.4
over 80	793015	2.8	1644341	5.4	2437356	4.1	47593	2.8	101461	5.5	149054	4.2
Total	28580396	100.0	30211471	100.0	58791867	100.0	1729915	100.0	1841117	100.0	3571032	100.0

According to the 2001 UK census, the population of the UK was 59 million, with 3.6 million in HMRN area. Both share a similar age and sex distribution (see Figure 3-6); the bars on the population pyramid show the age and sex distribution for the UK, and the lines show the distribution of HMRN region. This means that the prevalence rate calculated using HMRN data could be generalised to the whole of the UK without age standardization. Indeed, it could be applied to any well-characterised population to estimate the number of prevalent cases with age adjustment, with assumptions (details are shown in Chapter Five).

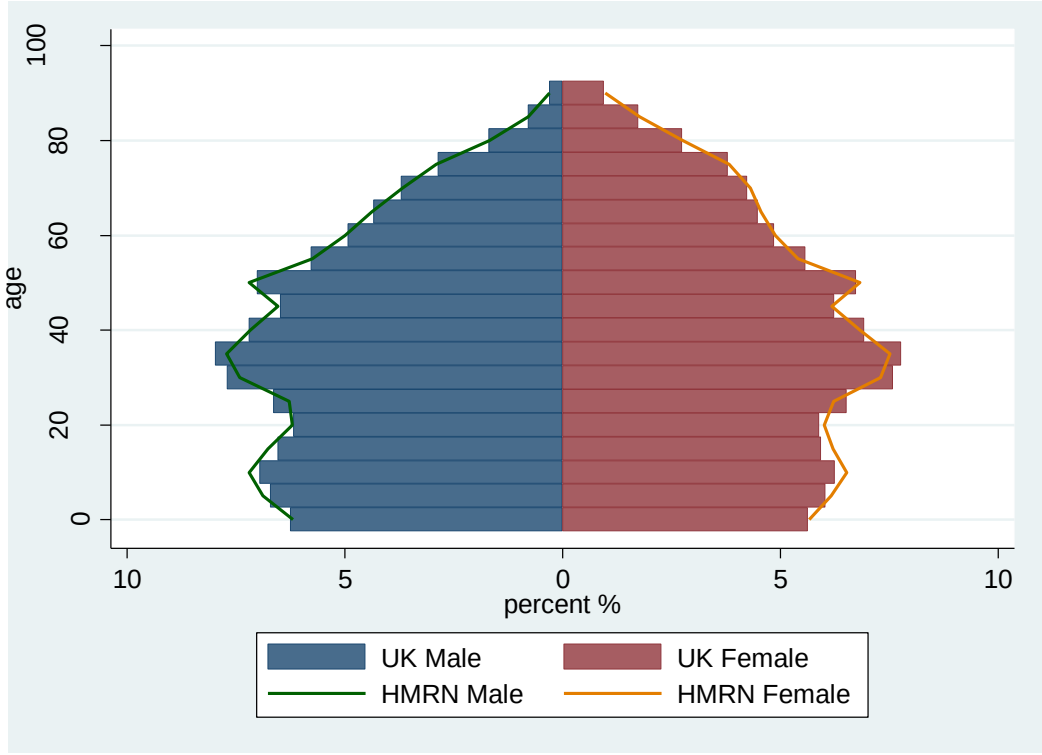


Figure 3- 6 Population age and sex structure of Haematological Malignancy Research Network (HMRN) region compared to the UK as a whole.

3.2 Descriptive statistics

Following the purpose of this study and the information available in the database, characteristics of the patients in HMRN are described first, including: diagnosis, gender, age at diagnosis, incidence and survival.

3.3 n-year prevalence

Although the main purpose of this work is to estimate total prevalence, it is necessary to first show the prevalence in the observed period. Furthermore, the observed prevalence calculated in this section is one of the steps in total prevalence estimation (details are shown in Section 3.4 and Chapter Five). Therefore n-year prevalence and observed prevalence calculation plays a transition role, and serves as a connecting link between the calculation from observed data and the estimation for the real disease burden.

HMRN includes newly diagnosed cases from 2004 to 2011. Figure 3-7 shows n-year prevalence and the corresponding calendar years.

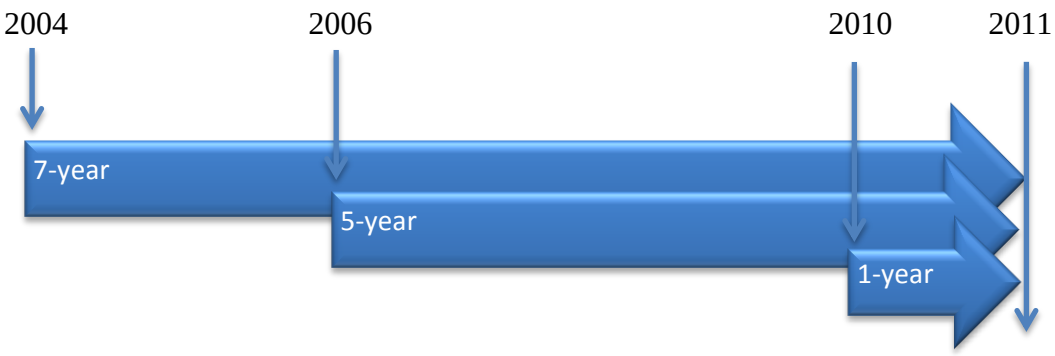


Figure 3- 7 n-year (1-year and 5-year) prevalence, observed prevalence (7-year prevalence) and the corresponding calendar years

3.3.1 1-year and 5-year prevalence

According to the definition of prevalence n-year prevalence can be calculated simply by counting the incidence cases that were still alive on a certain given date (31st, August, 2011) in the registry region, and then dividing by the population covered by HMRN.

$$prevalence = \frac{number\ of\ cases\ still\ alive\ on\ the\ index\ date}{total\ population} \quad (3.\ 1)$$

1-year prevalence counts the patients diagnosed within the most recent year before the index date (diagnosed between 1st, September 2010 and 31st, August 2011), and 5-year prevalence counts the patients diagnosed within recent five years before the index date (diagnosed between 1st, September 2006 and 31st, August 2011).

As described above, HMRN region population structure mirrors that of the UK as a whole in terms of age and sex. The number of prevalent cases of haematological malignancies for the UK could be estimated by applying HMRN prevalent rates to the UK population for both genders. The number of prevalent cases in the UK (N_{UK}) can be calculated by:

$$N_{UK} = P_{HMRN} * Population\ in\ the\ UK \quad (3.\ 2)$$

P_{HMRN} represents the prevalent rate in HMRN area.

3.3.2 Observed prevalence

Observed prevalence covers all patients diagnosed within the registry, and in fact is special n-year prevalence (n equals the length of registry). In this study, observed prevalence is 7-year prevalence. It is also extrapolated to estimate the number of prevalent cases in the UK.

3.3.3 Years of follow up

The maximum number of years of follow-up available is seven years, which may be long enough to show the burden for some subtypes with relatively short survival. For some subtypes with good prognosis, however, total prevalence is a more appropriate method to estimate prevalence. This section provides a visual representation of whether the length of the registry is sufficient or not to cover complete prevalent cases, and shows the necessity of estimating total prevalence.

The years of follow- up in HMRN may be sufficient for some diagnostic subtypes to show the disease burden. When the registry is long enough compared to the duration of the disease, the patients who are diagnosed before the start of registry and who died before the index date, do not contribute to prevalence. In this situation, the length of the registry will stop its effect on the observed prevalence, and observed prevalence will be stable if there is no change in incidence and survival.

The prevalence rate of each subtype can be calculated according to *n* (*n=1-7*) cumulative years before index date. Percentage changes in prevalence rate between successive years of cumulative prevalence (that is, n and n+1) within each subtype were calculated to indicate the number of years of follow- up required for complete prevalence. If we define prevalence as being sufficiently complete when the percentage change falls below 5%, this means that n+1-year

prevalence approaches to n-year prevalence, and the prevalence becomes stable after n years accumulated, therefore the length of the registry seems to be enough to show the burden of disease.

3.3.4 Move to "total prevalence"

Observed prevalence cannot show the real burden for some subtypes. Rather than showing separate haematological malignancies subtypes by survival, it is more convenient to develop a method to estimate the real burden for all subtypes either with short or long survival rates. Recalling the definitions in Chapter One, total prevalence (the expected complete prevalence) includes all those cases alive on a given date regardless of when they were diagnosed (those directly observed by a registry plus those that were diagnosed before the registry started). The characteristics of estimations of total prevalence should reflect the relationship between disease duration and registry length by itself. It goes without saying that diseases with shorter survival have a total prevalence approaching to observed prevalence, whilst diseases with longer survival have a total prevalence that is much higher than what can be observed in the data. Estimates of total prevalence, were calculated based on the mathematical relationships between prevalence, incidence, survival, and general mortality in the population. The method of calculation is presented in the next section.

3.4 Methods to estimate total prevalence

3.4.1 Definitions in the model

Cancer prevalence at a given time is the proportion of people in a population at a certain time diagnosed with cancer in the past and who are still alive. It may vary with calendar years, however in this study, it was assumed that it was constant with calendar years in calculation.

Consider the life history of an individual with cancer, whose life can be split into two parts according to the incidence of cancer: disease-free and survival with the disease (Figure 3-8).

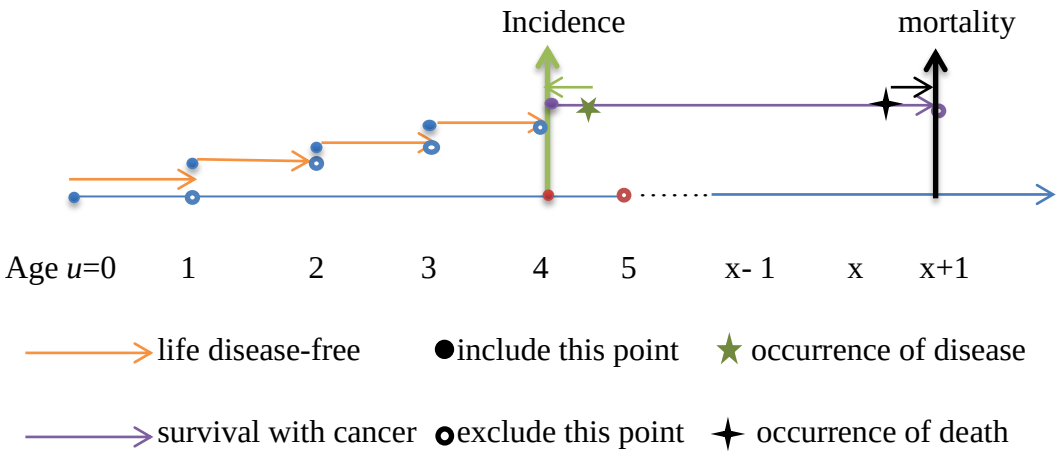


Figure 3- 8 the life of a patient (split into two parts according to the incidence of cancer: alive and disease- free, and survival with disease)

Ages are considered as discrete integer values in the calculation, however they are in fact continuous quantities. Therefore patients who die between their i^{th} and $(i+1)^{th}$ birthday are survival cases up to the end of their age interval $[i, i+1)$. Patients diagnosed between their i^{th} and $(i+1)^{th}$ birthdays (in the interval $[i, i+1)$) were incident cases on the i^{th} birthday. For example, suppose a patient diagnosed with cancer between his 4^{th} and 5^{th} birthday, and died between his 35^{th} and 36^{th} birthdays. In this calculation, the patient is considered as living disease-free up to the end of age interval $[3, 4)$, and becomes an incidence case on his 4^{th} birthday. After that, the patient was alive as a cancer patient up until the end of his age interval $[35, 36)$. This will be described as "incidence was at age 4" and "dies before age 36 (survival to age 35)".

3.4.5 Mathematical modelling of total prevalence

I. For the general population:

Let $G(u)$ be the general mortality in a reference year. Let u be an integer ($u \geq 0$), which is the age of death. This means that $G(u)$ is the conditional probability of an arbitrary person in the population dying between his u^{th} and $(u+1)^{st}$ birthdays, conditioned on surviving to his u^{th} birthday. $G(u)$ may vary with reference years, however in order to simplify the calculation, it is assumed that $G(u)$ is constant with the reference year chosen.

The probability of a person being alive at the end of their age interval $[0, 1)$ is $1 - G(0)$. The probability of a person being alive at the end of their age interval $[1, 2)$ using the definition of conditional probability is $(1 - G(0))(1 - G(1))$. Similarly the probability of a person surviving to the end of his age interval $[x, x+1)$ is

$$\prod_0^x (1 - G(u)) \qquad (3.\ 3)$$

II. For patients:

Suppose a patient was diagnosed at age t (t is an integer ($t \geq 0$)). Let $I(t)$ be the incidence at age t, which is the probability of a person being diagnosed with cancer in the age interval $[t, t+1)$. It was assumed that all cases between their t^{th} and $(t+1)^{th}$ birthdays were diagnosed on their t^{th} birthday. This means that the proportion of people diagnosed with disease in the age interval $[t, t+1)$ is considered as the estimated probability of people diagnosed on their t^{th} birthday $I(t)$. Recall from Figure 3- 8 that a patient's life can be divided into two parts.

In the first part, the case is disease free, which means there is no probability of transiting to death or disease. Let $G^*(u)$ be the non-disease mortality at age u, which means the conditional probability of death directly from disease free. (see Figure 3- 9). With the definition that incidence is considered as the beginning of an age interval, the probability of being considered disease free at the start of age interval $[0, 1)$ is $1 - I(0)$. With the definition that survival is considered up to the end of an age interval, the probability of being disease free at the end of age interval $[0, 1)$ is $\big(1 - I(0)\big) * (1 - G^*(0))$.

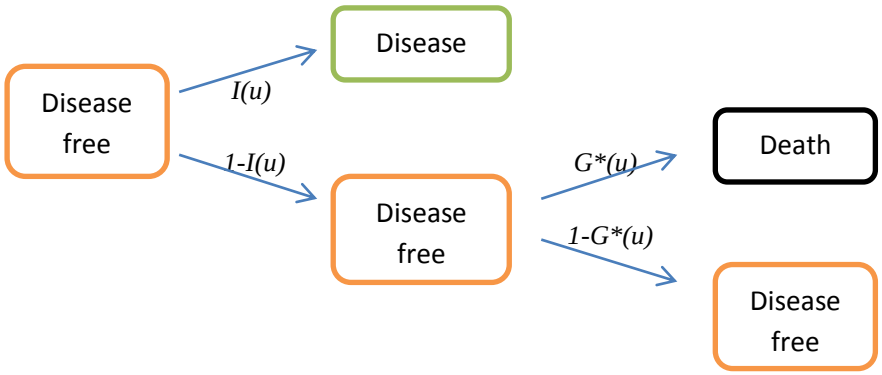


Figure 3- 9 The probability of being healthy at the end of an age interval.

Therefore, before the t^{th} birthday, $(t \geq 1)$, the probability at birth that an arbitrary person in the population will live disease free until age *t* is:

$$\prod_0^{t-1}\big(\big(1 - G^*(u)\big) * (1 - I(u))\big) \qquad (3.4)$$

In the second part of the patient's life, the person survives with cancer (Figure 3-10). Let $S(t, d)$ be the probability of a patient surviving *d* years after being diagnosed at age *t*. If the patient survives until age *x* $(x \geq t)$, then $d = x + 1 - t$. When $t = 0$, the probability that a person diagnosed with cancer at age 0 will survive with cancer until age *x* is:

$$I(0) * S(0, x+1) \qquad (3.5)$$

When $t \geq 1$ the probability that a person diagnosed with cancer at some age *t* will survive with cancer until age *x* is:

$$\sum_{t=1}^{x}\Big(\underbrace{\Big(\prod_{u=0}^{t-1}\big(\big(1-G^{*}(u)\big)*(1-I(u))\big)\Big)}_{1} * \underbrace{I(t)}_{2} * \underbrace{S(t, x+1-t)}_{3}\Big) \qquad (3.6)$$

Equation 3. 6 is comprised of two parts. The inner part of the equation is the probability that an arbitrary person in the population will be diagnosed at age *t* and will survive until *x*. This part describes the life of a person from being disease free (1) to the occurence of disease (2), then survives with the disease (3). The outer part sums this all up for the values of *t* to give the probability of some *x*. For a prevalent case at age *x*, the diagnosed age *t* can exist at any time between birth and age *x*, but can only appear once. Therefore all the possibilities of the value of *t* are mutually exclusive. The summation of the probabilities in the inner part of the equation makes the probability that a person diagnosed with cancer at some age *t* ($1 \leq t \leq x$) will survive with cancer until age *x*.

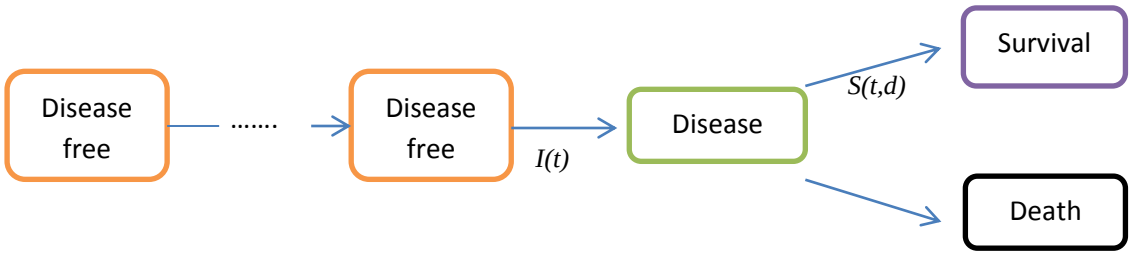


Figure 3- 10 The probability of a person diagnosed with cancer at age *t* surviving for *d* years

If the mortality rate of people who do not suffer from the disease $G^*(u)$ is approximated by the mortality rate of the general population $G(u)$ (see definitions of relative survival in Chapter One), the probability that a person diagnosed with cancer at some age *t* ($t \geq 1$) will survive with cancer until age *x* is:

$$\sum_{t=1}^{x}\left(\prod_{u=0}^{t-1}\big(1-G(u)\big)*\prod_{u=0}^{t-1}\big(1-I(u)\big)*I(t)*S(t,x+1-t)\right) \quad (3.\ 7)$$

Furthermore, in this study, the part $\prod_{0}^{t-1}\big(1-I(u)\big)$ contributes little to the results. This is because the diseases have an incidence rate of less than 10^{-4} (see incidence rates in Appendix 5). Considering the life span of a person is usually no more than 100 years, this makes:

$$\prod_{u=0}^{i}\big(1-I(u)\big)\approx 1\ (i\leq 100) \quad (3.\ 8)$$

This yields:

$$\sum_{t=1}^{x}\left(\prod_{u=0}^{t-1}\big(1-G(u)\big)*I(t)*S(t,x+1-t)\right) \quad (3.\ 9)$$

Therefore, the probability at birth that an arbitrary person diagnosed with cancer at age *t* and survives until age *x*, $P(x)$, can be divided into two parts: $t=0$ and $t\geq 1$. Both of them are calculated as a ratio with equation (3. 4) as the denominator

When $t=0$,

$$P(x|t=0) = \frac{I(0)*S(0,x+1)}{\prod_0^x (1-G(u))} \quad (3.10)$$

When $t \geq 1$,

$$P(x|t \geq 1) = \frac{\sum_{t=1}^{x}\left(\prod_{u=0}^{t-1}(1-G(u))*I(t)*S(t,x+1-t)\right)}{\prod_{u=0}^{x}(1-G(u))} \quad (3.11)$$

This can be simplified to:

$$P(x|t \geq 1) = \sum_{t=1}^{x}\left(\frac{I(t)*S(t,x+1-t)}{\prod_{u=t}^{x}(1-G(u))}\right) \quad (3.12)$$

Therefore the probability that an arbitrary person diagnosed with cancer at some age *t* will survive until age x, can be expressed as:

$$P(x) = \sum_{t=0}^{x}\left(\frac{I(t)*S(t,x+1-t)}{\prod_{u=t}^{x}(1-G(u))}\right) \quad (3.13)$$

Since $P(x)$ is the prevalence rate at age *x*, it is highly dependent on the model of incidence and survival. To make the results more closely to observed data, a method of estimating prevalence, from previous studies, is introduced in the next section. (Capocaccia and De Angelis 1997; Merrill, et al., 2000; Forman, et al., 2003; Gigli, et al., 2006; Simonetti, et al., 2008).

3.4.6 Completeness index of the observed prevalence

3.4.6.1 Model and definition of the completeness index

Suppose that a case is at age *x* on the index date, and the definition of age x is constant with that of survival. Patients alive at ages between their x^{th} and $(x+1)^{th}$ birthday on the index date are prevalent cases at age *x*, and are considered to live until the end of age interval *[x, x+1)*. For example, suppose a patient was diagnosed at age of 10, was at age 20 (between 20^{th} and 21^{st} birthday) on the index date. In the calculation, it is considered that he has been a survivor for 11 years $(20-10+1)$ up until the index date. Suppose that incidence is recorded on a registry for a time period of only L years, this means that a prevalent patient was at the age of $(x-L+1)$ when the registry started.

Here, the probability of a person being alive with cancer is used as the expected proportion of people being alive with cancer in the population. For ease of explanation, however, proportion was used for prevalence instead of probability in the following description.

Total prevalence in a population who are at age *x* on the index date can be separated into an observed part $P_o(x,L)$ and an unobserved part $P_u(x,L)$. The observed part derives from the incident cases observed between the age interval $[x-L+1, x]$, while the unobserved part refers to those cases diagnosed at previous a age and still living at *x* (see Figure 3- 11). That is:

$$P(x) = P_o(x,L) + P_u(x,L) \qquad (3.14)$$

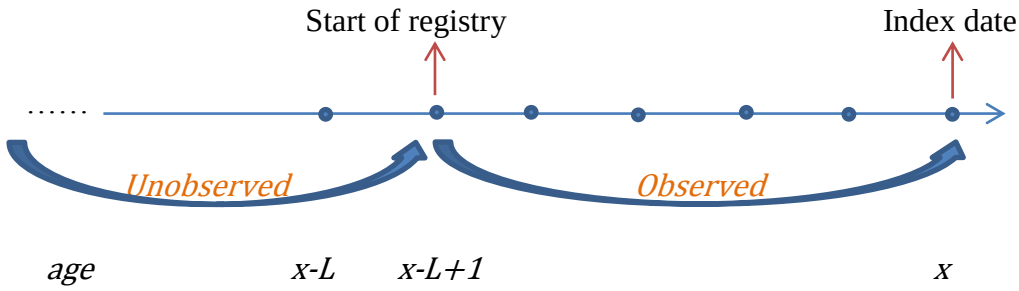


Figure 3- 11 Total prevalence can be separated into observed part and unobserved part

L is the time period of a registry; here it is seven years. The proportion of observed prevalence to the total prevalence is given by the ratio R:

$$R(x) = \frac{P_o(x,L)}{P(x)} = 1 - \frac{P_u(x,L)}{P(x)} \qquad (3.15)$$

R is called the **completeness index** and varies between 0 and 1. When all the prevalent cases have been diagnosed after the start of a registry, completeness index has the maximum value of 1. At the other extreme, the minimum value is 0 when all the prevalent cases were diagnosed before the beginning of the registry.

The completeness index R, is in turn used to inflate the observed prevalence to estimate the total prevalence:

$$N(x) = \frac{N_o(x,L)}{R} \qquad (3.16)$$

$N(x)$ is the number of prevalent cases at age x. $N_o(x,L)$ in function (3. 16) is the actual number of prevalent cases within the registry period based on the data, obtained using the direct method (counting the number of incident cases still alive

on the index date). The details have been described in Section 3.3.2. The total prevalence calculated in this way is closer to observed data, since it is estimated using a proportion of observed prevalence and total prevalence.

The completeness index R varies with age, and age-specific prevalence can be used to estimate the number of total prevalent cases for every age. For the prevalence of all ages on the index date, the algorithm is:

$$total\ prevalence = \frac{\sum total\ prevalent\ cases\ for\ an\ age}{total\ population} \qquad (3.17)$$

Similarly, R can be used to estimate the partial prevalence of a period longer than the observation period in the registry, for example, the 10-year and 20-year prevalence in the population observed only for 7 years (Capocaccia, et al., 2002).

3.4.6.2 How to calculate completeness index R

To get the completeness index, the probabilities $P(x)$ are involved in the equation 3.15. According to the equations in 3.4.5, $P(x)$ (observed part and unobserved part) can be expressed as:

$$\mathrm{P_u(x, L)} = \sum_{t=0}^{x-L} \frac{I(t)*S(t,x+1-t)}{\prod_{u=t}^{x}(1-G(u))} \qquad (3.18)$$

$$\mathrm{P_o(x, L)} = \sum_{t=x-L+1}^{x} \frac{I(t)*S(t,x+1-t)}{\prod_{u=t}^{x}(1-G(u))} \qquad (3.19)$$

The age-specific completeness index can be directly computed by means of the equation 3.20, when the incidence, survival, and general mortality functions are known:

$$R(x) = 1 - \frac{P_u(x,L)}{P(x)} = 1 - \frac{\sum_{t=0}^{x-L} \frac{I(t)*S(t,x+1-t)}{\prod_{u=t}^{x}(1-G(u))}}{\sum_{t=0}^{x} \frac{I(t)*S(t,x+1-t)}{\prod_{u=t}^{x}(1-G(u))}} \qquad (3.20)$$

Incidence may change over the years, and general mortality decreases with calendar year. The balance between these effects is difficult to predict in the calculation. For the purpose of simplicity of the model, for convenience it was assumed that those rates are constant with years. The details are described in the next section.

3.4.7 General mortality, incidence, and survival

General mortality and estimations of incidence and survival required for the total prevalence model are introduced in this section. These probabilities can be obtained either directly or by predictions. They are introduced one by one in the following sections. As the method is in discrete version, the effect factors (such as age and year) are truncated to integers (see Table 3- 3).

Table 3- 3 Probabilities used in estimating completeness index

	Mortality	Incidence	Survival
Data Source	Life table*	HMRN	HMRN
Effect factors	Age	Age	Age and Survival time
Available format	Discrete age	Age group	Continue age and duration
Use in the method	Direct use	Modelling	Modelling
Model	-	Non-parametric	Parametric

*See Appendix A2, and Section 3.4.7.1

An important assumption in the whole calculation is that general mortality, incidence, and survival probabilities are constant with calendar years. This means that the proportions and probabilities observed and estimated from the current information and data can be extrapolated to the years before the start of the registry. This assumption makes the estimates easier, and seems the most convenient way to calculate prevalence. This means that the calendar year component is omitted in the following models:

3.4.7.1 General mortality

General mortality is the fraction of the population of those living at the beginning of the age interval that died during the interval. It can be derived from certain data (number of deaths $D(u)$ in a year and population $T(u)$ at the beginning of a year, at age *u* by sex):

$$G(u) = \frac{D(u)}{T(u)} \qquad (3.21)$$

General mortality figures are obtained from life tables. They provide a summary of mortality for age and sex in a general population in an area. Life tables can be categorised further as either static or fluent life tables (Ederer, Axtell, and Cuter, 1961). A static life table shows the age-specific mortality rates at a given point time; this is also called a time-specific life table. However, if the observation time were longer, the mortality would change over calendar time. The fluent life table takes this factor into account and is referred to as a cohort life table. As stated earlier, to simplify the calculation for total prevalence, calendar year component is not considered in this model. In this case, general mortality should be consistent with the assumptions of incidence and survival. Thus, the static life table was chosen for the estimates and considered as being constant with years.

General mortality rates in this study were obtained from the London School of Hygiene & Tropical Medicine life tables (LSHTM, 2012). They provide general mortality from 1971-2009 for England and Wales for both genders, by one-year age stratum.

However, the index date (31st, August 2011) in this study is later than 2009, which is when this life table ended. In this situation, it is the standard practice to assume that the probabilities are the same as those most recently available (LSHTM, 2012).

The general mortality is obtained as a discrete version for every single age, therefore, estimates are not required and they can be introduced into the model directly.

3.4.7.2 Incidence

I. Model for incidence

For incidence, newly diagnosed cases at every age can be directly obtained from HMRN data. Theoretically, the incidence for every single age can be calculated and introduced into the model without estimations. However, for some subtypes, the number of cases is small. Incidence for every single age abstracted from data may be not identifiable. Therefore ages are grouped into every five years, and the corresponding incidence is:

$$Incidence = \frac{Incident\ cases\ in\ an\ age\ group\ per\ year}{population\ in\ corresponding\ age\ group} \qquad (3.22)$$

The exact rate values are taken to be the midpoint of the age group of possible values. For example, the incidence for age group 0-4 is considered as the certain incidence at age 2. Based on those scatters of age and incidence, the incidence model is built to calculate the estimated incidence of every single age. (Prevalent cases were aggregated into the corresponding age groups as used for incidence.)

Parametric incidence functions in the literature (Capocaccia and De Angelis 1997; Merrill, et al., 2000; Gigli, et al., 2006) could not provide a good fit for the data for some cancers. Haematological malignancies can occur at any age, and may have a different age distribution compared to other common cancers. To accommodate variation over age of a predictor's effect on incidence, a new model using regression splines was developed to model the incidence rate as a flexible function of age.

A spline is a function that is constructed piece-wise from polynomial functions. Cubic spline is a commonly used spline, which has linear, quadratic, and cubic terms. It makes a smooth curve composed of a linear combination of those terms:

$$I(t) = b_0 + b_1 t + b_2 t^2 + b_3 t^3 + \sum_{i=1}^{m} \beta_i (t - k_i)_+^3 \qquad (3.23)$$

In this cubic regression spline, *t* is age, whilst b_1, b_2, b_3, and β_i are coefficients; b_0 is a constant. In the function, $(t - k_i)_+$ are hinge functions $\max(0, (t - k_i))$, which equals 0 if $0 > (t - k_i)$, else $(t - k_i)$. In those hinge functions, k_i are called knots (k_i is the i^{th} knot). Usually, knots are equidistant, and we chose quartile knot sequence which disjoint the variable—age into equal intervals (Racine, 2011). Thus, there are five knots in total: three internal knots (*m* is the number of internal knots) and two end point knots.

$$I(t) = b_0 + b_1 t + b_2 t^2 + b_3 t^3 + \beta_1 (t - k_1)_+^3 + \beta_2 (t - k_2)_+^3 + \beta_3 (t - k_3)_+^3 \qquad (3.24)$$

This non-parametric method makes a smoothing curve, which is not sensitive to the assumptions made for a parametric incidence function. In this work, this flexible incidence function of age was used to calculate the total prevalence. The estimates of incidence by parametric and non-parametric methods are described in the next chapter.

II. Steps to predict incidence for every single age

As the method is in a discrete version, the algorithm requires estimation of the distribution of incidence by single year of age. However, as stated earlier, this could not be done directly. To predict incidence for every single age from data, we need to interpolate the rates specified per 5-year interval to 1-year age groups, using the spline method. There are five steps (see Figure 3- 12):

(1) Group the continuous diagnosis age into five years strata
(2) Calculate average incidence (7 years) for every 5- year age groups in HMRN (Figure 3-12-1)
(3) Plot the incidence with the midpoint age of every corresponding age group (Figure 3-12-2)
(4) Regression spline: 17 incidence value and 17 midpoint ages (Figure 3-12-3)
(5) Predict the incidence for every single age (Figure 3-12-4).

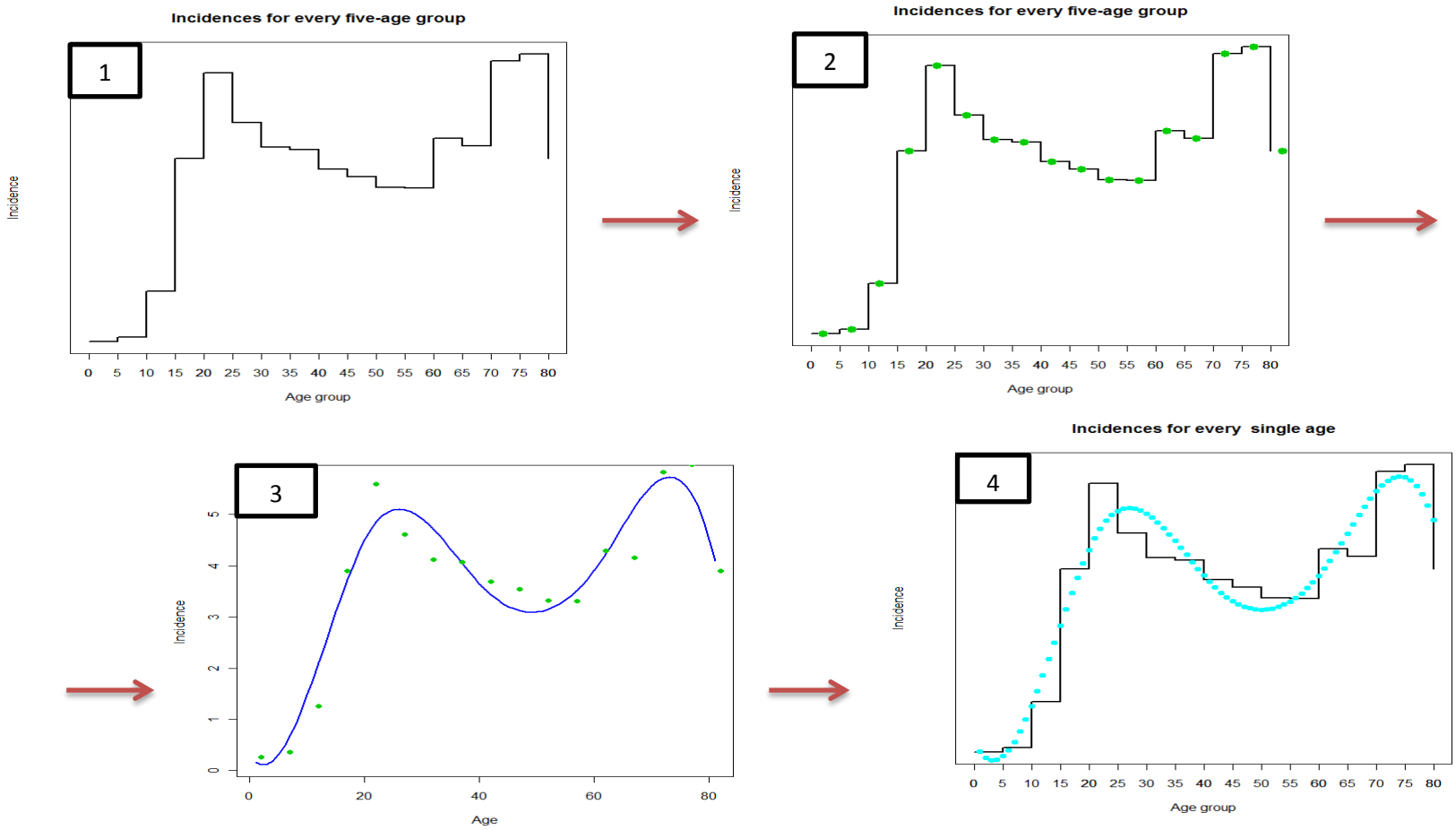


Figure 3- 12 Steps to predict incidence rate (per 100,000) for every single age

3.4.7.3 Survival

I. Model for survival

Unlike incidence, data on survival may not be adequate due to the limited length of the registry. The proportion of cases surviving decreases with time, but is unobserved when time since diagnosis becomes longer than the length of the registry (Capocaccia and De Angelis, 1997). With the data from HMRN, only seven years of observation can directly provide estimates of survival, and after that, model-based assumptions and estimations are required in predicting long-term survival proportions (Bray, et al., 2013). Unlike non-parametric approaches (such as the Kaplan-Meier survival analysis) to survival analysis, a certain distribution of survival time is assumed for parametric survival analysis. This makes the estimates for survival which are unobserved from data, follow a distribution. From this point of view, the parametric approach for survival analysis is more powerful in making estimates. Therefore a parametric model was used to estimate survival using the general Weibull distribution. Weibull function has previously been successfully applied to prevalence estimates (Capocaccia and De Angelis, 1997; Merrill, et al., 2000; Gigli, et al., 2006; Simonetti, et al., 2008). Unlike another common form used — exponential distribution that assumes constant hazard function with time, Weibull distribution assumes the hazard function will change monotonically over time. Exponential distribution can be considered as a special case of Weibull when the parameter that determines hazard rate trend equals 1. Due to this characteristic, Weibull distribution has broader application in research, and seems more suitable for survival analysis in this study (Golestan, et al., 2009).

Data from HMRN was used to fit the Weibull function and to estimate the survival pattern after seven years (the length of HMRN registry). According to

previous studies (Capocaccia and De Angelis, 1997; Merrill, et al., 2000; Gigli, et al., 2006; Simonetti, et al. 2008), survival is influenced by the age at diagnosis:

$$S(x-t,t) = \left[exp\left(-\lambda(x-t)^{\beta}\right)\right]^{exp(q*t)} \quad (3.25)$$

Where *t* is the age at diagnosis and *x* is the current age, then $(x-t)$ is the duration *d*. λ and β are scale and shape parameters of Weibull distribution. $exp(q*t)$ represents the effect of age at diagnosis, which means the relative risk of being diagnosed every one year older (Gigli, et al., 2006).

However, the log risk and age at diagnosis do not really have a linear relation. Some common cancers that the previous studies interested rarely occur at early age groups (such as lung cancer, stomach cancer, and colorectal cancer). It is reasonable to assume that survival deceases with age at diagnosis, and only calculate prevalence for adults (Merrill, et al., 2000). If a disease can occur at any age, the mortality may show different trends in childhood and adulthood. Indeed, for some subtypes of haematological malignancy, mortality decreases in younger age groups and increases in the old, for example with AML.

The effect of age on survival is difficult to be modeled for some subtypes of haematological malignancy, therefore a spline to model diagnosis age has been used.

$$S(x-t,t) = \left[exp\left(-\lambda(x-t)^{\beta}\right)\right]^{f(t)} \quad (3.26)$$

Equation (3.26) estimates survival probability under Weibull distribution. $f(t)$ represents a spline model of age at diagnosis. It allows for a smooth diagnosis age

effect. In this method, survival time is extrapolated using a parametric method, whilst the effect of age is a described using a spline method, which is flexible to capture the functional shape (Becher, et al., 2009).

The parametric method for survival is used for prevalence estimates in this study. It determined by both disease duration and age at diagnosis. The goodness of fit of survival models is checked, by comparing the Weibull estimated curves to Kaplan-Meier survival graphs. Survival was analyzed for males and females separately, since there are different survival figures for the two genders. However, for subtypes with a small number of cases, the survival analysis is not done separately for males and females in order to minimize the problems introduced by small numbers.

II. Steps to predict survival

Equation (3.26) provides survival for any continuous age and disease duration. To make predictions of survival in terms of integral age with integral years of duration that are involved in the discrete method, the survival probabilities at single ages and integral years are abstracted. The calculation process can be summarized into five steps (Figure 3-13):

(1) Obtain and format survival data from HMRN
(2) Fit a curve to data (7 years data) under Weibull distribution (Figure 3-13-1)
(3) Extrapolate the curve to estimate survival for longer disease duration (Figure 3-13-2)
(4) Introduce equation 3.26 to the method for prevalence estimation, and predict survival for every integral age with integral years of duration (Figure 3-13-3 shows an example at age 7).

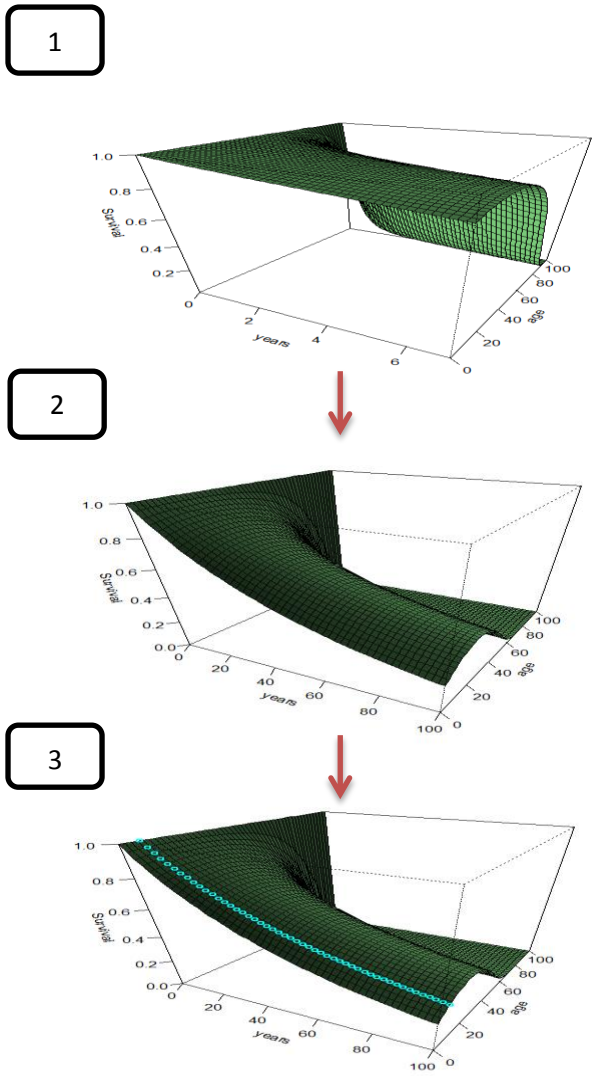


Figure 3- 13 Steps to predict survival for every integral age with integral years of duration

3.4.8.2 The process of calculation

In brief, total prevalence calculation has three main steps: (1) Calculate observed prevalent cases by age group. (2) Estimate completeness index, and apply it to observed prevalent cases by age group. (3) Estimate total prevalent cases by age group, and then calculate total prevalence rate for all ages together (see Figure 3-14). This section introduces a detailed calculation process as follows:

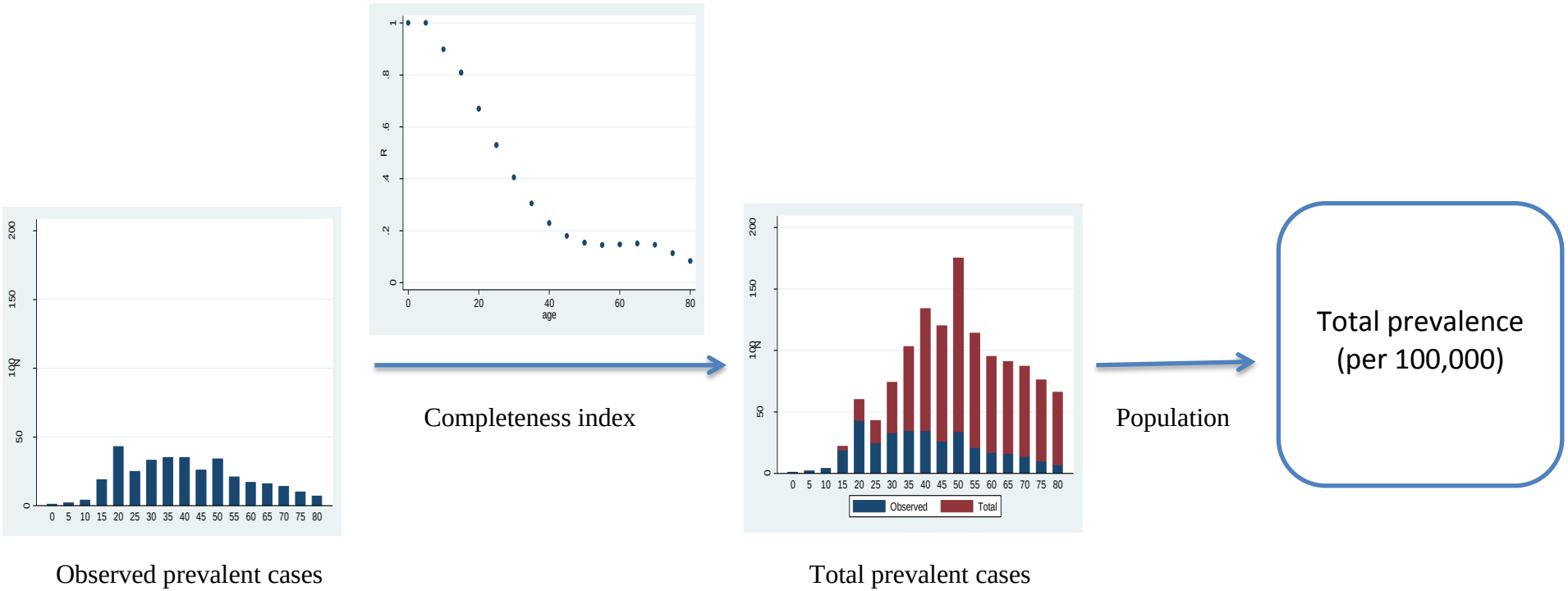


Figure 3- 14 Main steps for total prevalence calculation

Figure 3-15 shows the total prevalence calculation process using the method developed in this study. The basic data for the whole calculation are: HMRN data for all observed cases (HMRN, 2011), population from census (Office for National Statistics, 2001) , and general mortality from the London School of Hygiene and Tropical Medicine life table (LSHTM, 2012). Within the registry data, one can simply count the number of observed prevalent cases of every age group. We can also fit the data to a Weibull function and get parameters for survival. HMRN data combined with population in the area can provide the incidence for specific age groups. Thus, with parameters, it is possible to predict survival by given age and duration under Weibull function. A regression spline predicts incidence of every single age by smoothing the observed prevalence of every age group. Next, combined with data concerning general mortality, one can continue to get an age-specific completeness index. In order to keep the work coherent, the R-values of midpoint age in every age group are considered as the values for the specific whole age group, and then back- transformed to the calculation for the 5-year age groups. The number of observed prevalent cases from HMRN is divided by the R-value of the corresponding age group, and total prevalent cases for every age group are available until here. To get total prevalence, simply add up the total prevalent cases of every age group, then divide by the total population in the area.

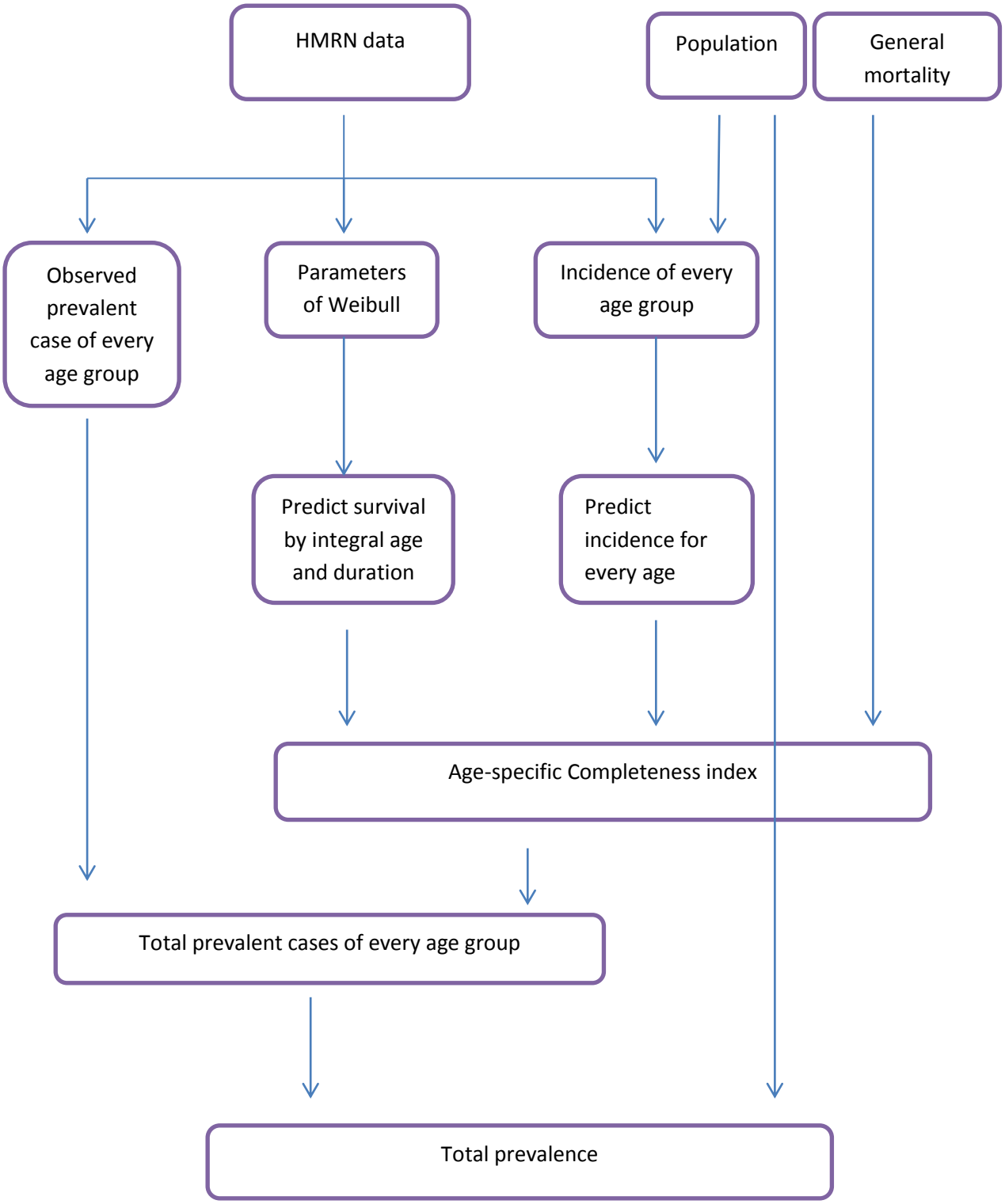


Figure 3- 15 Total prevalence calculation process using the method developed in this study

3.4.9 Validation and sensitivity analysis

The aim of this section is to validate the total prevalence estimated from HMRN. It should be noted that it is difficult to validate the method and results in this study, until the registry is long enough to cover all prevalent cases on the index date. The validation analysis was done from two aspects: (1) Check the goodness of fit, and (2) Predict the power of the method. These analyses identify whether the model is consistent with data, and whether the model has good predictive powers.

3.4.9.1 Goodness of fit

n-year prevalence (n=1, 2, 3, 4, 5, 6, 7) is estimated using the method in this study. It is then compared them to the actual n-year prevalence to check the goodness of fit.

Hodgkin lymphoma was chosen as an example for validation analysis, because it has uncommon age distribution on incidence and survival. In other words, it requires a more flexible method to fit the data than other common cancers. If the model in this work can provide suitable descriptions for Hodgkin lymphoma, it will be fine to make estimations for other subtypes with common distribution (such as monotone increasing incidence trend with age). There are also another two advantages to choose Hodgkin lymphoma. On the one hand, there is a relatively good sample size to support the estimations. On another, the prognosis of Hodgkin lymphoma is good (see Appendix A5) and so compared with subtypes with a poor survival rate, Hodgkin lymphoma can provide a better view of a trend.

3.4.9.2 Power to predict

According to the validation method used by Gigli et al. (2004), part of the data to estimate *L* –year prevalence is used, and then the estimated *L*-year prevalent cases are compared to the actual *L* -year prevalent cases.

The goodness of the total prevalence estimation for HMRN data is evaluated by comparing the observed 7-year prevalence with the estimated 7-year prevalence. The latter one is obtained by estimating total prevalence from recent five-year data (2006-2011), and then truncating the total prevalence to seven year prevalence (Gigli, et al., 2004):

$$N_7^{estimated}(x) = N_{total}(x) * R_7(x) = N_5^{observed}(x) * \frac{R_7(x)}{R_5(x)} \quad (3.27)$$

A plot of the estimated $N_7^{estimated}(x)$ and $N_7^{observed}(x)$ versus age highlights the difference between observed and estimated prevalence. This can help to identify whether the model fits the data well.

The difference between observed and estimated prevalent cases is calculated by:

$$\frac{|N_7^{observed} - N_7^{estiamted}|}{N_7^{observed}} * 100 \quad (3.28)$$

3.5 Subtypes where survival has changed greatly in the past

3.5.1 Method to estimate total prevalence range

For some diseases, the method in Section 3.4 may not accurately estimate total prevalence. This is because it assumes that the survival rate of a disease changes following the pattern observed over time in the data. However, the survival of some subtypes of haematological malignancy changed drastically in the past due to the introduction of new treatments. Therefore the survival pattern abstracted from a limited period of time cannot stand for the whole history of survival of the disease. In this study, the survival model was estimated using HMRN data from 2004 to 2011. If a new treatment were applied in clinical practice earlier than 2004, extrapolation of the survival trend to before 2004 would not reflect the poor survival before the new treatment was introduced.

Total prevalence range is a practical method to avoid the need for introducing another dataset to address the problems associated with changes in survival in the past. This method may be applied to some of the subtypes of haematological malignancies to complement the results estimated using the method in Section 3.4.

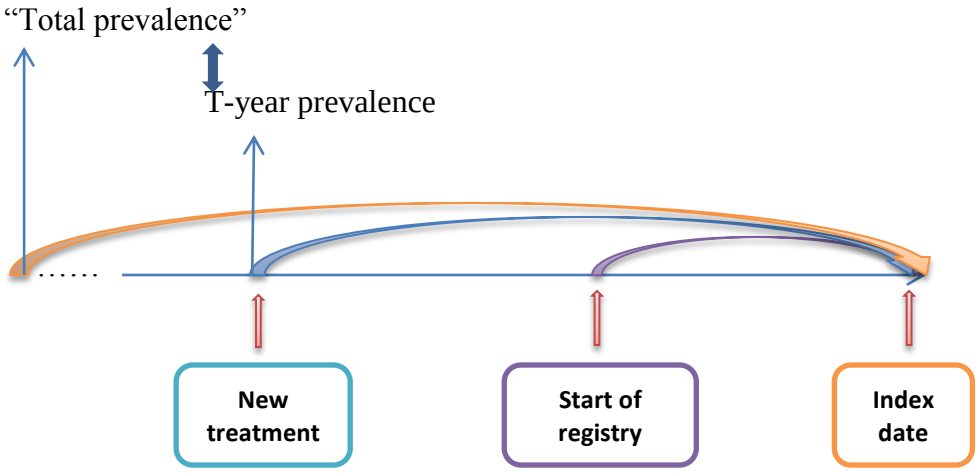


Figure 3- 16 Total prevalence range for a disease

Figure 3-16 shows the total prevalence range for a disease. Under the general method, "total prevalence" can be estimated based on observed data (from the start of the registry to the index date). However, as described above, it may be overestimated, since the survival rate may be much poorer before the application of new treatment. Observed data can only be extrapolated to the year when the new treatment applied. Prevalence for that period is called the T-year prevalence. Obviously, it is underestimated if T-year prevalence is considered as complete prevalence, because although the survival has previously been poorer, there still may be some cases alive on the index date. Therefore, a range is demonstrated with the "total prevalence" as an upper limit and T-year prevalence as a lower limit. The exact total prevalence cannot be estimated without bias, however the real complete prevalence must exist in this range (recall definitions about total prevalence and complete prevalence in Chapter One).

T-year prevalence is special n-year prevalence, when n equals the length of time (years) for which a new treatment (which improved survival greatly) is used on patients. It can be calculated by general method. The completeness index here is:

$$R_T = \frac{P_o(x,L)}{P_T(x)} \qquad (3.29)$$

Where $P_o(x, L)$ is observed prevalence in L years, and $P_T(x)$ is T-year prevalence $(L < T)$.

3.5.2 The choosing of "T"

In this method, data from HMRN is the only material used to make the estimates, except for the information used to choose "T", which is taken from the literature.

The choice of "T" is related to the calendar years of the application of the new treatments. 10-year prevalence for chronic myelogenous leukaemia (CML), 12-year prevalence for myeloma, 40-year prevalence for Hodgkin lymphoma, and 50-year prevalence for acute lymphocytic leukaemia (ALL) were calculated to show their total prevalence ranges with the "total prevalence" calculated using the method in section 3.4. The details of the chosen values of "T" for these conditions are described in Chapter Four.

3.5.3 The process of calculation for total prevalence range

Prevalence range is an easy way to show total prevalence and make suggestions for health resource allocation and survivorship planning. In this study, it is used to make estimations instead of trying to calculate an exact number for some subtypes. Thus in brief, there are two main steps in estimation: calculate "total prevalence", and calculate T-year prevalence (see Figure 3-17).

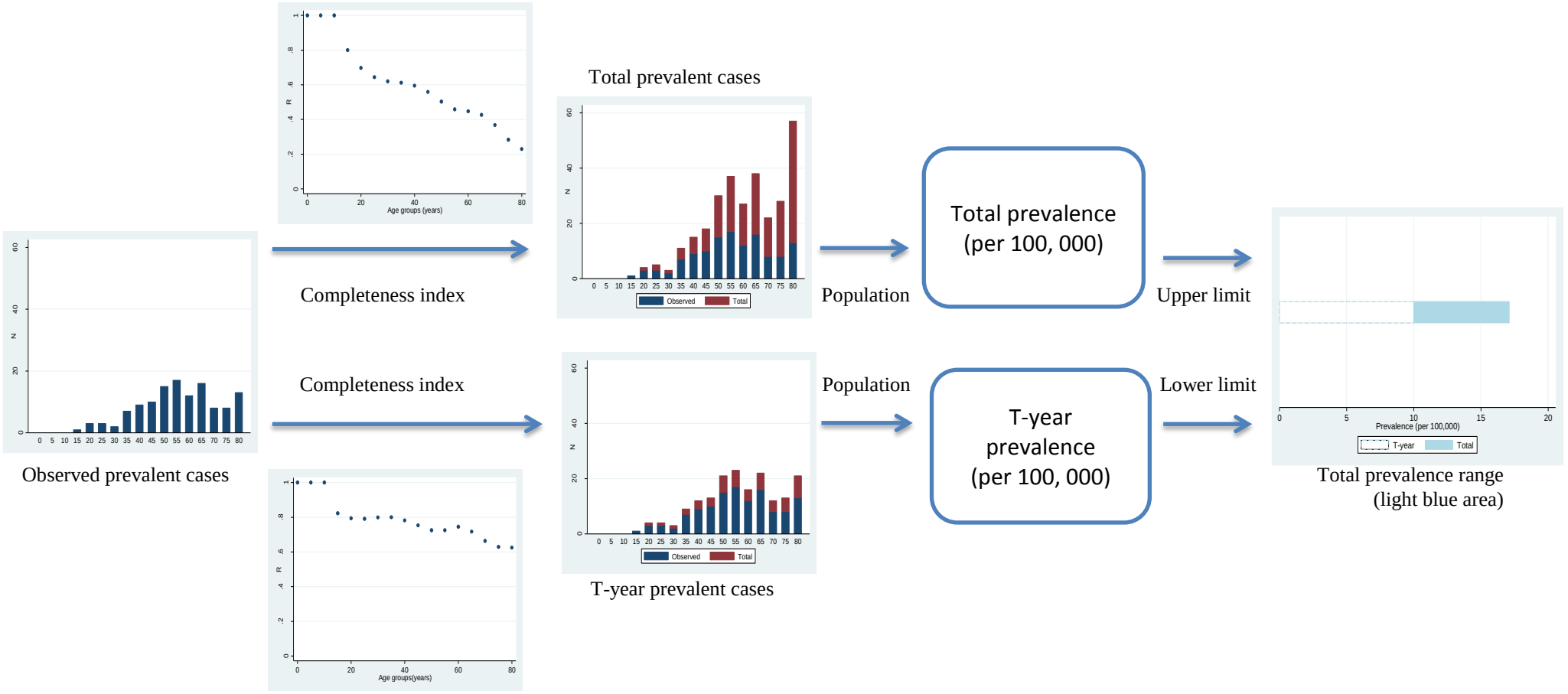


Figure 3- 17 Main steps for total prevalence calculation

3.6 Software

Data were obtained and formatted using Stata 11.0 software. The calculation for total prevalence was conduct using R 3.0.1 software. This included predicting incidence using regression splines, fitting data to Weibull function to find parameters, calculating observed prevalence, and calculating completeness index. Other mathematical calculations related to this work were implemented using Excel 2010.

R program for total prevalence calculation is an entire program (R codes are shown in Appendix A7). Observed prevalence, total prevalence and their ratios can be obtained directly by running this program for subtypes. However, it is necessary to show the full calculation progress to explain the method. Therefore calculations in this thesis have been done manually for some subtypes by way of example, whilst for other subtypes, automatic calculation using R software was used. The details of the results are shown in the next chapter.

Chapter 4 Results

4.1 Demographic characteristics

4.1.1 Diagnosis and gender

There were 15,810 diagnoses of haematological malignancies from 2004 to 2011 in HMRN, of which 8,799 were males (55.7%) and 7,011 were females (44.3%), (see Table 4- 1). The numbers of cases are shown in Figure 4-1, ordered by frequency. The most common subtype of haematological malignancies was diffuse large B-cell lymphoma (2,066 diagnosis), and the next most common one was chronic lymphocytic leukaemia (1,721 diagnosis).

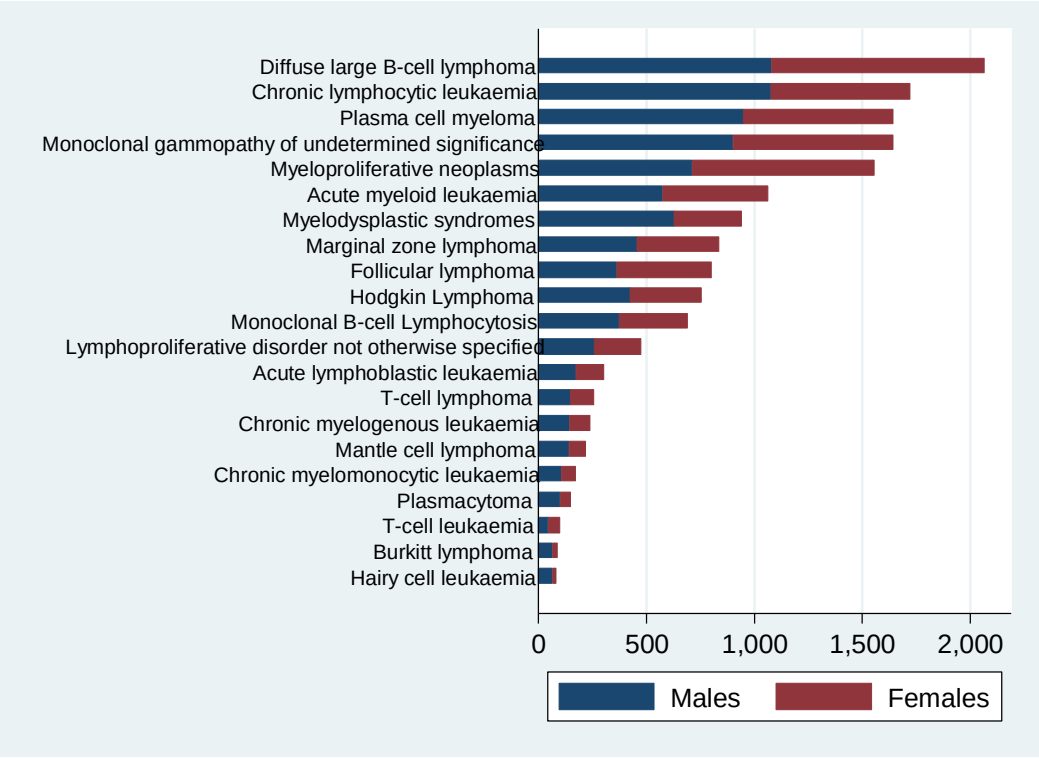


Figure 4- 1 Malignancy Research Network (HMRN), 2004-2011.

The proportions for both genders are shown in Figure 4-2 (in the order of the ratio of males and females). Males predominated for most subtypes, except T-cell leukaemia (46.0%), myeloproliferative neoplasms (45.8%), and follicular lymphoma (45.3%). This was most significant in the comparatively rare hairy cell leukaemia with 80.2% being male cases and 19.8% female cases. Some related conditions had similar proportions, for example, monoclonal gammopathy of undetermined significance and plasma cell myeloma were almost identical (males accounted for 54.9% and 57.7% respectively). However, others such as chronic lymphocytic leukaemia and monoclonal B-cell lymphocytosis showed different proportions (males accounted for 62.2% and 54.5% respectively). (Percentages for all subtypes were shown in Table 4-1). In fact, variations were also evident within some of these main subtypes. For example, for Hodgkin lymphoma, males accounted for 56.4% generally, but this ranged from 51.9% for nodular sclerosis classical Hodgkin lymphoma to 77.8% for lymphocyte-rich classical Hodgkin lymphoma.

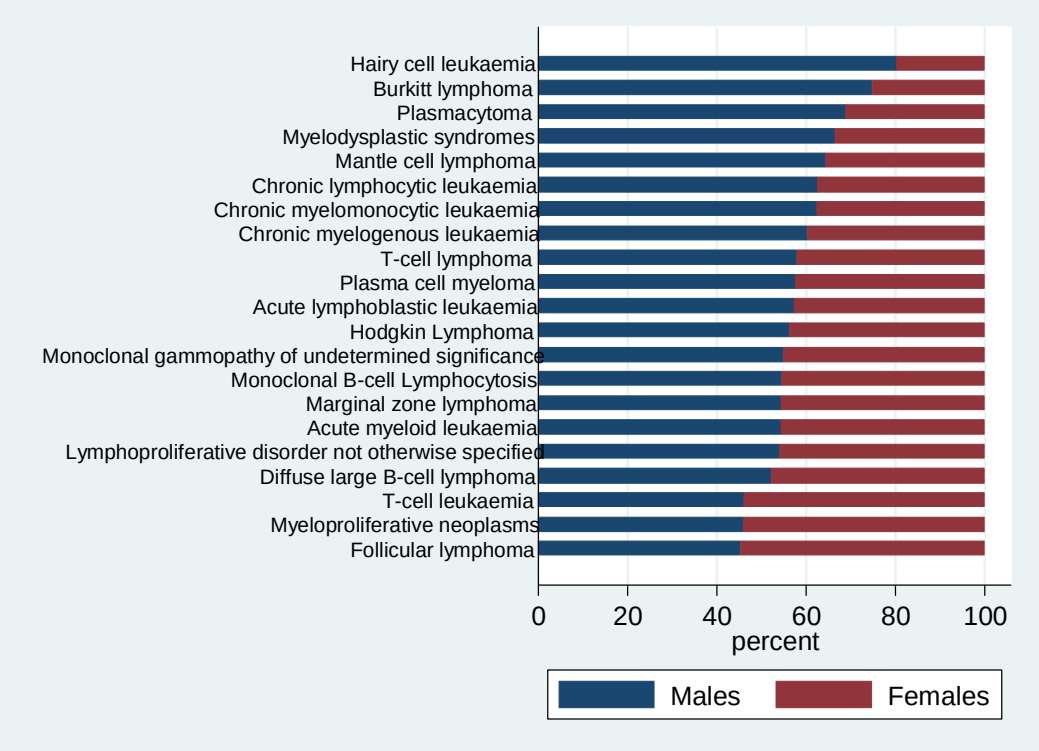


Figure 4- 2 Distribution by sex: The Haematological Malignancy Research Network (HMRN), 2004-2011

4.1.2 Age at diagnosis

Table 4-1 showed the number of cases, median age and age ranges for subtypes. Unlike many other common cancers, haematological malignancy can be diagnosed at any age and the range within the data was from one day to 100 years. Different subtypes dominated at different ages. The median age at diagnosis ranged from 15.3 years for acute lymphoblastic leukaemia to 77.3 years for chronic myelomonocytic leukaemia. Age similarities could be found within related conditions. Figure 4-3 showed similarities between precursor conditions and their more aggressive counterparts. For example, monoclonal B-cell lymphocytosis and chronic lymphocytic leukaemia had the same median ages at diagnosis of 71.6 year. Likewise, there were similar median ages at diagnosis for monoclonal gammopathy of undetermined significance and myeloma (72.6 and 73.1 years respectively).

Some subtypes, such as diffuse large B-cell lymphoma spanned the entire age range. It principally occurs at older ages, but sporadic cases arise at younger ages. Such wide age spans were not seen for all haematological malignancies. For example, monoclonal B-cell lymphocytosis, chronic lymphocytic leukaemia, mantle cell lymphoma, and myeloma seldom occurred below the age of 35. Variation could be found within some of the main subtypes. For example, acute myeloid leukaemia occurred at any age, but the median age of patients with MLL (11q23) rearrangement was 19.2 years, whilst the therapy– related acute myeloid leukaemia patients showed a median age at 73.0.

Although most subtypes had a median diagnostic age in old age (70.6 years for all haematological malignancies combined), some tended to be diagnosed at younger age, for example acute myeloid leukaemia, acute lymphoblastic leukaemia, Burkitt lymphoma, and Hodgkin lymphoma (see Figure 4-3). Paediatric cases may have significant effects on total prevalence estimates together with the

bimodal age distributions for Burkitt lymphoma and Hodgkin lymphoma. It was reasonable to suspect that they comprise several sub-subtypes with different features. However, for the purpose of estimation, the heterogeneities within the main subtypes were not considered, and the prevalence was only estimated for the main subtypes.

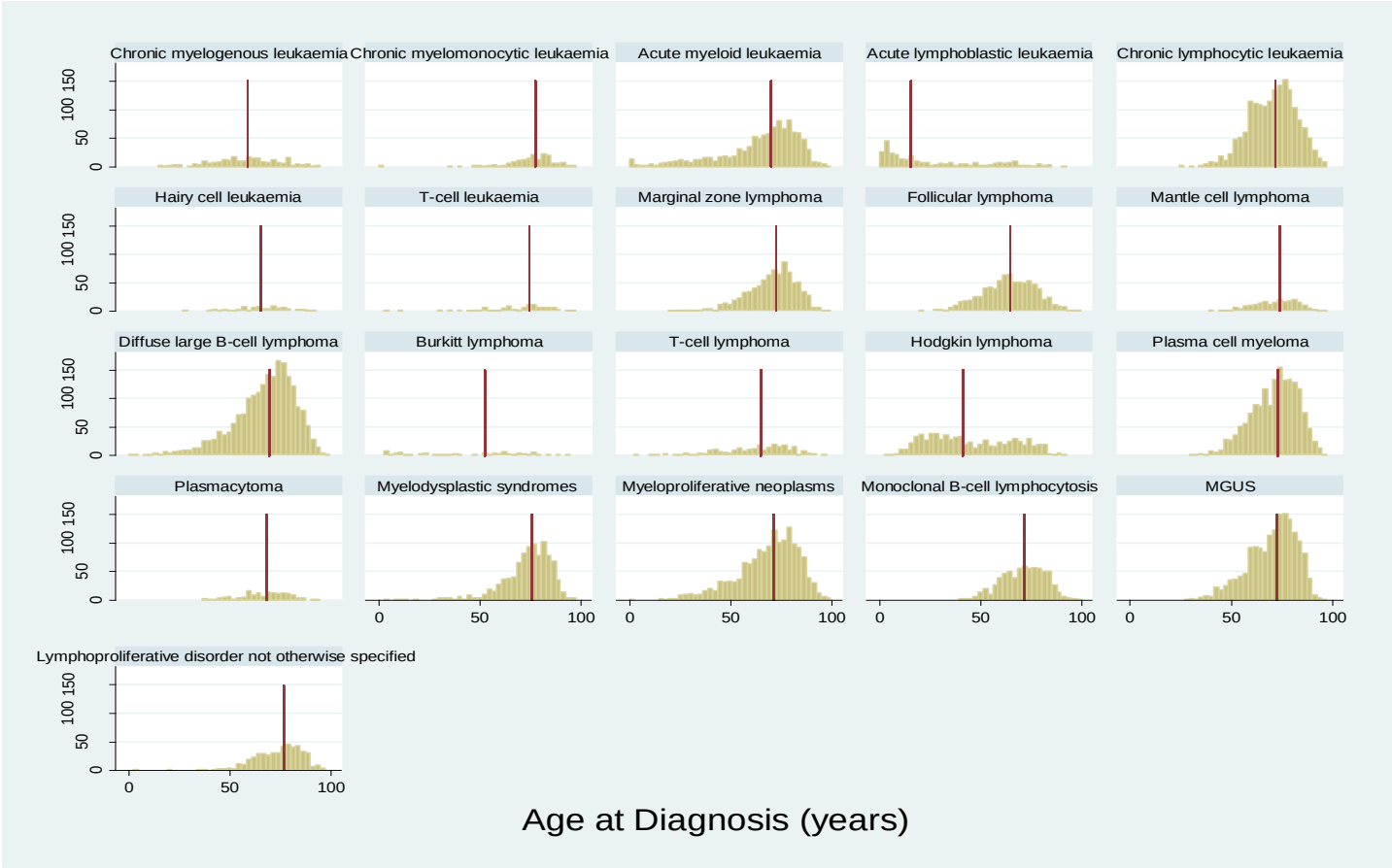


Figure 4- 3 Age (years) at diagnosis (with red lines indicating median ages), distributions for 2004-2011 (MGUS: monoclonal gammopathy of undetermined significance)

Table 4- 1 Demographic characteristics: The Haematological Malignancy Research Network (HMRN), 2004-2011

	Total		Male			Female		
	N	Median age (range)	N	%	Median age (range)	N	%	Median age (range)
Total	15,810	70.6 (0.003-99.7)	8,799	55.7	69.4 (0.003-99.7)	7,011	44.3	71.9 (0.05-99.0)
Leukaemia	3,683	69.3 (0.1-97.8)	2,193	59.5	67.8 (0.003-96.1)	1,490	40.5	71.5 (0.05-97.8)
Chronic myelogenous leukaemia	242	59.0 (15.1-94.7)	146	60.3	57.6 (15.7-94.7)	96	39.7	61.6 (15.1-92.6)
Chronic myelomonocytic leukaemia	173	77.2 (0.1-96.4)	108	62.4	76.4 (1.4-95.7)	65	37.6	78.5 (0.1-96.4)
Acute myeloid leukaemia	1,061	70.2 (0.2-97.8)	576	54.3	69.4 (0.2-94.3)	485	45.7	70.9 (0.2-97.8)
Acute lymphoblastic leukaemia	305	15.3 (0.003-90.5)	175	57.4	16.4 (0.003-84.6)	130	42.6	14.1 (0.05-90.5)
Chronic lymphocytic leukaemia	1,721	71.6 (25.0-97.2)	1,077	62.2	69.6 (25.0-96.1)	644	37.8	74.7 (26.1-97.2)
Hairy cell leukaemia	81	65.4 (28.9-90.9)	65	80.2	63.5 (28.9-88.5)	16	19.8	73.5 (46.9-90.9)
T-cell leukaemia	100	74.2 (3.4-95.2)	46	46.0	74.8 (3.4-94.0)	54	54.0	73.8 (30.7-95.2)
Non-Hodgkin lymphoma	4,271	69.0 (1.6-98.3)	2,254	52.8	67.9 (1.6-97.7)	2,017	47.2	70.5 (3.3-98.3)
Marginal zone lymphoma	839	72.4 (20.4-97.7)	456	54.4	71.3 (20.4-97.7)	383	45.6	73.6 (20.9-96.2)
Follicular lymphoma	804	64.6 (19.6-98.3)	364	45.3	62.9 (19.5-95.2)	440	54.7	65.9 (27.0-98.3)
Mantle cell lymphoma	219	73.9 (39.4-96.3)	141	64.4	71.2 (39.4-96.3)	78	35.6	75.6 (51.4-93.0)
Diffuse large B-cell lymphoma	2,066	69.8 (1.6-97.8)	1,080	52.3	68.2 (1.6-97.0)	986	47.7	71.4 (12.1-97.8)
Burkitt lymphoma	87	52.2 (3.1-95.6)	65	74.7	37.9 (3.1-88.2)	22	25.3	58.4 (3.3-93.4)
T-cell lymphoma	256	64.9 (2.9-95.6)	148	57.8	64.0 (2.9-91.1)	108	42.2	68.4 (3.7-95.6)
Hodgkin lymphoma	754	41.1 (3.6-90.9)	425	56.4	41.3 (3.6-88.0)	329	43.6	39.6 (9.4-90.9)
Myeloma	1,794	72.7 (30.6-95.5)	1,051	58.6	71.8 (30.6-94.4)	743	41.4	73.6 (36.0-95.5)
Plasma cell myeloma	1,646	73.1 (30.6-95.5)	949	57.7	72.2 (30.6-94.4)	697	42.3	73.8 (36.0-95.5)
Plasmacytoma	148	68.5 (36.6-94.5)	102	68.9	67.4 (36.6-87.3)	46	31.1	70.2 (38.7-94.5)
Myelodysplastic syndromes	944	75.6 (3.8-96.4)	627	66.4	75.6 (10.1-96.4)	317	33.6	75.6 (3.8-93.6)
Other Neoplasms of Uncertain or Unknown Behaviour	4,364	72.4 (1.8-99.7)	2,249	51.5	71.3 (1.8-99.7)	2,115	48.5	73.3 (4.0-99.0)
Myeloproliferative neoplasms	1,553	71.3 (1.8-99.7)	712	45.8	69.5 (1.8-99.7)	841	54.2	72.7 (17.0-99.0)
Monoclonal B-cell Lymphocytosis	690	71.6 (39.1-98.4)	376	54.5	70.9 (40.4-96.5)	314	45.5	72.9 (39.1-98.4)
Monoclonal gammopathy of undetermined significance	1,644	72.6 (27.7-95.7)	903	54.9	72.4 (27.7-94.3)	741	45.1	72.9 (29.9-95.7)
Lymphoproliferative disorder not otherwise specified	477	76.6 (4.1-96.6)	258	54.1	73.9 (21.2-96.3)	219	45.9	78.6 (4.1-96.6)

4.1.3 Incidence and survival

The incidence for all haematological malignancies combined was 63.2 per 100,000 per year. Subtypes showed different incidence: the incidence was as low as 0.3 per 100,000 for hairy cell leukaemia, and as high as 8.3 per 100,000 for diffuse large B-cell lymphoma. Fatal subtypes such as acute myeloid leukaemia showed a 5-year survival of 19.7%; in contrast, rare forms like hairy cell leukaemia and comparatively common subtypes like monoclonal B-cell lymphocytosis had 5-year survival estimates of 88.4% and 82.3% respectively. Table 4-2 shows grouped subtypes according to their incidence and 5-year survival rate combinations. Marginal zone lymphoma has a modest incidence (3.4 per 100, 000) and 5-year survival (62.5%). Mantle cell lymphoma may have lower prevalence estimates, due to its high mortality and low incidence. In contrast, myeloproliferative neoplasms with a relatively high incidence and survival, provides strong evidence for higher total prevalence estimates. The details of incidence and survival for every subtype were shown in Appendix A5. However, determination for total prevalence value cannot be made only based on Table 4-2, since incidence and survival varies with age. For example, although Hodgkin lymphoma and monoclonal B-cell lymphocytosis were grouped together with medium incidence and good survival, the higher incidence of Hodgkin lymphoma in childhood and young adulthood (see Figure 4-3) with good survival may result in much higher total prevalence estimates than monoclonal B-cell lymphocytosis that seldom occurs before 35 years old.

Table 4- 2 Subtypes considered in this study, according to their incidence and survival categories*

Incidence**	Survival**		
(per 100,000)	Poor (5-year prevalence<30%)	Medium (5-year survival 30-70%)	Good (5-year survival >70%)
Low (<2)	Chronic myelomonocytic leukaemia Mantle cell lymphoma	Acute lymphoblastic leukaemia T-cell leukaemia Burkitt lymphoma T-cell lymphoma Plasmacytoma Lymphoproliferative disorder not otherwise specified	Chronic myelogenous leukaemia Hairy cell leukaemia
Medium (2-5)	Acute myeloid leukaemia Myelodysplastic syndromes	Marginal zone lymphoma	Follicular lymphoma Hodgkin lymphoma Monoclonal B-cell lymphocytosis
High (>5)		Chronic lymphocytic leukaemia Diffuse large B-cell lymphoma Plasma cell myeloma	Myeloproliferative neoplasms Monoclonal gammopathy of undetermined significance

*Incidence and 5-year survival rates in HMRN from 2004 to 2011. Categories were made for this analysis only, and cannot be generalized to other diseases or other data

**Values of incidence rate and 5-year survival in the table can be found in Appendix A3

4.2 n-year prevalence

4.2.1 1-year, 5-year, and observed prevalence

The basic data of n-year prevalence are shown in Table 4-3 and Table 4-4, along with the calculation of the proportion of n-year prevalence over observed prevalence. For all haematological malignancies combined, there were 10,069 prevalent cases on the index date; 5,503 males and 4,566 females. Observed prevalence rate within the registry was 318.1 per 100,000 for males, and 248.0 per 100,000 for females.

Approximately 20% of observed prevalent cases were diagnosed in the last year, whilst about 80% were diagnosed in the last five years. These proportions varied with diagnostic subtypes. High proportions reflects that the diseases are frequently fatal; for example, 1- year prevalence of mantle cell lymphoma accounted for the largest proportion for both genders (31.3% and 36.7% respectively), whilst 5- year prevalence accounted for 94.0% in males, and for 96.7% in females. This implies that nearly all the alive patients were diagnosed in the last five years. On the other hand, those diseases with better survival, such as acute lymphoblastic leukaemia, accounted for a smaller proportion (19.7% for 1-year prevalence and 73.5% for 5-year prevalence).

Due to the similar age structure between the area covered by HMRN and the UK (see Chapter Three, section 3.1.3.3), prevalence in HMRN can be used to estimate the number of prevalent cases in the UK using equation 3.2. The number of n-year prevalent cases in the UK was shown in Table 4-5. For most subtypes, there were more prevalent cases in males than in females. For the two genders combined, it was estimated that about 35,679 of prevalent cases were diagnosed in the last year and still alive on the index date in the UK. The number of 5-year

prevalent cases was estimated to be 133, 565. Thus there were about 134 thousand persons living with haematological malignancies in 2011 who had been diagnosed within the last five years.

Figure 4-4 depicts the observed (7-year) prevalence counts (the two genders combined) in the UK by subtypes, and the proportion of cases surviving for one year and 5-year respectively. Within 165,841cases, chronic lymphocytic leukaemia was the most prevalent subtype, with 21,127 survivors on the index date diagnosed from 2004. The number of observed prevalence cases of myeloproliferative neoplasms ranked second, with similar 5-year prevalent cases to chronic lymphocytic leukaemia (about 17,000 cases each). Monoclonal gammopathy of undetermined significance, diffuse large B-cell lymphoma, and plasma cell myeloma ranked third to fifth. The five subtypes in combination were responsible for over half (56.5%) of the observed prevalence burden in the UK.

Table 4- 3 n-year prevalence rate per 100,000 population for males on 31st, August 2011 in HMRN

	Observed			1 year			5 years		
	N	Prevalence	%	N	Prevalence	%	N	Prevalence	%
Total	5,503	318.1	100.0	1,222	70.6	22.2	4,483	259.1	81.5
Leukaemia	1,326	76.7	100.0	314	18.2	23.7	1,063	61.4	80.2
Chronic myelogenous leukaemia	125	7.2	100.0	26	1.5	20.8	102	5.9	81.6
Chronic myelomonocytic leukaemia	36	2.1	100.0	10	0.6	27.8	34	2.0	94.4
Acute myeloid leukaemia	155	9.0	100.0	43	2.5	27.7	129	7.5	83.2
Acute lymphoblastic leukaemia	117	6.8	100.0	23	1.3	19.7	86	5.0	73.5
Chronic lymphocytic leukaemia	805	46.5	100.0	192	11.1	23.9	644	37.2	80.0
Hairy cell leukaemia	58	3.4	100.0	15	0.9	25.9	44	2.5	75.9
T-cell leukaemia	30	1.7	100.0	5	0.3	16.7	24	1.4	80.0
Non-Hodgkin lymphoma	1,408	81.4	100.0	309	17.9	21.9	1,145	66.2	81.3
Marginal zone lymphoma	328	19.0	100.0	70	4.0	21.3	275	15.9	83.8
Follicular lymphoma	304	17.6	100.0	58	3.4	19.1	244	14.1	80.3
Mantle cell lymphoma	67	3.9	100.0	21	1.2	31.3	63	3.6	94.0
Diffuse large B-cell lymphoma	596	34.5	100.0	135	7.8	22.7	481	27.8	80.7
Burkitt lymphoma	38	2.2	100.0	6	0.3	15.8	28	1.6	73.7
T-cell lymphoma	75	4.3	100.0	19	1.1	25.3	54	3.1	72.0
Hodgkin lymphoma	342	19.8	100.0	65	3.8	19.0	276	16.0	80.7
Myeloma	509	29.4	100.0	143	8.3	28.1	438	25.3	86.1
Plasma cell myeloma	445	25.7	100.0	124	7.2	27.9	382	22.1	85.8
Plasmacytoma	64	3.7	100.0	19	1.1	29.7	56	3.2	87.5
Myelodysplastic syndromes	214	12.4	100.0	55	3.2	25.7	192	11.1	89.7
Other Neoplasms of Uncertain or Unknown Behaviour	1,704	98.5	100.0	336	19.4	19.7	1,369	79.1	80.3
Myeloproliferative neoplasms	561	32.4	100.0	116	6.7	20.7	482	27.9	85.9
Monoclonal B-cell Lymphocytosis	315	18.2	100.0	54	3.1	17.1	243	14.0	77.1
Monoclonal gammopathy of undetermined significance	652	37.7	100.0	117	6.8	17.9	507	29.3	77.8
Lymphoproliferative disorder not otherwise specified	176	10.2	100.0	49	2.8	27.8	137	7.9	77.8

Table 4- 4 n-year prevalence rate per 100,000 population for females on 31st, August 2011 in HMRN

	Observed			1 year			5 years		
	N	Prevalence	%	N	Prevalence	%	N	Prevalence	%
Total	4,566	248.0	100.0	944	51.3	20.7	3,626	196.9	79.4
Leukaemia	849	46.1	100.0	186	10.1	21.9	685	37.2	80.7
Chronic myelogenous leukaemia	81	4.4	100.0	19	1.0	23.5	67	3.6	82.7
Chronic myelomonocytic leukaemia	28	1.5	100.0	7	0.4	25.0	25	1.4	89.3
Acute myeloid leukaemia	127	6.9	100.0	37	2.0	29.1	100	5.4	78.7
Acute lymphoblastic leukaemia	82	4.5	100.0	8	0.4	9.8	61	3.3	74.4
Chronic lymphocytic leukaemia	477	25.9	100.0	105	5.7	22.0	391	21.2	82.0
Hairy cell leukaemia	15	0.8	100.0	1	0.1	6.7	12	0.7	80.0
T-cell leukaemia	39	2.1	100.0	9	0.5	23.1	29	1.6	74.4
Non-Hodgkin lymphoma	1,259	68.4	100.0	283	15.4	22.5	1,004	54.5	79.7
Marginal zone lymphoma	282	15.3	100.0	70	3.8	24.8	238	12.9	84.4
Follicular lymphoma	355	19.3	100.0	63	3.4	17.7	264	14.3	74.4
Mantle cell lymphoma	30	1.6	100.0	11	0.6	36.7	29	1.6	96.7
Diffuse large B-cell lymphoma	528	28.7	100.0	121	6.6	22.9	425	23.1	80.5
Burkitt lymphoma	11	0.6	100.0	2	0.1	18.2	9	0.5	81.8
T-cell lymphoma	53	2.9	100.0	16	0.9	30.2	39	2.1	73.6
Hodgkin lymphoma	277	15.0	100.0	37	2.0	13.4	196	10.6	70.8
Myeloma	342	18.6	100.0	88	4.8	25.7	298	16.2	87.1
Plasma cell myeloma	316	17.2	100.0	82	4.5	25.9	278	15.1	88.0
Plasmacytoma	26	1.4	100.0	6	0.3	23.1	20	1.1	76.9
Myelodysplastic syndromes	126	6.8	100.0	33	1.8	26.2	116	6.3	92.1
Other Neoplasms of Uncertain or Unknown Behaviour	1,713	93.0	100.0	317	17.2	18.5	1,327	72.1	77.5
Myeloproliferative neoplasms	703	38.2	100.0	130	7.1	18.5	567	30.8	80.7
Monoclonal B-cell Lymphocytosis	275	14.9	100.0	39	2.1	14.2	207	11.2	75.3
Monoclonal gammopathy of undetermined significance	602	32.7	100.0	118	6.4	19.6	456	24.8	75.7
Lymphoproliferative disorder not otherwise specified	133	7.2	100.0	30	1.6	22.6	97	5.3	72.9

Table 4- 5 The number of n-year prevalent diagnoses of males and females in the UK on 31st, August,2011

	Total			Male			Female		
	Observed	1-year	5-year	Observed	1-year	5-year	Observed	1-year	5-year
Total	165,841	35,679	133,565	90,917	20,189	74,065	74,925	15,490	59,500
Leukaemia	35,839	8,240	28,802	21,907	5,188	17,562	13,932	3,052	11,240
Chronic myelogenous leukaemia	3,394	741	2,785	2,065	430	1,685	1,329	312	1,099
Chronic myelomonocytic leukaemia	1,054	280	972	595	165	562	459	115	410
Acute myeloid leukaemia	4,645	1,318	3,772	2,561	710	2,131	2,084	607	1,641
Acute lymphoblastic leukaemia	3,279	511	2,422	1,933	380	1,421	1,346	131	1,001
Chronic lymphocytic leukaemia	21,127	4,895	17,056	13,300	3,172	10,640	7,827	1,723	6,416
Hairy cell leukaemia	1,204	264	924	958	248	727	246	16	197
T-cell leukaemia	1,136	230	872	496	83	397	640	148	476
Non-Hodgkin lymphoma	43,921	9,749	35,392	23,262	5,105	18,917	20,659	4,644	16,475
Marginal zone lymphoma	10,046	2,305	8,449	5,419	1,156	4,543	4,627	1,149	3,905
Follicular lymphoma	10,848	1,992	8,363	5,022	958	4,031	5,825	1,034	4,332
Mantle cell lymphoma	1,599	527	1,517	1,107	347	1,041	492	181	476
Diffuse large B-cell lymphoma	18,511	4,216	14,921	9,847	2,230	7,947	8,664	1,986	6,974
Burkitt lymphoma	808	132	610	628	99	463	181	33	148
T-cell lymphoma	2,109	576	1,532	1,239	314	892	870	263	640
Hodgkin lymphoma	10,196	1,681	7,776	5,650	1,074	4,560	4,545	607	3,216
Myeloma	14,021	3,807	12,126	8,409	2,363	7,236	5,612	1,444	4,890
Plasma cell myeloma	12,537	3,394	10,873	7,352	2,049	6,311	5,185	1,346	4,562
Plasmacytoma	1,484	412	1,253	1,057	314	925	427	98	328
Myelodysplastic syndromes	5,603	1,450	5,076	3,536	909	3,172	2,068	542	1,903
Other Neoplasms of Uncertain or Unknown Behaviour	56,261	10,753	44,393	28,152	5,551	22,618	28,109	5,202	21,775
Myeloproliferative neoplasms	20,804	4,050	17,267	9,268	1,916	7,963	11,536	2,133	9,304
Monoclonal B-cell Lymphocytosis	9,717	1,532	7,411	5,204	892	4,015	4,513	640	3,397
Monoclonal gammopathy of undetermined significance	20,650	3,869	15,859	10,772	1,933	8,376	9,878	1,936	7,483
Lymphoproliferative disorder not otherwise specified	5,090	1,302	3,855	2,908	810	2,263	2,182	492	1,592

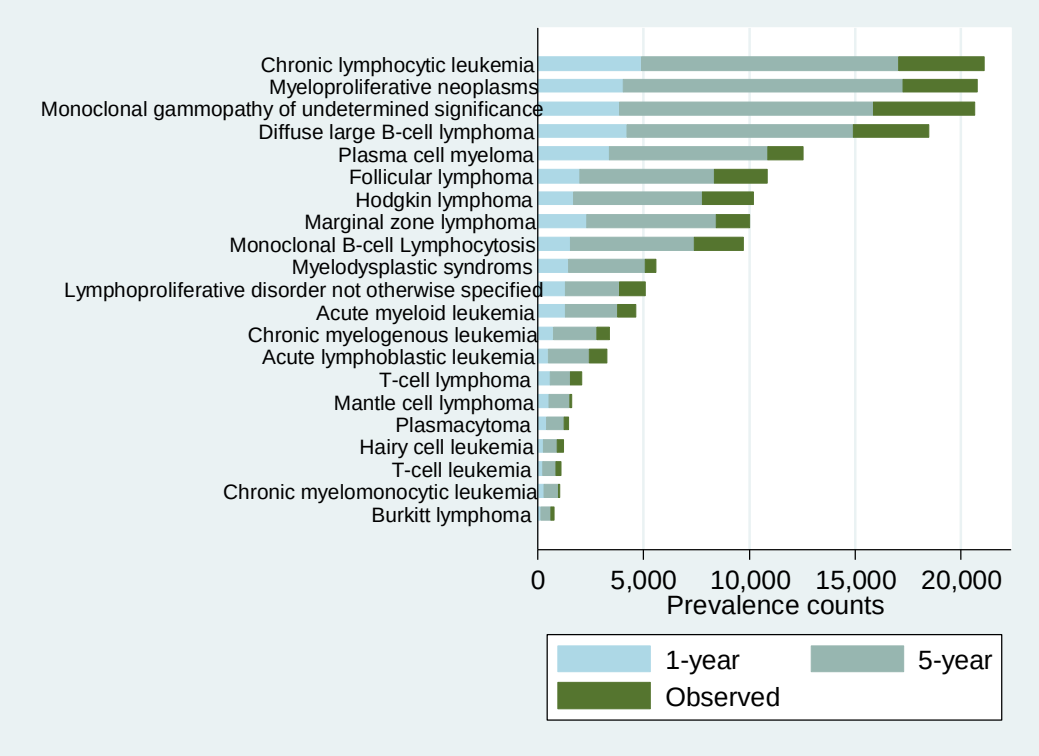


Figure 4- 4 Bar chart of observed prevalence counts in the UK by subtypes; stacked bars denote prevalence amongst patients alive on the index date who were diagnosed after 1st Sep. 2010, 1st Sep. 2006, and 1st Sep. 2004, respectively (two genders combined, and order sorted by observed prevalence counts)

For some diseases with a relatively high incidence rate and a good prognosis, n-year prevalence and observed prevalence within the registry may not provide accurate estimates, since there will be cases diagnosed before the start of the registry and still alive on the index date. Since HMRN is relatively young with only seven years of data, observed prevalence was only sufficient to show the burden of the subtypes with short survival, however for those with relatively longer survival, the bias due to the limited length of the registry cannot be ignored.

4.2.2 Sufficient years for complete prevalence

Table 4-6 showed n-year prevalence (per 100, 000) of haematological malignancies in HMRN on 31st, August 2011, according to number of years from diagnosis for both sexes. 1-year, 2-year, 3-year, 4-year, 5-year, 6-year, and 7-year prevalence were calculated separately. Percentage changes in prevalence between successive years (n and n+1) for every subtype, are shown by columns (change [%]). This information may be helpful to determine whether the years of follow up are sufficient to show complete prevalence. If the changes decrease and become very low, the prevalence tends to be stable with the accumulated years. It can therefore be said that observed prevalence was fine to show the burden of the disease, otherwise, the years of follow-up were not sufficient for complete prevalence.

Table 4- 6 N-year prevalence (per 100,000) and changes in HMRN for male and female subtypes

Diseases	n	Male		Female	
		Prevalence	Change (%)	Prevalence	Change (%)
Chronic myelogenous leukaemia	1	1.5		1.0	
	2	2.8	84.6	1.4	36.8
	3	3.5	27.1	2.2	53.8
	4	4.6	31.1	2.9	32.5
	5	5.9	27.5	3.6	26.4
	6	6.5	10.8	4.1	13.4
	7	7.2	10.6	4.4	6.6
Chronic myelomonocytic leukaemia	1	0.6		0.4	
	2	1.1	90.0	1.0	157.1
	3	1.6	42.1	1.1	11.1
	4	1.8	14.8	1.3	20.0
	5	2.0	9.7	1.4	4.2
	6	2.0	0.0	1.4	4.0
	7	2.1	5.9	1.5	7.7
Acute myeloid leukaemia	1	2.5		2.0	
	2	4.1	65.1	3.0	51.4
	3	5.7	39.4	3.8	25.0
	4	6.6	16.2	4.5	17.1
	5	7.5	12.2	5.4	22.0
	6	8.3	10.9	6.1	12.0
	7	9.0	8.4	6.9	13.4
Acute lymphoblastic leukaemia	1	1.3		0.4	
	2	2.5	87.0	1.1	162.5
	3	2.9	16.3	1.9	66.7
	4	3.8	30.0	2.7	42.9
	5	5.0	32.3	3.3	22.0
	6	5.8	16.3	3.7	13.1
	7	6.8	17.0	4.5	18.8
Chronic lymphocytic leukaemia	1	11.1		5.7	
	2	19.5	75.5	10.0	76.2
	3	26.8	37.7	14.6	44.9
	4	32.6	21.6	17.9	23.1
	5	37.2	14.2	21.2	18.5
	6	41.7	12.1	23.7	11.5
	7	46.5	11.5	25.9	9.4
Hairy cell leukaemia	1	0.9		0.1	
	2	1.2	40.0	0.1	100.0
	3	1.5	23.8	0.2	100.0
	4	1.8	23.1	0.5	150.0
	5	2.5	37.5	0.7	20.0
	6	2.8	11.4	0.7	8.3
	7	3.4	18.4	0.8	15.4

Table 4-6 continued

Diseases	n	Male		Female	
		Prevalence	Change (%)	Prevalence	Change (%)
T-cell leukaemia	1	0.3		0.5	
	2	0.6	120.0	1.0	100.0
	3	1.0	54.5	1.2	27.8
	4	1.0	0.0	1.4	8.7
	5	1.4	41.2	1.6	16.0
	6	1.5	8.3	2.0	27.6
	7	1.7	15.4	2.1	5.4
Marginal zone lymphoma	1	4.0		3.8	
	2	8.0	98.6	6.7	75.7
	3	11.8	46.8	9.0	34.1
	4	14.2	20.1	11.6	29.1
	5	15.9	12.2	12.9	11.7
	6	17.9	12.7	14.0	8.4
	7	19.0	5.8	15.3	9.3
Follicular lymphoma	1	3.4		3.4	
	2	6.2	84.5	7.1	106.3
	3	8.5	37.4	9.7	37.7
	4	11.6	36.1	12.6	29.6
	5	14.1	22.0	14.3	13.8
	6	16.0	13.5	16.8	17.4
	7	17.6	9.7	19.3	14.5
Mantle cell lymphoma	1	1.2		0.6	
	2	2.1	76.2	1.0	72.7
	3	2.8	32.4	1.2	15.8
	4	3.5	22.4	1.5	27.3
	5	3.6	5.0	1.6	3.6
	6	3.8	3.2	1.6	3.4
	7	3.9	3.1	1.6	0.0
Diffuse large B-cell lymphoma	1	7.8		6.6	
	2	14.0	80.0	11.0	66.9
	3	18.4	31.3	15.4	40.6
	4	23.6	27.9	19.5	26.4
	5	27.8	17.9	23.1	18.4
	6	31.4	12.9	26.0	12.7
	7	34.5	9.8	28.7	10.2
Burkitt lymphoma	1	0.3		0.1	
	2	0.7	100.0	0.2	100.0
	3	0.9	33.3	0.4	75.0
	4	1.3	37.5	0.5	28.6
	5	1.6	27.3	0.5	0.0
	6	1.9	17.9	0.5	11.1
	7	2.2	15.2	0.6	10.0

Table 4-6 continued

Diseases	n	Male		Female	
		Prevalence	Change (%)	Prevalence	Change (%)
T-cell lymphoma	1	1.1		0.9	
	2	1.6	47.4	1.1	25.0
	3	2.3	39.3	1.4	25.0
	4	2.7	20.5	1.7	24.0
	5	3.1	14.9	2.1	25.8
	6	3.7	18.5	2.6	23.1
	7	4.3	17.2	2.9	10.4
Hodgkin lymphoma	1	3.8		2.0	
	2	7.7	104.6	4.8	140.5
	3	11.0	43.6	7.1	46.1
	4	13.5	22.0	8.6	21.5
	5	16.0	18.5	10.6	24.1
	6	17.7	10.9	12.5	17.9
	7	19.8	11.8	15.0	19.9
Plasma cell myeloma	1	7.2		4.5	
	2	12.5	74.2	8.0	80.5
	3	16.8	34.3	10.6	32.4
	4	19.9	19.0	13.3	25.0
	5	22.1	10.7	15.1	13.5
	6	24.0	8.6	16.5	9.0
	7	25.7	7.2	17.2	4.3
Plasmacytoma	1	1.1		0.3	
	2	1.7	52.6	0.4	16.7
	3	2.5	51.7	0.7	85.7
	4	3.1	22.7	0.9	23.1
	5	3.2	3.7	1.1	25.0
	6	3.5	8.9	1.2	10.0
	7	3.7	4.9	1.4	18.2
Myelodysplastic syndromes	1	3.2		1.8	
	2	7.1	123.6	3.3	84.8
	3	8.7	22.0	4.8	44.3
	4	10.2	18.0	5.6	18.2
	5	11.1	8.5	6.3	11.5
	6	11.8	6.2	6.6	5.2
	7	12.4	4.9	6.8	3.3
Myeloproliferative neoplasms	1	6.7		7.1	
	2	12.7	89.7	12.8	80.8
	3	19.0	49.5	18.1	41.7
	4	23.4	23.1	24.6	35.7
	5	27.9	19.0	30.8	25.4
	6	30.8	10.4	35.7	15.9
	7	32.4	5.5	38.2	7.0

Table 4-6 continued

Diseases	n	Male		Female	
		Prevalence	Change (%)	Prevalence	Change (%)
Monoclonal B-cell lymphocytosis	1	3.1		2.1	
	2	6.6	113.0	4.6	117.9
	3	9.4	40.9	6.6	42.4
	4	11.5	22.8	9.3	41.3
	5	14.0	22.1	11.2	21.1
	6	16.5	17.7	13.3	18.4
	7	18.2	10.1	14.9	12.2
Monoclonal gammopathy of undetermined significance	1	6.8		6.4	
	2	13.3	96.6	12.2	90.7
	3	19.4	46.1	17.4	42.2
	4	25.3	30.4	21.3	22.8
	5	29.3	15.8	24.8	16.0
	6	33.6	14.6	28.8	16.4
	7	37.7	12.2	32.7	13.4
Lymphoproliferative disorder not otherwise specified	1	2.8		1.6	
	2	4.1	44.9	2.4	50.0
	3	5.0	22.5	3.2	28.9
	4	6.4	27.6	4.2	32.8
	5	7.9	23.4	5.3	26.0
	6	8.8	11.7	6.6	25.8
	7	10.2	15.0	7.2	9.0

From Table 4-6, it is easy to see that the percentage changes fall to under 5% for mantle cell lymphoma and myelodysplastic syndromes. This indicated that the seven year follow up in HMRN registry seemed sufficient for those subtypes with poor survival such as mantle cell lymphoma and myelodysplastic syndromes. Observed data for these cases was sufficient to estimate the true prevalence. However, subtypes such as Hodgkin lymphoma and acute lymphoblastic leukaemia still had percentage changes up to 19% between 6-year and 7-year prevalence. Therefore the length of HMRN registry was not enough to cover all patients who were alive with those diseases on the index date. Prevalence estimates of haematological malignancies in these ways (n-year prevalence or observed prevalence) may be of little value for those subtypes with better survival, and therefore for these subtypes it was necessary to estimate the total prevalence on the index date. It should be mentioned here that this presentation may not be appropriate for subtypes with small numbers of cases. For example, hairy cell leukaemia showed up to a 150% change for females (only 16 diagnoses in HMRN), which seems unreasonable compared with males. For similar reasons,

for some subtypes, the percentage change falls down to 0 in two certain n-year prevalence but increases again. Usually the 0% in the middle length of follow-up appears in the subtypes with a small number of cases. For example, the percentage change was 0% between 4-year and 3-year prevalence for T-cell leukaemia, but went up to 41.2% between 5-year and 4-year prevalence. This is because there were no newly diagnosed cases of T-cell leukaemia in the 4th year after the start of the registry (see Figure 4-5).

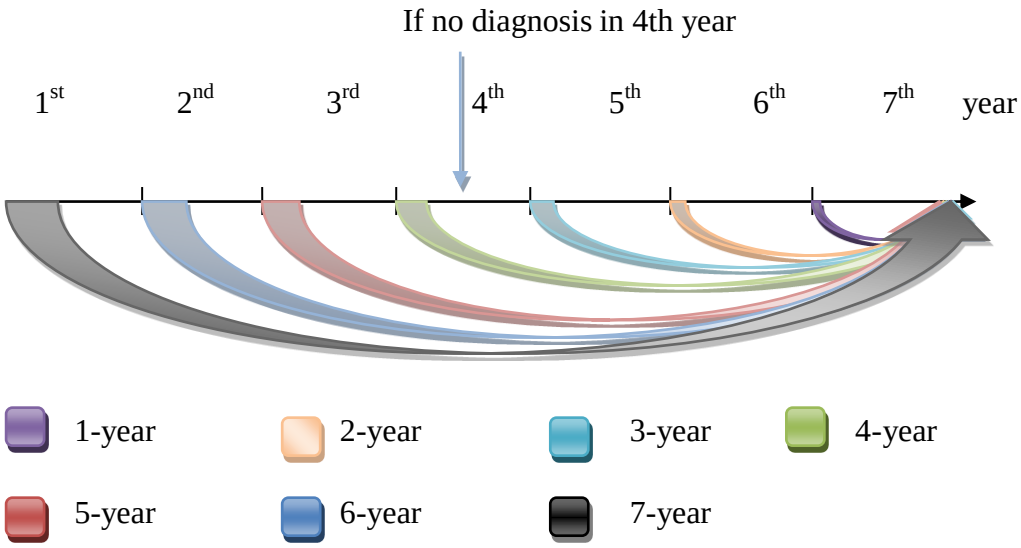


Figure 4- 5 The ith year of diagnosis and n-year prevalence

4.2.3 Summary

Using HMRN data, the figures shown in section 4.2 were all within the registry (information from 2004 to 2011). These statistics may provide limited insight regarding the longer survivor population and its needs. Seven years follow-up may not be sufficient to show the real burden of most of the subtypes. Therefore it is necessary to estimate total prevalence for all the patients in the population alive on the index date who previously had a diagnosis of haematological malignancies.

4.3 Total prevalence

The results for all subtypes of haematological malignancies calculated using the method for total prevalence are explored in this section. However, results for acute myeloid leukaemia and Hodgkin lymphoma have been used as examples to illustrate the method, as they represented two typical diseases with different incidence and survival characteristics.

4.3.1 Acute myeloid leukaemia (AML)

Generally, the incidence rate of AML increases with age, however it can occur at any age, in these data ranging from 0.2 to 97.8 years old with a median age of 70.2 years. There was a male predominance, and males had a higher incidence rate of AML than females in all age groups. The divergence between the two genders' rates became more marked as age increases (see Figure 4- 6). Crude incidence of AML by age and gender (per 100,000 population) was shown in Table 4- 7.

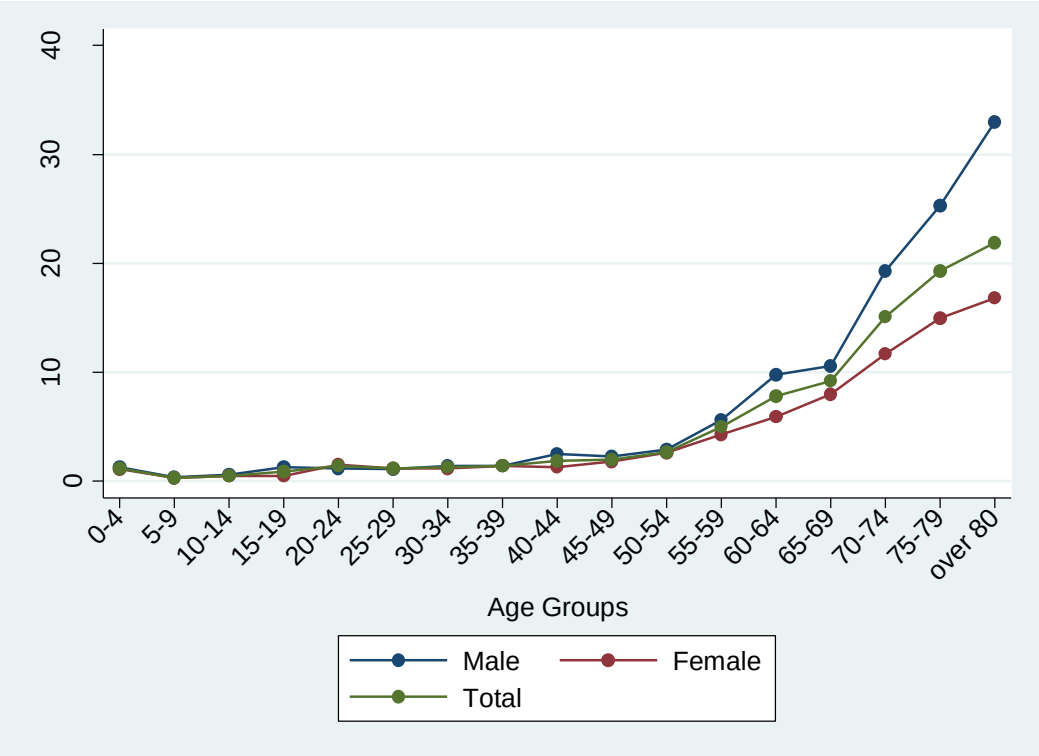


Figure 4- 6 Incidence of AML per 100,000 for males, females, and total

Table 4- 7 Crude incidence of AML rate per 100,000 by age and gender

Age group	Total		Male		Female	
(Years)	N	Incidence	N	Incidence	N	Incidence
0-4	18	1.2	10	1.3	8	1.1
05-Sep	5	0.3	3	0.4	2	0.3
Oct-14	9	0.5	5	0.6	4	0.5
15-19	15	0.9	11	1.3	4	0.5
20-24	21	1.4	9	1.2	12	1.5
25-29	18	1.2	8	1.1	10	1.2
30-34	24	1.3	13	1.4	11	1.2
35-39	27	1.4	13	1.4	14	1.4
40-44	33	1.9	22	2.5	11	1.3
45-49	32	2	18	2.3	14	1.8
50-54	48	2.7	25	2.9	23	2.6
55-59	69	5	39	5.6	30	4.3
60-64	96	7.8	59	9.8	37	5.9
65-69	103	9.2	56	10.6	47	8
70-74	151	15.1	86	19.3	65	11.7
75-79	163	19.3	89	25.3	74	15
over 80	229	21.9	110	33	119	16.8
Total	1,061	4.2	576	4.8	485	3.8

The survival for AML was shown in the Figure 4- 7. AML had a poor survival (regardless of age), and there were 779 deaths in HMRN from 2004 to 2011. Men and women had a similar survival (log rank test: p=0.845).

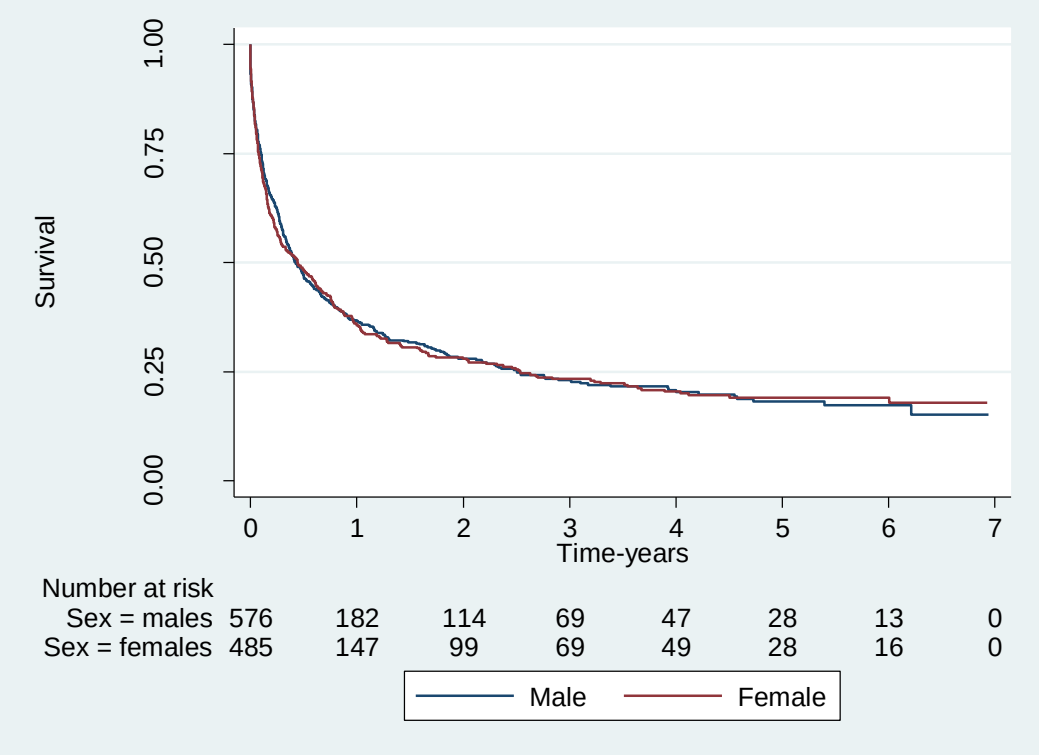


Figure 4- 7 Kaplan-Meier survival estimates for AML patients by gender

4.3.1.1 The modelling of incidence

As an example of the calculation process, the predictions of incidence were shown in this section. The incidence for males has been taken as the example to show the benefit of my method (the comparison with the method to estimate incidence in the literature are shown in Chapter Five):

In the literature, the parametric incidence model, $I = ax^b$ (I is incidence and x is age), has been widely used (Capocaccia and De Angelis, 1997; Merrill, et al., 2000; Gigli, et al., 2006). For a general class of cancers that rarely occur at an early age, this model could provide a reasonable fit for the data. In this study, if

the incidence cases at an early age (age before 35 years) could be ignored, this parametric model could be useful and fit the incidence data (see Figure 4- 8). However, according to the incidence calculations (see Appendix A5), haematological malignancies were a group of diseases that usually include earlier ages. The parametric incidence model did not fit the incidence data well, especially for the older age groups, if all ages were considered (see Figure 4- 9). Thus, the predictions of this incidence function would bring large bias to the results.

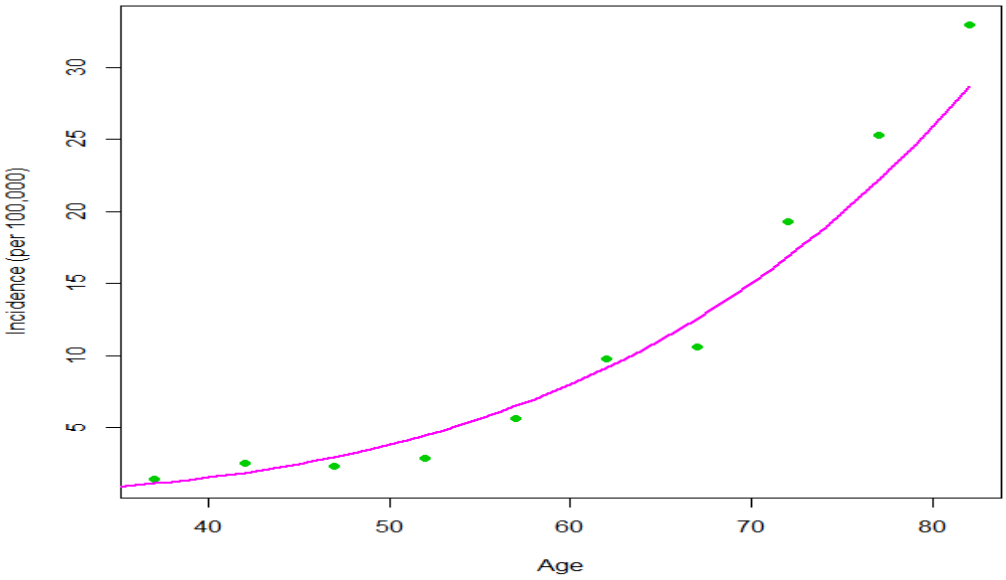


Figure 4- 8 The modelling of incidence using a log linear model for age after 35 years for AML

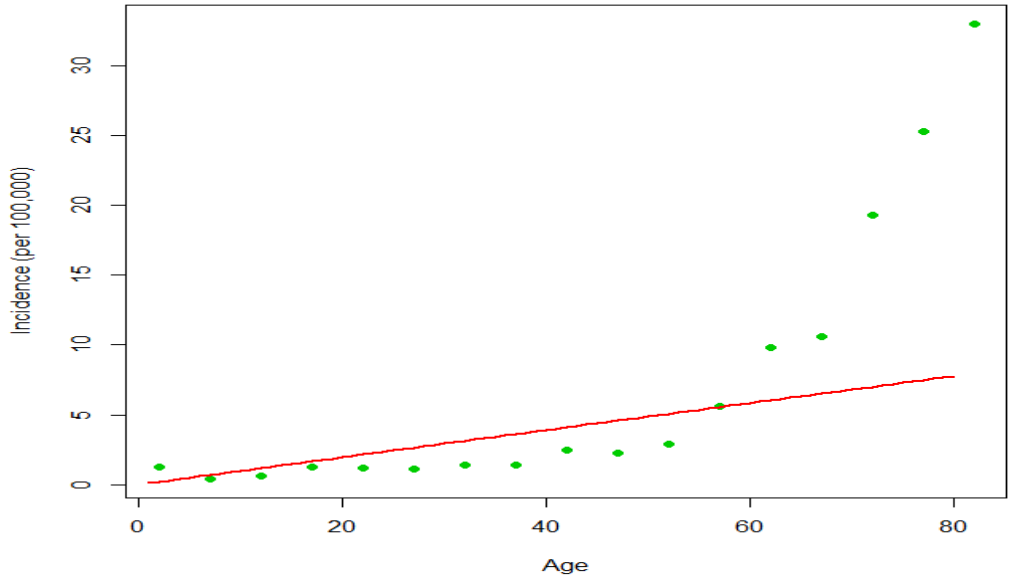


Figure 4- 9 The modelling of incidence using a log linear model for all ages for AML

The regression spline demonstrates good fit of incidence (Figure 4- 10), and this method of modelling incidence was used in the whole work to calculate the total prevalence of hematological malignancies.

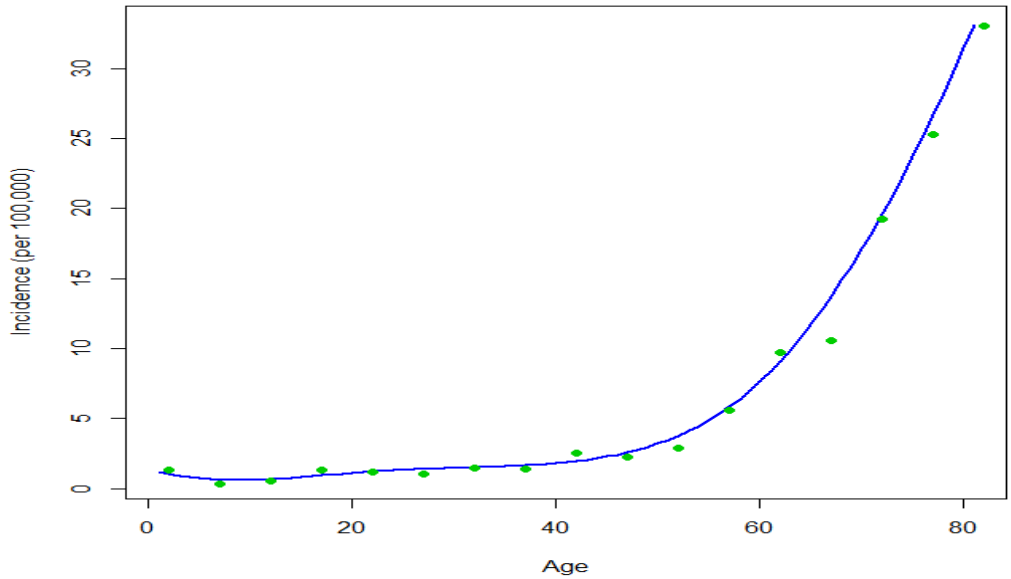


Figure 4- 10 The modelling of incidence by spline regression for AML

4.3.1.2 The modelling of survival

The estimates of survival in the calculation involved parametric functions (see Chapter Three, Section 3.4.6.3). They depended on both the duration of the disease (years since diagnosis) and age. Survival probability decreased with disease duration, however the hazard may vary according to age at diagnosis.

Figure 4- 11 shows AML survival of the model in this study as a 3D version of age (years) and duration (survival years) for males in HMRN, to achieve a better visual impression of the probability of survival. The observed data was limited to 7 years (the green lines on the curve show the survival with the disease duration at 7 years). Graph A was the one seen from a 30-degree angle, and graph B was from a 120-degree angle. Both indicate that survival of AML is determined by both age and duration from diagnosis.

There was an arch around age 20 years in Figure 4- 11. However, the curve at early ages may lack precision due to the small number of cases in the younger age groups. The confidential interval for younger age groups may be wider than for the older age groups (details were shown in Section 4.3.4), and there may be no evidence to show the existence of this arch when the number of cases becomes large.

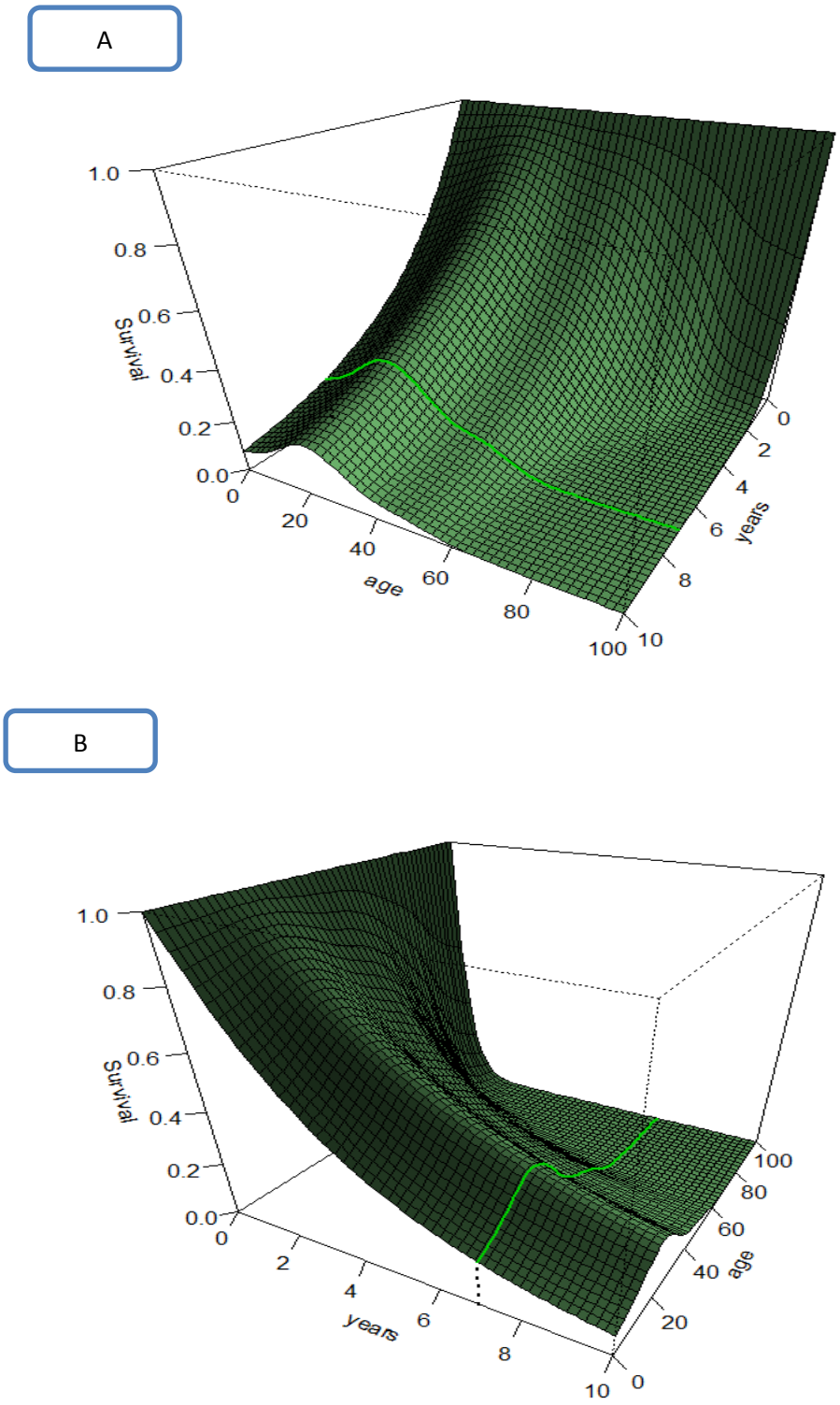


Figure 4- 11 3D Version for survival curve of AML by age and duration (A: 30 degree angle; B: 120 degree angle)

4.3.1.3 Total prevalence

The completeness indices for AML based on 7 years of follow up for both genders in HMRN was reported by age (see Figure 4- 12). The values of R (completeness index) vary according to age and gender. For AML, it ranged from 0.53 to 1. Such high values demonstrated the poor survival for AML. The values of completeness index of males were a little higher than those of females in most age groups. This may be because males have a higher incidence of AML than females.

The values of R shown in Table 4- 8 were the completeness index values of the middle age in every age group. Because the length of the registry was 7 years (from 2004 to 2011), in the age group 0-7years, there was no case diagnosed before the start of the registry and the completeness index values in those age groups were 1. Generally, for both genders, the values of R had a decreasing trend in younger age, and increased after 40 years old. Details are shown in Table 4- 9.

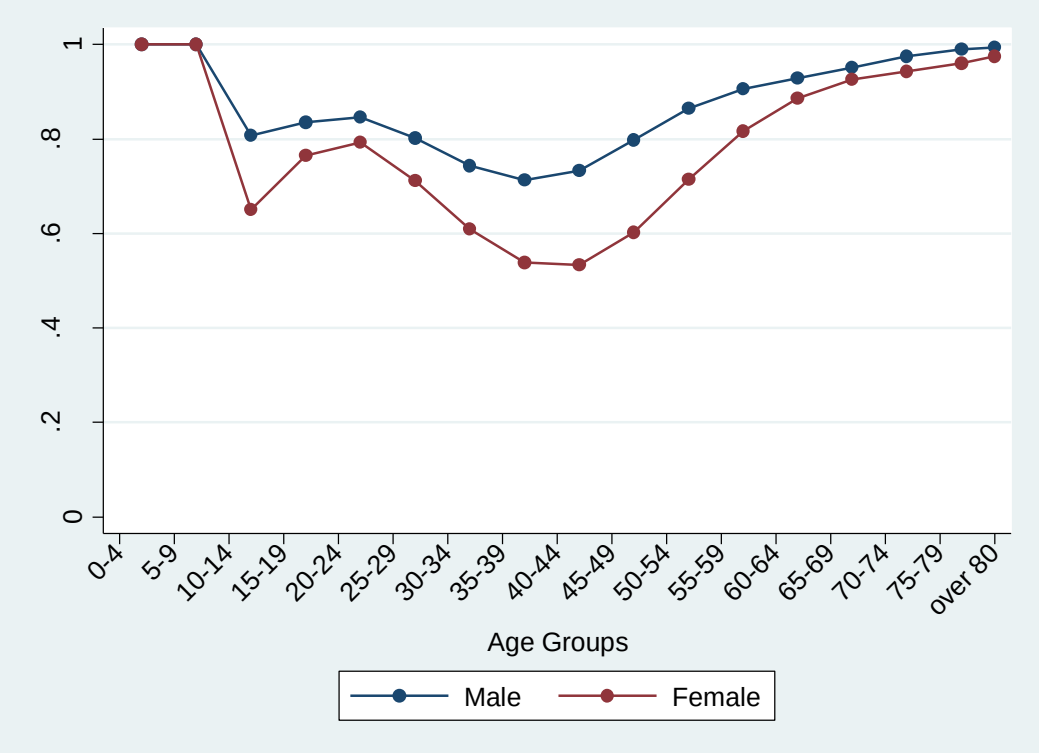


Figure 4- 12 The completeness index of AML for males and females

Total prevalence was 10.2 per 100,000 for males and 9.0 per 100, 000 for females. Figure 4- 13 shows age specific observed and total prevalence (dark blue bars are observed prevalence and light blue bars are total prevalence). The total prevalence was slightly higher than observed prevalence in the middle age groups, but similar to observed prevalence in the young (due to the length of the registry) and older age groups (due to the high mortality in older patients).

Table 4- 8 Calculation process for Total prevalence of AML by age group and gender

Age group (Years)	Male			Female		
	No	R	Nt	No	R	Nt
0-4	4	1	4	1	1	1
5-9	3	1	3	3	1	3
10-14	1	0.80708	1	5	0.650215	8
15-19	2	0.835425	2	1	0.765125	1
20-24	9	0.845672	11	6	0.793124	8
25-29	8	0.801608	10	9	0.711563	13
30-34	6	0.74334	8	8	0.608586	13
35-39	9	0.712666	13	5	0.539299	9
40-44	8	0.732622	11	7	0.533163	13
45-49	9	0.797788	11	9	0.602103	15
50-54	9	0.865273	10	8	0.714961	11
55-59	12	0.905976	13	7	0.816199	9
60-64	17	0.929141	18	14	0.885529	16
65-69	18	0.951631	19	10	0.926413	11
70-74	16	0.975479	16	13	0.943646	14
75-79	13	0.990472	13	11	0.961207	11
over 80	11	0.994419	11	10	0.975478	10
Total	155	0.88	176	127	0.77	166
Prevalence	9.0		10.2	6.9		9.0

No: number of observed prevalent cases Nt: number of total prevalent cases
R: completeness index
*Per 100, 000 population

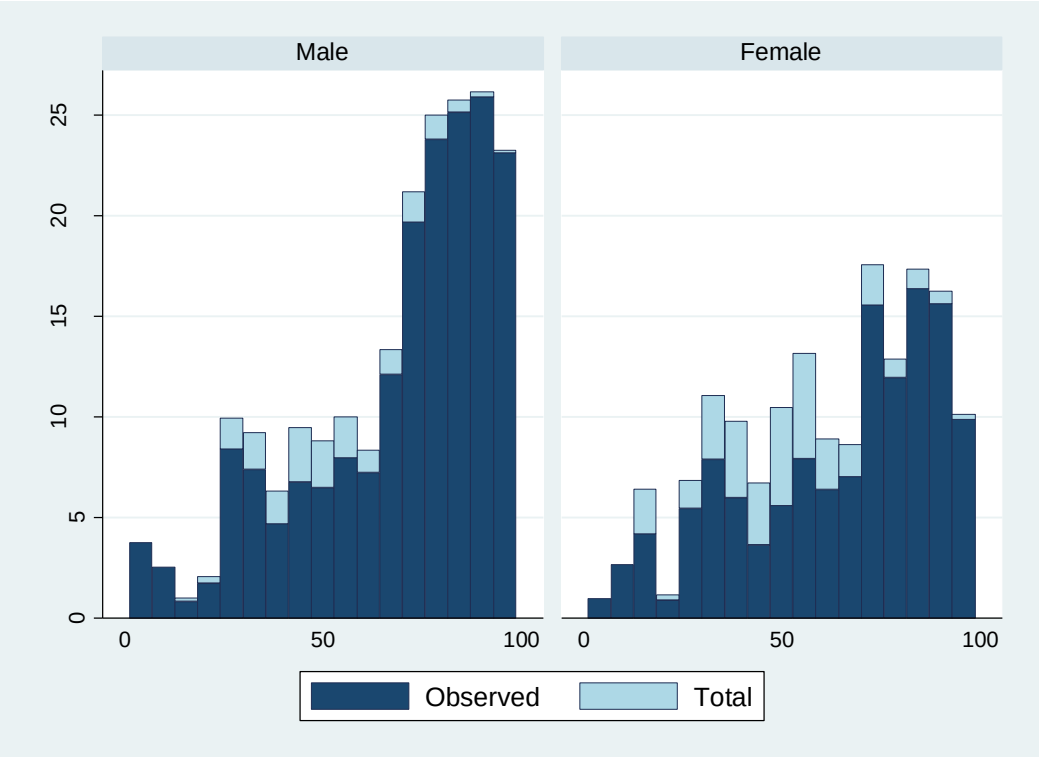


Figure 4- 13 Age-specific observed and total prevalence (per 100 000) for males and females with AML in HMRN on the index date of 31st, August 2011.

4.3.2 Hodgkin lymphoma

Unlike some other cancers, whose incidence increases with age, Hodgkin lymphoma has a bimodal incidence curve. Patients diagnosed with Hodgkin lymphoma had a median age of 41.3 years (ranging from 3.6 to 90.9 years). It occurred most frequently in HMRN in two separate age groups, the first being young adulthood (age 15–35 years) and the second being in those over 60 years old (see Figure 4- 14). The annual incidence was 3.5 per 100,000 and 2.6 per 100,000 for men and women respectively. Overall, it was more common in males in adult years, and females had a deeper trough (differences in age and sex patterns had been described in Section 4.1). Crude incidence of Hodgkin lymphoma by age and gender (per 100,000 population) was shown in Table 4- 9.

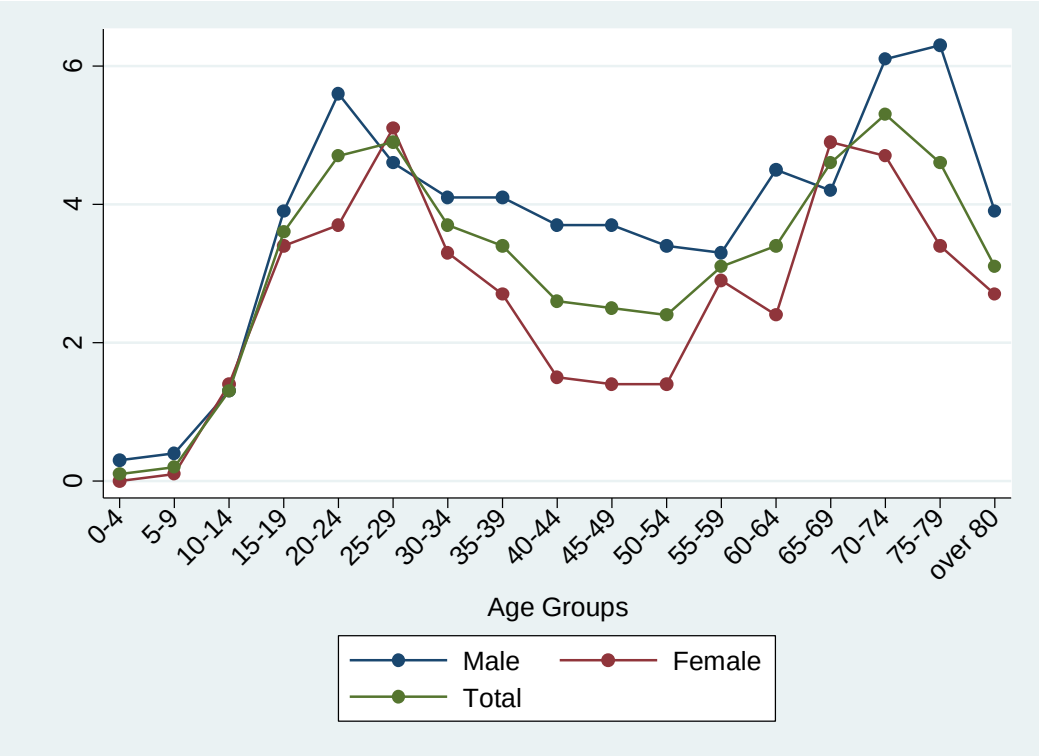


Figure 4- 14 Incidence of Hodgkin lymphoma per 100,000 for males, females, and total

Table 4- 9 Crude incidence of Hodgkin lymphoma rate per 100,000 by age and gender

Age group	Total		Male		Female	
(Years)	N	Incidence	N	Incidence	N	Incidence
0-4	2	0.1	2	0.3	0	0
05-Sep	4	0.2	3	0.4	1	0.1
Oct-14	23	1.3	11	1.3	12	1.4
15-19	59	3.6	32	3.9	27	3.4
20-24	71	4.7	42	5.6	29	3.7
25-29	76	4.9	35	4.6	41	5.1
30-34	68	3.7	37	4.1	31	3.3
35-39	64	3.4	38	4.1	26	2.7
40-44	45	2.6	32	3.7	13	1.5
45-49	40	2.5	29	3.7	11	1.4
50-54	42	2.4	30	3.4	12	1.4
55-59	43	3.1	23	3.3	20	2.9
60-64	42	3.4	27	4.5	15	2.4
65-69	51	4.6	22	4.2	29	4.9
70-74	53	5.3	27	6.1	26	4.7
75-79	39	4.6	22	6.3	17	3.4
Over 80	32	3.1	13	3.9	19	2.7
Total	754	3	425	3.5	329	2.6

The survival of Hodgkin lymphoma is shown in Figure 4-15. Hodgkin lymphoma had a good survival; there were only 135 deaths in HMRN from 2004 to 2011. The median age of death in HMRN was 72.2 years (ranging from 5.8 to 89.5 years). The difference in survival between males and females was not significant (log rank test: p=0.087). Both incidence and survival determine the total prevalence of Hodgkin lymphoma.

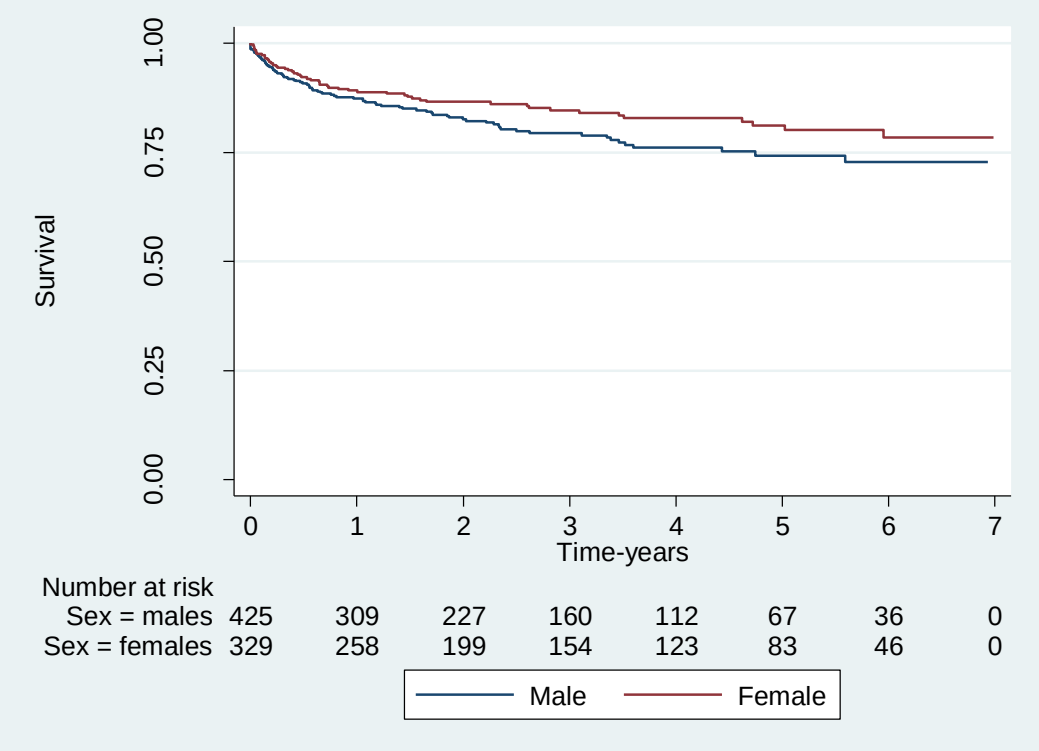


Figure 4- 15 Kaplan-Meier survival estimates for Hodgkin lymphoma patients by gender

4.3.2.1 The modelling of incidence

For Hodgkin lymphoma, the disadvantage of a log linear incidence model and the advantage of a regression spline incidence model were more significant.

Obviously, the exponential shape for age of incidence failed to describe the bimodal incidence data of Hodgkin lymphoma (Figure 4-16).

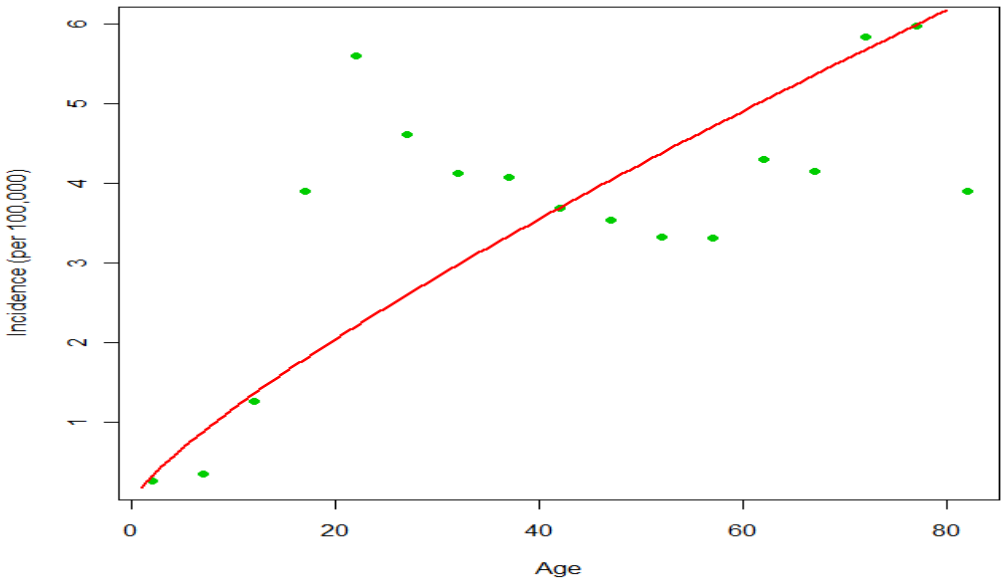


Figure 4- 16 The modelling of incidence using a log linear model for Hodgkin lymphoma

The regression spline model aptly described the bimodal incidence data of Hodgkin lymphoma. Figure 4- 17 shows Hodgkin lymphoma incidence for men as the example to demonstrate the benefit of my method in this study.

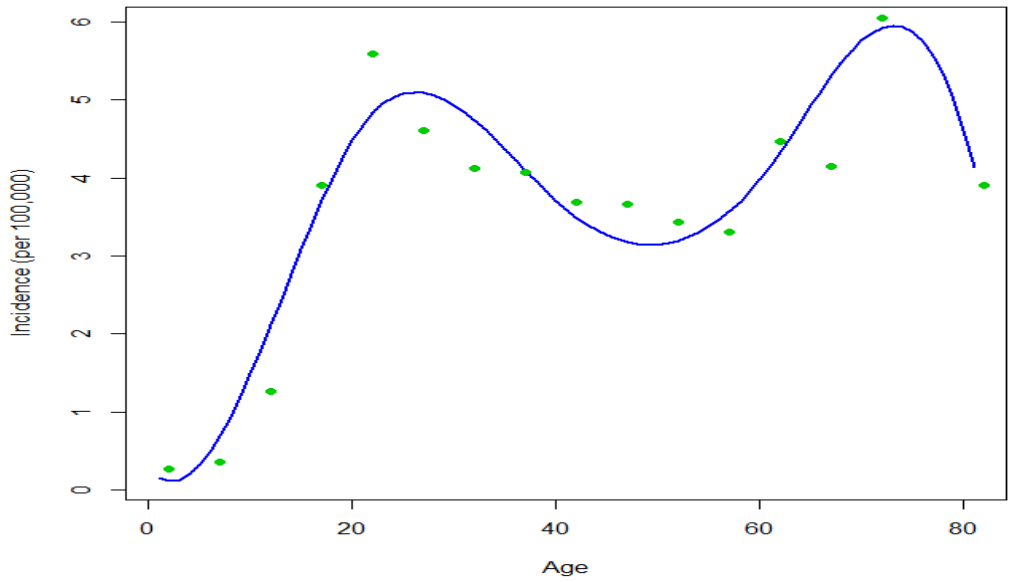


Figure 4- 17 The modelling of incidence by spline regression for Hodgkin lymphoma

4.3.2.2 The modelling of survival

Compared with AML, Hodgkin lymphoma had good survival (see Figure 4- 7 and Figure 4- 15). However, the age at diagnosis had a non- linear effect on the survival of Hodgkin lymphoma. Figure 4- 18 shows a 3D version of the estimated survival curve by age and duration of Hodgkin lymphoma for males in HMRN. It is more complicated than the survival curves for AML. However, again, the precision may be low in some age groups due to the small number of cases. The observed data was limited to 7 years (the green lines on the curve showed the duration of survival with the disease at 7 years). Graph A was the one seen from a 30-degree angle, and graph B was from a 120-degree angle.

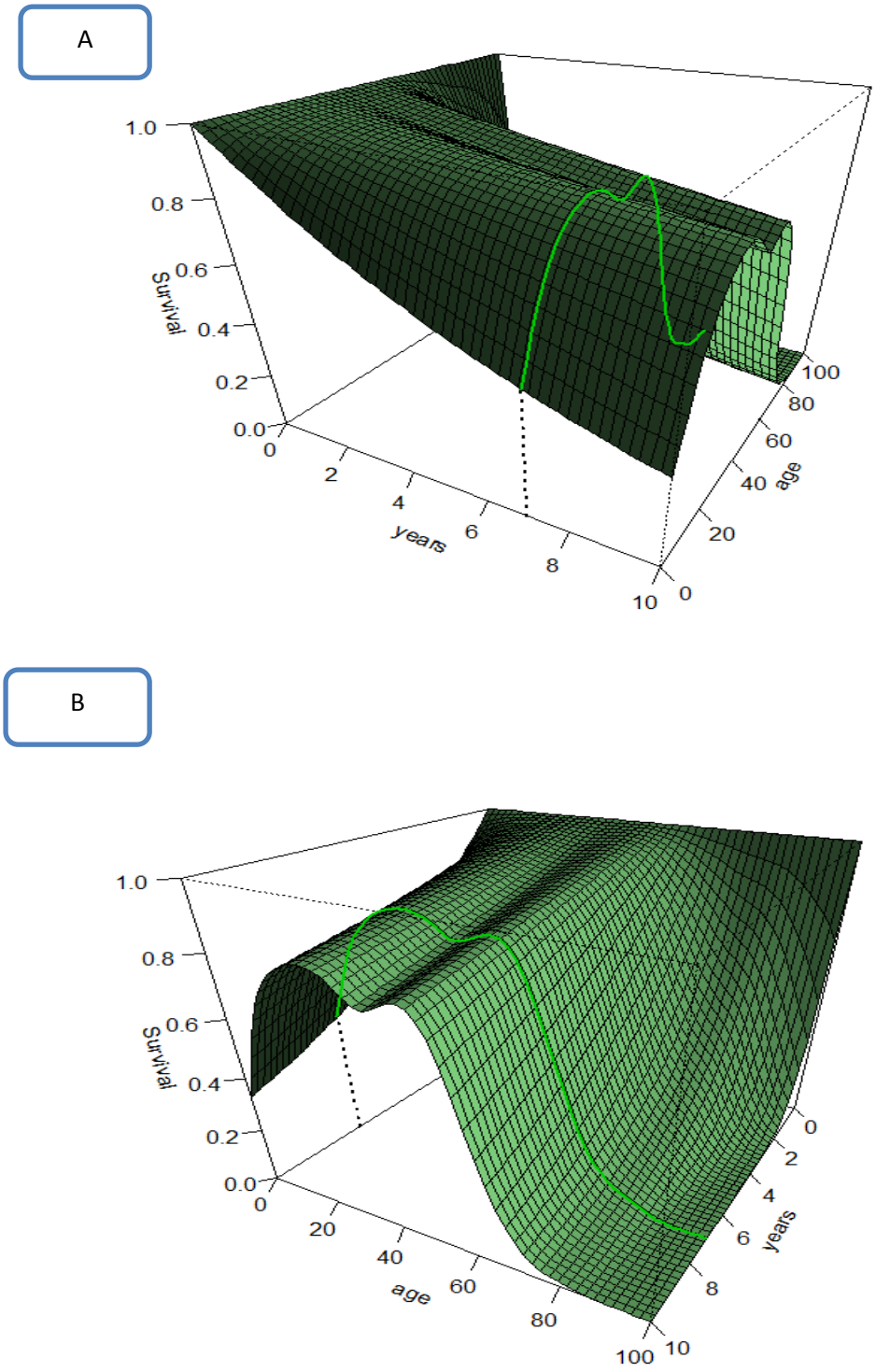


Figure 4- 18 3D-version for survival curve of Hodgkin lymphoma by age and duration (A: 30 degree angle; B: 120 degree angle)

4.3.2.3 Total prevalence

The completeness indices for Hodgkin lymphoma based on 7 years of follow up for both genders in HMRN was reported by age (see Figure 4- 19). The values of R (completeness index) varied according to age and gender. For Hodgkin lymphoma, it had a range from 0.1 to 1; such low values demonstrated the good prognosis for Hodgkin lymphoma.

The balance of incidence and survival determined the pattern of completeness index. The higher incidence of males made the completeness index lower, but the better survival of females resulted in more cumulative survivors on the index date. Generally, for both genders, the values of R had a significant decreasing trend before 65 years old, and slightly increased after that. Details are shown in Table 4- 10.

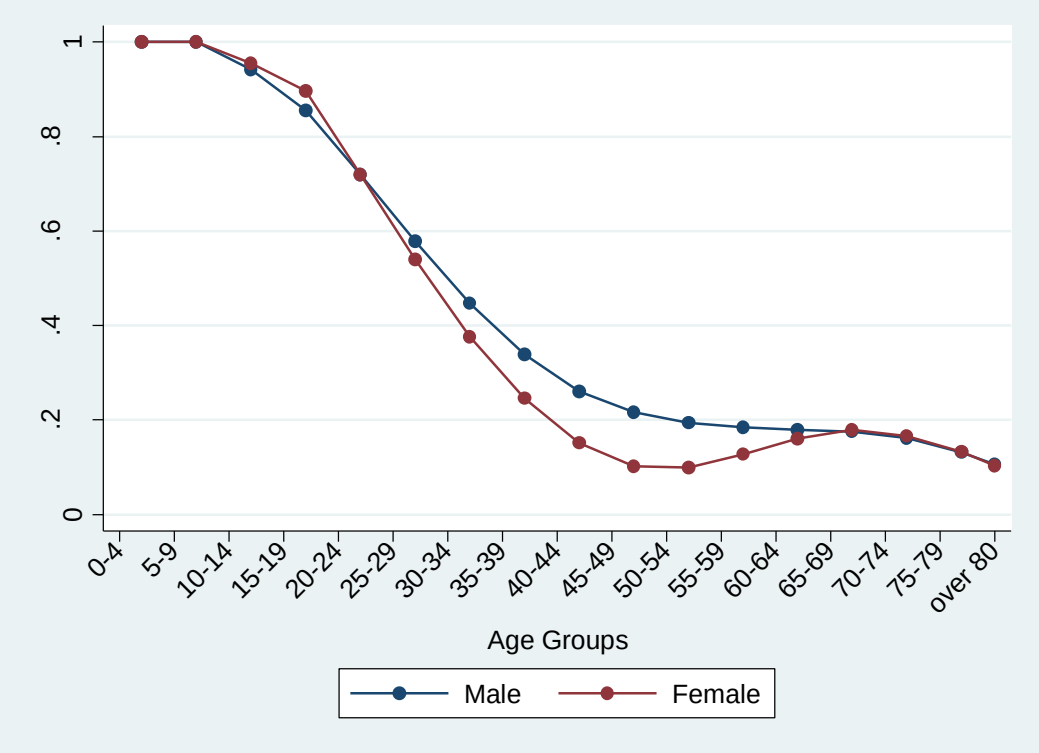


Figure 4- 19 The completeness index of Hodgkin lymphoma for males and females

Total prevalence was estimated to be 73.3 per 100,000 for males and 71.4 per 100, 000 for females. There were no prevalent cases in the first age group for females, however there were more cumulative prevalent cases in the last age group for females (see Table 4-10). This may be because women generally have a longer life span, and female Hodgkin lymphoma patients had a slightly better survival rate than male patients. Figure 4- 20 showed age-specific total prevalence of Hodgkin lymphoma for both genders. Unlike incidence, the total prevalence curve of Hodgkin lymphoma did not show a bimodal distribution. This is because the high incidence and good prognosis in young adulthood (age 15–35 years) resulted in a large amount of cumulative prevalent cases in middle age. On the other hand, although there was a second incidence peak over the age of 60, the survival became worse in the older age groups.

Table 4- 10 Total prevalence calculation process of Hodgkin lymphoma by age group and gender

Age group (Years)	Male			Female		
	No	R	Nt	No	R	Nt
0-4	1	1	1	0	1	0
5-9	2	1	2	1	1	1
10-14	4	0.942242	4	1	0.954945	1
15-19	19	0.85558	22	19	0.896376	21
20-24	43	0.719678	60	25	0.719359	35
25-29	25	0.578142	43	33	0.539985	61
30-34	33	0.447276	74	36	0.376416	96
35-39	35	0.339477	103	32	0.24576	130
40-44	35	0.261082	134	17	0.151837	112
45-49	26	0.216264	120	12	0.102336	117
50-54	34	0.194131	175	12	0.100002	120
55-59	21	0.184077	114	11	0.127203	86
60-64	17	0.179281	95	16	0.161136	99
65-69	16	0.17492	91	19	0.179138	106
70-74	14	0.161693	87	19	0.16585	115
75-79	10	0.131549	76	11	0.132758	83
Over 80	7	0.105524	66	14	0.102926	136
Total	342	0.27	1268	278	0.21	1319
Prevalence	19.8		73.3	15.1		71.7

No: number of observed prevalent cases Nt: number of total prevalent cases
R: completeness index
*Per 100, 000 population

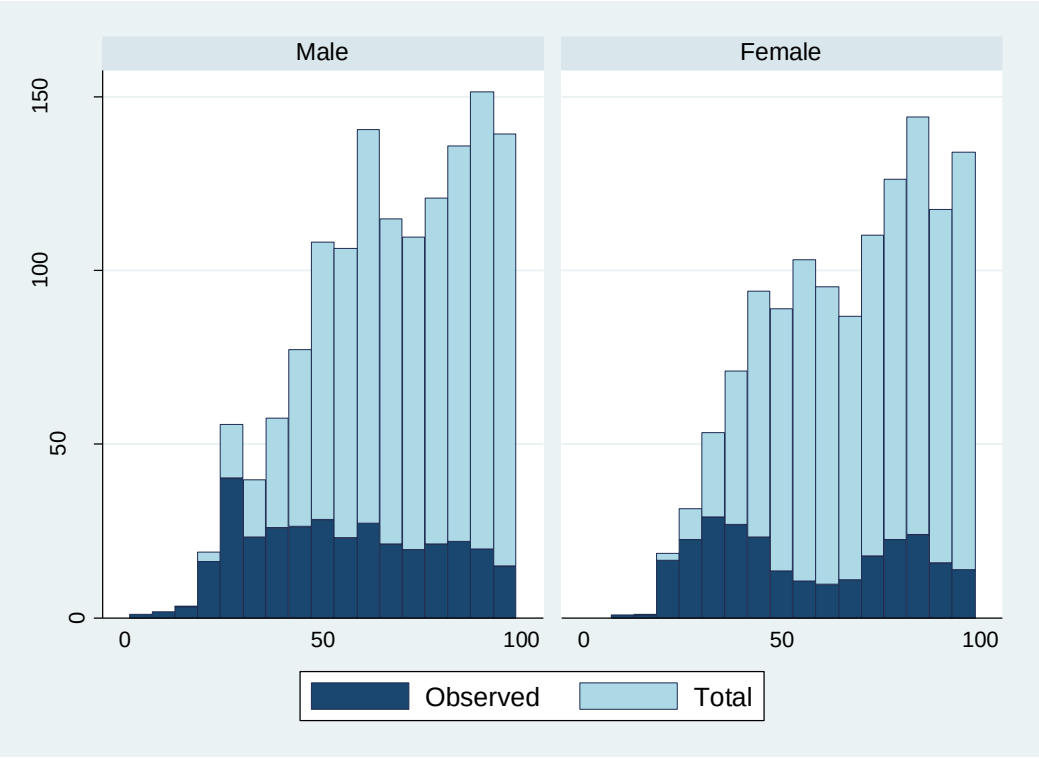


Figure 4- 20 Age-specific observed and total prevalence of Hodgkin lymphoma for males and females (per 100,000)

4.3.3 Dependence of R on registry and validation of R

To analyse the sensitivity and validation of R, a set of parameters was chosen, and R computed varying one parameter (length of the registry) at a time, taking the remaining ones as fixed. Hodgkin lymphoma was chosen as the example in this section (reasons were described in Section 3.4.9.1). To avoid repetitive computation, Hodgkin lymphoma for males was chosen as the standard value in this section and extrapolated to the past (because there is a male predominance for most subtypes, and the incidence and survival patterns of Hodgkin lymphoma are similar for males and females [Section 4.3.2]). Hodgkin lymphoma had a good prognosis, so the value of completeness index would vary greatly according to the length of the registry.

4.3.3.1 Dependence of completeness index on the registry

Figure 4- 21 showed the R-L (completeness index- length of follow-up) curve family obtained by varying the age at diagnosis. The lowest curve was the one used in the previous section to calculate total prevalence for Hodgkin lymphoma when the registry (HMRN) is 7 years old. Other R (completeness index) was computed varying the parameter *L* (the length of the registry), taking the remaining ones (incidence, survival, and general mortality) as fixed.

As expected, the incompleteness bias decreased with the length of follow- up. For any certain age, the values of R increase with the length of the registry. If the length of the registry reached 50 years, the observed prevalence would account for the majority of prevalent cases. Then, the completeness index adjustment for the bias due to the unobserved part would be limited. If the registry time were to go on to infinity, observed prevalence should converge to the total prevalence and the values of R would be 1, if all other things remained the same.

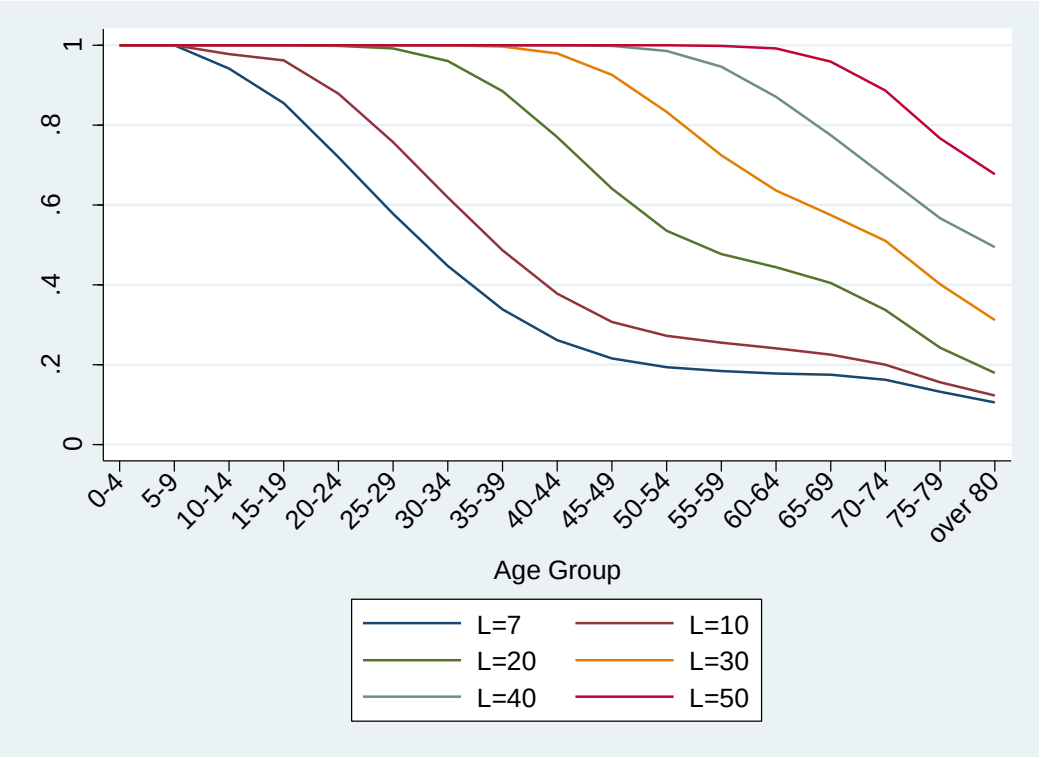


Figure 4- 21 Prevalence completeness index R as a function of age for various lengths of registry follow-up (L). (Hodgkin lymphoma for males)

The effect of the length of the registry (L) was straightforward. The longer the follow-up was, the lower the underestimation bias. Table 4- 11 showed a matrix of completeness index for age and length of registry. When $age \leq the\ length\ of\ registry$, there was no case diagnosed before the start of registry and who was still alive on the index date. The value of R was 1, which means the observed prevalence was exactly the same as the total prevalence in those age groups.

Table 4- 11 Completeness index of Hodgkin lymphoma for men by age group for various lengths of registry follow-up (L)

Age group (Years)	Length of registry follow-up (L)					
	L=7	L=10	L=20	L=30	L=40	L=50
0-4	1	1	1	1	1	1
5-9	1	1	1	1	1	1
10-14	0.942	0.979	1	1	1	1
15-19	0.856	0.962	1	1	1	1
20-24	0.720	0.880	0.999	1	1	1
25-29	0.578	0.758	0.993	1	1	1
30-34	0.447	0.620	0.961	1.000	1	1
35-39	0.339	0.487	0.885	0.998	1	1
40-44	0.261	0.378	0.770	0.980	1.000	1
45-49	0.216	0.308	0.641	0.926	0.999	1
50-54	0.194	0.273	0.536	0.833	0.987	1.000
55-59	0.184	0.255	0.477	0.725	0.947	0.999
60-64	0.179	0.242	0.445	0.637	0.872	0.992
65-69	0.175	0.226	0.405	0.576	0.775	0.960
70-74	0.162	0.200	0.338	0.511	0.672	0.887
75-79	0.132	0.157	0.243	0.402	0.568	0.768
Over 80	0.106	0.123	0.180	0.312	0.495	0.678

The results presented above were derived based on certain simplifying assumptions on the morbidity modelling. For example, the last column of Table 4-11 may not reflect the truth, since the survival of Hodgkin lymphoma was much poorer 40 years ago (details were in Section 4.4.2). As expected, for the much poorer survival in the past, the R curves tended to move towards the top of the figure, indicating a lower proportion of prevalent cases were diagnosed before the start of the registry. At 50 years follow-up, for example, the higher mortality rate before the 1960s may lead to an increase of R, even approaching to 1.

4.3.3.2 Validation of the analysis

I. Goodness of fit

Table 4-12 and Figure 4-22 used Hodgkin lymphoma to compare n-year prevalence estimated using the method with the actual n-year prevalence for both genders. The similar values suggest this is a relatively good fit for the data, and Chi square test shows there was no difference between actual and estimated prevalence. The longer the period was, the smaller the differences between estimated and actual values. For 7-year prevalence, the estimated prevalence was exactly the same as the actual one, because the estimation was made using the completeness index, which must be 1 in that instance.

Table 4- 12 n-year prevalence estimated using the method and the actual n-year prevalence for both genders (Hodgkin lymphoma)

n-year Prevalence	Period of Diagnosis	Estimated		Actual		
		Prevalence	Cases	Prevalence	Cases	
Male						
1-year	2010-2011	3.5	61	3.8	65	
2-year	2009-2011	6.6	114	7.7	133	
3-year	2008-2011	9.6	166	11.0	191	$\chi2$ =2.1(df=6)
4-year	2007-2011	12.3	213	13.5	233	p=0.908
5-year	2006-2011	14.9	258	16.0	276	
6-year	2005-2011	17.4	301	17.7	306	
7-year	2004-2011	19.8	342	19.8	342	
Female						
1-year	2010-2011	2.4	44	2.0	37	
2-year	2009-2011	4.7	87	4.8	88	
3-year	2008-2011	6.9	127	7.1	131	$\chi2$=1.0(df=6)
4-year	2007-2011	9.0	166	8.5	156	p=0.986
5-year	2006-2011	11.1	204	10.6	195	
6-year	2005-2011	13.1	241	12.5	230	
7-year	2004-2011	15.0	277	15.0	277	

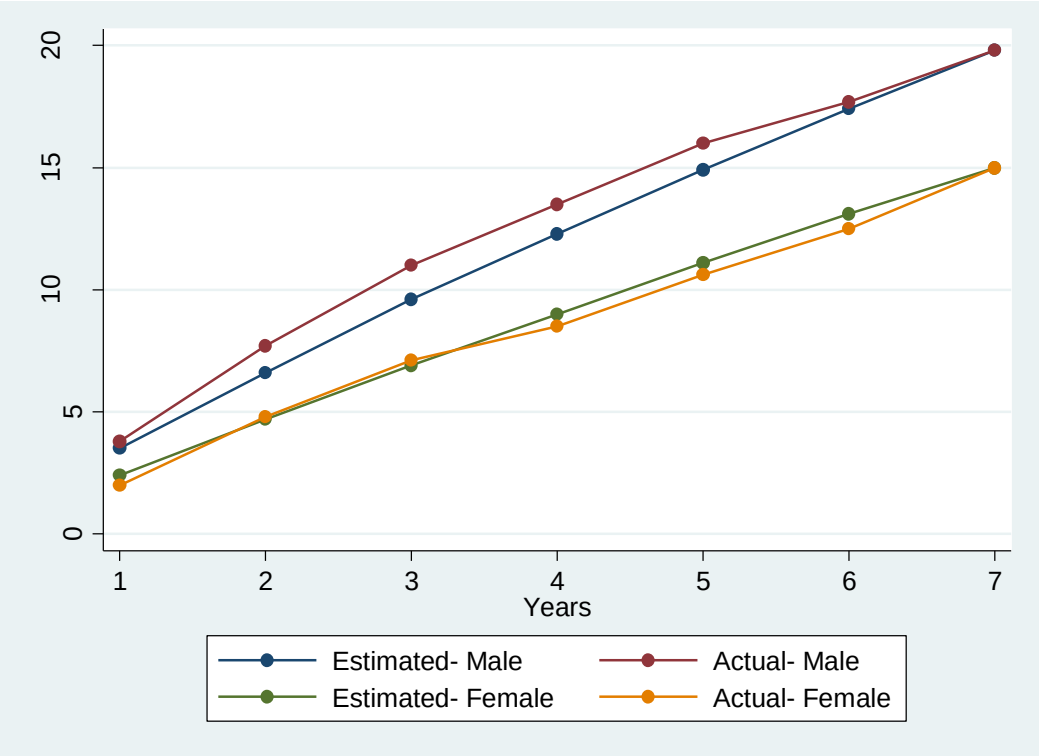


Figure 4- 22 n-year prevalence estimated using the method and the actual n-year prevalence for both genders (Hodgkin lymphoma)

II. Predictive power

The method developed in this study was applied to estimate the expected 7-year prevalence in order to compare them to the actual values observed in HMRN data. Figure 4-23 and Figure 4-24 summarize the information; the plots were satisfactory for both sexes (p=0.842 for males and p=0.852 for females). Generally, the estimated 7-year prevalent cases followed the age distribution of the observed ones. For all age groups combined, the number of estimated cases was slightly lower than the number of observed cases with a difference of 2.9% for males (332 estimated cases, and 342 observed cases), and 9.0% for females (253 estimated cases, and 278 observed cases). The differences for both sexes were less than 10% and did not exceed the limit (Gigli, et al., 2004). Thus, the predictive power of the method was acceptable.

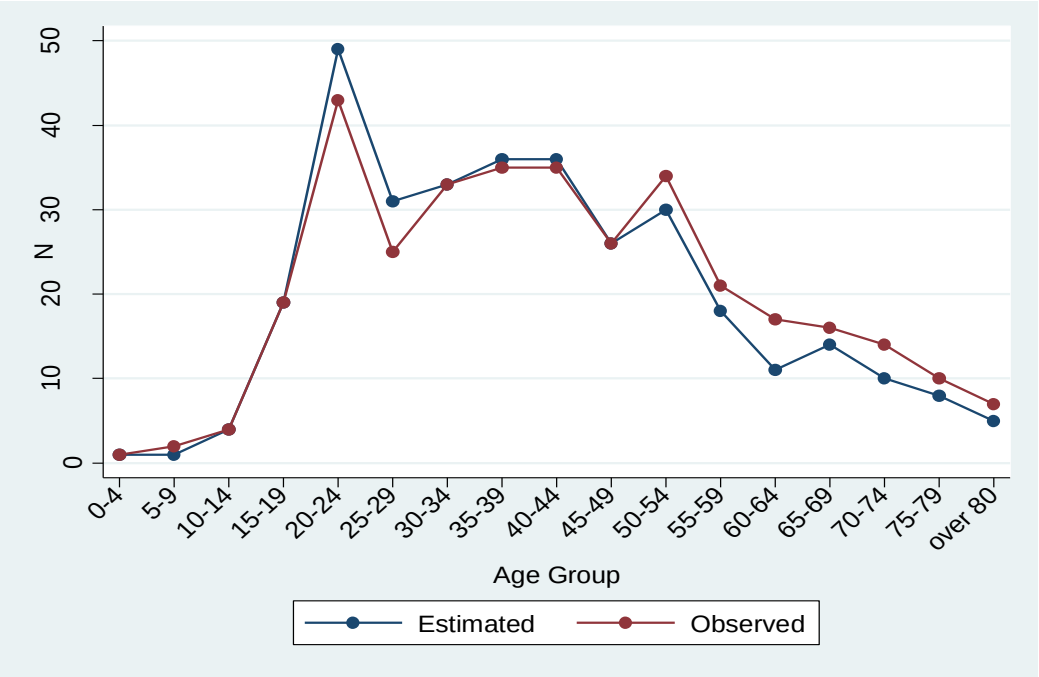


Figure 4- 23 The number of 7-year prevalent cases estimated using the method and the number of 7-year prevalent cases observed in HMRN for males of Hodgkin lymphoma

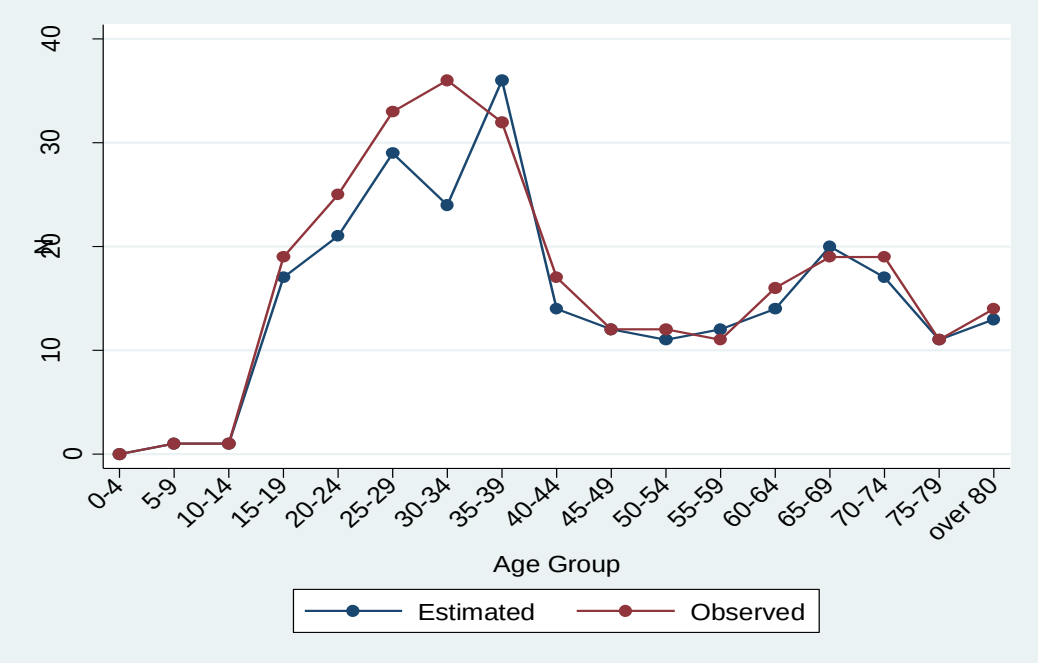


Figure 4- 24 The number of 7-year prevalent cases estimated using the method and the number of 7-year prevalent cases observed in HMRN for females of Hodgkin lymphoma

4.3.4 Total prevalence of all subtypes of haematological malignancies

Age- specific R-values for every subtype were calculated following the general method, and age-specific total prevalence was calculated as the ratio of observed prevalence over R. The basic calculation of incidence and survival for all subtypes were shown in Appendix A5. Age specific incidence was simply estimated using a non-parametric regression spline regardless of the distributions. Survival probabilities were estimated using the parametric model. However, the effect of age was difficult to model for some subtypes. Figure 4- 25 shows the log relative hazard of age at diagnosis for all subtypes. Age appeared to have an overall relevant effect on hazard.

Most subtypes that occur in adulthood showed an approximate linear age effect on hazard. The non-linear effects can be found in subtypes that occur at any age, such as acute myeloid leukaemia and acute lymphoblastic leukaemia. For Hodgkin lymphoma with bimodal age distribution, age showed more complicated effects on survival. Wide confidence interval appeared in the subtypes with small numbers of cases, such as hairy cell leukaemia, T-cell leukaemia, and Burkitt lymphoma. This indicated that they were too small in numbers to be able to show a clear effect therefore the results of prevalence may be not robust. Furthermore, for some subtypes, the small number of diagnoses in early age groups usually resulted in a wider confidence interval than in older ones. All this may suggest a lower precision for the model and the results.

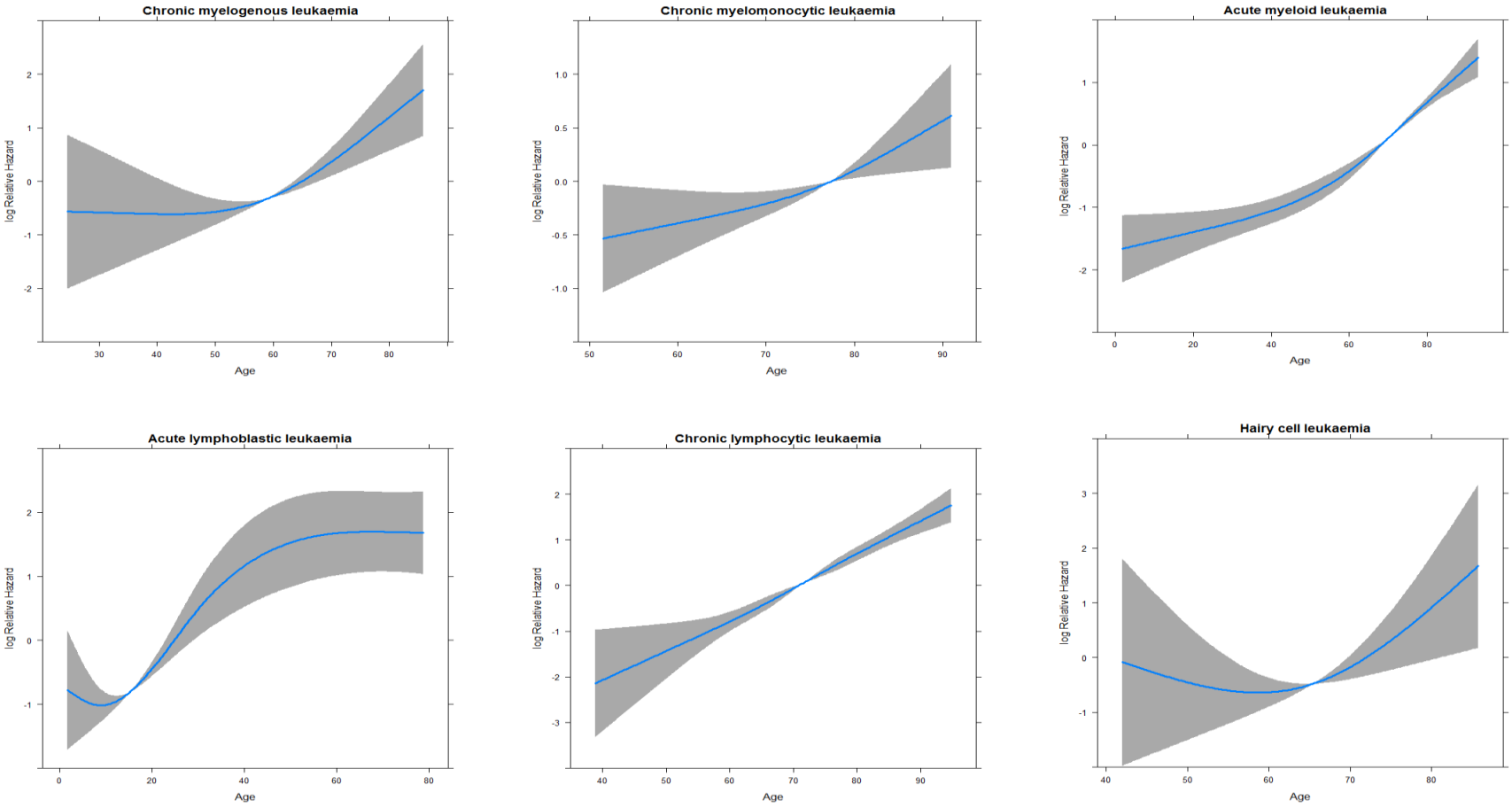


Figure 4- 25 Fitted age effects with confidence interval

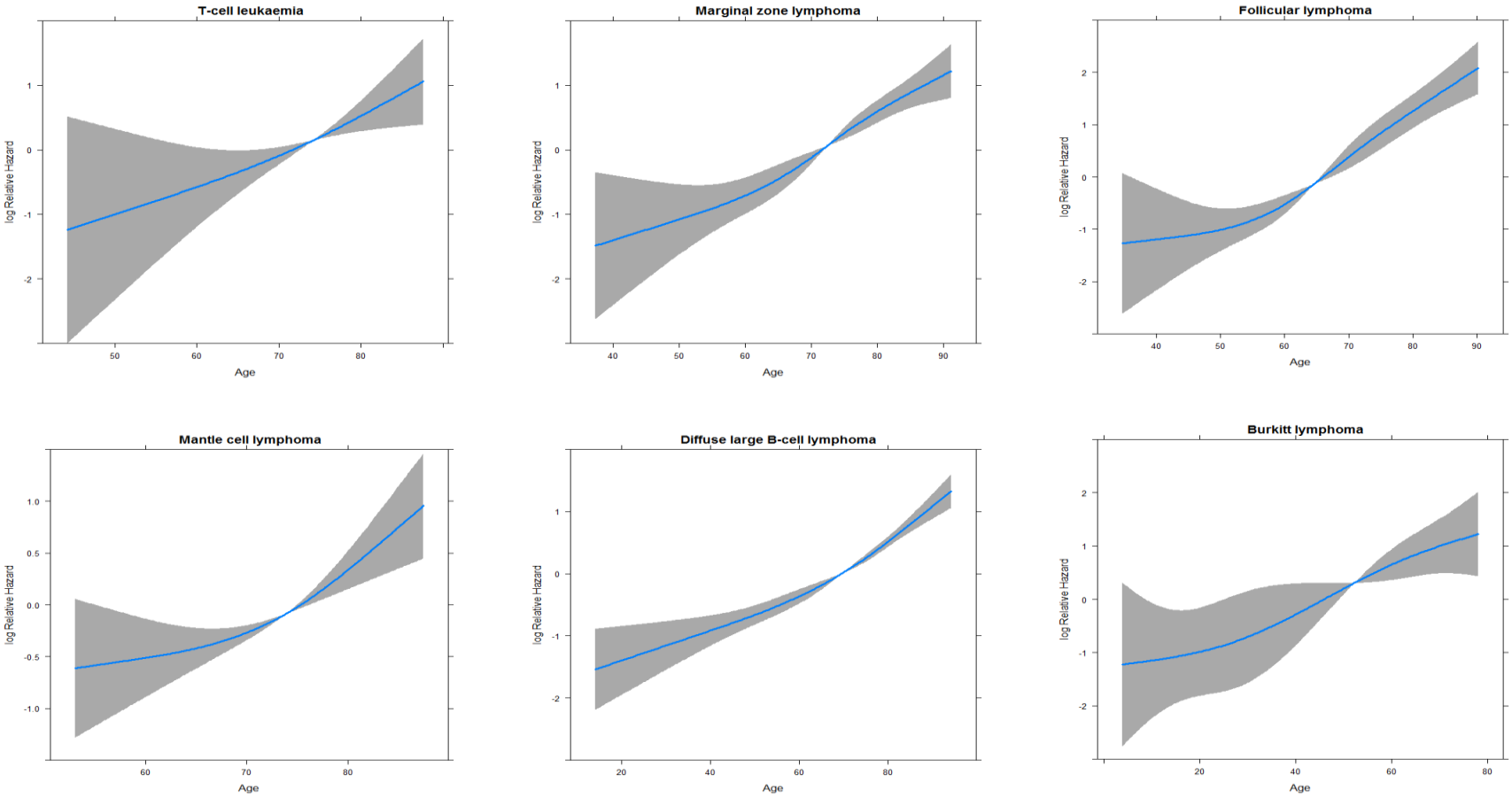


Figure 4- 25 Continued

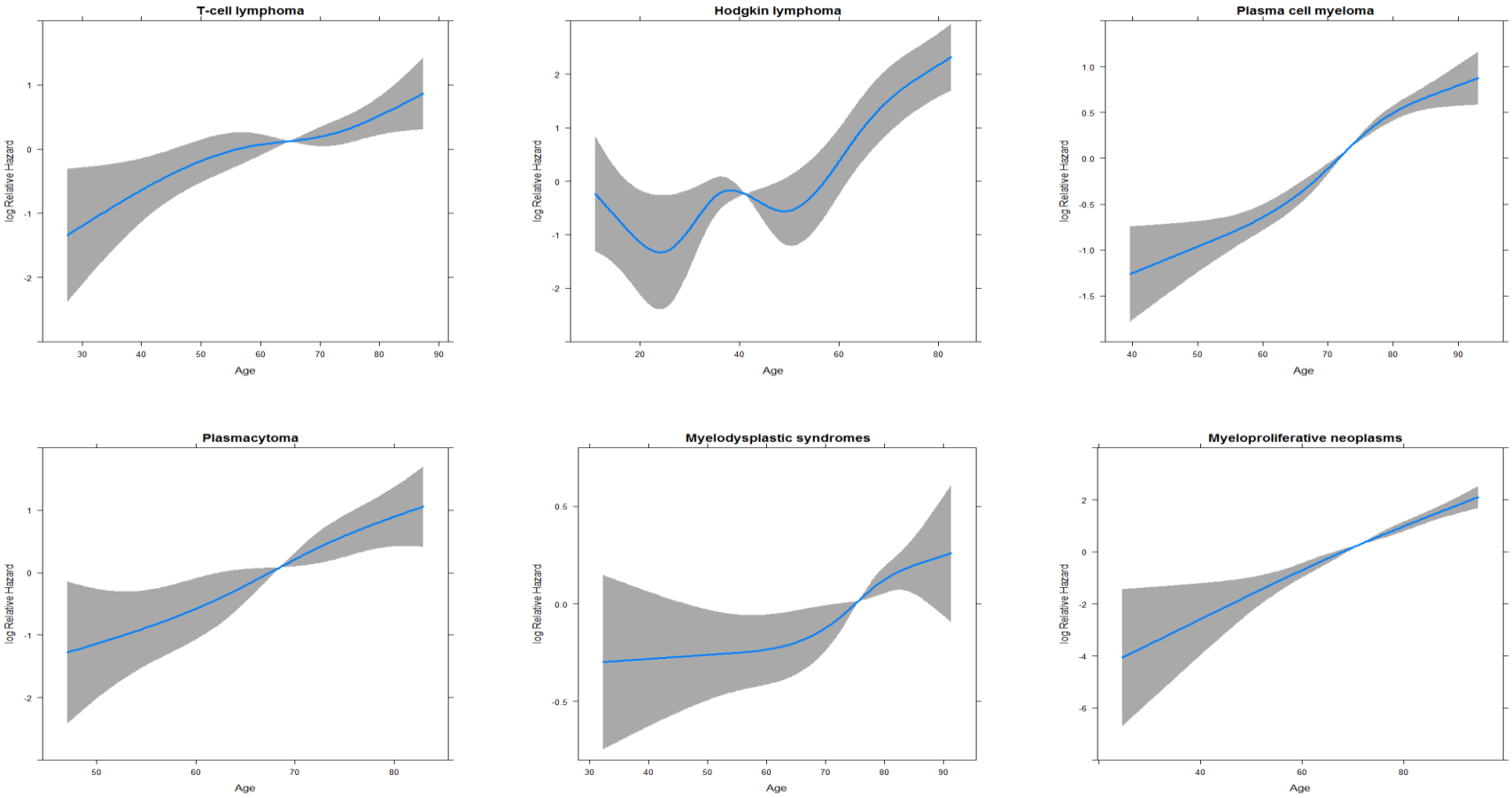
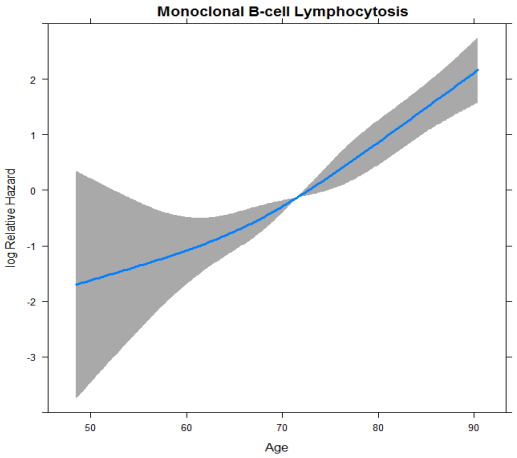
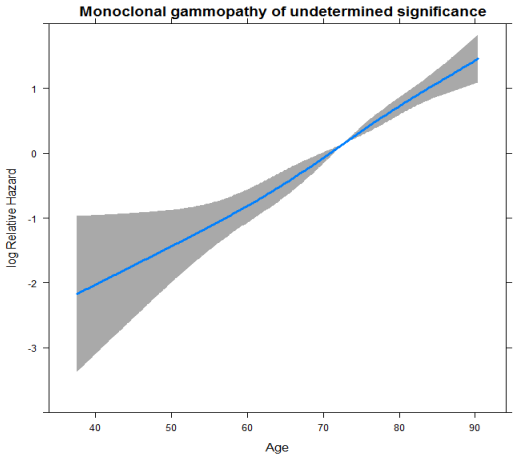


Figure 4- 25 Continued





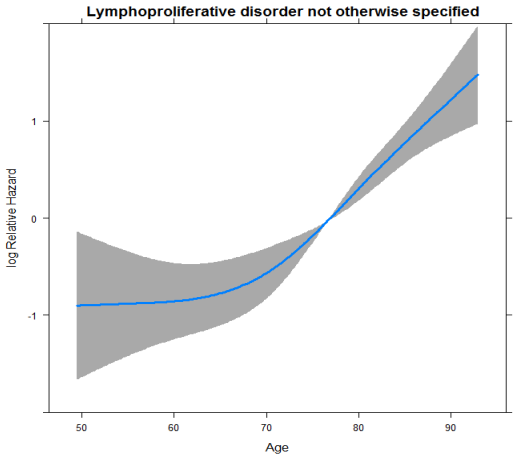


Figure 4-25 Continued

For all subtypes combined, observed prevalence made up about half of the total prevalence. Table 4- 13 presents the completeness index, observed prevalence, and total prevalence (per 100,000) for all haematological malignancies under WHO classification. The five subtypes with the greatest prevalence rate were Hodgkin lymphoma, monoclonal gammopathy of undetermined significance, myeloproliferative neoplasms, chronic lymphocytic leukaemia, and diffuse large B-cell lymphoma, comprising 60% of haematological malignancies survivors in HMRN area. They varied in rank order between observed prevalence and total prevalence estimates.

The proportion of observed prevalence to total prevalence ranged from 0.2 for Hodgkin lymphoma to nearly 0.9 for mantle cell lymphoma. Compared to subtypes with typically short survival duration, subtypes of greater survival duration tended to have greater difference between observed and total prevalence. For example, according to Table 4-2, acute myeloid leukaemia, mantle cell lymphoma, chronic myelomonocytic leukaemia, and myelodysplastic syndromes were associated with poor survival, and in these data, the completeness indexes were higher, with 0.83, 0.90, 0.93, 0.95 respectively. This implied that total prevalence included less than 20% of patients diagnosed before the start of the registry for those subtypes. By comparison, Hodgkin Lymphoma, chronic myelogenous leukaemia, hairy cell leukaemia, and follicular lymphoma had good survival and the completeness indexes were 0.24, 0.39, 0.41, 0.48 respectively.

Subtypes with more cases in childhood usually showed greater differences between observed and total prevalence. For example, although both Hodgkin lymphoma and monoclonal B-cell lymphocytosis had medium incidence (2-5 per 100, 000) and good survival (5-year survival> 70%) (See Section 4.1.3, Table 4-2), the completeness index of monoclonal B-cell lymphocytosis was 50% higher than that of Hodgkin lymphoma. This is because younger patients of Hodgkin lymphoma with good survival resulted in a large amount of cumulative prevalent

cases after middle age, whilst there was no monoclonal B-cell lymphocytosis diagnosed before the age of 35 years. Likewise, total prevalence included more than 60% of patients diagnosed before the start of the registry for Burkitt lymphoma and acute lymphoblastic leukaemia. It was also the reason why acute myeloid leukaemia had a lower completeness index than myelodysplastic syndromes. The completeness index usually increased to some degree, for diseases that could transform from more indolent to aggressive subtypes, for example follicular lymphoma/ diffuse large B-cell lymphoma (0.48 to 0.57), monoclonal B-cell lymphocytosis/ chronic lymphocytic leukaemia (0.50 to0.58), and monoclonal gammopathy of undetermined significance / myeloma (0.49 to 0.79). However, there was a decrease from myelodysplastic syndromes to acute myeloid leukaemia (0.95 to 0.83). This may be because the survival of acute myeloid leukaemia in childhood was relatively good, whilst the diagnosis of myelodysplastic syndromes is rare in younger age groups.

Using these prevalence estimates, the number of prevalent cases in the UK can be estimated due to the similar age and gender structures. Table 4- 14 lists the observed and total prevalence of the top five most common haematological malignancies in the UK. This analysis demonstrated that relying on observed prevalence alone would result in a significant underestimation of the relative burden of some diseases such as Hodgkin lymphoma. It identified Hodgkin lymphoma as only ranking as 6^{th} and 8^{th} most prevalent of all heamatological malignancies amongst men and women in the UK, whereas total prevalence calculation in this data would present it as second for both genders. In other words, compared with observed prevalence, the relative contribution of Hodgkin lymphoma increased when longer prevalence periods were considered. Differences between observed prevalence and total prevalence estimates also pushed chronic myelogenous leukemia, acute lymphoblastic leukemia, monoclonal B-cell lymphocytosis, and lymphoproliferative disorder not otherwise specified, slightly up in rank. It indicated that observed prevalence only cannot show disease burden correctly, whilst total prevalence was a better guide to inform population needs.

Table 4- 13 Observed and total prevalence (per 100 000) for males, females, and total in HMRN on the index date of 31st, August 2011

	Total			Male			Female		
	R	Observed	Total	R	Observed	Total	R	Observed	Total
Total	0.51	281.9	548.8	0.54	318	587.7	0.48	248.1	512.3
Leukaemia	0.55	60.9	111.3	0.55	76.6	138.8	0.54	46.2	85.5
Chronic myelogenous leukaemia	0.39	5.8	14.7	0.42	7.2	17.1	0.36	4.5	12.5
Chronic myelomonocytic leukaemia	0.93	1.8	1.9	0.95	2.1	2.2	0.91	1.5	1.7
Acute myeloid leukaemia	0.83	7.9	9.6	0.88	9	10.2	0.77	6.9	9
Acute lymphoblastic leukaemia	0.39	5.6	14.5	0.41	6.8	16.5	0.35	4.5	12.6
Chronic lymphocytic leukaemia	0.58	35.9	62.1	0.57	46.5	81.3	0.59	25.9	44.1
Hairy cell leukaemia	0.41	2	4.9	0.39	3.4	8.6	0.54	0.8	1.5
T-cell leukaemia	0.53	1.9	3.6	0.58	1.7	3	0.51	2.1	4.2
Non-Hodgkin Lymphoma	0.55	74.7	136.9	0.55	81.3	147.4	0.54	68.4	127.1
Marginal zone lymphoma	0.59	17.1	28.9	0.59	19	32.1	0.59	15.3	26
Follicular lymphoma	0.48	18.5	38.5	0.53	17.6	33.4	0.45	19.3	43.3
Mantle cell lymphoma	0.9	2.7	3	0.89	3.9	4.3	0.93	1.6	1.8
Diffuse large B-cell lymphoma	0.57	31.5	55.1	0.56	34.5	61.2	0.58	28.7	49.4
Burkitt lymphoma	0.29	1.4	4.8	0.26	2.2	8.3	0.41	0.6	1.5
T-cell lymphoma	0.54	3.6	6.6	0.53	4.3	8.1	0.56	2.9	5.2
Hodgkin Lymphoma	0.24	17.3	72.4	0.27	19.8	73.3	0.21	15	71.5
Myeloma	0.79	23.8	30.1	0.78	29.4	37.5	0.8	18.6	23.1
Plasma cell myeloma	0.8	21.3	26.5	0.79	25.7	32.4	0.82	17.2	21
Plasmacytoma	0.71	2.5	3.5	0.72	3.7	5.1	0.69	1.4	2.1
Myelodysplastic syndromes	0.95	9.5	10	0.95	12.4	13	0.94	6.8	7.3
Other Neoplasms of Uncertain or Unknown Behaviour	0.51	95.7	188.1	0.55	98.5	177.7	0.47	93	197.9
Myeloproliferative neoplasms	0.53	35.4	67.2	0.51	32.4	63	0.54	38.2	71.2
Monoclonal B-cell Lymphocytosis	0.5	16.5	32.9	0.58	18.2	31.3	0.43	14.9	34.5
Monoclonal gammopathy of undetermined significance	0.49	35.1	72	0.59	37.7	63.9	0.41	32.7	79.5
Lymphoproliferative disorder not otherwise specified	0.54	8.7	16	0.52	10.2	19.5	0.57	7.2	12.6

Table 4- 14 Comparison of observed (7-year) and total prevalence of the top 5 haematological malignancies by gender*

Observed		Total	
Disease	Prevalence	Disease	Prevalence
Male		Male	
Chronic lymphocytic leukaemia	13,300	Chronic lymphocytic leukaemia	23,222
Monoclonal gammopathy of undetermined significance	10,772	Hodgkin Lymphoma	20,950
Diffuse large B-cell lymphoma	9,847	Monoclonal gammopathy of undetermined significance	18,274
Myeloproliferative neoplasms	9,268	Myeloproliferative neoplasms	18,007
Plasma cell myeloma	7,352	Diffuse large B-cell lymphoma	17,483
Female		Female	
Myeloproliferative neoplasms	11,536	Monoclonal gammopathy of undetermined significance	24,020
Monoclonal gammopathy of undetermined significance	9,878	Hodgkin Lymphoma	21,608
Diffuse large B-cell lymphoma	8,664	Myeloproliferative neoplasms	21,515
Chronic lymphocytic leukaemia	7,827	Diffuse large B-cell lymphoma	14,924
Follicular lymphoma	5,825	Chronic lymphocytic leukaemia	13,316

*Total prevalent cases in the UK for all other subtypes are shown in Appendix A4

Figure 4- 26 and Figure 4- 27 illustrated the differences in observed prevalence (blue bars) and total prevalence (red bars- additional cases added by observed prevalence) for all subtypes in the UK, ranked in order of descending total prevalence.

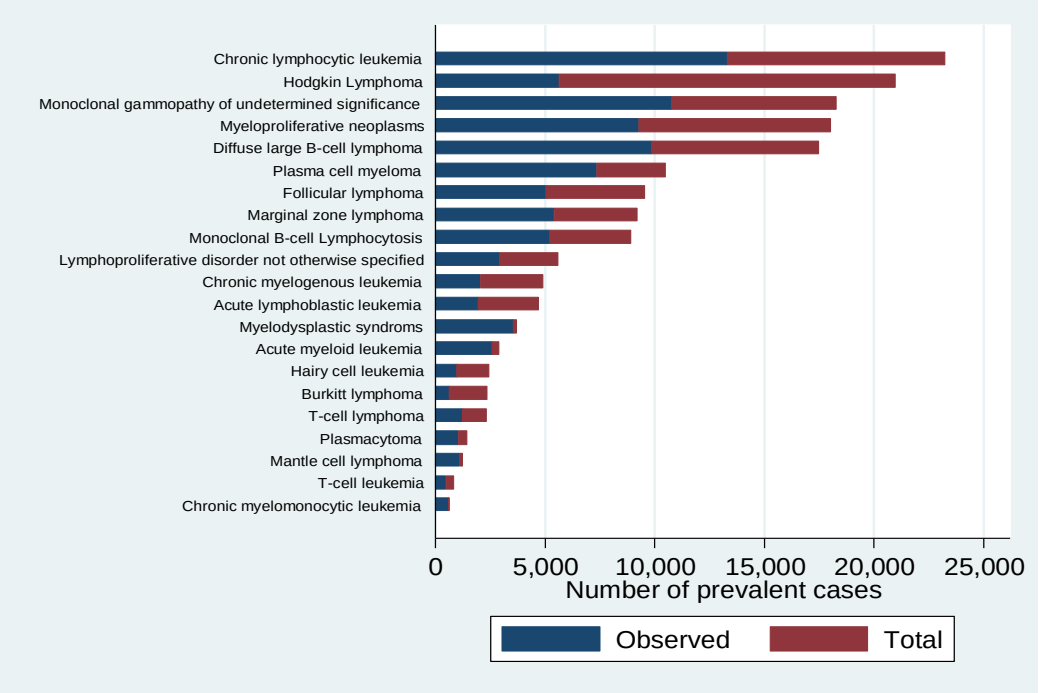


Figure 4- 26 Observed and total prevalence cases for males in the UK on 31st, August 2011

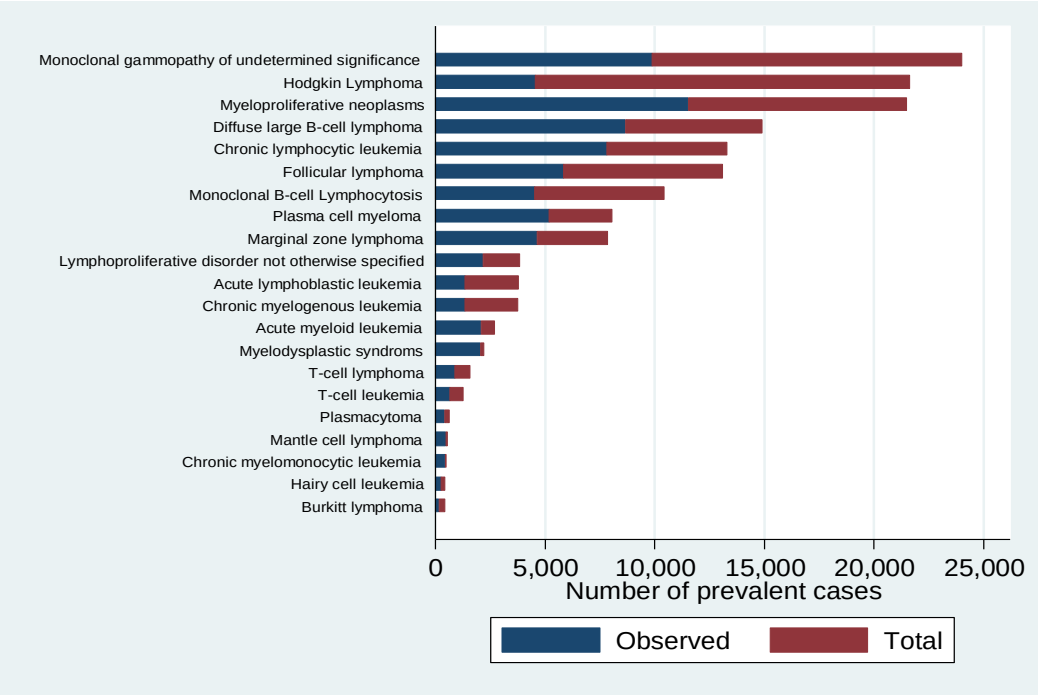


Figure 4- 27 Observed and total prevalence cases for females in the UK on 31st, August 2011

4.4 Total prevalence range

4.4.1 Chronic Myelogenous Leukaemia (CML)

4.4.1.1 Diagnostic characteristics of CML

CML was used as an example to show the calculation process details in this section. It is very rare in children, however according to the data in HMRN incidence increases with age. Patients diagnosed with CML had a median age of 59.0 years (range from 15.1 to 94.7 years). There was a male predominance, and men had a higher incidence of CML than women in nearly all age groups (see Figure 4-28). Crude incidence of CML by age and gender (per 100,000 population) was shown in Table 4-15.

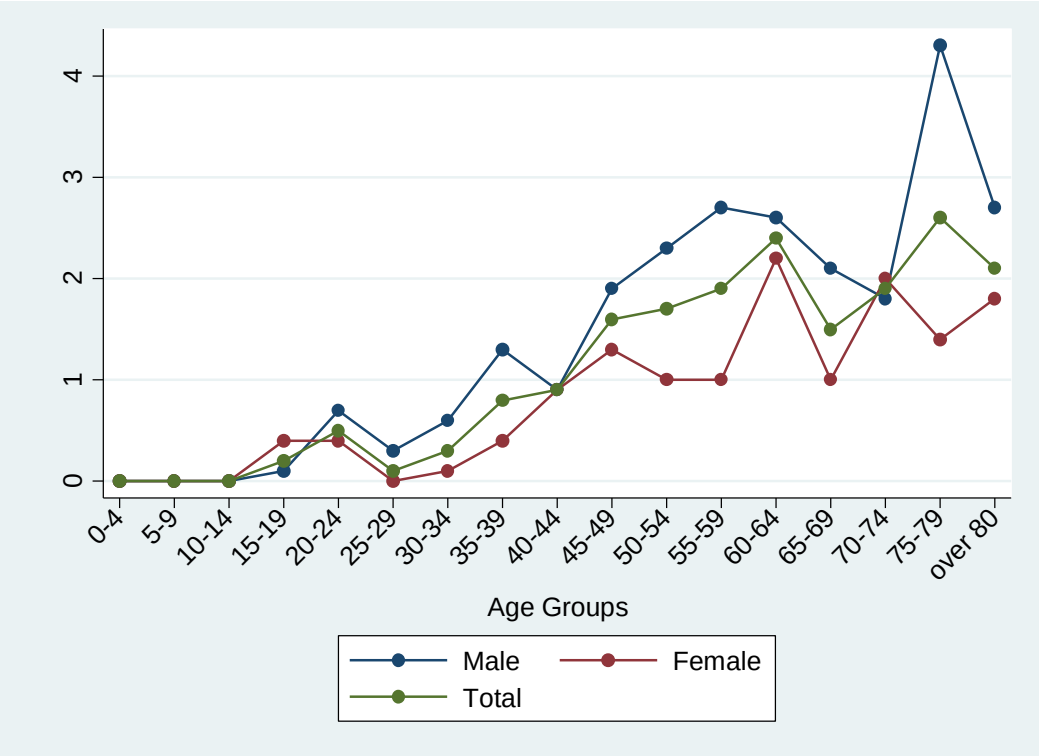


Figure 4- 28 Incidence of CML per 100,000 for males, females, and total

Table 4- 15 Crude incidence of CML by age and gender (per 100,000 population)

Age group	Total		Male		Female	
(Years)	N	Incidence	N	Incidence	N	Incidence
0-4	0	0	0	0	0	0
05-Sep	0	0	0	0	0	0
Oct-14	0	0	0	0	0	0
15-19	4	0.2	1	0.1	3	0.4
20-24	8	0.5	5	0.7	3	0.4
25-29	2	0.1	2	0.3	0	0
30-34	6	0.3	5	0.6	1	0.1
35-39	16	0.8	12	1.3	4	0.4
40-44	16	0.9	8	0.9	8	0.9
45-49	25	1.6	15	1.9	10	1.3
50-54	29	1.7	20	2.3	9	1
55-59	26	1.9	19	2.7	7	1
60-64	30	2.4	16	2.6	14	2.2
65-69	17	1.5	11	2.1	6	1
70-74	19	1.9	8	1.8	11	2
75-79	22	2.6	15	4.3	7	1.4
Over 80	22	2.1	9	2.7	13	1.8
Total	242	1.0	146	1.2	96	0.7

The survival of CML is shown in Figure 4-29. There were only 36 deaths and the median age at time of death was 74.2 years (range 25.1 to 92.6 years). There were no differences between males and females in terms of survival (logrank test: p=0.972).

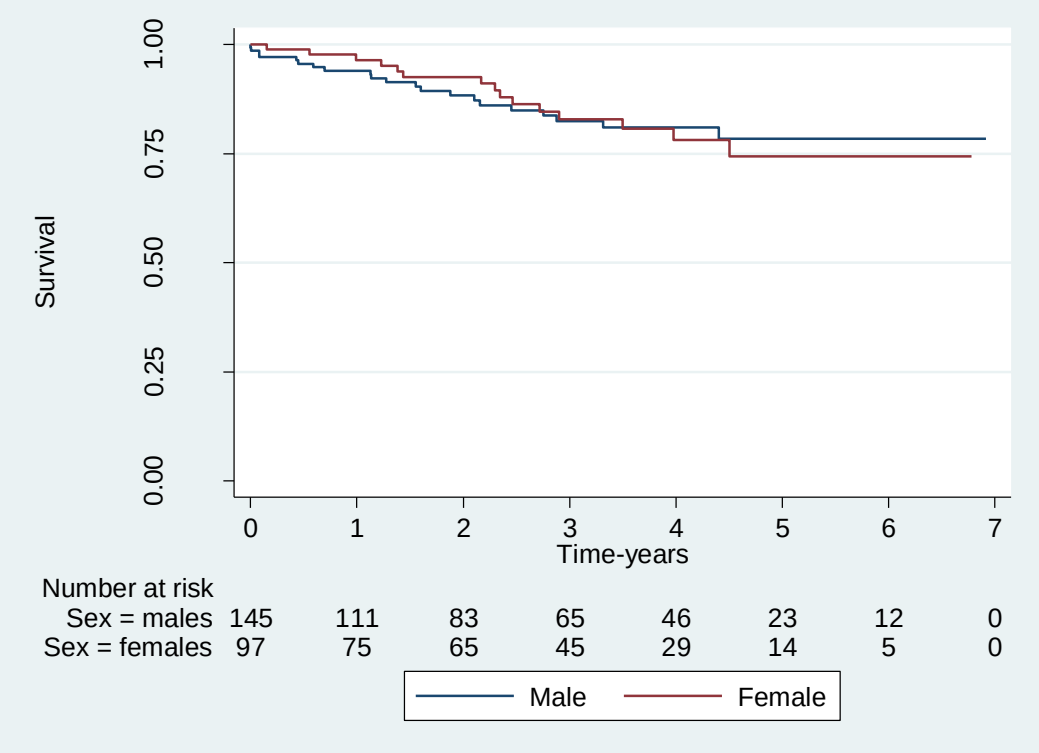


Figure 4- 29 Kaplan-Meier survival estimates for CML patients in HMRN by gender

4.4.1.2 Choosing "T" for CML

The treatment of CML has experienced dramatic progress in recent years. The previous therapies for CML consisted of interferon alpha based treatments (IFN-α), hemopoietic stem cell transplantation (HSCT), and simple cell reduction treatment with hydroxyurea (HU). However, the introduction of tyrosine kinase inhibitors (TKIs) has proved to be highly effective in the treatment of CML (Hehlmann, Hochhaus, and Baccarani, 2007) (see Figure 4-30). The first clinical

trial of imatinib (a kind of TKIs) was conducted in CML patients in 1998, and was approved by the Food and Drug Administration in 2001. Thereafter it rapidly became considered as front-line therapy for CML (Wang, et al., 2010; Wiggins, et al., 2010). Recently, two additional novel kinase inhibitors, dasatinib and nilotinib have become available as treatment options for patients who have developed resistance to, or those who have shown intolerance to imatinib (Hehlmann, Hochhaus, and Baccarani, 2007).

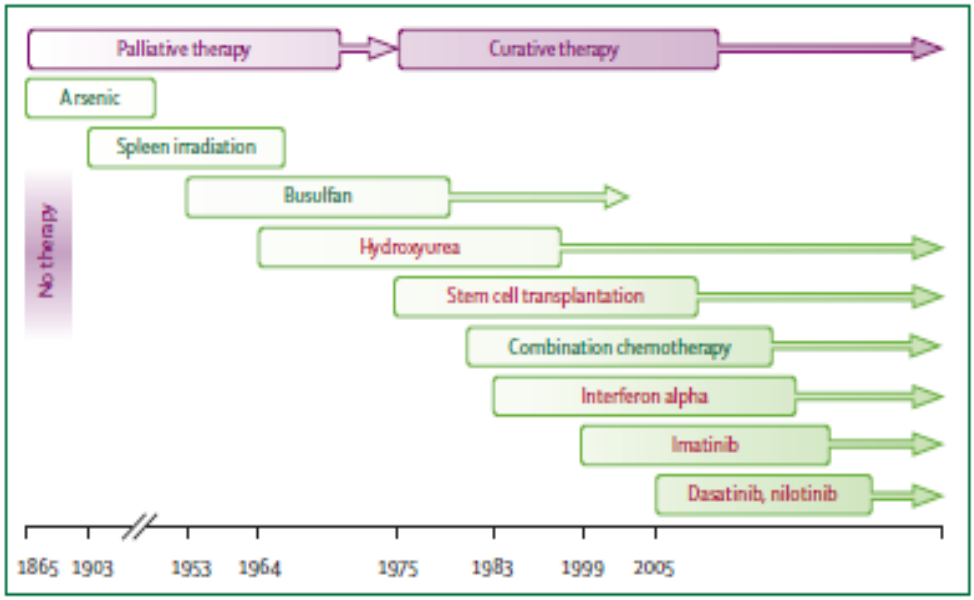


Figure 4- 30 Development of treatment for CML (Hehlmann, Hochhaus, and Baccarani, 2007)

In this study, the registry data began in 2004, which means that all observed cases are in the TKIs treatment era. The period of time covered by HMRN data, which is used for estimation of survival trends, was not sufficiently adequate to correctly estimate the survival trends in the past for CML. So, for CML, T-year prevalence is 10-year prevalence ($T = 10$), since the new treatment TKIs had been used in clinical practice in the UK for 10 years up to the index date.

4.4.1.2 Total prevalence range for CML

Figure 4-31 showed in one graph the completeness index R and R_T for both "total prevalence" and 10-year prevalence. They had a similar decline pattern with age, since they were calculated using the same parameters of incidence, survival, and general mortality. Generally, R_T was higher than R.

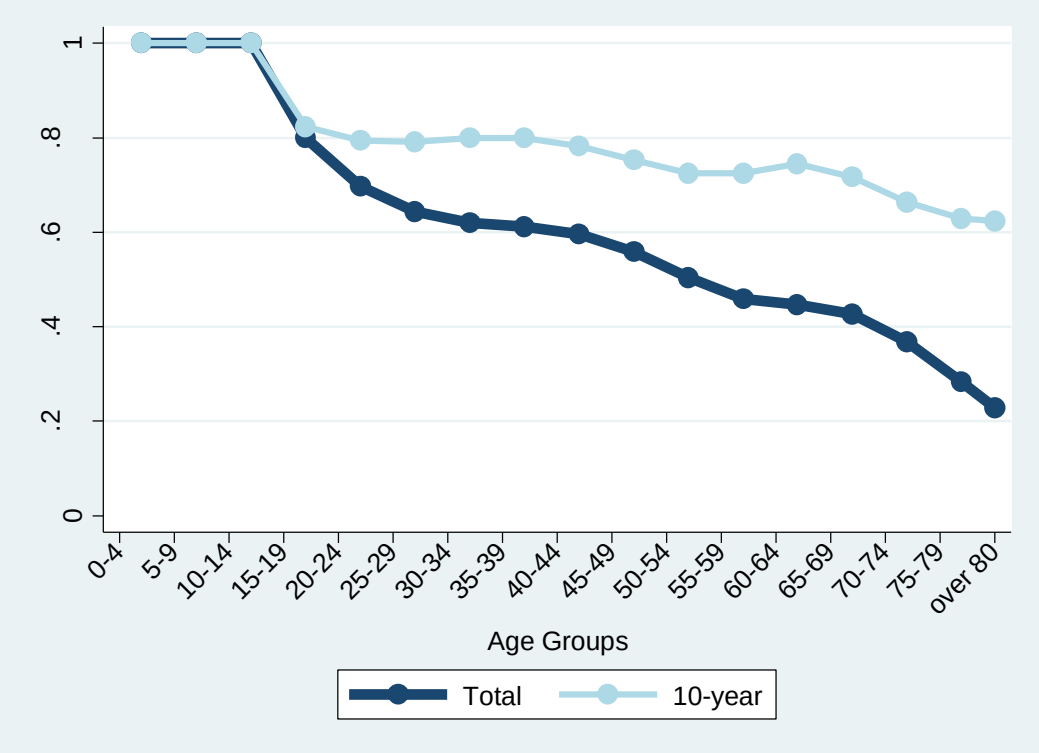


Figure 4- 31 Completeness index to calculate "total prevalence" and 10-year prevalence of CML for men

Total prevalence was 17.1 per 100,000 and 10-year prevalence was 10.0 per 100,000 for males. The process of calculation was shown in Table 4-16.

The total prevalence range of CML was 10.0-17.1 per 100,000 for males. The real complete prevalence may approach the lower limit of the range. This is because

compared to those diagnosed and actually observed within the registry, one could expect that fewer people diagnosed in the old treatment era were alive on the index date (since most of patients in the old treatment era may have died under the poor survival and prognosis). Thus, the prevalent patients who were diagnosed after the application of a new drug account for the majority of prevalent cases, and the "total prevalence's" adjustment for those who were diagnosed before the application of new treatment will be limited.

Table 4- 16 Total prevalence and 10-year prevalence of CML by age group for men (per 100,000)

Age group (Years)	N_o	"Total"		10-year	
		R	N_t	R_T	N_{10}
0-4	0	1	0	1	0
5-9	0	1	0	1	0
10-14	0	1	0	1	0
15-19	1	0.799403	1	0.823032	1
20-24	3	0.697204	4	0.794594	4
25-29	3	0.643452	5	0.791698	4
30-34	2	0.619998	3	0.799097	3
35-39	7	0.611804	11	0.799761	9
40-44	9	0.595212	15	0.781837	12
45-49	10	0.55898	18	0.75344	13
50-54	15	0.503745	30	0.725226	21
55-59	17	0.45856	37	0.725046	23
60-64	12	0.447061	27	0.745082	16
65-69	16	0.426584	38	0.716842	22
70-74	8	0.367705	22	0.663847	12
75-79	8	0.28374	28	0.629169	13
over 80	13	0.229152	57	0.624437	21
Total	124	0.42	296	0.72	173
Prevalence	7.2		17.1		10.0

No: number of observed prevalent cases Nt: number of total prevalent cases
N_{10}: number 10-year prevalent cases
R: completeness index for "total prevalence"
R_T : completeness index for 10-year prevalence

4.4.2 Total prevalence range of some other subtypes of haematological malignancies

In this section, total prevalence ranges of some other subtypes of haematological malignancies (myeloma, Hodgkin lymphoma, and ALL) are presented.

Myeloma is a neoplasm of plasma cells. Its survival characteristics and treatments have changed and developed in recent decades. Melphalan, introduced in the 1960s, improved the poor survival (median survival was less than a year) of myeloma patients (Alexanian, et al., (1968)). In the 1980s, high dose chemotherapy and autologous stem cell transplant (ASCT) was given to patients, and it again improved survival. However, the median duration of response after ASCT does not exceed 3 years, and relapse of disease is common in patients (Attal, et al., 2006). In 1999, the introduction of immunnodulatory drugs (thalidomide and lenalidomide) and proteasome inhibitor (bortezomib) represented major milestones in the treatment of myeloma (Attal, et al., 2006; Kumar, et al., 2008). It is believed that the duration of response is prolonged and salvages relapsed disease. The survival changed from the cut-off of year 1999, contemporaneous with the availability of the new drug (Attal, et al., 2006; Kumar, et al., 2008). 12-year prevalence is used as the lower limit of total prevalence range for myeloma.

The use of new more effective therapies such as new types of chemotherapy (such as mechlorethamine, vincristine, procarbazine, and prednisone (MOPP); doxorubicin, bleomycin, and vinblastine (ABV) [DeVita, Serpick & Carbone, 1970; Ferm é, et al., 2007]) became widespread for Hodgkin lymphoma, and have improved survival for the past decades (Capocaccia, et al., 2002). The great improvements in treatment took place in the 1960s and 1970s, which decreased the mortality of Hodgkin lymphoma by about over two thirds (Swerdlow, et al., 2001; Levi, et al., 2002; Swerdlow, 2003). Therefore in this study, 40-year prevalence is estimated as the lower limit of total prevalence range for Hodgkin lymphoma.

ALL is shown as an example to demonstrate long period prevalence and total prevalence. The survival of ALL changed drastically around 1960 due to the introduction of innovative treatments (Mauer and Simone 1976; Simonetti, et al., 2008). The improvement of treatment was for children and not adult (Pui ,Campana, and Evans, 2001; Pui and Evans, 2006; Simonetti et al., 2008). Nowadays, the treatment of ALL includes chemotherapy, steroids, radiation therapy, and intensive combined treatments (including bone marrow or stem cell transplants) (The Mount Sinai Hospital, 2012). 50-year prevalence is calculated for ALL, and used as the lower limit of its total prevalence range. Since 50 years is a long period that may cover most of the patients alive on the index date, the total prevalence range will be narrow. This is because the patients diagnosed 50 years earlier have a high probability of death before the index date. It worth noting that although the improvements of treatment for ALL were only for children (Pui ,Campana, and Evans, 2001; Pui and Evans, 2006; Simonetti et al., 2008), it is considered it had the effects on all age groups. This is because the model in this section ignores the differences of the survival improvement among different age groups, and fortunately, most of the cases of ALL occur in childhood which brings less bias to the estimates.

Table 4-17 shows the estimated total prevalence range for CML, myeloma, Hodgkin lymphoma, and ALL, by gender. The survival changed greatly recently for CML (10 years before index date) and myeloma (12 years before index date), while, much longer ago for Hodgkin lymphoma and ALL (about 40 and 50 years respectively before the index date). Amongst them, 40-year prevalence was estimated as the lower limit of total prevalence range for Hodgkin lymphoma. This was also the reason why the sensitivity analysis based on Hodgkin lymphoma (Section 4.3) was only a theoretical analysis. If the length of the registry were longer than 40 years, the fixed parameters of the survival model for Hodgkin lymphoma could not reflect the truth, and the completeness index might have gone up even faster due to the high mortality rate of the disease.

Consistent with expectation, the later the new treatment appeared, the wider the range was. When the new treatment was introduced more than 50 years ago for ALL, the T-year prevalence was being predicted as close to "total prevalence". This is because, according to the incidence and survival of ALL, as well as general mortality in the population, the probability that a patient diagnosed with ALL 50 years ago being still alive was very low. This means that a 50 year period may cover nearly all ALL patients who are alive on the index date. However, the total prevalence ranges of myeloma are also narrow, although the new treatment for myeloma was applied relatively late on (12 years ago). This can be explained by its survival. The maintenance treatment thalidomide improved the survival of myeloma, but did not make it a "curable" disease that has a good prognosis (Attal, et al., 2006; Kumar, et al., 2008). The survival of myeloma shown in Appendix A5 indicated the disease duration. The survivor function declined to about 0.25 within the registry period. It was therefore reasonable to imagine that most patients diagnosed earlier than the introduction of thalidomide would not be able to live up to the index date, due to even poorer survival in the past.

Table 4- 17 Total prevalence and T-year prevalence for chronic myelogenous leukaemia, myeloma, Hodgkin lymphoma, acute lymphoblastic leukaemia by gender

		T(years)	"Total"*	T-year	Range
Prevalence (per 100,000)					
Chronic myelogenous leukaemia	Male	10	17.1	10.0	10.0-17.1
	Female		12.5	5.8	5.8-12.5
Myeloma	Male	12	37.4	33.4	33.4-37.4
	Female		23.3	21.1	21.1-23.3
Hodgkin lymphoma	Male	40	73.3	66.1	66.1-73.3
	Female		71.5	58.7	58.7-71.5
Acute lymphoblastic leukaemia	Male	50	16.5	16.3	16.3-16.5
	Female		12.6	11.9	11.9-12.6
Prevalent cases in the UK					
Chronic myelogenous leukaemia	Male	10	4,887	2,858	2,858-4,887
	Female		3,776	1,752	1,752-3,776
Myeloma	Male	12	10,689	9,546	9,546-10,689
	Female		7,039	6,375	6,375-7,039
Hodgkin lymphoma	Male	40	20,949	18,892	18,892-20,949
	Female		21,601	17,734	17,734-21,601
Acute lymphoblastic leukaemia	Male	50	4,716	4,659	4,659-4,716
	Female		3,807	3,595	3,595-3,807

T: the number of years from the application of new treatment

* “total prevalence” is calculated by general method (See Section 4.3)

Figure 4-32 shows the estimated prevalence range for the subtypes. The complete prevalence was shown in the light blue area that is composed of "total prevalence" and T –year prevalence. The ranges were wide for CML and Hodgkin lymphoma for different reasons: there was a great improvement in survival for CML but a larger number of diagnoses for Hodgkin lymphoma. The ranges were also narrow for myeloma and ALL for different reasons: a less significant change in survival for myeloma but a much earlier appearance of new treatment for ALL.

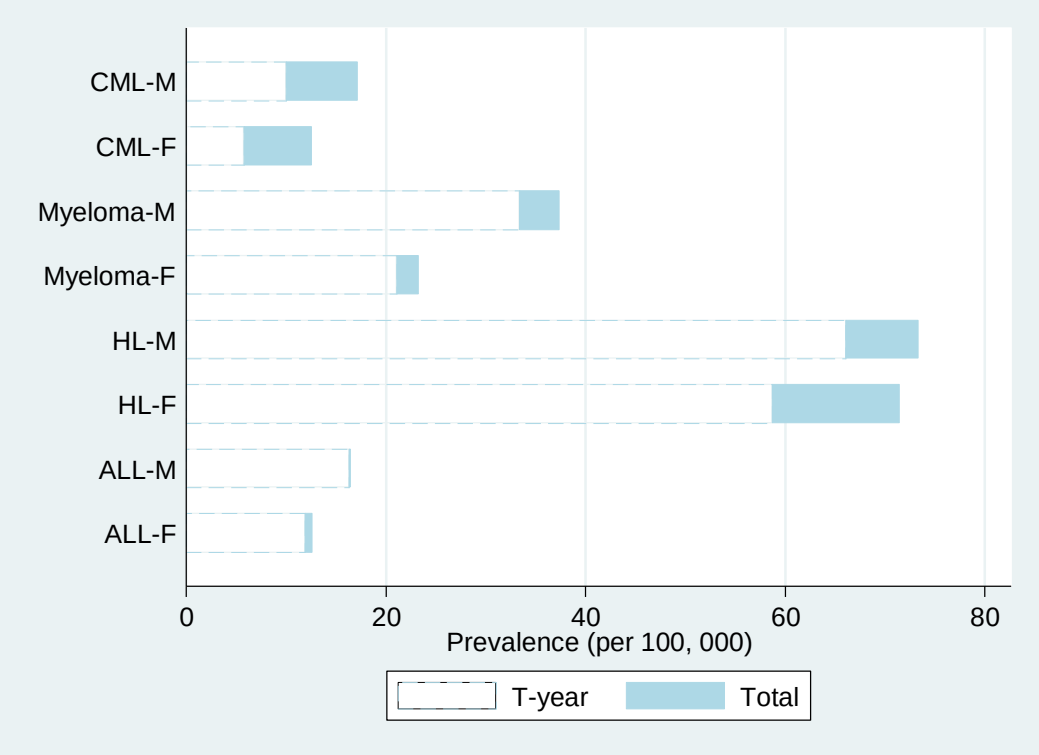


Figure 4- 32 Prevalence range for the subtypes (per 100,000) (CML: chronic myelogenous leukaemia, HL: Hodgkin lymphoma, ALL: acute lymphoblastic leukaemia, M: males, F: females)

As described above, if one considers that the complete prevalence approaches the lower limit of the range, the ranks of subtypes may vary slightly. Figure 4-33 and Figure 4-34 show the total prevalent cases for males and females in the UK on

31st, August 2011. The maximum estimates in the figures were obtained using the general method, and the minimum estimates were derived from the "T-year prevalence". Figures were sorted according to the minimum estimates. Details of the values can be found in Table 4-18, which showed the estimated counts of observed and total prevalence/ total prevalence ranges, ranking in order of descending total prevalence/ total prevalence range for both genders combined. Compared to the results in section 4.3, the ranks were pushed down a little for the subtypes with survival changed greatly in the past. For example, the rank of Hodgkin lymphoma (for the two genders combined) dropped from first to third after introducing the total prevalence range.

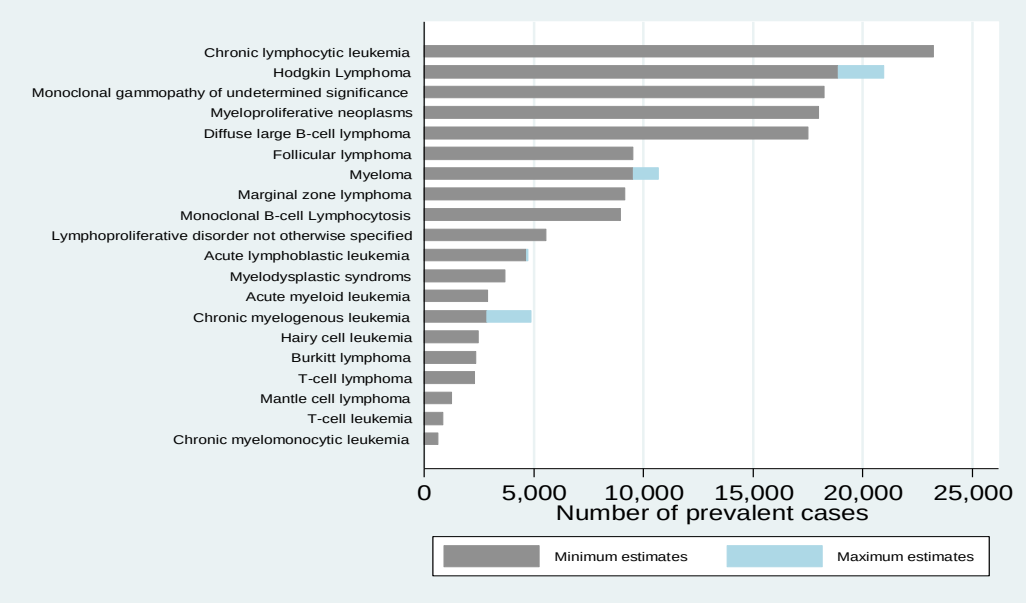


Figure 4- 33 Total prevalent cases for males in the UK on 31st, August 2011 (the maximum estimates were obtained using the general method, and the minimum estimates were derived from "T-year prevalence") sorted according to the minimum estimates

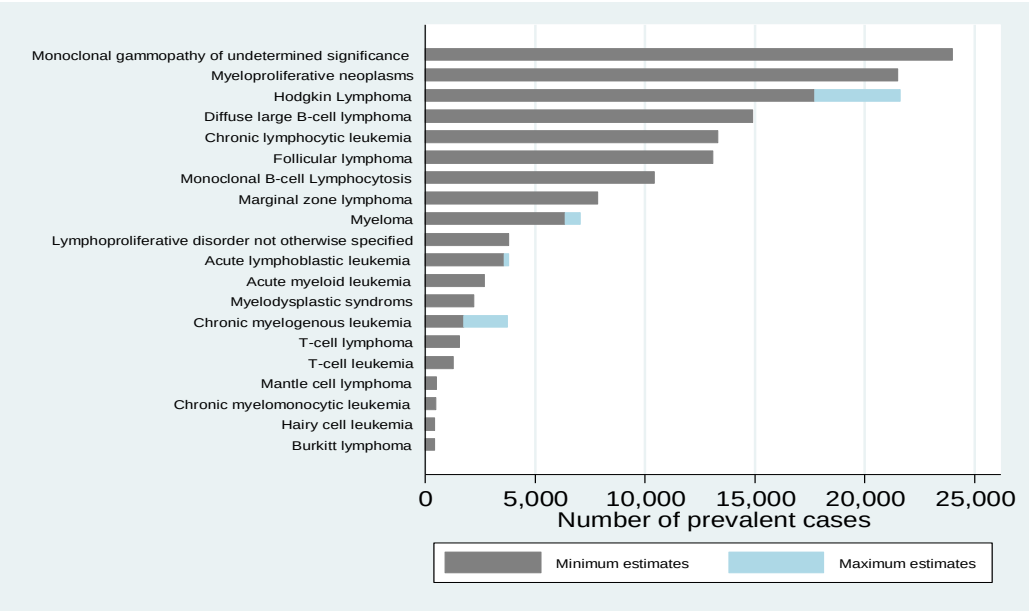


Figure 4- 34 Total prevalent cases for females in the UK on 31st, August 2011 (the maximum estimates were obtained using the general method, and the minimum estimates were derived from "T-year prevalence") sorted according to the minimum estimates

Table 4- 18 The estimated counts of observed and total prevalence /range in the UK on 31st, August 2011, ranked in order of descending total prevalence for both genders.

	Total		Male		Female	
	Observed	Total/Range	Observed	Total/Range	Observed	Total/Range
Monoclonal gammopathy of undetermined significance	20,645	42,310	10,772	18,274	9,878	24,020
Myeloproliferative neoplasms	20,810	39,530	9,268	18,007	11,536	21,515
Hodgkin Lymphoma	10,191	36,626-42,550	5,650	18,892-20,949	4,545	17,734-21,601
Chronic lymphocytic leukaemia	21,106	36,500	13,300	23,222	7,827	13,316
Diffuse large B-cell lymphoma	18,505	32,396	9,847	17,483	8,664	14,924
Follicular lymphoma	10,849	22,641	5,022	9,549	5,825	13,082
Monoclonal B-cell Lymphocytosis	9,713	19,364	5,204	8,937	4,513	10,424
Marginal zone lymphoma	10,043	17,018	5,419	9,178	4,627	7,847
Myeloma	14,005	15,921-17,728	8,404	9,546-10,689	5,612	6,375-7,039
Lymphoproliferative disorder not otherwise specified	5,087	9,380	2,908	5,568	2,182	3,819
Acute lymphoblastic leukaemia	3,276	8,254-8,523	1,933	4,659-4,716	1,346	3,595-3,807
Myelodysplastic syndromes	5,598	5,904	3,536	3,704	2,068	2,205
Acute myeloid leukaemia	4,643	5,617	2,561	2,901	2,084	2,716
Chronic myelogenous leukaemia	3,391	4,610-8,663	2,049	2,858-4,887	1,346	1,752-3,776
T-cell lymphoma	2,091	3,866	1,223	2,307	870	1,562
Hairy cell leukaemia	1,202	2,903	958	2,462	246	448
Burkitt lymphoma	807	2,813	628	2,378	181	442
T-cell leukaemia	1,136	2,124	496	858	640	1,265
Mantle cell lymphoma	1,597	1,768	1,107	1,240	492	530
Chronic myelomonocytic leukaemia	1,054	1,132	595	627	459	505

Chapter 5 Discussion

5.1 Conclusion and main findings

5.1.1 Key findings and conclusion

This thesis is the first time that prevalence for haematological malignancies has been calculated using current disease classification (ICD-O-3). The methods used estimated that around 19,700 people in HMRN region are living with a prior diagnosis of a haematological malignancy; this equates to about 327,800 people in the UK. After calculating total prevalence, the top five prevalent subtypes, monoclonal gammopathy of undetermined significance, myeloproliferative neoplasms, Hodgkin lymphoma, chronic lymphocytic leukaemia, and diffuse large B-cell lymphoma were found to comprise about 60% of prevalent haematological malignancies in HMRN area.

The importance of estimating "total" prevalence instead of observed prevalence was evident for some subtypes. In this thesis using HMRN data provided an estimate of haematological malignancy prevalence that is about 95% (with completeness index of 0.51) greater than observed prevalence for all subtypes combined. Out of all the diagnoses, about 9,600 cases were diagnosed before the establishment of HMRN registry.

The relative burden presented by each subtype does not parallel that of observed prevalence. Consistent with expectations, the greatest differences between total prevalence and observed prevalence estimates were typically seen in less fatal

cancers that are commonly diagnosed at a younger age. For example, Hodgkin lymphoma typically has a good survival, and total prevalence estimates exceed those of observed prevalence greatly, whilst the difference between observed prevalence and total prevalence is slight for mantle cell lymphoma which has a short survival duration. Subtypes that occurred at an early age also lead to more accumulative prevalent cases. For example, ALL has an incidence peak under the age of five years old, and its completeness index was estimated as low as 0.38. However, this pattern may not be suitable to be applied to those cancers that have experienced a great improvement in treatment. This is because the survival observed in the registry cannot reflect the poor survival rate before the new treatment was introduced. In this study, HMRN started to accrue cases in 2004, therefore changes in treatment that have led to a change in survival cannot be extrapolated from observed data. Total prevalence ranges can help to give some information about those conditions. For CML, for example, the total prevalence range indicated that at least 1,752 units of health resource were needed (such as hospital beds, medicine for diseases, and doctors dedicated to a clinic), and no more than 3,776 units. Therefore the higher limit and lower limit avoided the chance of a resource shortage and surplus for CML. Thus, to some degree, the ranges are instructive for society to meet the population's needs.

5.1.2 Importance of HMRN data

HMRN provides high quality data as described in Chapter Three. The advantages of HMRN can be summarised into three aspects: (1) the confirmation of diagnosis of haematological malignancies, (2) the completeness of data, and (3) the percentage of lost- to- follow- up. Firstly, all cases are ascertained by a centralised laboratory (HMDS) that contains all relevant expertise and technologies to provide an integrated diagnostic service including histology, cytology, immunophenotyping and molecular cytogenetics. This provides the integrated confirmation of diagnosis of haematological malignancies in HMRN. Secondly, a list of newly diagnosed patients is downloaded on a weekly basis, and a group of trained nurses in ECSG abstract clinical data from patients' medical records. They

collect relevant information that includes demographic details, prognostic factors, and treatment and response to treatment for all patients. These data extracted by the ECSG are input into HILIS linking patients' diagnostic information with their clinical data, which makes high completeness of data. The nurses also confirm from the notes that the patients are newly diagnosed, so the diagnosis is truly an incident case. Lastly, like National cancer registries (UKACR, 2013), all cases in HMRN are flagged by the NHS Central Register, so it is known the status of patients, survival or died. With the high quality data in HMRN, prevalence of haematological malignancies can be calculated under WHO classification in this work rather than broad categories. Furthermore, it is not necessary to make adjustment for those lost-to-follow-up, since its percentage is small and can be ignored (Gigli et al., 2006). All in all, high quality data in HMRN is the foundation to estimate more accurate prevalence of haematological malignancies than other previous report (Ferlay, et al., 2010; NORDCAN, 2010; NCIN, 2012; SEER; 2012).

5.1.3 Importance of estimates of prevalence

The basic importance of prevalence estimates is to gain a better understanding of the size of the survivor population who received a diagnosis of a certain disease. The survivors may require treatment, monitoring for recurrence, and screening for other cancers (Capocaccia, et al., 2002). So the estimates of prevalence should be useful to agencies charged with planning for health care, such as the treatment, medical consultation, and long term counselling and support. It helps doctors and cancer care providers to know the cost of diseases management, and aid health resources allocation by governments to improve the quality of life for people with cancer who survive. For haematological malignancies, some previous sporadic reports of prevalence in the literature have not proved to be very useful due to data limitations and the lack of completeness, old broad classification, or a standard methodology (Capocaccia and De Angelis, 1997; Merrill, et al., 2000; Capocaccia, et al., 2002; Verdecchia, et al., 2002; Forman, et al., 2003; Lutz, et al., 2003; Möller, et al., 2003; Gigli, et al., 2006). This work is to develop a flexible

method to estimate prevalence of haematological malignancies under WHO classification using data from HMRN, and show their disease burden in HMRN area and the UK. In fact, it can be also applied to other defined populations to estimate prevalence. For example, health insurance coverage for survivors in the US (Carpenter et al., 2011) can be estimated using total prevalence ranges when the exact total prevalence in unavailable.

5.2 Methods in this thesis and the possible shortcomings

5.2.1 Model and calculation of prevalence

One of the main purposes of this work was to develop a suitable method to estimate total prevalence for haematological malignancies. Besides age, prevalence is also related to the time period of observation. Within registry data, only a limited number of years of observation data are available for prevalence calculation. Factors influencing the number of years of follow- up to capture the majority of prevalent cases include the age at which the disease is common and the survival of the disease. For example, the registration period was essentially sufficient for mantle cell lymphoma because it only occurs in later adulthood and the survival is relatively poor. On the other hand, for acute lymphoblastic leukaemia where the diagnosis is generally made at a young age and survival is generally good, many more years of follow- up are required to capture prevalence. In general, females need more years of follow- up than males. This may be because of better survival and longer life expectancy in females than in males.

Haematological malignancies have the characteristic ability to transform (Davies et al., 2007; Kyle et al., 2010; Landgren et al., 2009; Lossos et al., 2002; Shanafelt et al., 2010; Shi et al., 2004), so the effects of multiple cancers cannot be ignored. One single patient may suffer from more than one cancer; prevalence can either

count the number of patients or refer to the number of cancers in the population. Previous studies of prevalence estimation only included the first primary cancer diagnosed in each person (Capocaccia, et al., 2002). Such studies only count a person once, which is useful if the reader would like to summarise various prevalence estimates across cancer sites without double counting individuals (i.e., the prevalence of specific cancer sites adds up to the prevalence of all cancers combined) (SEER, 2012). However, the ability to transform and process in haematological malignancy determines that diagnosis prevalence is more appropriate in this study.

Amongst the 14,901 patients in HMRN diagnosed from 2004 to 2011, there were 821 (5.5%) cases who had a second diagnosis, and 88 (<1%) cases who had a third diagnosis, resulting in 15,810 diagnoses in this study. Many followed the expected pattern, with either a precursor disease or a more indolent diagnosis progressing to one that is considered a more aggressive diagnosis (this has been described in Chapter One) (Bagguley et al., 2012; Davies et al., 2007; Landgren et al., 2009; Lossos et al., 2002; Shanafelt et al., 2010; Shi et al., 2004). For example, 12% of MDS cases had progressed to AML. 17.5% of MBL cases had transformed to CLL, 7.8% of follicular lymphoma had transformed to DLBCL, and 4.1% of MGUS to myeloma.

.

HMRN also solves the problems associated with classification of haematological malignancies. It provides unbiased ascertainment and accurate capture of detailed diagnostic data (HMDS, 2011) and provides a solid foundation for research into haematological malignancies in the area covered by HMRN (HMRN, 2011; Smith et al., 2010). The high quality functional data makes the calculation easy to move to the next step to estimate total prevalence.

A flexible analytical model was developed for use in estimating the total prevalence of haematological malignancies. The analytical framework presented

is an adaption and an extension of the method used by Capocaccia and De Angelis (1997), which involves a mathematical model and statistical estimation.

To estimate total prevalence, the method is presented by fitting data from the population-based cancer registry, HMRN.

I. Completeness index (R)

R is calculated using age, and it varies according to age group. Adults who were diagnosed in childhood may not be registered within the registry. The older age groups have more probability that the patients diagnosed in the younger age before the start of the registry and do not have records. For younger age groups, born closer to the time of the establishment of the registry, the likelihood of having had a diagnosis before the start of the registry is smaller. For the first age group, the value of R must be 1, since all of them were born after the start of the registry and are registered in HMRN.

II. Validation of the method

There is no gold standard for measuring or estimating cancer prevalence (Carpenter et al., 2011). Therefore it is impossible to validate the estimates of total prevalence against the actual proportion of the population living with a haematological malignancy. The validation analysis used in this study was conducted according to two aspects; the goodness of fit of the data was checked, along with the predictive power of the method. Both of them provided assurance for the methods and results in this thesis. As the length of the registry was limited, this work could only use the data in the last five years to estimate 7-year prevalence. When the registry becomes more mature, the validation analysis may show better results (for example, use 15 years data to estimate 25-year prevalence) (Gigli, Simonetti, and Capocaccia, 2004).

As more information about survivor populations continues to become available, R becomes more robust with respect to the estimation of parameters. When the period of registration is long, the difference between total prevalence and observed prevalence decreases. As the registry time tends to increase, observed prevalence should converge to the total prevalence, which means the observational time is long enough to use observed prevalence instead of total prevalence to show the burden of cancer patients in a population. As HMRN continues to add new cases, increased duration of haematological malignancies registration will allow continued examination of the validity of this method and stability of these estimates over longer year periods. The limitations of this method are described in the later sections.

III. Using AML and Hodgkin lymphoma as examples

AML and Hodgkin lymphoma were chosen as examples to show the calculation process in this work, while for other subtypes, the results of total prevalence and completeness index were shown directly. This is because AML and Hodgkin lymphoma represented two typical diseases with different incidence and survival characteristics. AML is an easy example to show the calculation details, as there have been no significant changes in incidence and survival in the past years. It can occur at any age, and the survival is relatively poor (See Appendix A5). On the contrary, the survival of Hodgkin lymphoma is good, which means there may be more cases of Hodgkin lymphoma who were diagnosed before the start of HMRN and still alive on the index date than AML. Furthermore, Hodgkin lymphoma has an unusual age distribution for incidence and is therefore, a good example to show the requirement for a more flexible model, since the model in the literature may fail to fit the data (See Section 4.3.2).

Hodgkin lymphoma was chosen as an example to demonstrate the validation analysis. First, Hodgkin lymphoma can provide a better view of a trend due to its

good prognosis compared to AML (See Section4.3.3). Second, there is a relatively good sample size to support the estimation. The most important reason is that if the method is flexible enough for Hodgkin lymphoma (with its unusual age distribution for incidence), it will be fine for other subtypes with common age distributions on incidence and survival.

IV. Comparability of HMRN and UK

The prevalence estimated using HMRN data was generalised to the whole of the UK. This is because HMRN region population structure mirrors that of the UK in terms of age and sex (See Section 3.1.3). The prevalence rates were applied on the population of UK without age adjustment. In this estimation, ethnicity was not included since it was not available in data. In fact, different ethnic composition may bring bias to the results. According to literature (National Cancer Intelligence Network and Cancer Research UK, 2009; SEER, 2012), incidence of Hodgkin lymphoma and non- and Hodgkin lymphoma in the black ethnic group was not significantly different from the white ethnic group, and the incidence of leukaemia is slightly higher in white than black (the ethnic comparisons were only available for broad categories but not for subtypes in literature). However, it should be noting that myeloma has much lower incidence in white than black (National Cancer Intelligence Network and Cancer Research UK, 2009; SEER, 2012). So, if there are more black people in some areas such as London than HMRN region (Office for National Statistics, 2012), the generalization in this work may underestimate the prevalence counts in the UK for myeloma due to its higher incidence in black than white.

V. Total prevalence range

For those subtypes with a change in survival, a range of estimates of total prevalence were made, based on the maximum estimates calculated using the general method; the minimum estimates were derived from “T-year prevalence”.

This is consistent with the aim of this study—estimating prevalence of haematological malignancies using data from HMRN, and avoids borrowing complementary information from other datasets. Total prevalence range estimation is a practical method to show the burden of disease, and to our knowledge, there has been no similar report made up to this point. .

In the literature, limited duration prevalence is usually used to substitute for estimating total prevalence in this situation (Capocaccia, et al., 2002; Forman, et al., 2003). However, when the length of the registry is short (seven years in this study), the surrogate—observed prevalence, does not appear reasonable. As described above, complete prevalence in the real world is difficult to estimate until the registration period becomes very long. Total prevalence is only an estimate for it. When the exact figures are difficult to estimate, finding a range for total prevalence seems a convenient way to show a reasonable result for a certain subtype.

5.2.2 Improvements and differences from previous methods

This method is an adaption of the "completeness index method" to estimate total prevalence from limited duration prevalence (Capocaccia and De Angelis 1997; Gigli, et al., 2006). However, unlike the methods used in the literature, in this thesis some adaptations were used for haematological malignancy according to its characteristics of incidence and survival.

5.2.2.1 Parametric and non-parametric

Capocaccia and De Angelis (1997) used a parametric model to estimate incidence and survival. In this thesis, however, the incidence model is a non-parametric model. To accommodate variation over time of a predictor's effect on incidence,

regression splines were used to model the incidence rate as a flexible function of age, without having to specify a particular functional form.

As described in Section 4.3, the predictions of this incidence function in the literature (Capocaccia and De Angelis, 1997; Merrill, et al., 2000; Gigli, et al., 2006) could potentially bias the results. However, the regression spline showed a much better fit for incidence. This is because the log linear model used in the literature for incidence was designed mainly for common cancers, which occur after adulthood. Their incidence can be simply described by exponential shape. However, for some subtypes of haematological malignancy, they can occur at any age or have unusual age distributions. The log linear model could not fit the data for them. However, regression spline is a non- parametric method, so it can describe the data regardless of the distributions. Section 4.3 in Chapter Four took incidence of AML and Hodgkin lymphoma as the examples to compare the incidence models and to show the benefit of regression spline method in this study.

If the unusual incidence patterns of Hodgkin lymphoma are ignored, for example, there will be an underestimation of total prevalence using the log linear model as the model will not describe the bimodal incidence curve, leading to an underestimation of incidence before the age 35 (See Figure 4- 16). The survival before the age of 35 is relatively good, therefore the prevalent cases on the index date may be underestimated due to the incorrect incidence description.

For survival, a parametric model was used. Survival probability depends on both the duration of disease (years since diagnosis) and age at diagnosis. Non-parametric models can provide a better fit for the data. However, observed survival in a short- lived registry cannot model the whole life of patients. In this instance, a parametric model provides a best guess to extrapolate beyond the survival observed in the sample data. A Weibull model has been applied previously and successfully for cancer survival (Capocaccia and De Angelis, 1997;

Merrill, et al., 2000; Capocaccia, et al., 2002; Verdecchia, et al., 2002; Forman, et al., 2003; Lutz, et al., 2003; Möller, et al., 2003; Gigli, et al., 2006). However, the age effect on survival of haematological malignancies is difficult to describe, and does not follow the linear assumptions used in previous studies for some subtypes (see Figure 4- 24). For some subtypes that can occur at any age, the survival pattern may be different in children and adult. The survival of AML, for example, increases with age in childhood and then decreases after adulthood (See Section 4.3.1, Figure 4-11). The survival model in the literature (Capocaccia and De Angelis, 1997; Gigli, et al., 2006) could not be used on this non-monotonic trend. In this study it was described using splines. They are a useful tool for analysing survival especially for subtypes which can occur at any age (Becher, et al., 2009). The 3D survival curves estimated using this method are shown in Figure 4-10 and Figure 4-17, using AML and Hodgkin lymphoma as examples. All in all, this study used more appropriate statistical methods to fit the data (details about the comparisons were shown in Chapter Four).

5.2.2.2 Continuous and discrete model

Compared to Capocaccia and De Angelis's (1997) method, the method in this thesis was formulated using discrete time instead of continuous time as used in previous studies. This was mainly because practical applications usually deal with discrete data (Verdecchia, et al., 2007).

Capocaccia and De Angelis (1997) framed their method in continuous time, and modelled the incidence and survival functions parametrically to facilitate the necessary integrations. However, the quantities that were available to this research were quite naturally framed in terms of discrete time. Therefore it made sense to look for the discrete version of the fundamental equations and to perform the calculations on them numerically. More attention should be paid in building models using a discrete version. Some approximation is necessarily involved in the model even when one-year age classes are in use. This is because, for example,

if a patient was diagnosed and died within the same year, it is difficult to show the survival time in a discrete version. In Chapter Three, assumptions were made in the first section to help build the model. Diagnoses were assumed to have been made at the beginning of the age groups, whilst deaths were all assumed to have occurred at the end of the age groups. In addition this approach led to discrete survival times. This assumption avoided the zero survival time in some special cases (events which occurred at the same age), however there were overestimates of survival time. Fortunately the overestimation in this method could be mitigated, by calculating the proportion of observed prevalence over total prevalence.

The model relating prevalence, incidence, mortality, and survival was developed in a discrete- time version in Chapter Three. However it should be noted that the method stands as a mixed approach, since the equations were given in discrete form, whilst incidence and survival were modelled before being included in the calculation. Ideally it would have been good to use incidence and survival data directly, however presumably sample sizes are too small to avoid noisy estimates due to sample variation. Errors can occur when abstracting corresponding values at single ages or integer disease durations from incidence and survival models. Both incidence and survival models provide smooth curves, therefore there may be underestimations and overestimations for incidence at certain ages and survival probability of certain survival time. However these uncertainties are likely to be small and would not affect the results greatly.

5.3 Limitations and weaknesses of the study

Despite the improvements of the method in this study, estimates may be still affected by the method chosen, since not all the characteristics of subtypes can be captured using the modelling techniques employed here. Inaccuracies in estimation may have also occurred due to data limitations and assumptions employed to model prevalence estimates. A general limitation of prevalence estimation is the incompleteness bias from the limited length of the registry.

Therefore in this study, the prevalence was estimated using age-specific incidences that were relatively current and survival that was observed only for seven years. However, current prevalence is not only based on current incidence and survival, but also on past values as well; this is lacking in this thesis due to the assumption of constant probabilities with calendar years. The assumptions of stable incidence and survival may inaccurately estimate prevalence, because both measures increased over time (Cancer Research, UK 2011). Using the incidence rate from 2004 to 2011 for all years prior to 2004 would overestimate the total prevalence. In addition, owing to the improvement of survival and increasing life expectancy, there may be overestimations of the true complete prevalence. Briefly, in the model, if there is an increase in incidence and survival, the model will overestimate the prevalence. These uncertainties from calendar years may be weakened by calculating proportions of observed prevalence over total prevalence.

The incidence that can be observed from these seven years was used, and considered as the constant incidence for the past years. This was not only because it was not reliable to estimate the trend of incidence from seven years of data, but also due to the purpose of the work. For the purpose of estimating, the most appealing choice was not to try to adopt the best hypothesis that can be taken, but to provide a plausible calculation of estimated probabilities (Verdecchia, et al., 2002). Therefore the assumption about constant probabilities in this study was the most convenient way to build the model. Although there was an assumption that the survival under Weibull distribution and the hazard function would change monotonically over time, year at diagnosis was not included in the function. General mortality rates were taken from the most recent available life table. The London School of Hygiene and Tropical Medicine (LSHTM, 2012) offers life table from 1971 to 2009. Although it changes over these 38 years, this work only used the latest one (2009) and considered it to be constant with years. All in all, to simplify the calculation for total prevalence, the calendar year component was not considered in the model.

A further limitation may be the restriction of the length of the registry. HMRN has an advantage in providing high quality data, yet also has an obvious disadvantage in its limited length. The estimates of prevalence dated to 2011, suffer from a delay (it is 2013 this year) that may limit its use for making decisions for health resource planning. This is because estimations were made based on data from a cancer registry, and it should be stressed that preparations of several years are needed to fit criteria of completeness and accuracy of data. Additionally, for some subtypes with small numbers of cases in the registry, the estimations may be not reliable and robust enough to show total prevalence. Sample size is an important feature of the study in which the goal was to make estimations of prevalence from the observed incidence and survival. Generally, larger sample sizes lead to increased precision when estimating unknown parameters. The numbers of observations are quite different for each subtype. Therefore the precision is higher in those diseases with more cases in the registry, whilst there is inevitable inaccuracy for subtypes with fewer cases. The method is not recommended for diseases with small number of cases. For example it does not perform very well for rare subtypes (RARECARE, 2013) such as hairy cell leukaemia, T-cell leukaemia, Burkitt lymphoma, and T-cell lymphoma. The results for those subtypes can only provide a suggestion, but are not sensitive enough to show the real burdens. In the calculation, in fact, there may be diversities within one subtype. For example acute promyelocytic myeloid leukaemia (APML) shows better survival than other AML due to the introduction of all-trans retinoic acid (ATRA) and arsenic trioxide (ATO) in its treatment, which turns acute promyelocytic myeloid leukaemia from being highly fatal to having a good prognosis (Wang and Chen, 2008). It seems reasonable to exclude APML from AML for prevalence estimates. However, being different from the descriptive analyses, the main purpose of this study was to make estimations, for which sample size is an important factor. Therefore APML was not estimated separately from AML in this thesis.

5.4 Comparisons with other published knowledge

Prevalence for most of subtypes cannot be compared with other reports in the literature, because data are not coded in the same way. Fortunately, as described in section 3.2.1 in Chapter Three, it can be confirmed that the prevalence estimates for the conditions in which the bridge coding can provide a reasonable approximation. For example, although for haematological malignancies (all subtypes combined), n-year prevalence are not in line with expectations as compared with national program (NCIN, 2006), the estimates of total survivors of Hodgkin lymphomas in the U.S. using my prevalence rates are broadly similar to the most recent reports by SEER (SEER, 2012).

The appropriate classification for the disease is important for haematological malignancy epidemiological research (Smith, et al., 2009). However, many prevalence reports about haematological malignancy have aggregated their data into broad groups (Ferlay, et al., 2010; NORDCAN, 2010; NCIN, 2012; SEER; 2012). The results in Table 4-5 showed that both overall 1-year prevalence and 5-year prevalence estimates were not consistent with the national published figures (NCIN, 2006). They doubled the frequencies in the UK reported in 2006 (35,679 vs.16, 432 for 1-year prevalence, and 133,565 vs. 61,755 for 5-year prevalence), which were shown in broad categories. The double counting of patients due to multiple cancers increased the estimates in this study. A more meaningful reason may be the different way of coding, that not all of the subtypes in Table 4-5, such as myelodysplastic syndromes, were uniformly compiled (Smith, et al., 2009). The national figures do not include conditions such as MGUS, MBL, MDS and MPNs, which account for a large proportion of prevalent cases. However, for Hodgkin lymphoma alone, the comparison shows a more reasonable result: a slight in increase for both 1-year prevalence and 5-year prevalence (1681 vs.1437 for 1-year prevalence, and 7776 vs.6190 for 5-year prevalence).

Prevalence estimated in this study can be extrapolated not only to the UK as a whole, but also to other populations, by making certain assumptions. If one assumes that incidence, survival and general mortality in HMRN area are similar in the target population, the number of prevalent cases can be generated from my data with adjustment for age structure (applying age- specific rates on the target population). For example, it is estimated that in the U.S., the number of prevalent cases of Hodgkin lymphoma ranged from 88,147 to 96,040 for males, and 82,932 to 97,778 for females, based on age-specific prevalence rates in HMRN and populations in the U.S. (the population was obtained from IARC [IARC, 2013b]). This is consistent with the SEER reports (SEER, 2012) in the U.S. (93,890 for males and 88,038 for females). However, the estimates for myeloma were lower than their reports. In fact, after making adjustments for age, the incidence rate of myeloma in HMRN was slightly lower than in U.S., which leads to underestimations of total prevalence. This may be because of the different ethnic composition, since for myeloma the incidence in the black population (14.4 per 100, 000 for males and 10.2 per100, 000 for females) is much higher than in the white population (7.1 per 100, 000 for males and 4.2 per 100, 000 for females) (SEER, 2012).

Another example is the comparison to prevalence rates of Hodgkin lymphoma reported in Denmark. NORDCAN showed the prevalence in Denmark from 1963 to 2011 (Engholm, et al., 2013), which may be long enough to cover all live patients in the country. The age-adjusted incidences were estimated to be 3.5 and 2.5 per 100, 000 for males and females respectively, which were higher than the reports in NORDCAN (2.8 and 2.1 per 100, 000 for males and females). Thus, as might have been expected, the total prevalence estimated in HMRN (66.1 to 73.3 and 58.7 to 71.5 per 100, 000 for males and females respectively) was much higher than in the NORDCAN reports (53.8 and 40.9 per 100, 000 for males and females in 2011). This indicates the Hodgkin lymphoma prevalence rate estimated in HMRN cannot reasonably be applied to the Danish population. However, my estimates of the ratio of total prevalence to incidence (P: I) were reasonably consistent with those observed in NORDCAN. The P: I estimated in HMRN for Hodgkin lymphoma was 18.9 to 20.9 in males, which includes the value reported

in the NORDCAN reports (19.2), and 22.6-27.5 in females, which is slightly higher than in the NORDCAN reports (19.5). This comparison is not only an explanation of the assumption (similar incidence and survival between local population and target population), but also reassuring validation of the method in this study.

5.5 Contributions

Total prevalence entails much more than just a measure of the percentage of the population still alive following a diagnosis of disease, frequently represented as a single statistic reflecting the proportion of the population alive in n-year (usually 1-year or 5-year) post-diagnosis. Rather, total prevalence includes all patients diagnosed in the past and still alive, and related issues of getting health care and follow-up treatment, late effects of treatment, and quality of life. It is therefore necessary to produce methods and statistics to inform those comprehensive cancer programs, for ensuring aspects such as health insurance coverage for survivor needs, facilitating basic healthy behaviours, and informing plans for long-term care. One important goal of estimating prevalence is to develop a more complete and accurate characterization of the survivor population, and to provide better estimates of their burden. That allows cancer networks and countries to better meet a population's needs all along the survivorship spectrum.

The prevalence in this work was calculated under WHO classification. It avoids high level of clinical diversity among the subtypes contained within each of the traditional groupings, and has better value for epidemiological (Smith et al., 2010). This is the first time to estimate prevalence of haematological malignancies under WHO classification. The estimates of survivors in the UK in this study demonstrate the importance of understanding the details of the prevalent haematological malignancies population when prioritizing survivorship services for each subtype. For example, Hodgkin lymphoma changed largely in rank order between observed prevalence and total prevalence estimates (moving from 7^{th} to

third for two sexes combined). The results in this study (see Figure 4-34 and Figure 4-35) may provide not only a ranking and corresponding prioritization of haematological malignancies of prevalence for the UK health system, but also an estimate of the number of individuals with a history of each subtypes of haematological malignancies under new WHO classification. The contributions of this study can be summarized into as follows:

I. Adapted and developed a more flexible model to estimate prevalence
II. Estimated prevalence of haematological malignancies under WHO classification, which previously has not been reported.
III. Estimated national prevalence counts and provided rank and periodization for subtypes.
IV. Showed prevalence ranges for subtypes whose patterns of survival had changed greatly in the past due to new treatment for the first time.

5.6 Recommendations for future research

5.6.1 Cure

In the calculations, all haematological malignancy patients are included, from diagnosis to the end of life. In other words, recovery and cure are not considered in this method. In fact, even if a cancer patient becomes a long-term survivor after treatment, he or she usually still needs extra medical care due to the psychological and physical consequences of the disease (Simonetti, et al., 2008). Sometimes, the risk of subsequent cancers can increase because of the aggressiveness of the treatment (Simonetti, et al., 2008). Furthermore, the patient may suffer disability and impairments arising from the cancer treatment and may make more demands on health resources than the age-matched general population (Verdecchia, et al., 2002). From this perspective, the method and the estimates in this study could better inform health planning for long term and end-of-life care.

In fact, the conception of recovery and cure requires careful definition (Gras et al., 2006). The purest definition of recovery should be based on the complete eradication of the disease in the individual. However, for cancers, it is not always possible to determine this, since people who appear to be "cured" according to clinical criteria often have recurrences (Gras et al., 2006). Therefore, in some previous calculations, there was an assumption that the disease is irreversible (see Chapter Two). That means, the disease, once diagnosed, is irreversible and both fatal cases and "cured (long-term survivors)" contribute to prevalence estimates.

After 5-year disease free survival a patient is often considered cured (Hoffbrand, et al., 2006; Howard and Hamilton, 2007; Hughes-Jones, et al., 2008). Although the "cure" time is not considered in the model, one can also show a proportion of those with "high consumption of health resources" by the differences between total prevalence and 5-year prevalence (Möller et al., 2003). The details of the results are shown in Appendix A6. However, the definition of "cure" of haematological malignancies and the involvement of remission rates in the models needs further research.

5.6.2 Prevalence in the future

In the future, prevalence may increase. This can be caused by many factors. Firstly, the better the cancer registration is, the higher the prevalence might be. This can to some degree explain the higher prevalence in developed countries than in developing countries (Pisani, Bray and Parkin, 2002). Secondly, increasing incidence and better survival (due to lead- time bias in screening detected cancer and earlier diagnosis of cancer as well as the improved treatment being available to patients [Möller, et al., 2003]) lead to increasing prevalence. In these cases, the prevalence will inevitably markedly increase. Thirdly, increasing life expectancy will cause an increase in prevalence (Möller, et al., 2003). In a Swedish study, from 1961 to 1995, it was estimated that of the increase in prevalence, 40-47% could be attributed to population dynamics (ageing and

growth), and 30% and 23-29% to better survival and increasing incidence respectively (Stenbeck, et al., 1999).

For HMRN, there is only seven years data available now, so it is impossible to obtain robust estimates of the trend of incidence and survival to calculate prevalence in the future with such limited data. When the registry is more mature and more cases registered in HMRN, calendar year can be added into the model, and the robust trends of probabilities with time may be abstracted from data, which make the estimation of future prevalence possible.

5.7 Summary

In this work, for the first time, the prevalence of haematological malignancies has been estimated for clinically meaningful diagnostic groups. Whilst additional research is necessary to continue improving prevalence measures and validating them, this study demonstrates the value of understanding total prevalence, as it allows more informed planning for health services and resource allocation in both HMRN area and in the UK. It illustrates the use of this method for converting observed prevalence to total prevalence using limited length of data from HMRN rather than based on other registries and their populations. Furthermore, it provides total prevalence rates under WHO classification, which can be extrapolated to the national or even worldwide level to estimate the number of survivors with age structure adjustment and other assumptions.

Appendices

Appendix A1 Cancer Network

Table A- 1 Cancer Networks and their codes in the UK

Codes	Cancer Network
N01	Lancashire and South Cumbria Cancer Network
N02	Greater Manchester and Cheshire Cancer Network
N03	Merseyside and Cheshire Cancer Network
N06	Yorkshire Cancer Network
N07	Humber and Yorkshire Coast Cancer Network
N08	North Trent Cancer Network
N11	Pan Birmingham Cancer Network
N12	Arden Cancer Network
N20	Mount Vernon Cancer Network
N21	West London Cancer Network
N22	North London Cancer Network
N23	North East London Cancer Network
N24	South East London Cancer Network
N25	South West London Cancer Network
N26	Peninsula Cancer Network
N27	Dorset Cancer Network
N28	Avon, Somerset and Wiltshire Cancer Network
N29	3 Counties Cancer Network
N30	Thames Valley Cancer Network
N31	Central South Coast Cancer Network
N32	Surrey, West Sussex and Hampshire Cancer Network
N33	Sussex Cancer Network
N34	Kent and Medway Cancer Network
N35	The Greater Midlands Cancer Network
N36	North of England Cancer Network
N37	Anglia Cancer Network
N38	Essex Cancer Network
N39	East Midlands Cancer Network
N96	North Wales Cancer Network
N97	South West Wales Cancer Network
N98	South East Wales Cancer Network

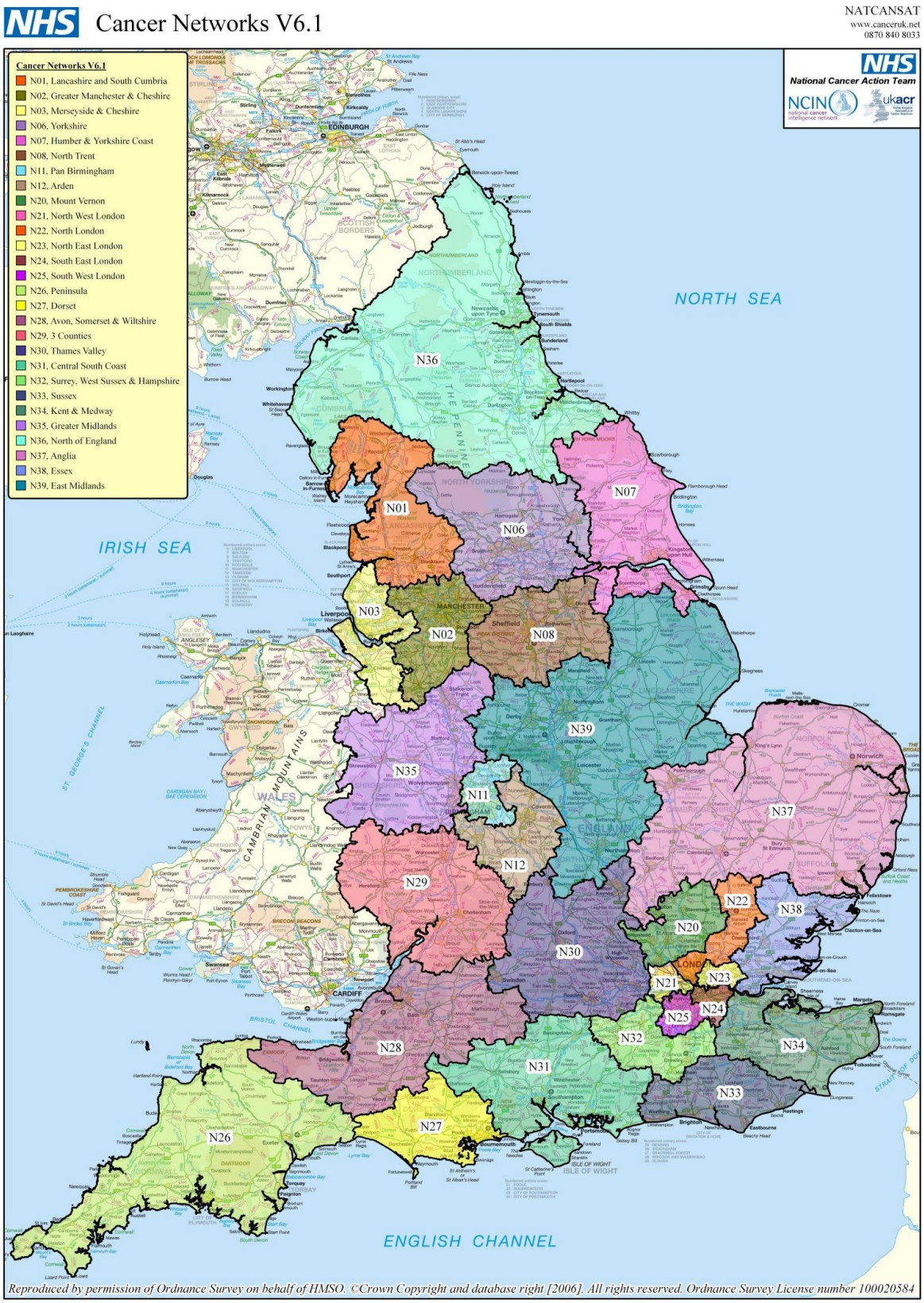


Figure A- 1 Map of Cancer Networks in England (NHS 2011)

Appendix A2 General Mortality by Age and Gender in England in 2009

Table A- 2 Life table in England (2009) (LSHTM, 2012)

Age	Male	Female	Age	Male	Female	Age	Male	Female	Age	Male	Female
0	0.004902	0.004014	25	0.000621	0.000277	50	0.003744	0.002338	75	0.039341	0.025064
1	0.000326	0.000314	26	0.000638	0.000296	51	0.004076	0.002578	76	0.043715	0.028205
2	0.000165	0.000154	27	0.000659	0.000317	52	0.004431	0.002845	77	0.048552	0.031759
3	0.000141	0.000124	28	0.000683	0.000339	53	0.004814	0.003145	78	0.053891	0.035773
4	0.000123	0.000106	29	0.000713	0.000364	54	0.005229	0.003485	79	0.059767	0.040297
5	0.000109	0.0000945	30	0.000747	0.000391	55	0.005678	0.003871	80	0.066216	0.045381
6	0.0000993	0.0000882	31	0.000787	0.000421	56	0.006168	0.004296	81	0.073274	0.051078
7	0.0000931	0.0000854	32	0.000832	0.000453	57	0.006703	0.004688	82	0.080972	0.057440
8	0.0000902	0.0000855	33	0.000885	0.000489	58	0.007295	0.004987	83	0.089337	0.064518
9	0.0000905	0.0000879	34	0.000944	0.000529	59	0.007954	0.005300	84	0.098393	0.072362
10	0.0000945	0.0000925	35	0.001012	0.000573	60	0.008688	0.005665	85	0.108156	0.081016
11	0.000103	0.000099	36	0.001089	0.000622	61	0.009510	0.006088	86	0.118633	0.090516
12	0.000117	0.000107	37	0.001176	0.000677	62	0.010430	0.006577	87	0.129823	0.100888
13	0.000140	0.000117	38	0.001275	0.000738	63	0.011464	0.007139	88	0.141712	0.112146
14	0.000174	0.000128	39	0.001386	0.000807	64	0.012625	0.007786	89	0.154270	0.124287
15	0.000220	0.000140	40	0.001511	0.000885	65	0.013928	0.008527	90	0.167456	0.137287
16	0.000278	0.000152	41	0.001651	0.000971	66	0.015389	0.009376	91	0.181208	0.151100
17	0.000349	0.000164	42	0.001807	0.001067	67	0.017028	0.010348	92	0.195447	0.165653
18	0.000428	0.000176	43	0.001980	0.001175	68	0.018864	0.011459	93	0.210075	0.180840
19	0.000503	0.000188	44	0.002171	0.001294	69	0.020919	0.012728	94	0.224973	0.196527
20	0.000559	0.000201	45	0.002382	0.001428	70	0.023217	0.014177	95	0.240034	0.212543
21	0.000585	0.000214	46	0.002613	0.001575	71	0.025785	0.015829	96	0.255289	0.228734
22	0.000592	0.000228	47	0.002865	0.001739	72	0.028650	0.017712	97	0.270832	0.245153
23	0.000598	0.000243	48	0.003138	0.001920	73	0.031843	0.019857	98	0.286793	0.261954
24	0.000608	0.000259	49	0.003431	0.002119	74	0.035396	0.022295	99	0.303338	0.279351

Appendix A3 Incidence and 5-year Survival for Subtypes of Haematological Malignancies (details for Table 4-2)

Table A- 3 Incidence and 5-year survival for subtypes

	Incidence (per 100,000)	5-year Survival
Chronic myelogenous leukaemia	1.0	78.4%
Chronic myelomonocytic leukaemia	0.7	19.9%
Acute myeloid leukaemia	4.2	19.7%
Acute lymphoblastic leukaemia	1.2	60.8%
Chronic lymphocytic leukaemia	6.9	65.3%
Hairy cell leukaemia	0.3	88.4%
T-cell leukaemia	0.4	62.7%
Marginal zone lymphoma	3.4	62.5%
Follicular lymphoma	3.2	75.9%
Mantle cell lymphoma	0.9	23.7%
Diffuse large B-cell lymphoma	8.3	48.2%
Burkitt lymphoma	0.3	54.1%
T-cell lymphoma	1.0	42.9%
Hodgkin Lymphoma	3.0	78.8%
Plasma cell myeloma	6.6	34.1%
Plasmacytoma	0.6	51.1%
Myelodysplastic syndromes	3.8	21.4%
Myeloproliferative neoplasms	6.2	74.6%
Monoclonal B-cell Lymphocytosis	2.8	82.3%
Monoclonal gammopathy of undetermined significance	6.6	69.4%
Lymphoproliferative disorder not otherwise specified	1.9	58.4%

Appendix A4 Observed and Total Prevalence in the UK

Table A- 4 Observed and total prevalence cases in the UK on 31st, August 2011, ranked in order of descending total prevalence for both genders.

	Male		Female		Total	
	Observed	Total	Observed	Total	Observed	Total
Hodgkin Lymphoma	5,650	20,950	4,545	21,608	10,191	42,556
Monoclonal gammopathy of undetermined significance	10,772	18,274	9,878	24,020	20,645	42,310
Myeloproliferative neoplasms	9,268	18,007	11,536	21,515	20,810	39,530
Chronic lymphocytic leukaemia	13,300	23,222	7,827	13,316	21,106	36,500
Diffuse large B-cell lymphoma	9,847	17,483	8,664	14,924	18,505	32,396
Follicular lymphoma	5,022	9,549	5,825	13,082	10,849	22,641
Monoclonal B-cell Lymphocytosis	5,204	8,937	4,513	10,424	9,713	19,364
Plasma cell myeloma	7,352	10473	5,185	8066	12,528	18530
Marginal zone lymphoma	5,419	9,178	4,627	7,847	10,043	17,018
Lymphoproliferative disorder not otherwise specified	2,908	5,568	2,182	3,819	5,087	9,380
Chronic myelogenous leukaemia	2,049	4,887	1,346	3,775	3,391	8,657
Acute lymphoblastic leukaemia	1,933	4,711	1,346	3,794	3,276	8,501
Myelodysplastic syndromes	3,536	3,704	2,068	2,205	5,598	5,904
Acute myeloid leukaemia	2,561	2,901	2,084	2,716	4,643	5,617
T-cell lymphoma	1,223	2,303	870	2,106	2,091	4,407
T-cell leukaemia	496	1,277	640	2,014	1,136	3,294
Burkitt lymphoma	628	2,139	181	1,096	807	3,231
Hairy cell leukaemia	958	2,497	246	431	1,202	2,920
Plasmacytoma	1,052	1,376	427	635	1,476	2,008
Mantle cell lymphoma	1,107	1,408	492	517	1,597	1,922
Chronic myelomonocytic leukaemia	595	627	459	505	1,054	1,132

Appendix A5 Age- specific Incidence and Survival of Subtypes of Haematological Malignancy

1 Chronic Myelogenous Leukaemia

Table A- 5 Crude incidence of chronic myelogenous leukaemia by age and gender (per 100,000 population)

Age group (Years)	Total		Male		Female	
	N	Incidence	N	Incidence	N	Incidence
0-4	0	0.0	0	0.0	0	0.0
5-9	0	0.0	0	0.0	0	0.0
10-14	0	0.0	0	0.0	0	0.0
15-19	4	0.2	1	0.1	3	0.4
20-24	8	0.5	5	0.7	3	0.4
25-29	2	0.1	2	0.3	0	0.0
30-34	6	0.3	5	0.6	1	0.1
35-39	16	0.8	12	1.3	4	0.4
40-44	16	0.9	8	0.9	8	0.9
45-49	25	1.6	15	1.9	10	1.3
50-54	29	1.7	20	2.3	9	1.0
55-59	26	1.9	19	2.7	7	1.0
60-64	30	2.4	16	2.6	14	2.2
65-69	17	1.5	11	2.1	6	1.0
70-74	19	1.9	8	1.8	11	2.0
75-79	22	2.6	15	4.3	7	1.4
Over 80	22	2.1	9	2.7	13	1.8
Total	242	1.0	146	1.2	96	0.7

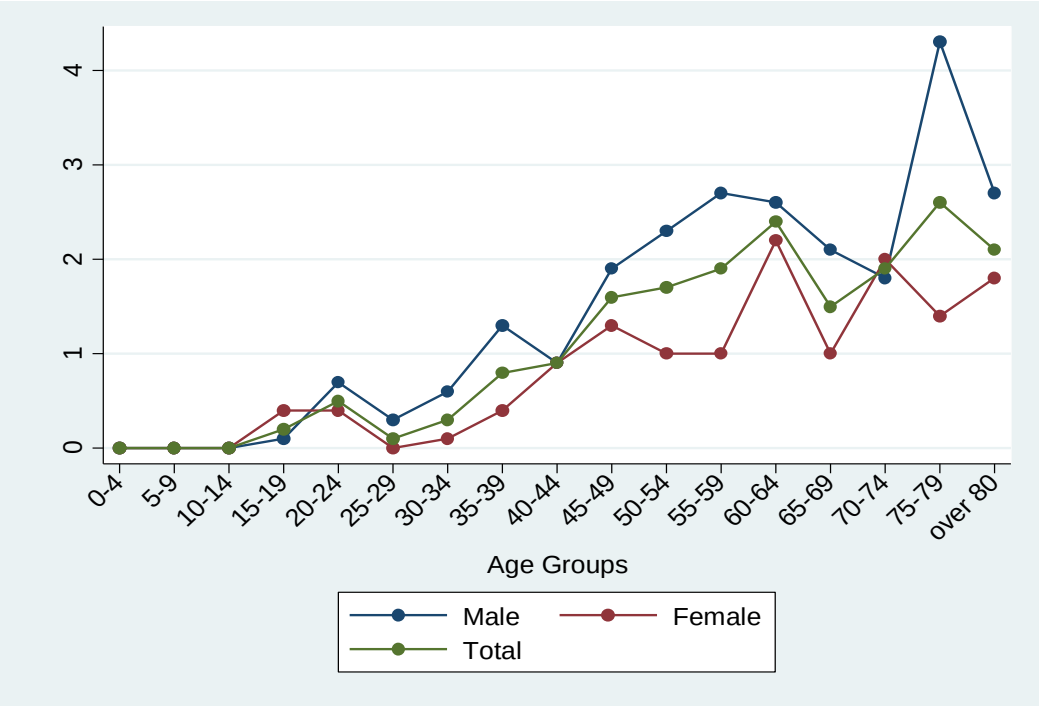


Figure A- 2 Incidence of chronic myelogenous leukaemia per 100,000 for males females, and total

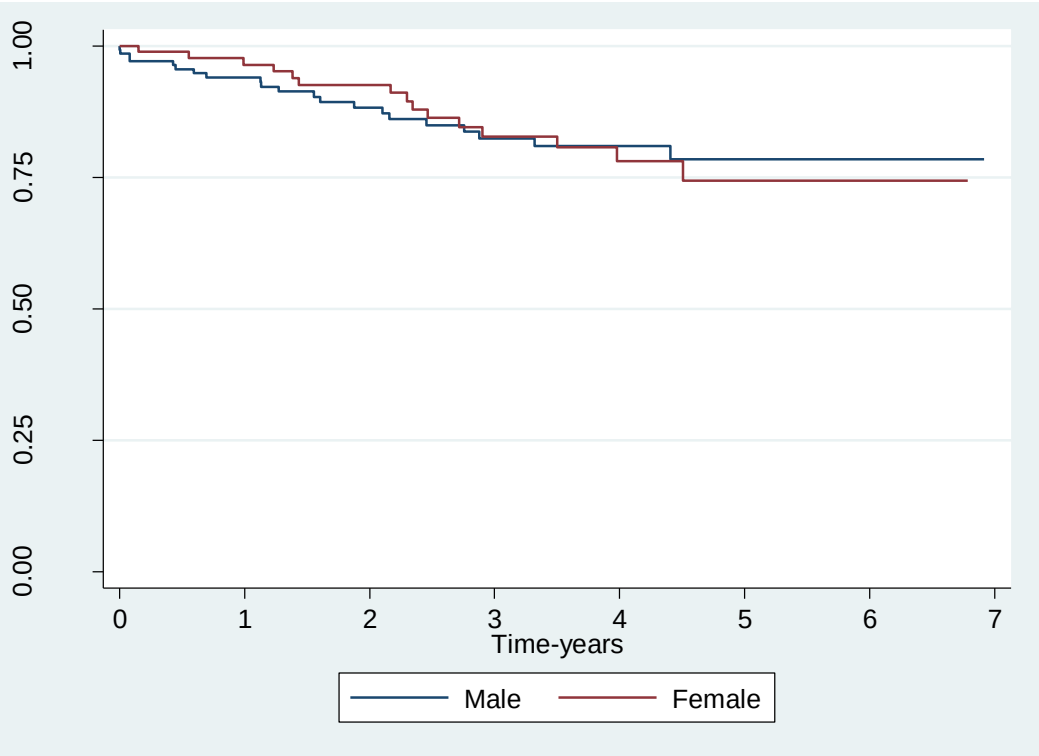


Figure A- 3 Kaplan-Meier survival estimates for chronic myelogenous leukaemia patients by gender

2 Chronic Myelomonocytic-Leukaemi

Table A- 6 Crude incidence of chronic myelomonocytic leukaemia by age and gender (per 100,000 population)

Age group (Years)	Total		Male		Female	
	N	Incidence	N	Incidence	N	Incidence
0-4	3	0.2	1	0.1	2	0.3
5-9	0	0.0	0	0.0	0	0.0
10-14	0	0.0	0	0.0	0	0.0
15-19	0	0.0	0	0.0	0	0.0
20-24	0	0.0	0	0.0	0	0.0
25-29	0	0.0	0	0.0	0	0.0
30-34	0	0.0	0	0.0	0	0.0
35-39	1	0.1	0	0.0	1	0.1
40-44	1	0.1	1	0.1	0	0.0
45-49	3	0.2	2	0.3	1	0.1
50-54	4	0.2	3	0.3	1	0.1
55-59	5	0.4	4	0.6	1	0.1
60-64	5	0.4	5	0.8	0	0.0
65-69	23	2.1	14	2.6	9	1.5
70-74	25	2.5	15	3.4	10	1.8
75-79	33	3.9	25	7.1	8	1.6
Over 80	70	6.7	38	11.4	32	4.5
Total	173	0.7	108	0.9	65	0.5

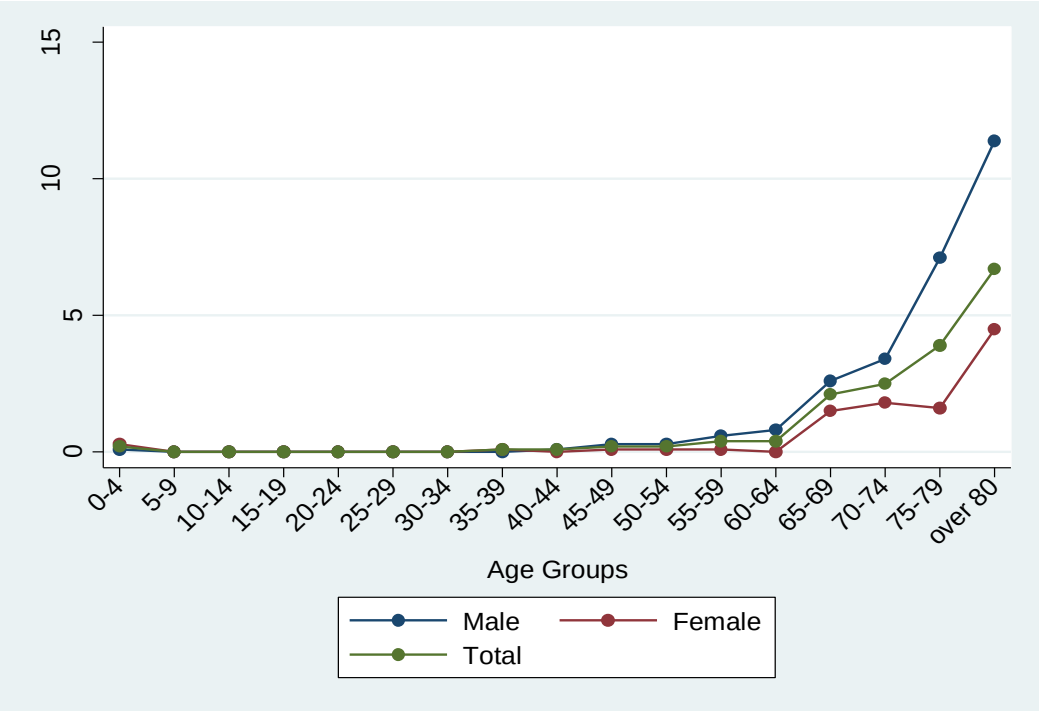


Figure A- 4 Incidence of chronic myelomonocytic leukaemia per 100,000 for males, females, and total

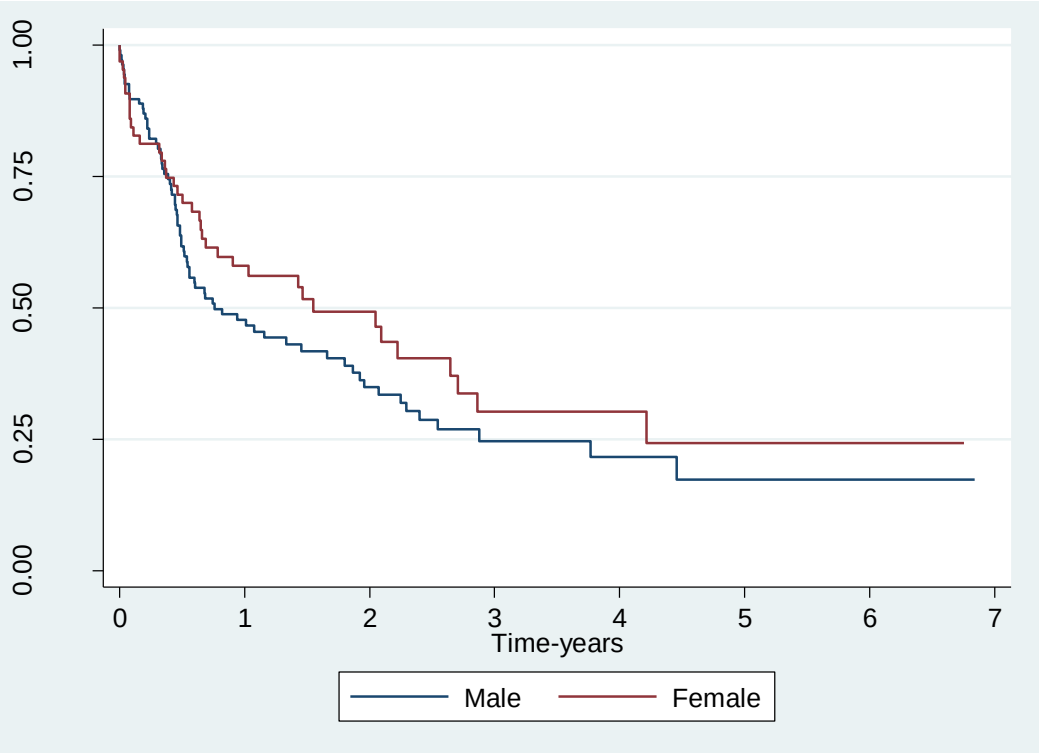


Figure A- 5 Kaplan-Meier survival estimates for chronic myelomonocytic leukaemia patients by gender

3 Acute Myeloid-Leukaemia

Table A- 7 Crude incidence of acute myeloid leukaemia by age and gender (per 100,000 population)

Age group (Years)	Total		Male		Female	
	N	Incidence	N	Incidence	N	Incidence
0-4	18	1.2	10	1.3	8	1.1
5-9	5	0.3	3	0.4	2	0.3
10-14	9	0.5	5	0.6	4	0.5
15-19	15	0.9	11	1.3	4	0.5
20-24	21	1.4	9	1.2	12	1.5
25-29	18	1.2	8	1.1	10	1.2
30-34	24	1.3	13	1.4	11	1.2
35-39	27	1.4	13	1.4	14	1.4
40-44	33	1.9	22	2.5	11	1.3
45-49	32	2.0	18	2.3	14	1.8
50-54	48	2.7	25	2.9	23	2.6
55-59	69	5.0	39	5.6	30	4.3
60-64	96	7.8	59	9.8	37	5.9
65-69	103	9.2	56	10.6	47	8.0
70-74	151	15.1	86	19.3	65	11.7
75-79	163	19.3	89	25.3	74	15.0
Over 80	229	21.9	110	33.0	119	16.8
Total	1061	4.2	576	4.8	485	3.8

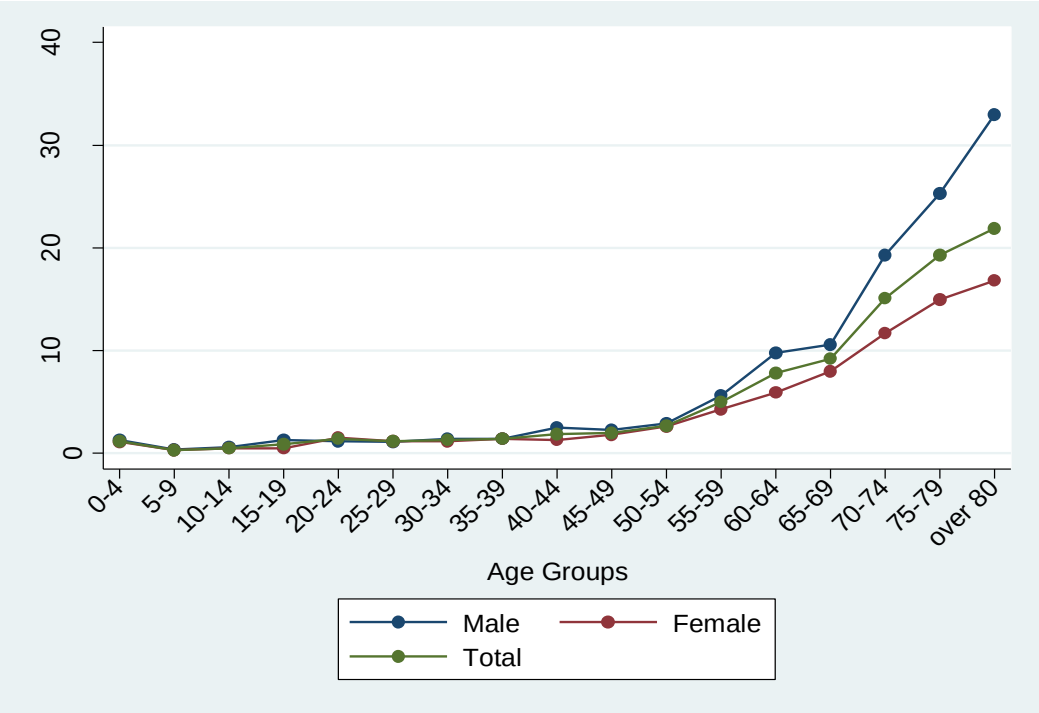


Figure A- 6 Incidence of acute myeloid leukaemia per 100,000 for males, females, and total

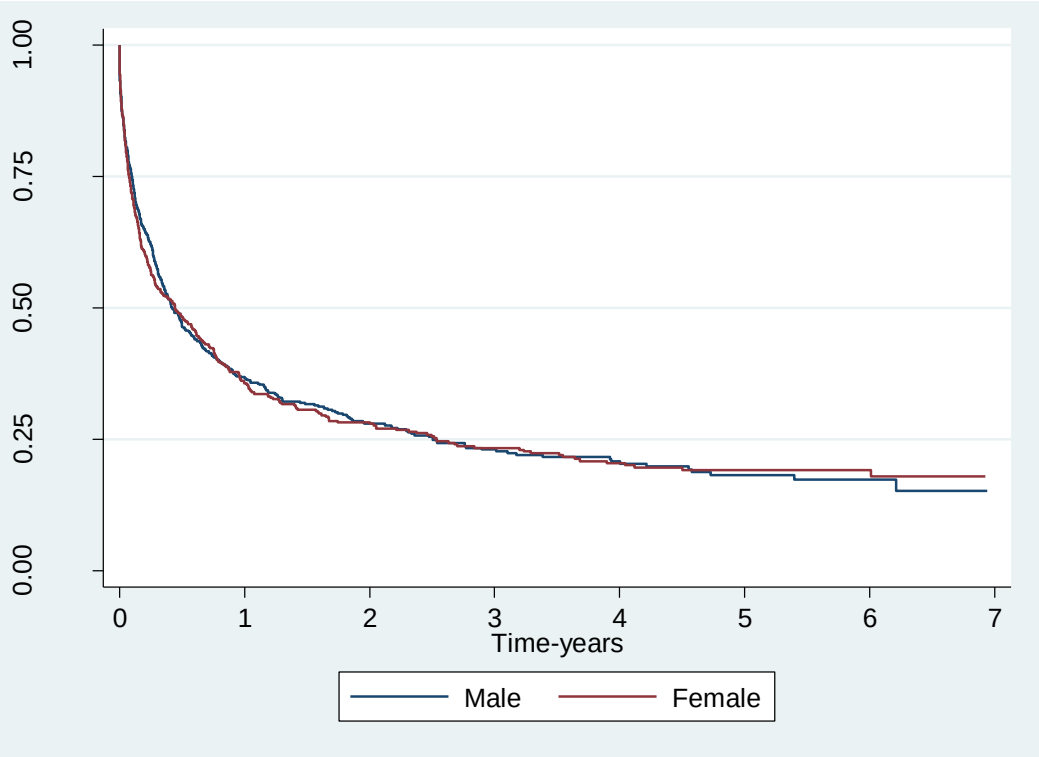


Figure A- 7 Kaplan-Meier survival estimates for acute myeloid leukaemia patients by gender

4 Acute Lymphoblastic Leukaemia

Table A- 8 Crude incidence of acute lymphoblastic leukaemia by age and gender (per 100,000 population)

Age group (Years)	Total		Male		Female	
	N	Incidence	N	Incidence	N	Incidence
0-4	72	4.9	40	5.3	32	4.4
5-9	47	2.9	24	2.9	23	2.9
10-14	29	1.7	16	1.8	13	1.5
15-19	31	1.9	22	2.7	9	1.1
20-24	12	0.8	9	1.2	3	0.4
25-29	7	0.4	1	0.1	6	0.7
30-34	11	0.6	9	1.0	2	0.2
35-39	11	0.6	5	0.5	6	0.6
40-44	7	0.4	3	0.3	4	0.5
45-49	12	0.8	7	0.9	5	0.6
50-54	7	0.4	5	0.6	2	0.2
55-59	13	0.9	7	1.0	6	0.9
60-64	11	0.9	5	0.8	6	1.0
65-69	16	1.4	13	2.5	3	0.5
70-74	5	0.5	3	0.7	2	0.4
75-79	6	0.7	4	1.1	2	0.4
Over 80	8	0.8	2	0.6	6	0.8
Total	305	1.2	175	1.4	130	1.0

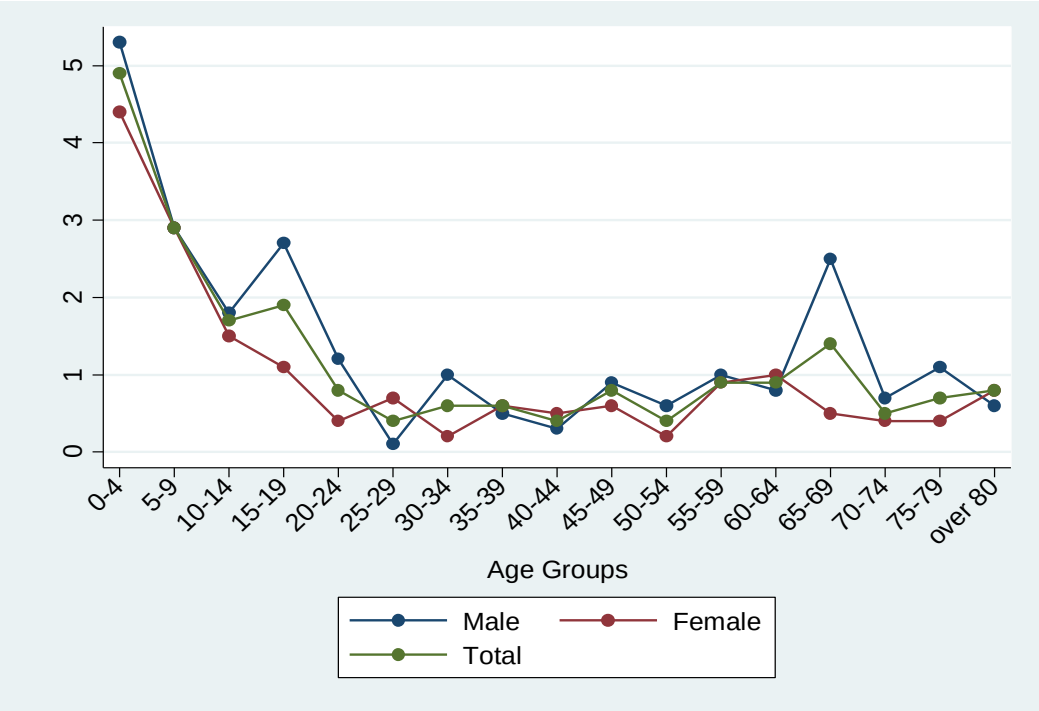


Figure A- 8 Incidence of acute lymphoblastic leukaemia per 100,000 for males females, and total

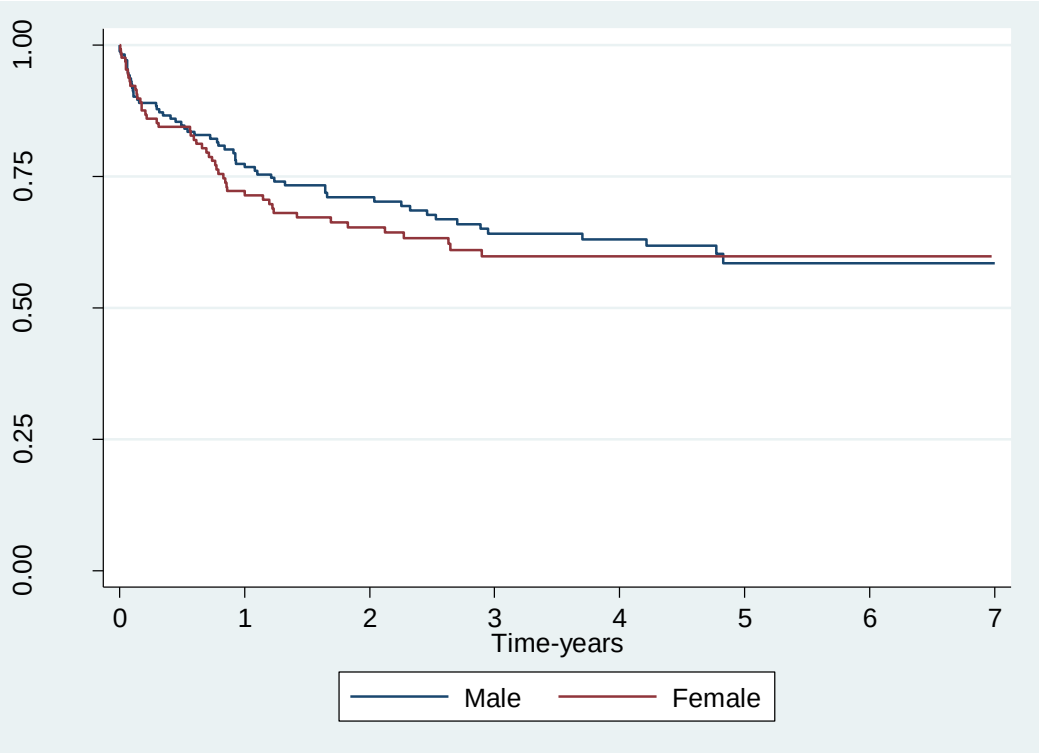


Figure A- 9 Kaplan-Meier survival estimates for acute lymphoblastic leukaemia patients by gender

5 Chronic Lymphocytic Leukaemia

Table A- 9 Crude incidence of chronic lymphocytic leukaemia by age and gender (per 100,000 population)

Age group (Years)	Total		Male		Female	
	N	Incidence	N	Incidence	N	Incidence
0-4	0	0.0	0	0.0	0	0.0
5-9	0	0.0	0	0.0	0	0.0
10-14	0	0.0	0	0.0	0	0.0
15-19	0	0.0	0	0.0	0	0.0
20-24	0	0.0	0	0.0	0	0.0
25-29	2	0.1	1	0.1	1	0.1
30-34	1	0.1	0	0.0	1	0.1
35-39	13	0.7	10	1.1	3	0.3
40-44	22	1.3	15	1.7	7	0.8
45-49	32	2.0	23	2.9	9	1.1
50-54	84	4.8	58	6.6	26	3.0
55-59	160	11.5	108	15.5	52	7.5
60-64	237	19.2	169	27.9	68	10.8
65-69	225	20.1	161	30.4	64	10.9
70-74	273	27.2	174	39.0	99	17.8
75-79	305	36.1	177	50.4	128	25.9
Over 80	367	35.2	181	54.3	186	26.2
Total	1721	6.9	1077	8.9	644	5.0

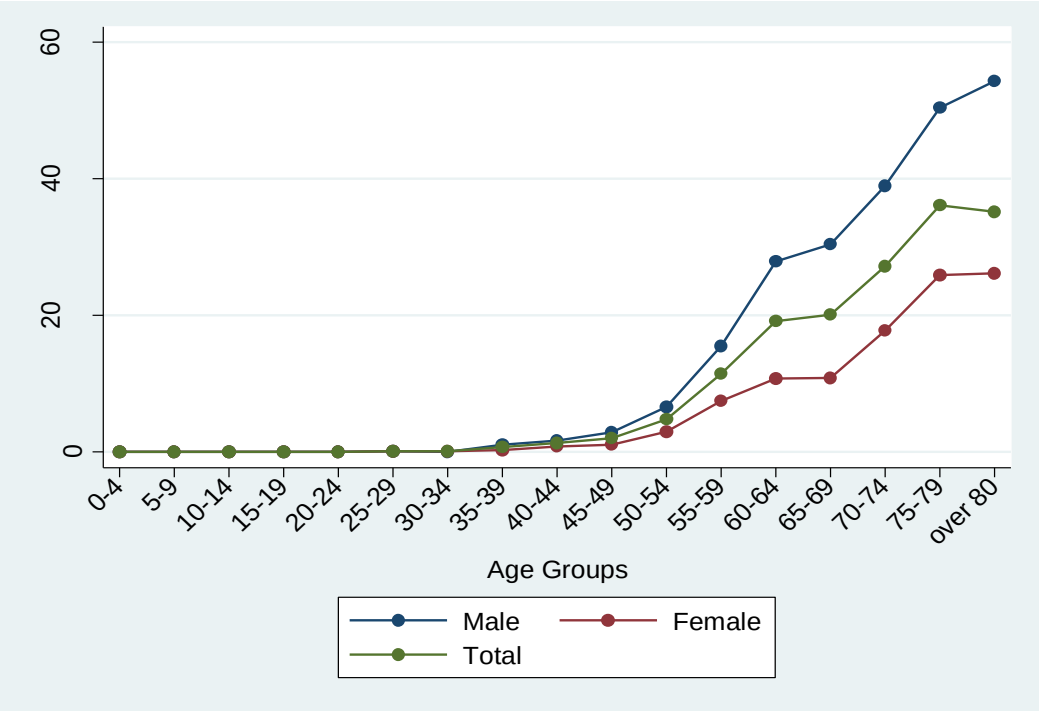


Figure A- 10 Incidence of chronic lymphocytic leukaemia per 100,000 for males females, and total

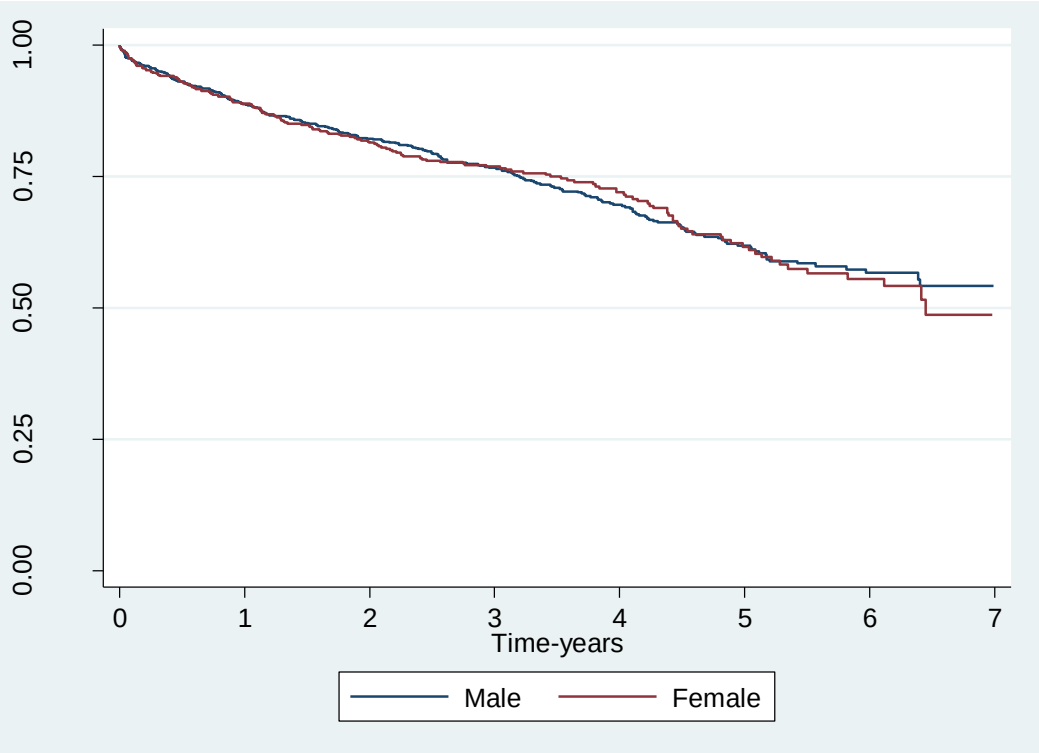


Figure A- 11 Kaplan-Meier survival estimates for chronic lymphocytic leukaemia patients by gender

6 Hairy Cell Leukaemia

Table A- 10 Crude incidence of hairy cell leukaemia by age and gender (per 100,000 population)

Age group (Years)	Total		Male		Female	
	N	Incidence	N	Incidence	N	Incidence
0-4	0	0.0	0	0.0	0	0.0
5-9	0	0.0	0	0.0	0	0.0
10-14	0	0.0	0	0.0	0	0.0
15-19	0	0.0	0	0.0	0	0.0
20-24	0	0.0	0	0.0	0	0.0
25-29	2	0.1	2	0.3	0	0.0
30-34	0	0.0	0	0.0	0	0.0
35-39	1	0.1	1	0.1	0	0.0
40-44	3	0.2	3	0.3	0	0.0
45-49	6	0.4	4	0.5	2	0.3
50-54	3	0.2	3	0.3	0	0.0
55-59	11	0.8	11	1.6	0	0.0
60-64	11	0.9	10	1.7	1	0.2
65-69	10	0.9	8	1.5	2	0.3
70-74	16	1.6	12	2.7	4	0.7
75-79	10	1.2	6	1.7	4	0.8
Over 80	8	0.8	5	1.5	3	0.4
Total	81	0.3	65	0.5	16	0.1

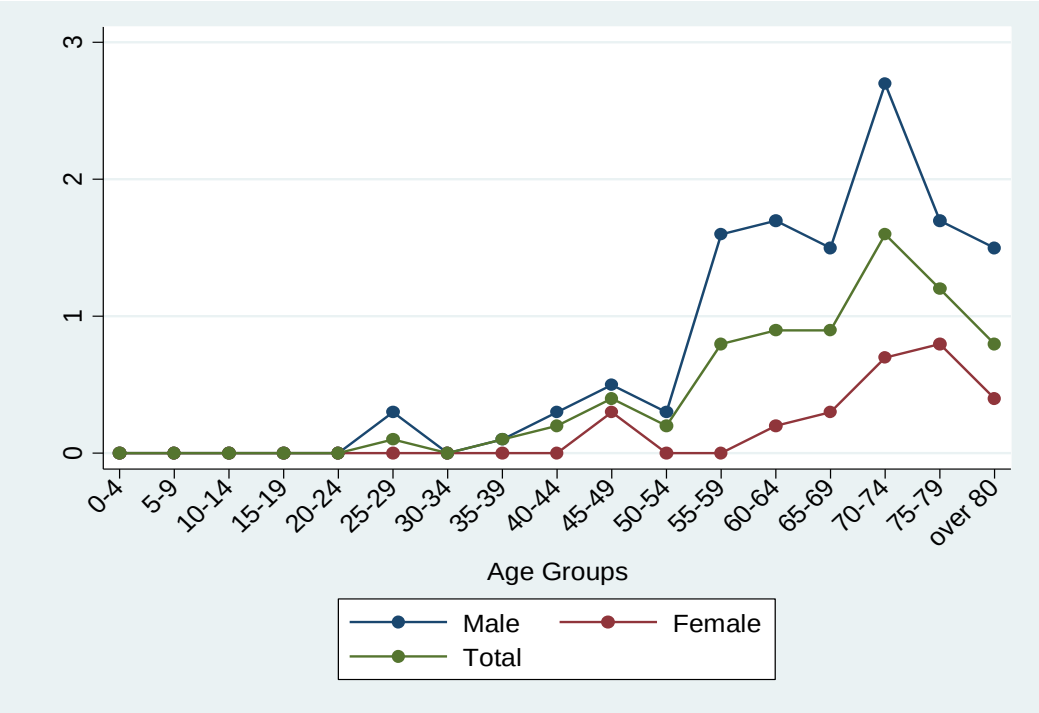


Figure A- 12 Incidence of hairy cell leukaemia per 100,000 for males, females, and total

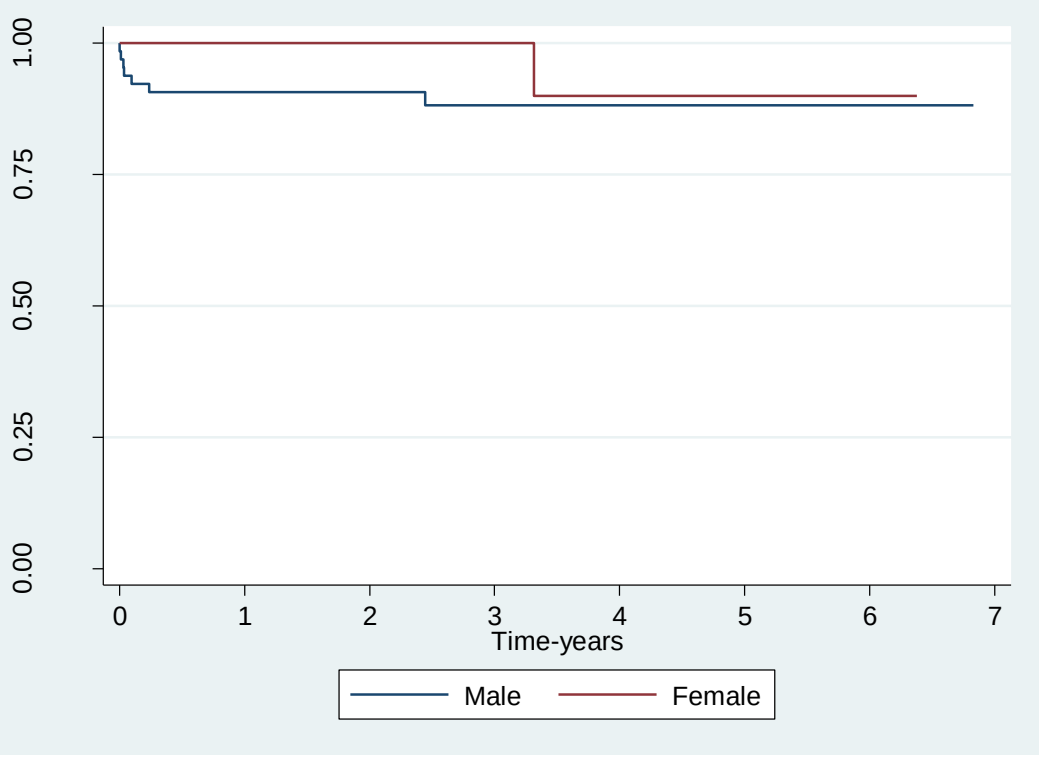


Figure A- 13 Kaplan-Meier survival estimates for hairy cell leukaemia patients by gender

7 T-cell Leukaemia

Table A- 11 Crude incidence of T-cell leukaemia by age and gender (per 100,000 population)

Age group (Years)	Total		Male		Female	
	N	Incidence	N	Incidence	N	Incidence
0-4	1	0.1	1	0.1	0	0.0
5-9	0	0.0	0	0.0	0	0.0
10-14	1	0.1	1	0.1	0	0.0
15-19	0	0.0	0	0.0	0	0.0
20-24	0	0.0	0	0.0	0	0.0
25-29	0	0.0	0	0.0	0	0.0
30-34	1	0.1	0	0.0	1	0.1
35-39	2	0.1	1	0.1	1	0.1
40-44	1	0.1	1	0.1	0	0.0
45-49	2	0.1	1	0.1	1	0.1
50-54	8	0.5	5	0.6	3	0.3
55-59	4	0.3	0	0.0	4	0.6
60-64	12	1.0	4	0.7	8	1.3
65-69	6	0.5	2	0.4	4	0.7
70-74	16	1.6	7	1.6	9	1.6
75-79	20	2.4	12	3.4	8	1.6
Over 80	26	2.5	11	3.3	15	2.1
Total	100	0.4	46	0.4	54	0.4

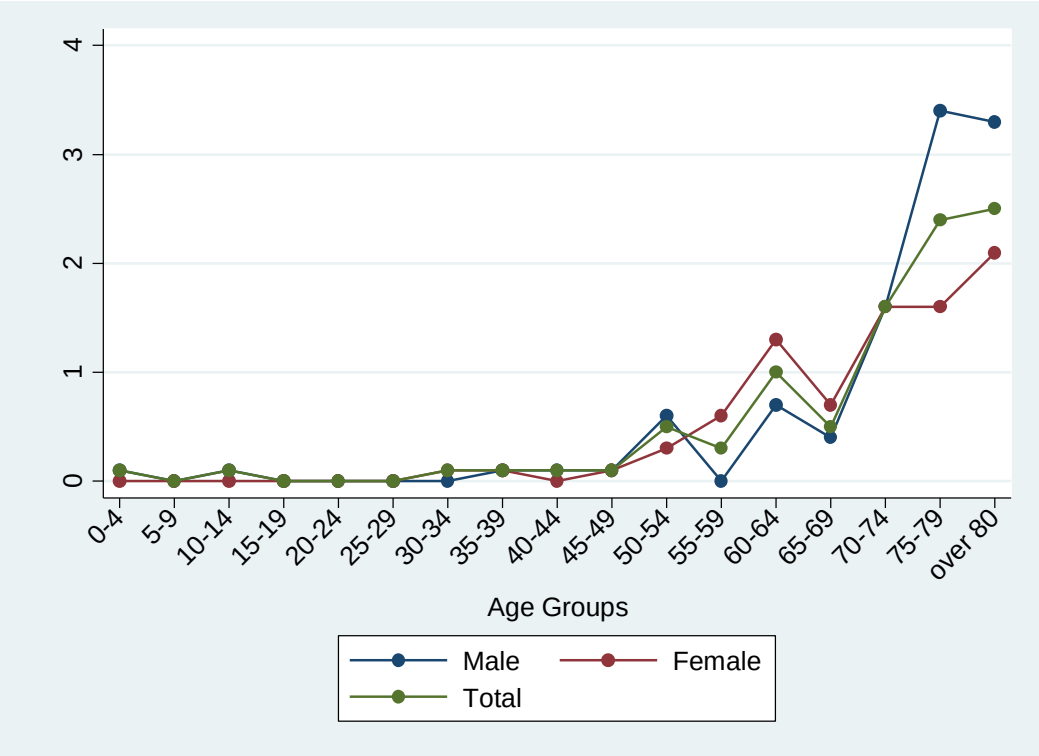


Figure A- 14 Incidence of T-cell leukaemia per 100,000 for males, females, and total

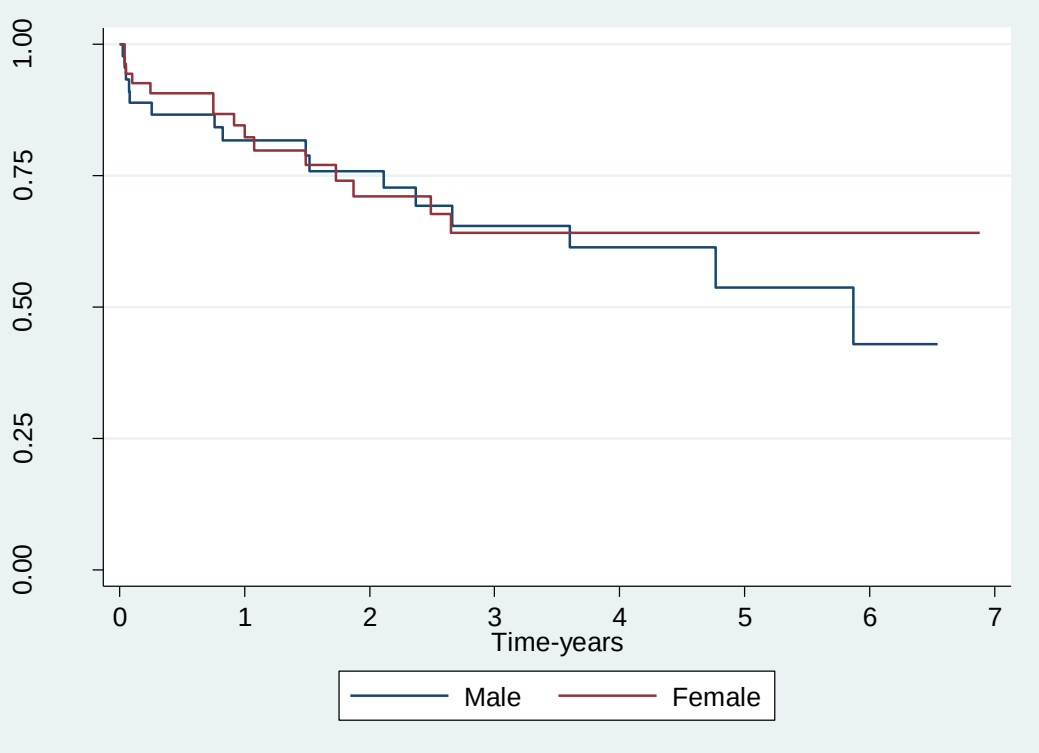


Figure A- 15 Kaplan-Meier survival estimates for T-cell leukaemia patients by gender

8 Marginal Zone Lymphoma

Table A- 12 Crude incidence of marginal zone lymphoma by age and gender (per 100,000 population)

Age group (Years)	Total		Male		Female	
	N	Incidence	N	Incidence	N	Incidence
0-4	0	0.0	0	0.0	0	0.0
5-9	0	0.0	0	0.0	0	0.0
10-14	0	0.0	0	0.0	0	0.0
15-19	0	0.0	0	0.0	0	0.0
20-24	4	0.3	2	0.3	2	0.3
25-29	2	0.1	1	0.1	1	0.1
30-34	2	0.1	2	0.2	0	0.0
35-39	6	0.3	4	0.4	2	0.2
40-44	10	0.6	3	0.3	7	0.8
45-49	25	1.6	13	1.6	12	1.5
50-54	40	2.3	20	2.3	20	2.3
55-59	63	4.5	36	5.2	27	3.9
60-64	89	7.2	56	9.3	33	5.2
65-69	115	10.3	69	13.0	46	7.8
70-74	139	13.9	81	18.2	58	10.4
75-79	161	19.1	79	22.5	82	16.6
Over 80	183	17.5	90	27.0	93	13.1
Total	839	3.4	456	3.8	383	3.0

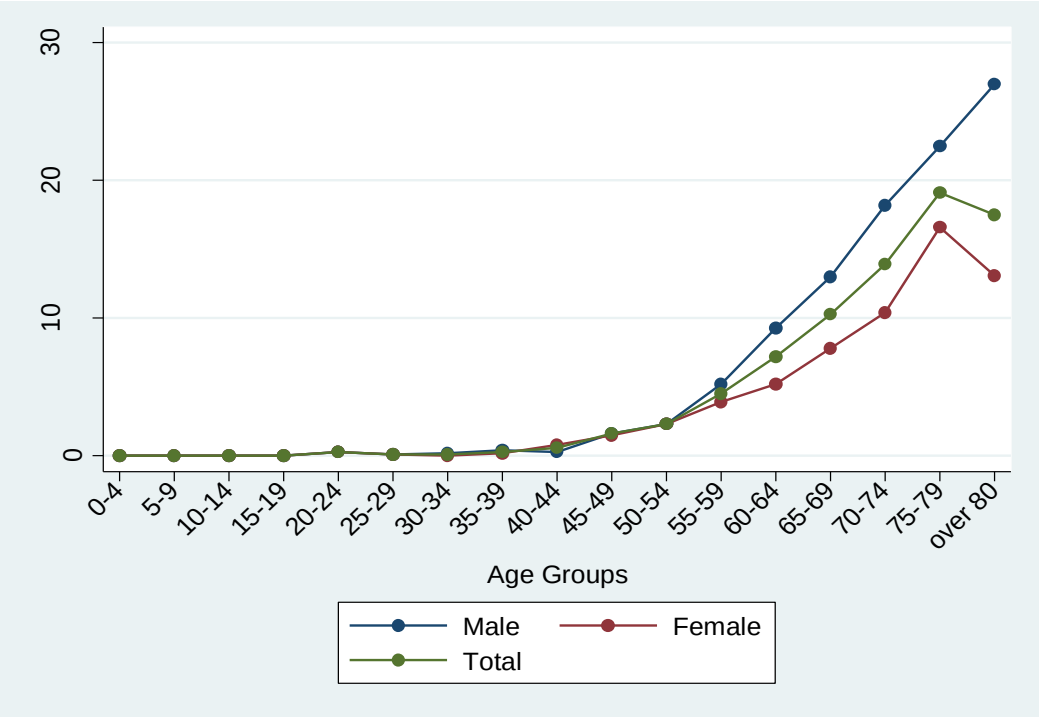


Figure A- 16 Incidence of marginal zone lymphoma per 100,000 for males, females, and total

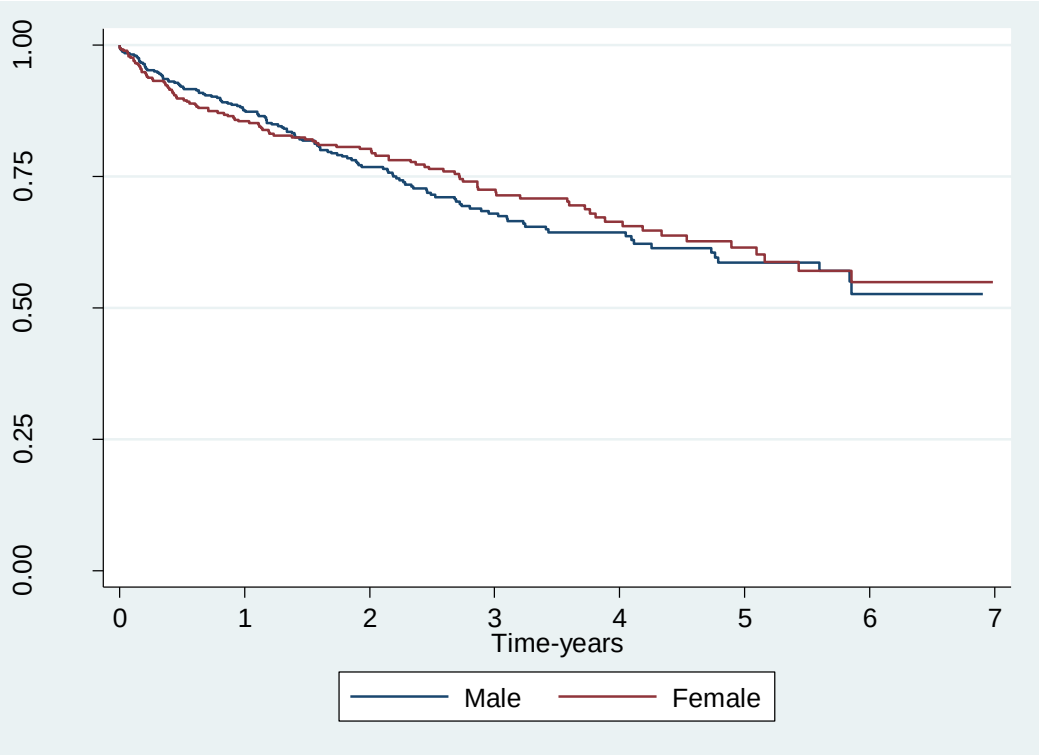


Figure A- 17 Kaplan-Meier survival estimates for marginal zone lymphoma patients by gender

9 Follicular Lymphoma

Table A- 13 Crude incidence of follicular lymphoma by age and gender (per 100,000 population)

Age group (Years)	Total		Male		Female	
	N	Incidence	N	Incidence	N	Incidence
0-4	0	0.0	0	0.0	0	0.0
5-9	0	0.0	0	0.0	0	0.0
10-14	0	0.0	0	0.0	0	0.0
15-19	1	0.1	1	0.1	0	0.0
20-24	1	0.1	1	0.1	0	0.0
25-29	3	0.2	2	0.3	1	0.1
30-34	7	0.4	5	0.6	2	0.2
35-39	19	1.0	11	1.2	8	0.8
40-44	36	2.1	21	2.4	15	1.7
45-49	50	3.1	25	3.2	26	3.3
50-54	74	4.2	30	3.4	45	5.1
55-59	87	6.2	46	6.6	46	6.6
60-64	121	9.8	63	10.4	64	10.1
65-69	104	9.3	42	7.9	67	11.4
70-74	108	10.8	53	11.9	60	10.8
75-79	74	8.8	37	10.5	46	9.3
Over 80	87	8.3	27	8.1	60	8.4
Total	772	3.1	364	3.0	440	3.4

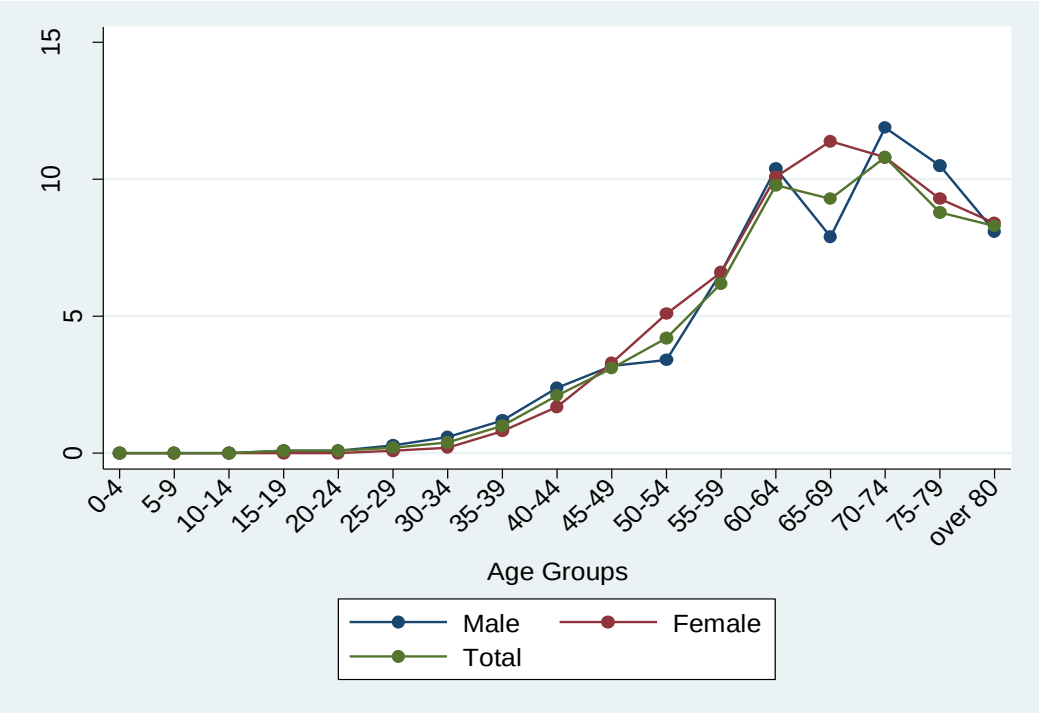


Figure A- 18 Incidence of follicular lymphoma per 100,000 for males, females, and total

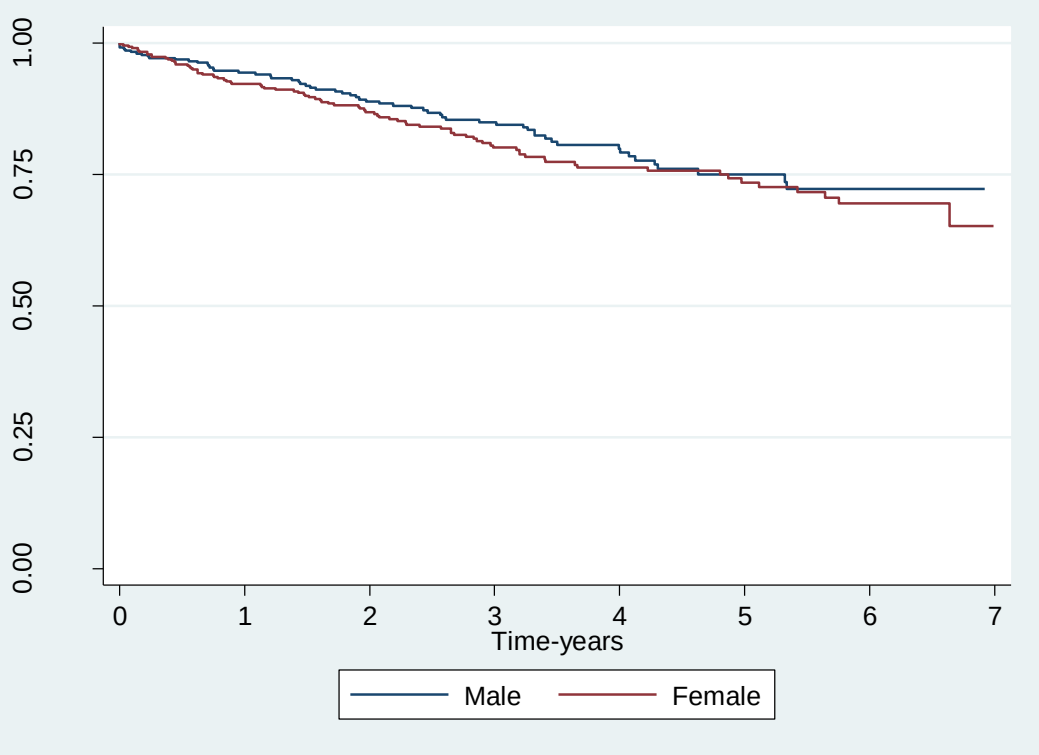


Figure A- 19 Kaplan-Meier survival estimates for follicular lymphoma patients by gender

10 Mantle Cell Lymphoma

Table A- 14 Crude incidence of mantle cell lymphoma by age and gender (per 100,000 population)

Age group (Years)	Total		Male		Female	
	N	Incidence	N	Incidence	N	Incidence
0-4	0	0.0	0	0.0	0	0.0
5-9	0	0.0	0	0.0	0	0.0
10-14	0	0.0	0	0.0	0	0.0
15-19	0	0.0	0	0.0	0	0.0
20-24	0	0.0	0	0.0	0	0.0
25-29	0	0.0	0	0.0	0	0.0
30-34	0	0.0	0	0.0	0	0.0
35-39	2	0.1	2	0.2	0	0.0
40-44	1	0.1	1	0.1	0	0.0
45-49	4	0.3	4	0.5	0	0.0
50-54	9	0.5	6	0.7	3	0.3
55-59	19	1.4	17	2.4	2	0.3
60-64	21	1.7	12	2.0	9	1.4
65-69	33	3.0	23	4.3	10	1.7
70-74	33	3.3	19	4.3	14	2.5
75-79	40	4.7	25	7.1	15	3.0
Over 80	57	5.5	32	9.6	25	3.5
Total	219	0.9	141	1.2	78	0.6

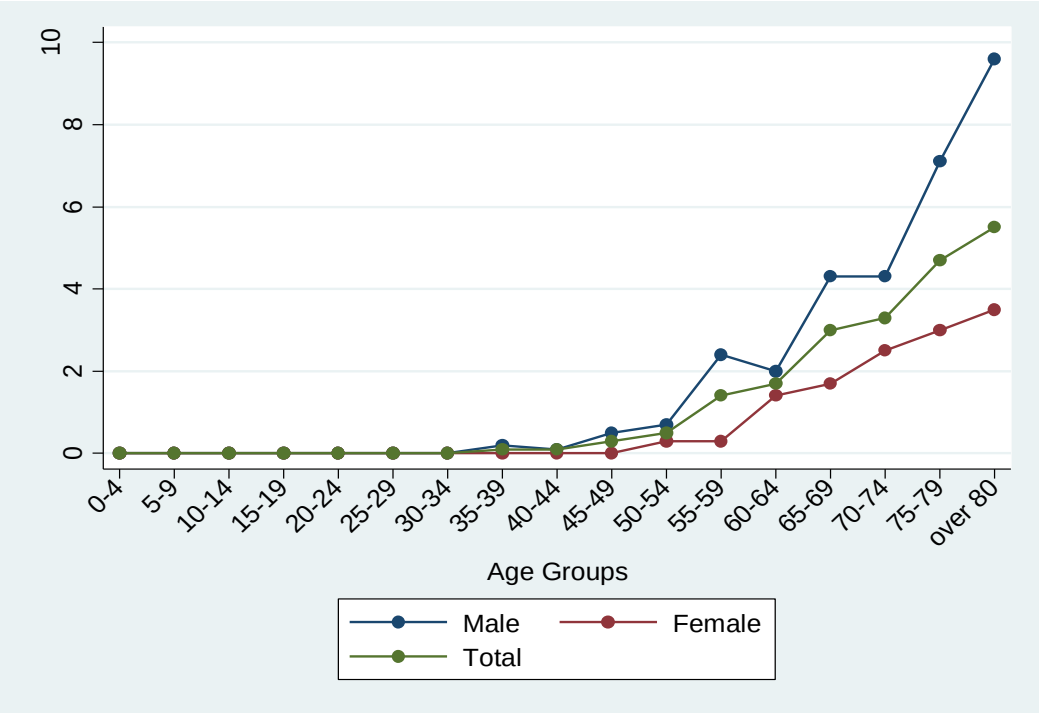


Figure A- 20 Incidence of mantle cell lymphoma per 100,000 for males, females, and total

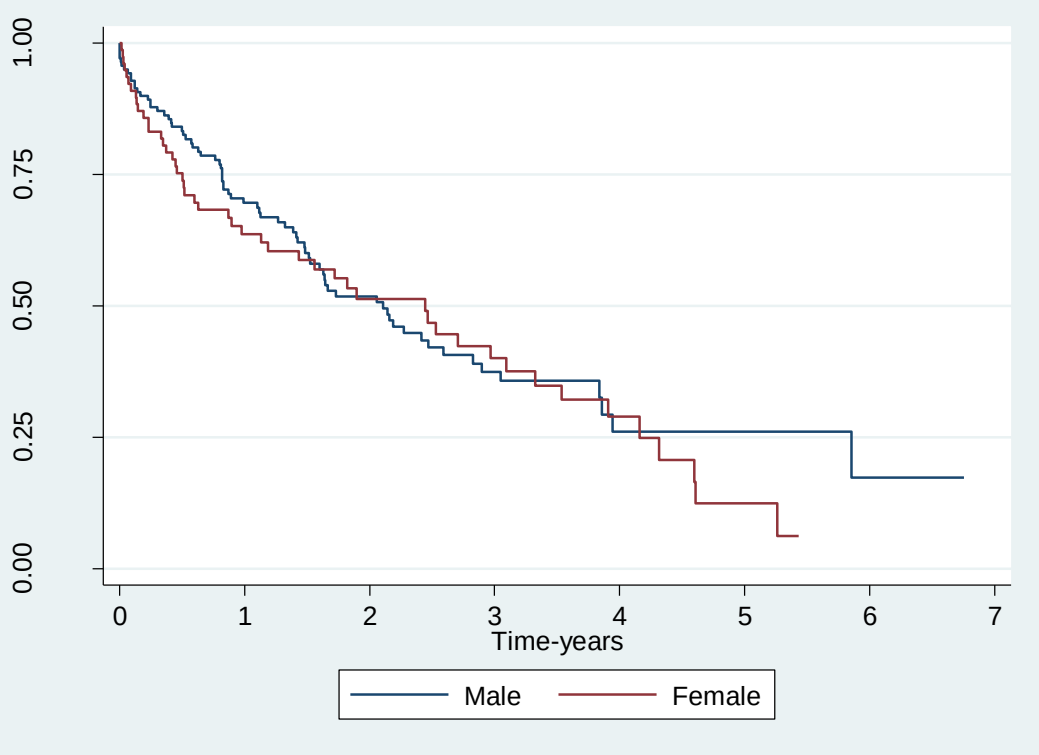


Figure A- 21 Kaplan-Meier survival estimates for mantle cell lymphoma patients by gender

11 Diffuse Large B-cell Lymphoma

Table A- 15 Crude incidence of diffuse large B-cell lymphoma by age and gender (per 100,000 population)

Age group (Years)	Total		Male		Female	
	N	Incidence	N	Incidence	N	Incidence
0-4	3	0.2	3	0.4	0	0.0
5-9	2	0.1	2	0.2	0	0.0
10-14	8	0.5	4	0.5	4	0.5
15-19	8	0.5	5	0.6	3	0.4
20-24	13	0.9	8	1.1	5	0.6
25-29	20	1.3	13	1.7	7	0.9
30-34	26	1.4	16	1.8	10	1.1
35-39	40	2.1	20	2.1	20	2.1
40-44	70	4.0	43	5.0	27	3.1
45-49	78	4.9	47	5.9	31	3.9
50-54	116	6.6	66	7.6	50	5.7
55-59	168	12.1	103	14.8	65	9.3
60-64	210	17.0	99	16.4	111	17.6
65-69	280	25.1	156	29.4	124	21.1
70-74	312	31.1	165	37.0	147	26.4
75-79	313	37.1	164	46.7	149	30.2
Over 80	399	38.2	166	49.8	233	32.8
Total	2066	8.3	1080	8.9	986	7.7

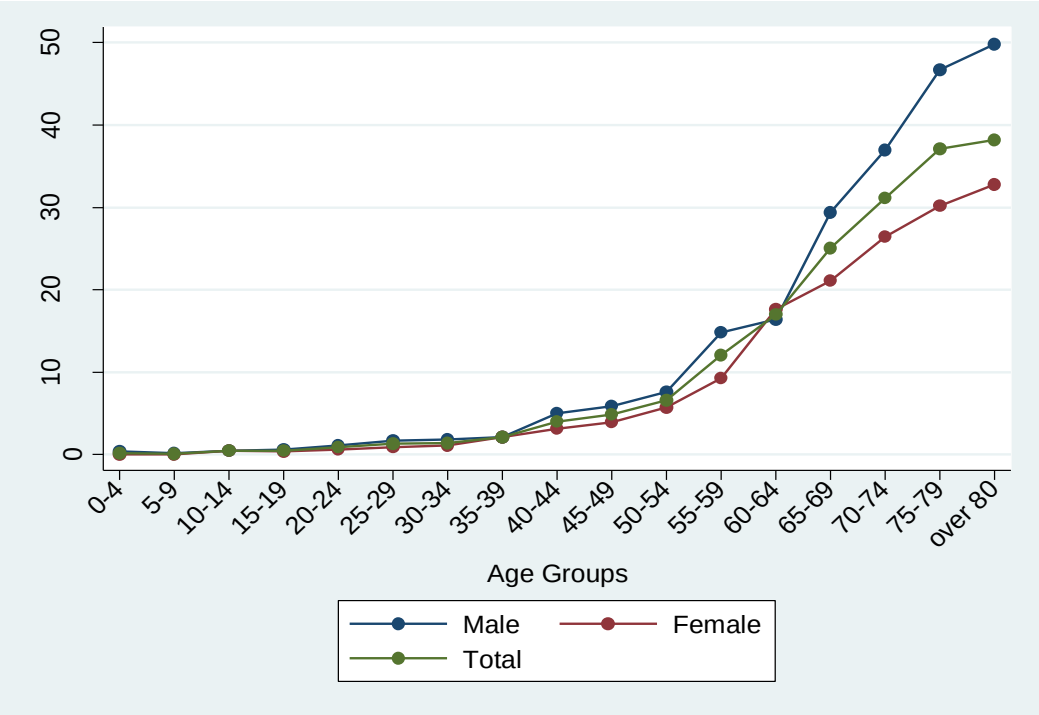


Figure A- 22 Incidence of diffuse large B-cell lymphoma per 100,000 for males, females, and total

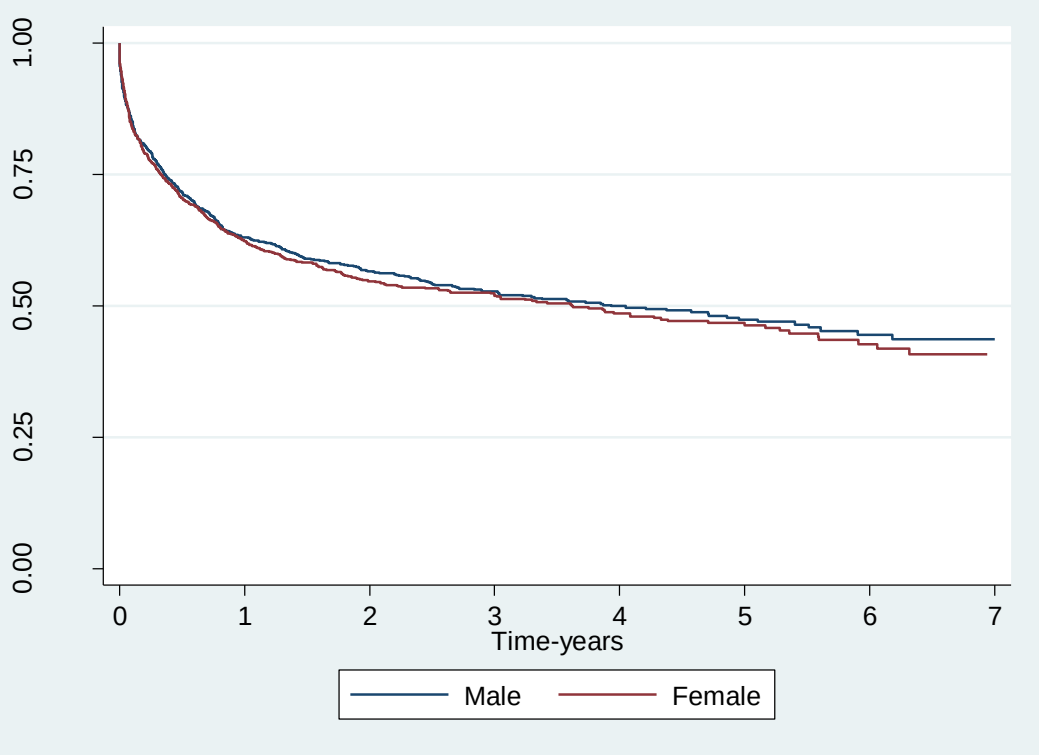


Figure A- 23 Kaplan-Meier survival estimates for diffuse large B-cell lymphoma patients by gender

12 Burkitt Lymphoma

Table A- 16 Crude incidence of Burkitt lymphoma by age and gender (per 100,000 population)

Age group (Years)	Total		Male		Female	
	N	Incidence	N	Incidence	N	Incidence
0-4	8	0.5	5	0.7	3	0.4
5-9	6	0.4	6	0.7	0	0.0
10-14	8	0.5	8	0.9	0	0.0
15-19	2	0.1	2	0.2	0	0.0
20-24	6	0.4	5	0.7	1	0.1
25-29	3	0.2	3	0.4	0	0.0
30-34	3	0.2	3	0.3	0	0.0
35-39	3	0.2	1	0.1	2	0.2
40-44	1	0.1	1	0.1	0	0.0
45-49	2	0.1	1	0.1	1	0.1
50-54	5	0.3	3	0.3	2	0.2
55-59	8	0.6	5	0.7	3	0.4
60-64	11	0.9	7	1.2	4	0.6
65-69	4	0.4	3	0.6	1	0.2
70-74	6	0.6	3	0.7	3	0.5
75-79	8	0.9	7	2.0	1	0.2
Over 80	3	0.3	2	0.6	1	0.1
Total	87	0.3	65	0.5	22	0.2

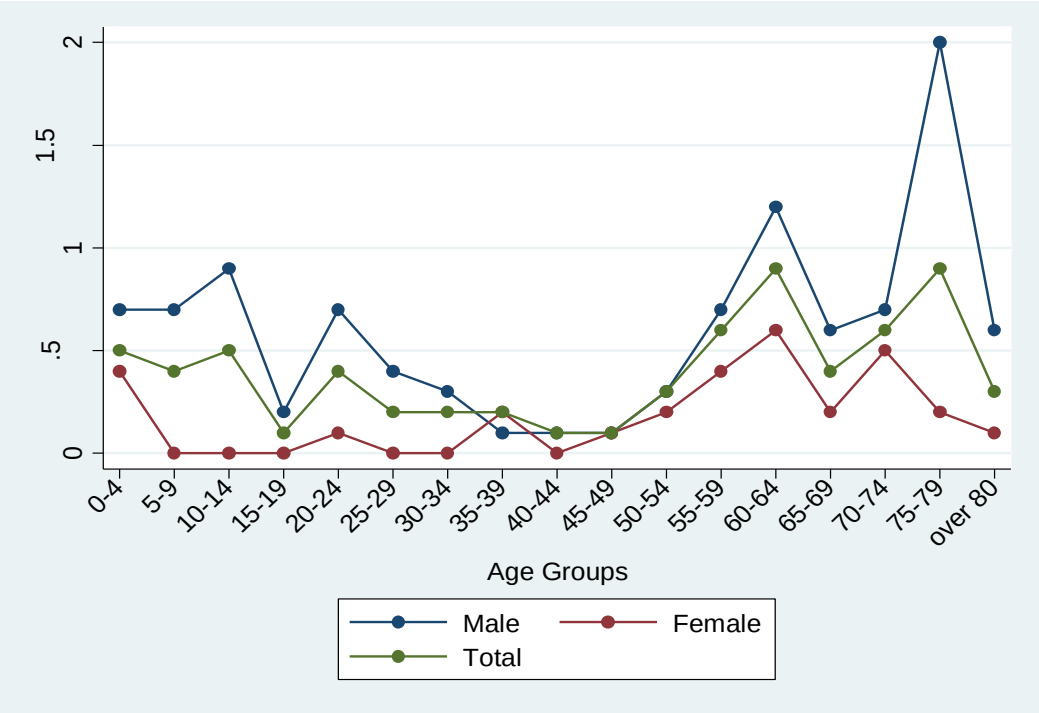


Figure A- 24 Incidence of Burkitt lymphoma per 100,000 for males, females, and total

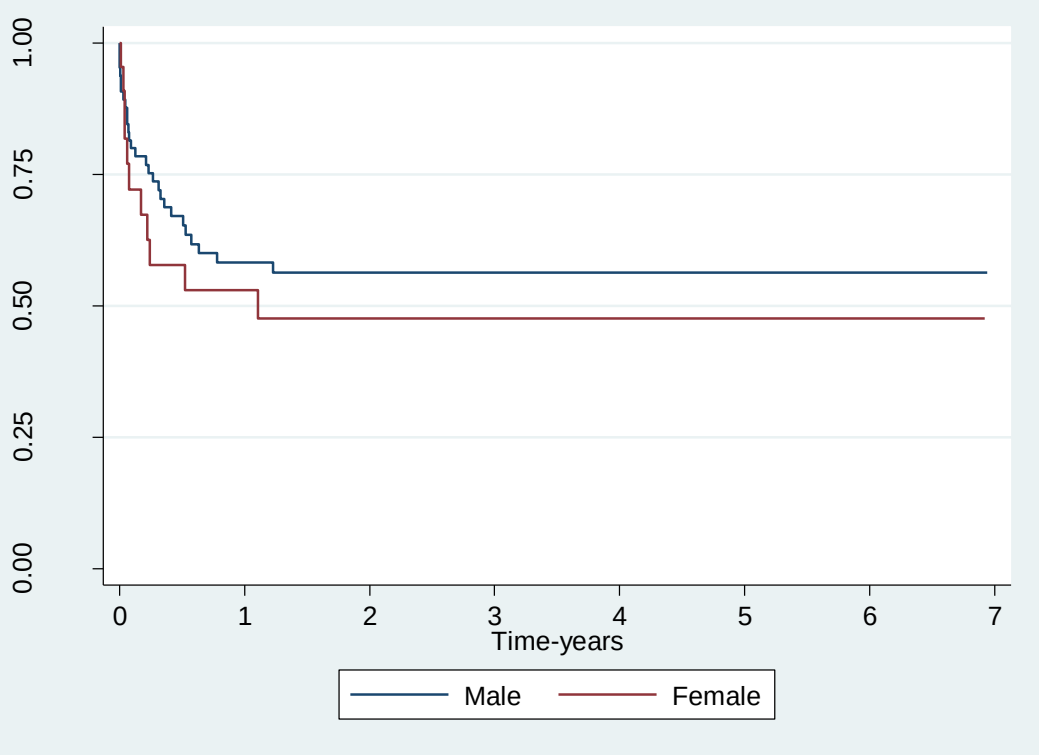


Figure A- 25 Kaplan-Meier survival estimates for Burkitt lymphoma patients by gender

13 T-cell Lymphoma

Table A- 17 Crude incidence of T-cell lymphoma by age and gender (per 100,000 population)

Age group (Years)	Total		Male		Female	
	N	Incidence	N	Incidence	N	Incidence
0-4	2	0.1	1	0.1	1	0.1
5-9	0	0.0	0	0.0	0	0.0
10-14	2	0.1	1	0.1	1	0.1
15-19	3	0.2	1	0.1	2	0.2
20-24	1	0.1	0	0.0	1	0.1
25-29	8	0.5	6	0.8	2	0.2
30-34	3	0.2	3	0.3	0	0.0
35-39	10	0.5	6	0.6	4	0.4
40-44	20	1.1	14	1.6	6	0.7
45-49	13	0.8	8	1.0	5	0.6
50-54	18	1.0	9	1.0	9	1.0
55-59	21	1.5	14	2.0	7	1.0
60-64	28	2.3	16	2.6	12	1.9
65-69	27	2.4	18	3.4	9	1.5
70-74	37	3.7	22	4.9	15	2.7
75-79	27	3.2	14	4.0	13	2.6
Over 80	36	3.5	15	4.5	21	3.0
Total	256	1.0	148	1.2	108	0.8

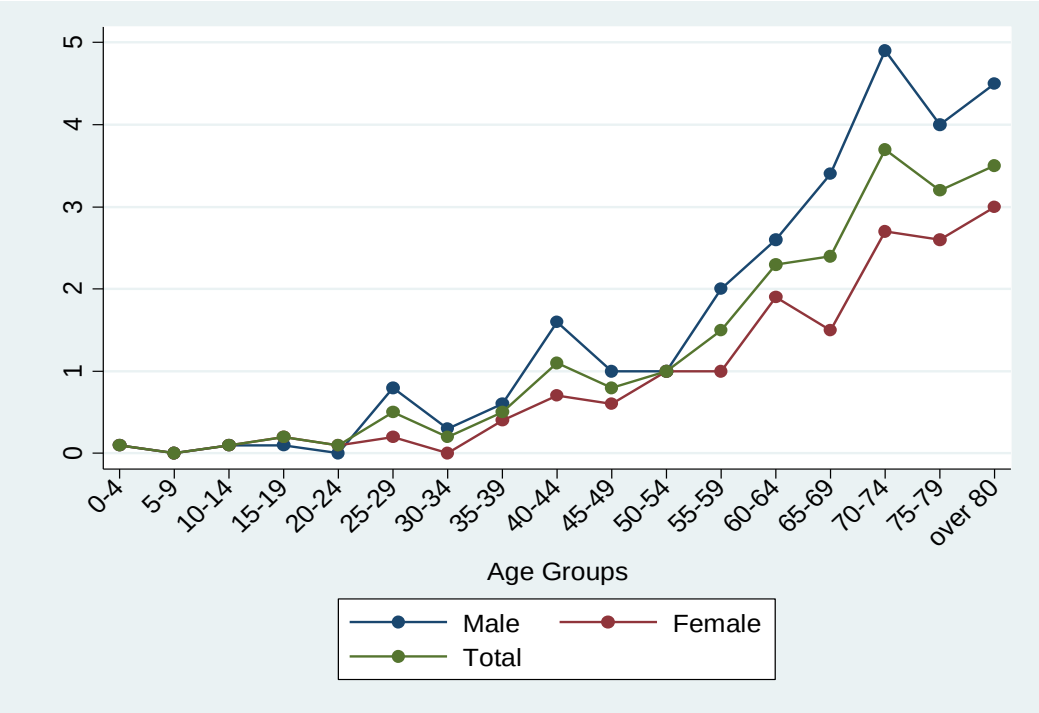


Figure A- 26 Incidence of T-cell lymphoma per 100,000 for males, females, and total

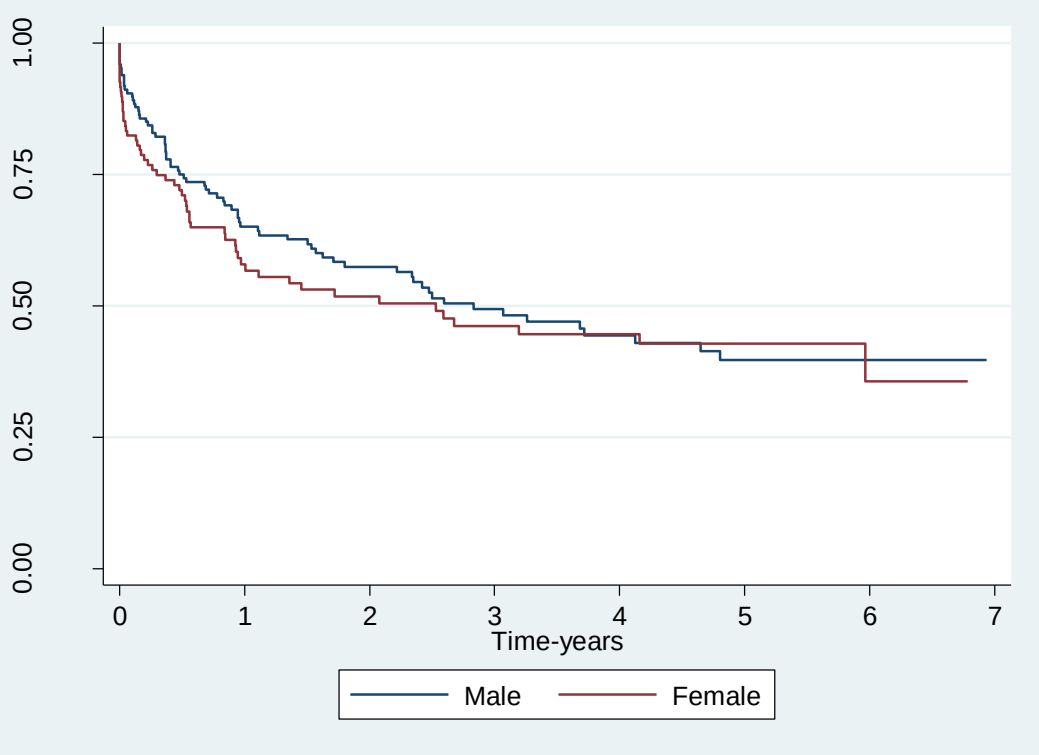


Figure A- 27 Kaplan-Meier survival estimates for T-cell lymphoma patients by gender

14 Hodgkin Lymphoma

Table A- 18 Crude incidence of Hodgkin lymphoma by age and gender (per 100,000 population)

Age group (Years)	Total		Male		Female	
	N	Incidence	N	Incidence	N	Incidence
0-4	2	0.1	2	0.3	0	0.0
5-9	4	0.2	3	0.4	1	0.1
10-14	23	1.3	11	1.3	12	1.4
15-19	59	3.6	32	3.9	27	3.4
20-24	71	4.7	42	5.6	29	3.7
25-29	76	4.9	35	4.6	41	5.1
30-34	68	3.7	37	4.1	31	3.3
35-39	64	3.4	38	4.1	26	2.7
40-44	45	2.6	32	3.7	13	1.5
45-49	40	2.5	29	3.7	11	1.4
50-54	42	2.4	30	3.4	12	1.4
55-59	43	3.1	23	3.3	20	2.9
60-64	42	3.4	27	4.5	15	2.4
65-69	51	4.6	22	4.2	29	4.9
70-74	53	5.3	27	6.1	26	4.7
75-79	39	4.6	22	6.3	17	3.4
Over 80	32	3.1	13	3.9	19	2.7
Total	754	3.0	425	3.5	329	2.6

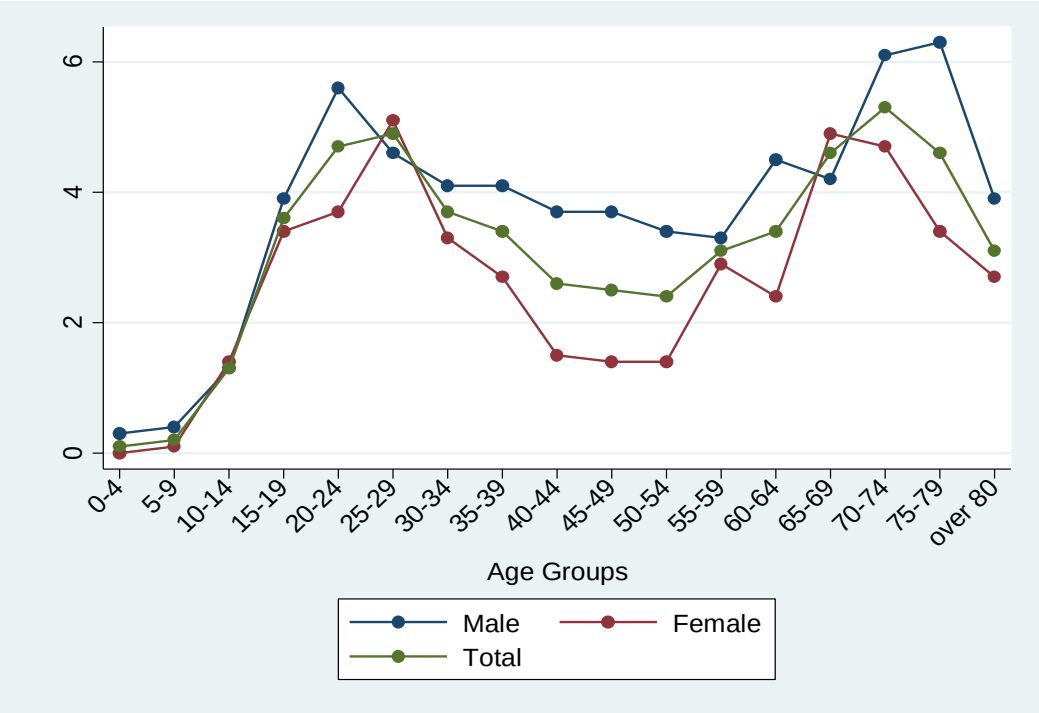


Figure A- 28 Incidence of Hodgkin lymphoma per 100,000 for males, females, and total

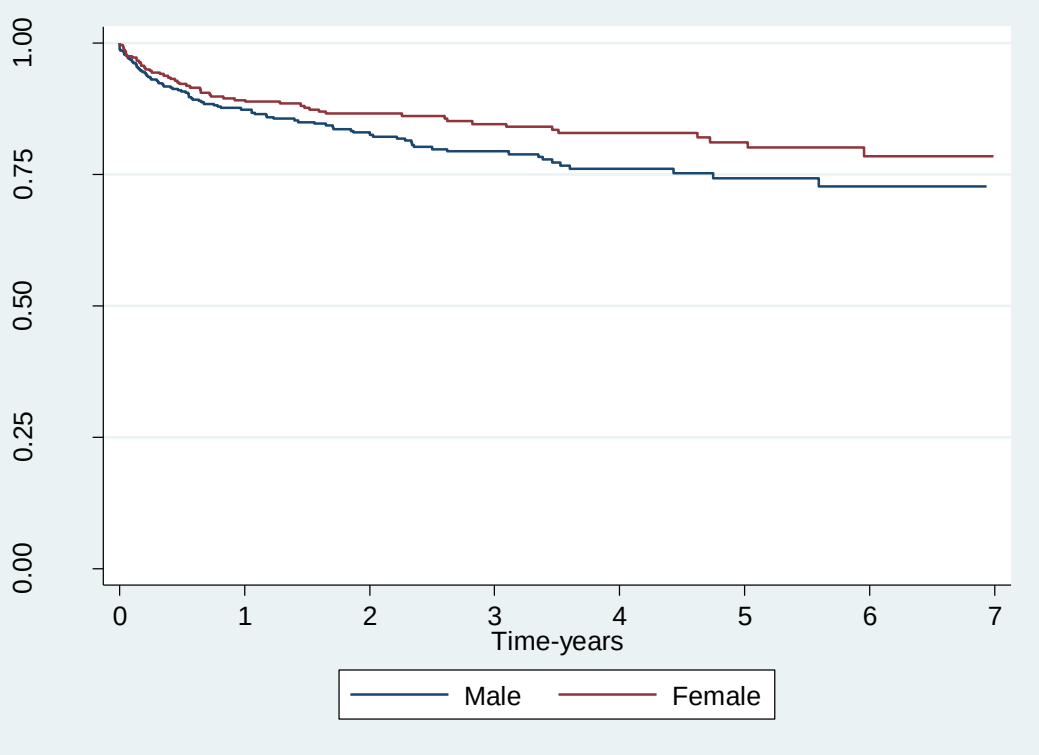


Figure A- 29 Kaplan-Meier survival estimates for Hodgkin lymphoma patients by gender

15 Plasma Cell Myeloma

Table A- 19 Crude incidence of plasma cell myeloma by age and gender (per 100,000 population)

Age group (Years)	Total		Male		Female	
	N	Incidence	N	Incidence	N	Incidence
0-4	0	0.0	0	0.0	0	0.0
5-9	0	0.0	0	0.0	0	0.0
10-14	0	0.0	0	0.0	0	0.0
15-19	0	0.0	0	0.0	0	0.0
20-24	0	0.0	0	0.0	0	0.0
25-29	0	0.0	0	0.0	0	0.0
30-34	3	0.2	3	0.3	0	0.0
35-39	10	0.5	6	0.6	4	0.4
40-44	12	0.7	6	0.7	6	0.7
45-49	50	3.1	36	4.5	14	1.8
50-54	85	4.8	54	6.2	31	3.5
55-59	117	8.4	68	9.8	49	7.0
60-64	183	14.8	126	20.8	57	9.0
65-69	204	18.3	115	21.7	89	15.1
70-74	284	28.3	152	34.1	132	23.7
75-79	277	32.8	164	46.7	113	22.9
Over 80	421	40.3	219	65.7	202	28.4
Total	1646	6.6	949	7.8	697	5.4

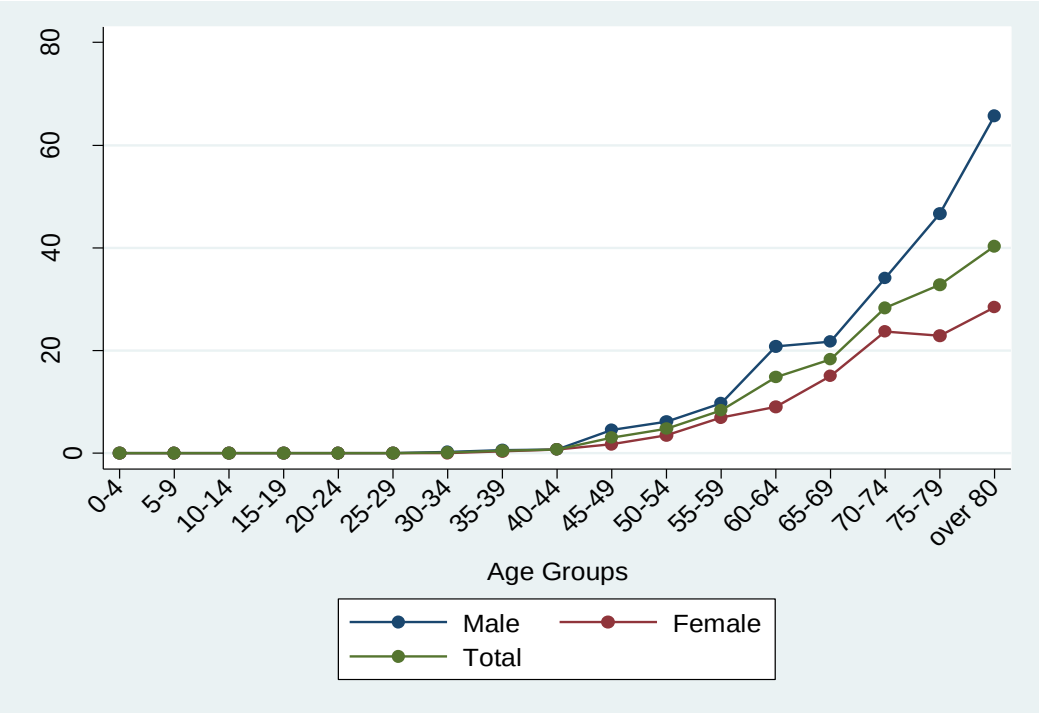


Figure A- 30 Incidence of plasma cell myeloma per 100,000 for males, females, and total

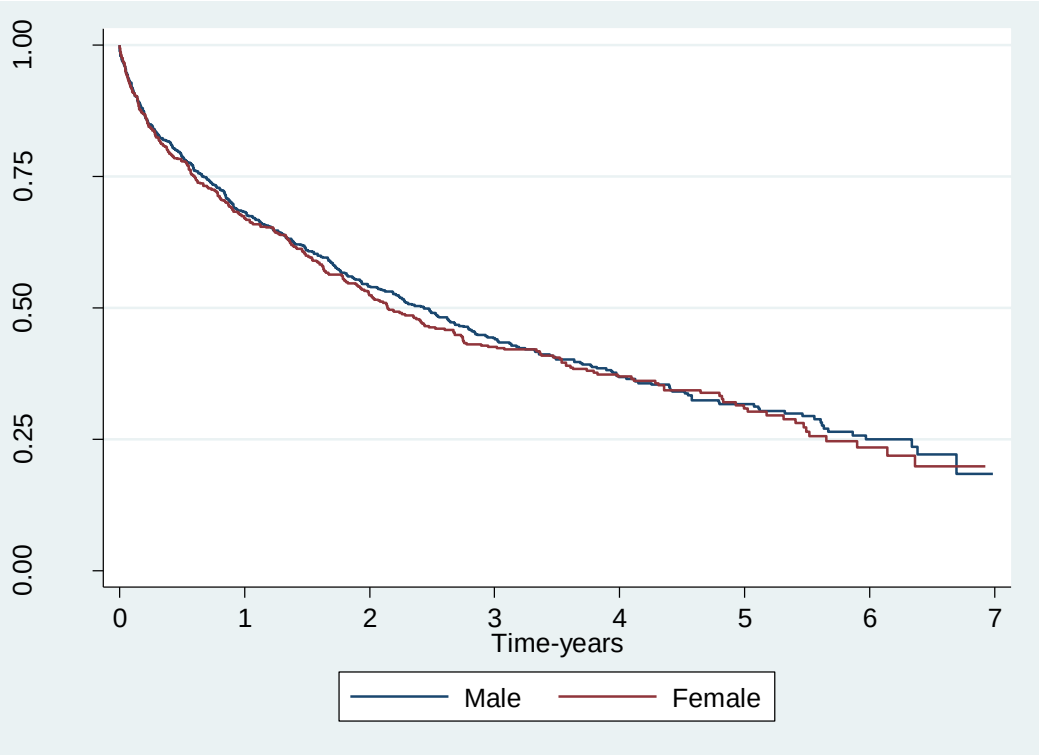


Figure A- 31 Kaplan-Meier survival estimates for plasma cell myeloma patients by gender

16 Plasmacytoma

Table A- 20 Crude incidence of plasmacytoma by age and gender (per 100,000 population)

Age group (Years)	Total		Male		Female	
	N	Incidence	N	Incidence	N	Incidence
0-4	0	0.0	0	0.0	0	0.0
5-9	0	0.0	0	0.0	0	0.0
10-14	0	0.0	0	0.0	0	0.0
15-19	0	0.0	0	0.0	0	0.0
20-24	0	0.0	0	0.0	0	0.0
25-29	0	0.0	0	0.0	0	0.0
30-34	0	0.0	0	0.0	0	0.0
35-39	2	0.1	1	0.1	1	0.1
40-44	4	0.2	2	0.2	2	0.2
45-49	11	0.7	7	0.9	4	0.5
50-54	4	0.2	2	0.2	2	0.2
55-59	16	1.1	13	1.9	3	0.4
60-64	23	1.9	18	3.0	5	0.8
65-69	22	2.0	17	3.2	5	0.9
70-74	26	2.6	17	3.8	9	1.6
75-79	24	2.8	18	5.1	6	1.2
Over 80	16	1.5	7	2.1	9	1.3
Total	148	0.6	102	0.8	46	0.4

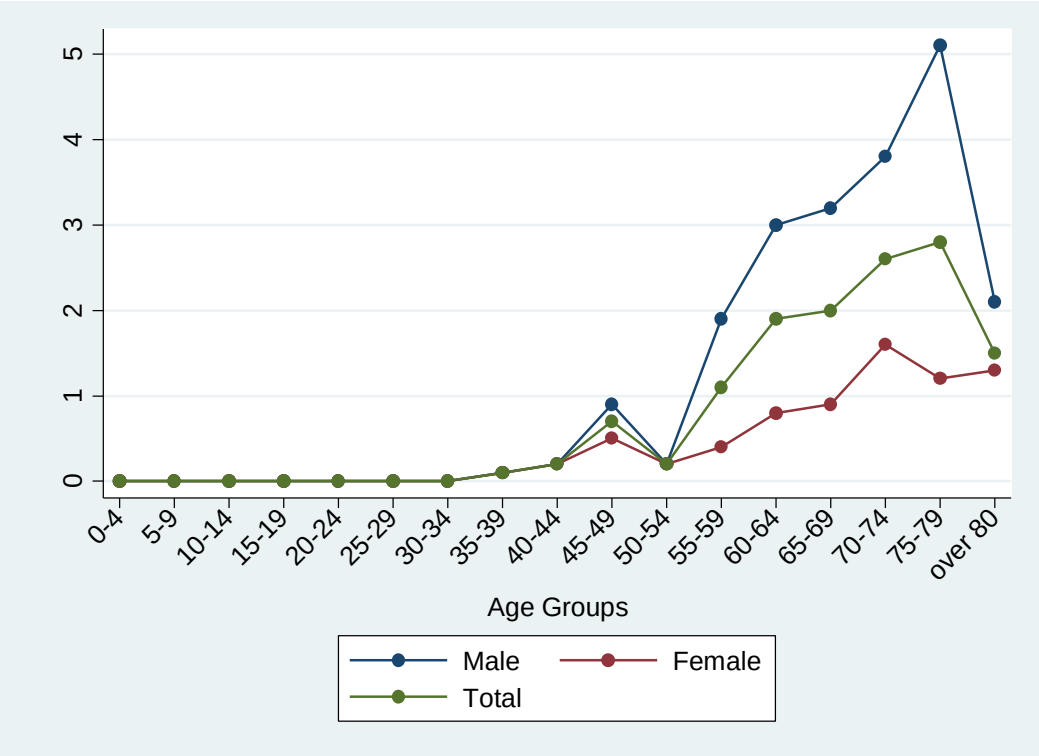


Figure A- 32 Incidence of plasmacytoma per 100,000 for males, females, and total

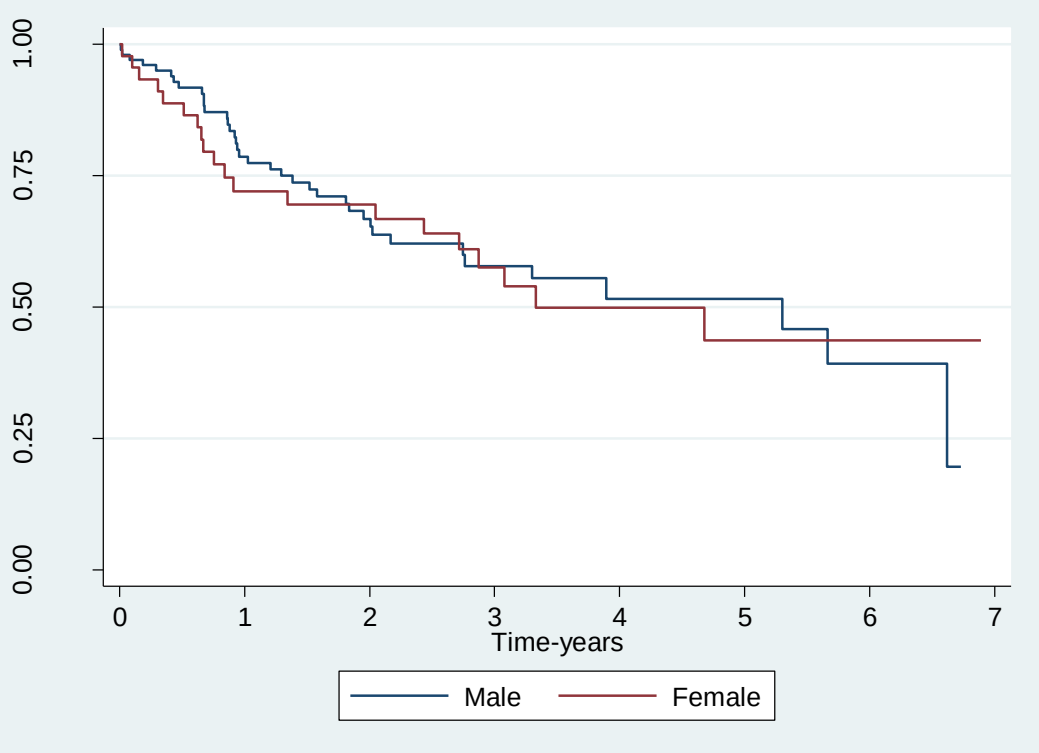


Figure A- 33 Kaplan-Meier survival estimates for plasmacytoma patients by gender

Table A- 21 Crude incidence of myelodysplastic syndromes by age and gender (per 100,000 population)

Age group (Years)	Total		Male		Female	
	N	Incidence	N	Incidence	N	Incidence
0-4	1	0.1	0	0.0	1	0.1
5-9	1	0.1	0	0.0	1	0.1
10-14	2	0.1	1	0.1	1	0.1
15-19	1	0.1	0	0.0	1	0.1
20-24	0	0.0	0	0.0	0	0.0
25-29	3	0.2	0	0.0	3	0.4
30-34	7	0.4	5	0.6	2	0.2
35-39	3	0.2	2	0.2	1	0.1
40-44	10	0.6	3	0.3	7	0.8
45-49	9	0.6	6	0.8	3	0.4
50-54	24	1.4	15	1.7	9	1.0
55-59	37	2.7	26	3.7	11	1.6
60-64	72	5.8	50	8.3	22	3.5
65-69	106	9.5	73	13.8	33	5.6
70-74	172	17.2	117	26.2	55	9.9
75-79	185	21.9	137	39.0	48	9.7
Over 80	311	29.8	192	57.6	119	16.8
Total	944	3.8	627	5.2	317	2.5

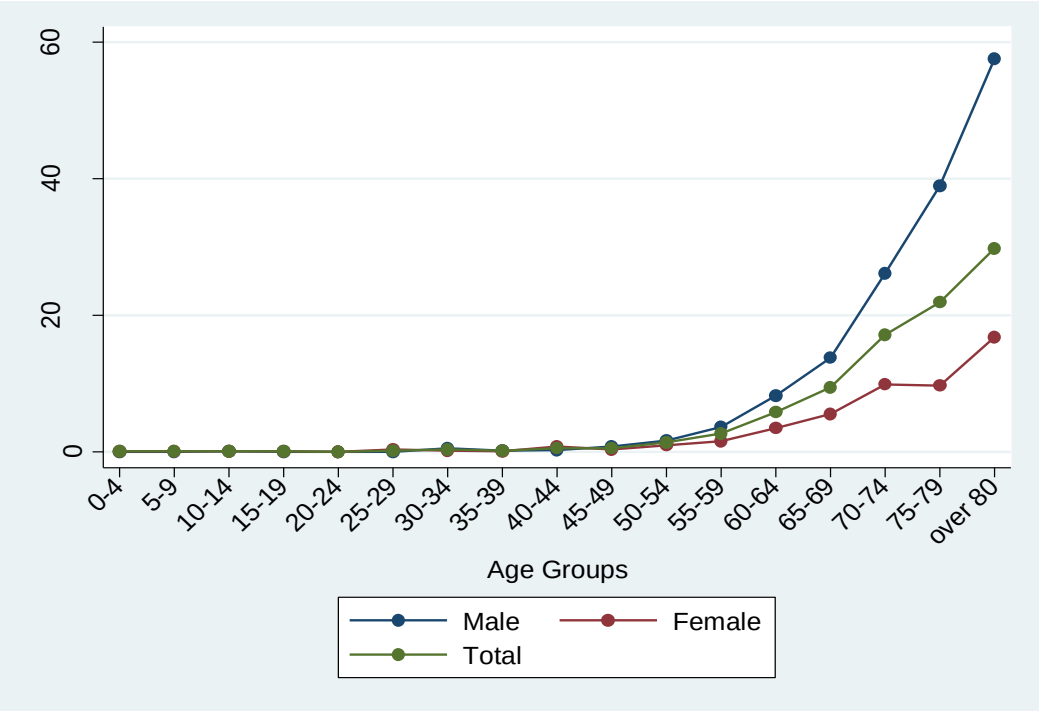


Figure A- 34 Incidence of myelodysplastic syndromes per 100,000 for males, females, and total

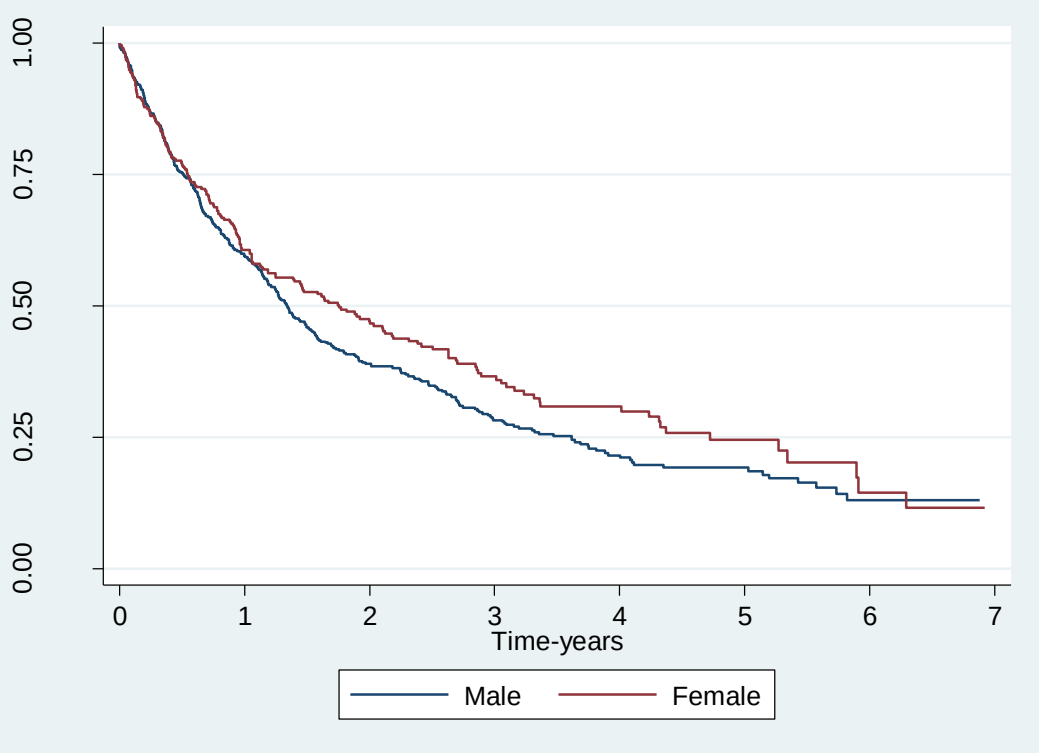


Figure A- 35 Kaplan-Meier survival estimates for myelodysplastic syndromes patients by gender

18 Myeloproliferative Neoplasms

Table A- 22 Crude incidence of myeloproliferative neoplasms by age and gender (per 100,000 population)

Age group (Years)	Total		Male		Female	
	N	Incidence	N	Incidence	N	Incidence
0-4	1	0.1	1	0.1	0	0.0
5-9	0	0.0	0	0.0	0	0.0
10-14	0	0.0	0	0.0	0	0.0
15-19	5	0.3	3	0.4	2	0.2
20-24	7	0.5	4	0.5	3	0.4
25-29	19	1.2	5	0.7	14	1.7
30-34	19	1.0	7	0.8	12	1.3
35-39	34	1.8	16	1.7	18	1.9
40-44	36	2.1	16	1.8	20	2.3
45-49	71	4.5	33	4.2	38	4.8
50-54	65	3.7	31	3.6	34	3.9
55-59	122	8.8	66	9.5	56	8.0
60-64	155	12.5	90	14.9	65	10.3
65-69	182	16.3	97	18.3	85	14.5
70-74	231	23.1	102	22.9	129	23.2
75-79	238	28.2	98	27.9	140	28.4
Over 80	368	35.3	143	42.9	225	31.7
Total	1553	6.2	712	5.9	841	6.5

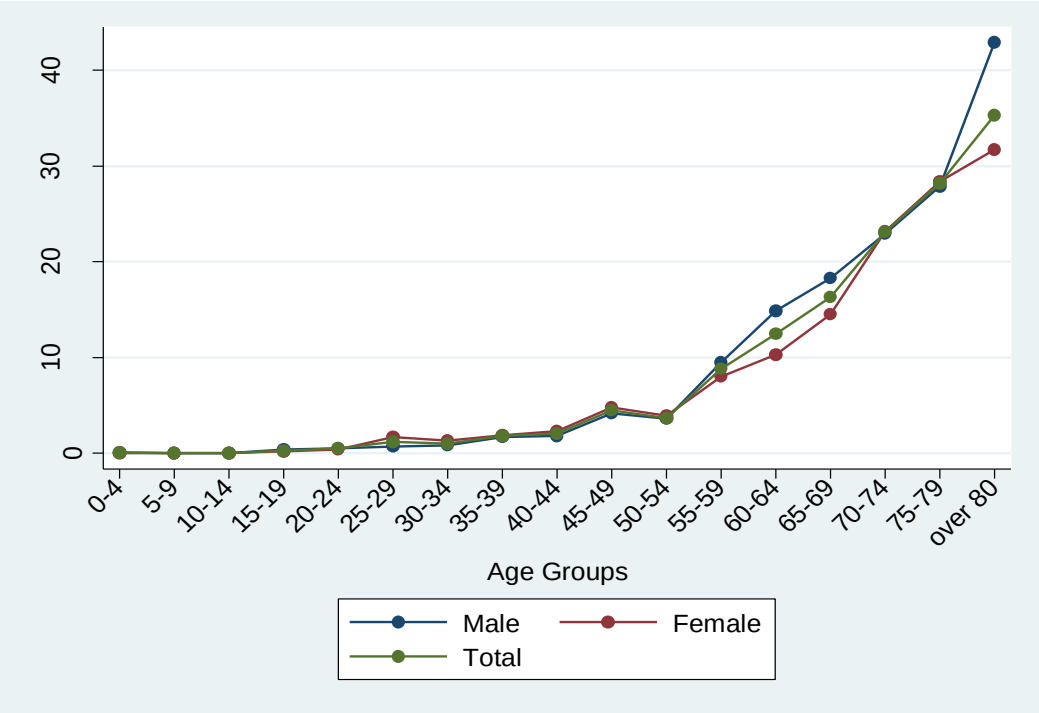


Figure A- 36 Incidence of myeloproliferative neoplasms per 100,000 for males females, and total

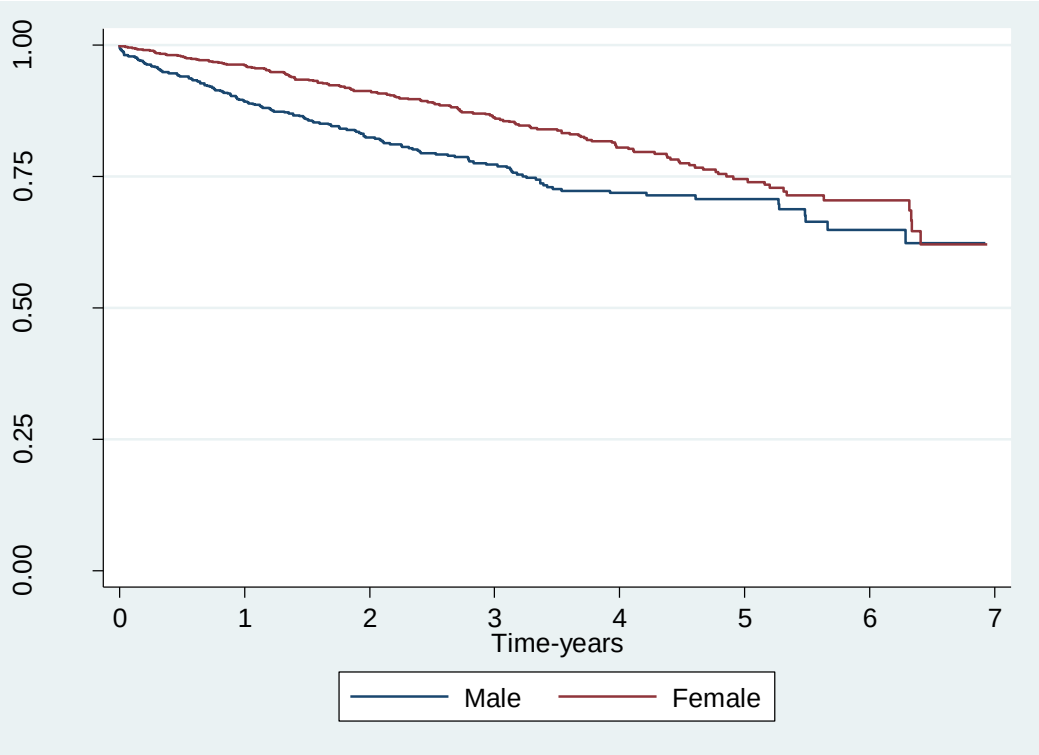


Figure A- 37 Kaplan-Meier survival estimates for myeloproliferative neoplasms patients by gender

19 Monoclonal B-cell Lymphocytosis

Table A- 23 Crude incidence of monoclonal B-cell Lymphocytosis by age and gender (per 100,000 population)

Age group (Years)	Total		Male		Female	
	N	Incidence	N	Incidence	N	Incidence
0-4	0	0.0	0	0.0	0	0.0
5-9	0	0.0	0	0.0	0	0.0
10-14	0	0.0	0	0.0	0	0.0
15-19	0	0.0	0	0.0	0	0.0
20-24	0	0.0	0	0.0	0	0.0
25-29	0	0.0	0	0.0	0	0.0
30-34	0	0.0	0	0.0	0	0.0
35-39	1	0.1	0	0.0	1	0.1
40-44	3	0.2	3	0.3	0	0.0
45-49	15	0.9	11	1.4	4	0.5
50-54	32	1.8	16	1.8	16	1.8
55-59	53	3.8	32	4.6	21	3.0
60-64	99	8.0	57	9.4	42	6.7
65-69	95	8.5	55	10.4	40	6.8
70-74	118	11.8	61	13.7	57	10.3
75-79	114	13.5	55	15.6	59	12.0
Over 80	160	15.3	86	25.8	74	10.4
Total	690	2.8	376	3.1	314	2.4

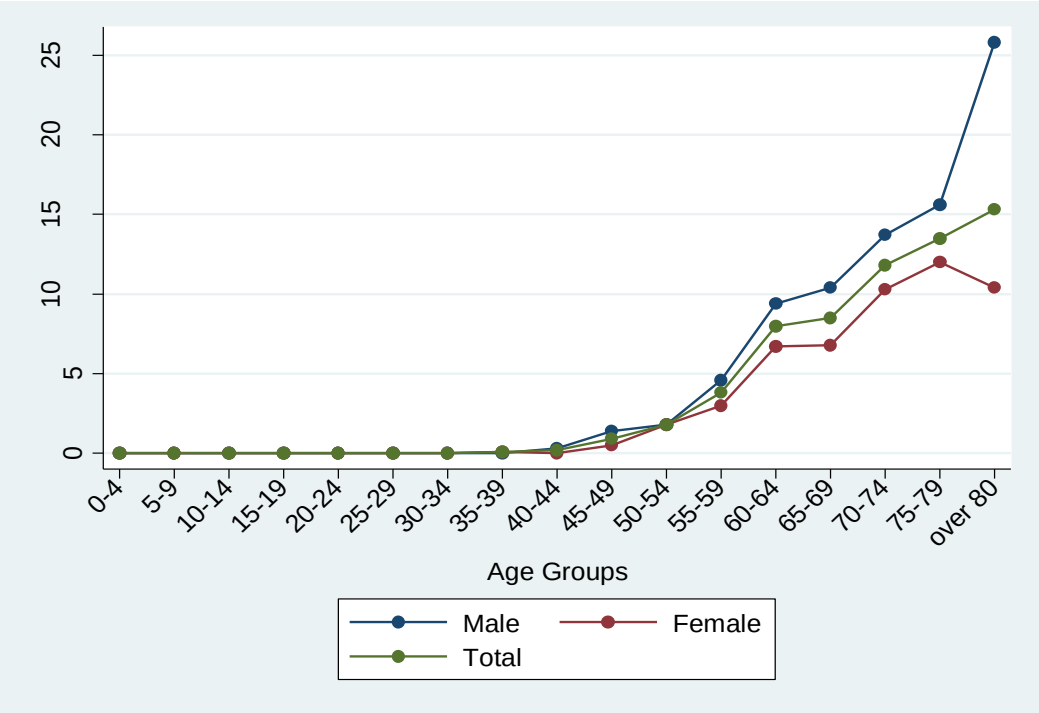


Figure A- 38 Incidence of monoclonal B-cell Lymphocytosis per 100,000 for males, females, and total

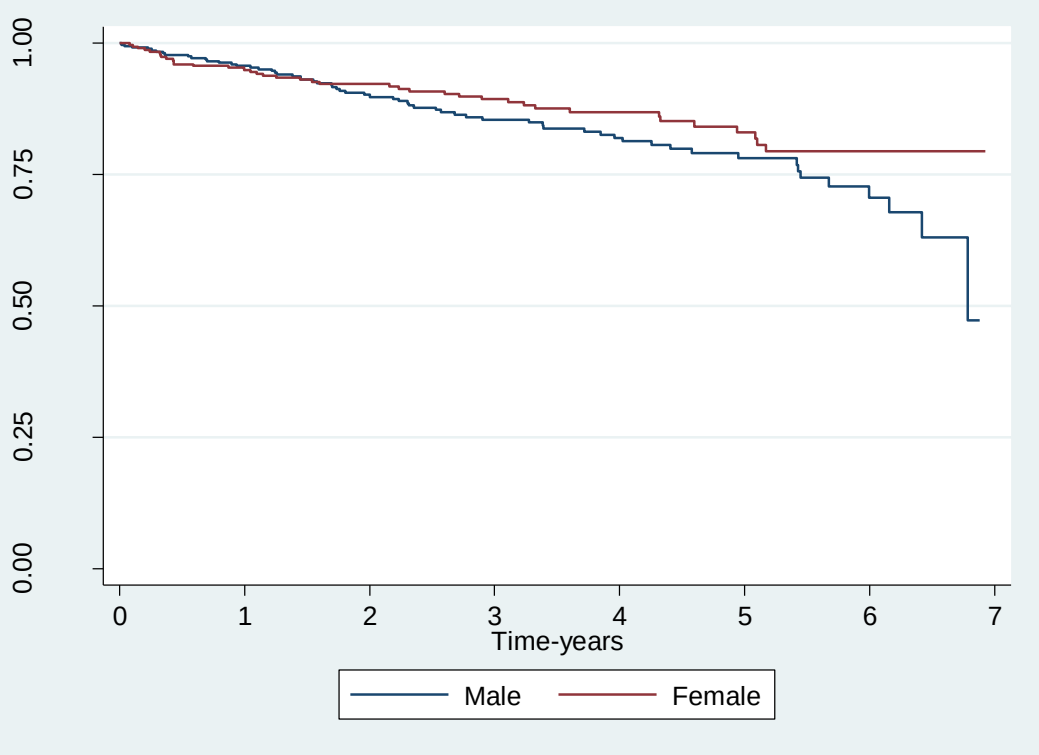


Figure A- 39 Kaplan-Meier survival estimates for monoclonal B-cell Lymphocytosis patients by gender

20 Monoclonal Gammopathy of Undetermined Significance

Table A- 24 Crude incidence of monoclonal gammopathy of undetermined significance by age and gender (per 100,000 population)

Age group (Years)	Total		Male		Female	
	N	Incidence	N	Incidence	N	Incidence
0-4	0	0.0	0	0.0	0	0.0
5-9	0	0.0	0	0.0	0	0.0
10-14	0	0.0	0	0.0	0	0.0
15-19	0	0.0	0	0.0	0	0.0
20-24	0	0.0	0	0.0	0	0.0
25-29	2	0.1	1	0.1	1	0.1
30-34	5	0.3	2	0.2	3	0.3
35-39	11	0.6	3	0.3	8	0.8
40-44	36	2.1	17	2.0	19	2.2
45-49	52	3.3	30	3.8	22	2.8
50-54	75	4.3	35	4.0	40	4.5
55-59	132	9.5	62	8.9	70	10.0
60-64	181	14.6	108	17.8	73	11.6
65-69	198	17.7	124	23.4	74	12.6
70-74	279	27.8	161	36.1	118	21.2
75-79	307	36.3	169	48.1	138	28.0
Over 80	366	35.1	191	57.3	175	24.6
Total	1644	6.6	903	7.5	741	5.7

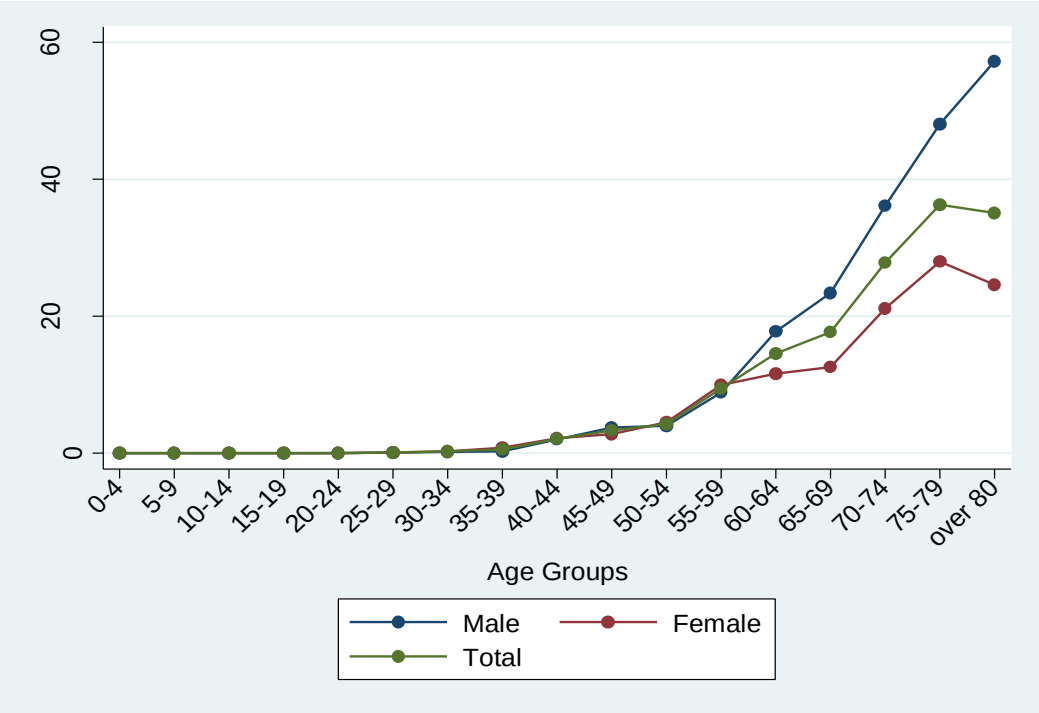


Figure A- 40 Incidence of monoclonal gammopathy of undetermined significance per 100,000 for males, females, and total

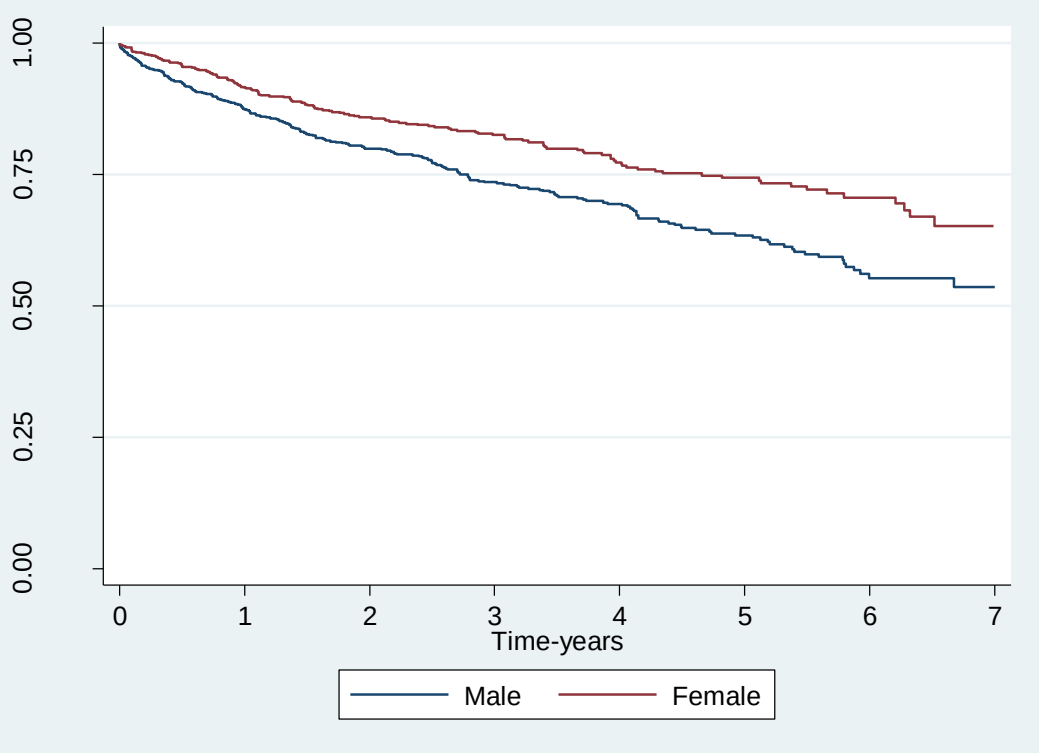


Figure A- 41 Kaplan-Meier survival estimates for monoclonal gammopathy of undetermined significance patients by gender

21 Lymphoproliferative Disorder Not Otherwise Specified

Table A- 25 Crude incidence of lymphoproliferative disorder not otherwise specified by age and gender (per 100,000 population)

Age group (Years)	Total		Male		Female	
	N	Incidence	N	Incidence	N	Incidence
0-4	1	0.1	0	0.0	1	0.1
5-9	0	0.0	0	0.0	0	0.0
10-14	0	0.0	0	0.0	0	0.0
15-19	0	0.0	0	0.0	0	0.0
20-24	1	0.1	1	0.1	0	0.0
25-29	0	0.0	0	0.0	0	0.0
30-34	1	0.1	1	0.1	0	0.0
35-39	1	0.1	1	0.1	0	0.0
40-44	1	0.1	1	0.1	0	0.0
45-49	6	0.4	6	0.8	0	0.0
50-54	7	0.4	4	0.5	3	0.3
55-59	28	2.0	16	2.3	12	1.7
60-64	52	4.2	35	5.8	17	2.7
65-69	60	5.4	39	7.4	21	3.6
70-74	60	6.0	34	7.6	26	4.7
75-79	86	10.2	41	11.7	45	9.1
Over 80	173	16.6	79	23.7	94	13.2
Total	477	1.9	258	2.1	219	1.7

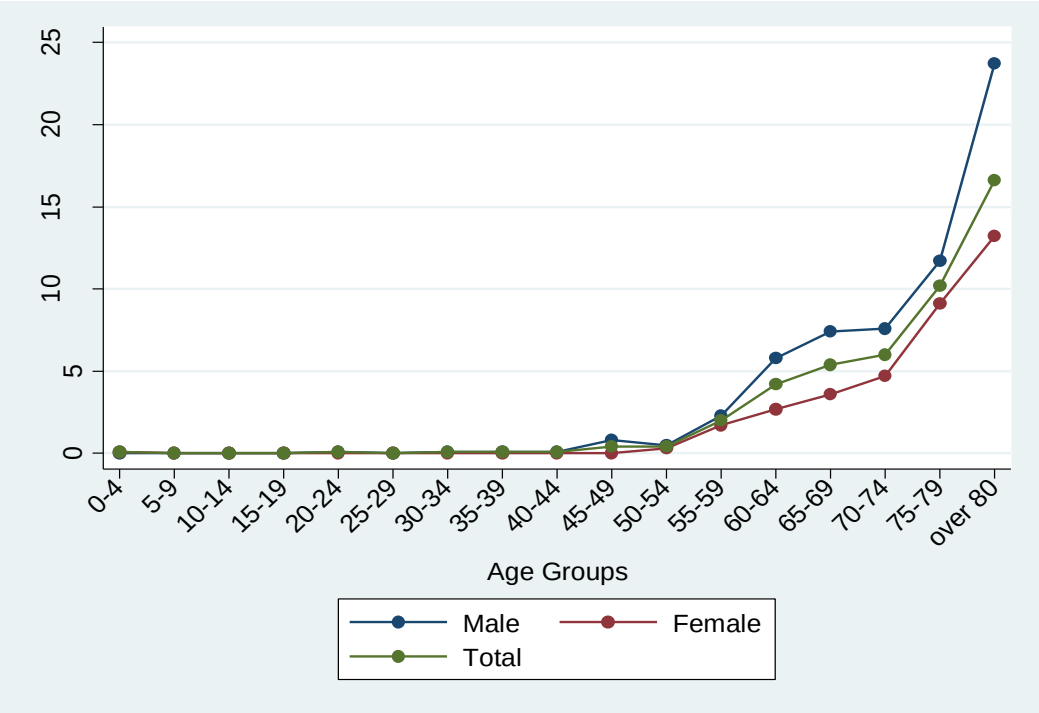


Figure A- 42 Incidence of lymphoproliferative disorder not otherwise specified per 100,000 for males, females, and total

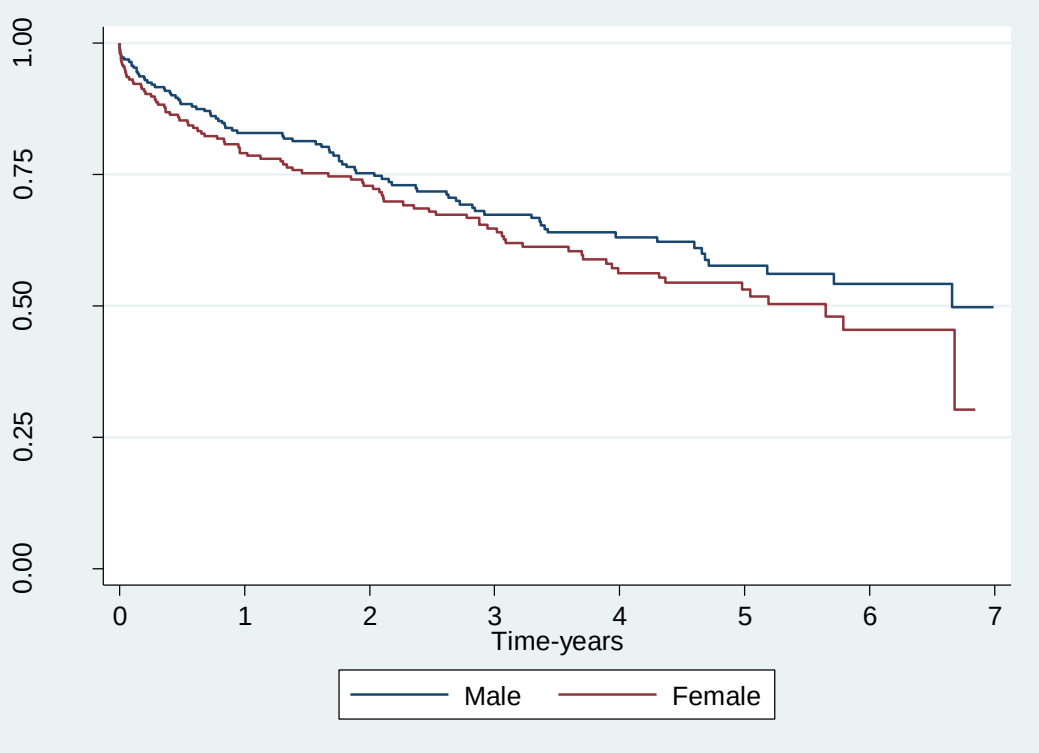


Figure A- 43 Kaplan-Meier survival estimates for lymphoproliferative disorder not otherwise specified patients by gender

Appendix A6 The Notion of "Cure"

If five years is considered as the "period of maximum consumption of health resources" (Colonna, et al., 2000), and if the patients that survive with cancer longer than five years are considered to be "cured" patients, the proportion of prevalent subjects who are considered to be cured can be computed as the difference between total prevalence (estimated in Chapter Six) and observed 5-year prevalence on the index date (calculated in Chapter Four):

$$P_{cured} = P_{total} - P_{5-year} \quad \text{(A. 1)}$$

The results are shown in Table A-26. These cured prevalent patients may require fewer health resources compared to patients who have been diagnosed recently. More than half of the patients diagnosed with haematological malignancies have survived for over five years on the index date. However, for different subtypes, these percentages vary due to the varying prognoses of these diseases. For example, 81.7% of Hodgkin lymphoma patients live for longer than five years, whilst only 14.3% of patients diagnosed with mantle cell lymphoma survive for over five years.

Table A- 26 5-year, total, "cured" prevalence (per 100,000) for males and females in HMRN on the index date 31st, August 2011

	Total				Male				Female			
	5-year	Total	"cured"		5-year	Total	"cured"		5-year	Total	"cured"	
			Prevalence	%			Prevalence	%			Prevalence	%
Total	227.1	550.9	323.8	58.8	259.1	591.9	332.8	56.2	196.9	512.3	315.4	61.6
Leukaemia	48.9	111.3	62.3	56	61.4	138.8	77.3	55.7	37.2	85.5	48.3	56.5
Chronic myelogenous leukaemia	4.7	14.7	10	67.9	5.9	17.1	11.2	65.5	3.6	12.5	8.9	70.9
Chronic myelomonocytic leukaemia	1.7	1.9	0.3	14.2	2	2.2	0.2	10.5	1.4	1.7	0.3	18.8
Acute myeloid leukaemia	6.4	9.6	3.1	32.9	7.5	10.2	2.7	26.5	5.4	9	3.6	39.6
Acute lymphoblastic leukaemia	4.1	14.5	10.3	71.5	5	16.5	11.5	69.8	3.3	12.6	9.2	73.6
Chronic lymphocytic leukaemia	29	62.1	33.1	53.3	37.2	81.3	44	54.2	21.2	44.1	22.8	51.8
Hairy cell leukaemia	1.6	4.9	3.4	68.2	2.5	8.6	6.1	70.5	0.7	1.5	0.8	56
T-cell leukaemia	1.5	3.6	2.1	58.9	1.4	3	1.6	53.8	1.6	4.2	2.6	62.4
Non-Hodgkin Lymphoma	60.2	136.9	76.7	56.1	66.2	147.4	81.2	55.1	54.5	127.1	72.5	57.1
Marginal zone lymphoma	14.4	28.9	14.6	50.4	15.9	32.1	16.2	50.5	12.9	26	13	50.2
Follicular lymphoma	14.2	38.5	24.3	63.1	14.1	33.4	19.3	57.8	14.3	43.3	29	66.9
Mantle cell lymphoma	2.6	3	0.4	14.3	3.6	4.3	0.7	16	1.6	1.8	0.2	10.3
Diffuse large B-cell lymphoma	25.4	55.1	29.7	54	27.8	61.2	33.4	54.5	23.1	49.4	26.3	53.3
Burkitt lymphoma	1	4.8	3.7	78.3	1.6	8.3	6.7	80.5	0.5	1.5	1	66.6
T-cell lymphoma	2.6	6.6	4	60.4	3.1	8.1	4.9	61.3	2.1	5.2	3.1	59
Hodgkin Lymphoma	13.2	72.4	59.2	81.7	16	73.3	57.3	78.2	10.6	71.5	60.9	85.1
Myeloma	20.6	32.1	11.5	35.8	25.3	41.7	16.4	39.3	16.2	23.1	6.9	29.9
Plasma cell myeloma	18.5	28.6	10.1	35.4	22.1	36.6	14.6	39.7	15.1	21	5.9	28.2
Plasmacytoma	2.1	3.5	1.4	39.6	3.2	5.1	1.9	36.4	1.1	2.1	1	47.1
Myelodysplastic syndromes	8.6	10	1.4	14.1	11.1	13	1.9	14.4	6.3	7.3	1	13.7
Other Neoplasms of Uncertain or Unknown Behaviour	75.5	188.1	112.6	59.9	79.1	177.7	98.6	55.5	72.1	197.9	125.8	63.6
Myeloproliferative neoplasms	29.4	67.2	37.9	56.3	27.9	63	35.1	55.8	30.8	71.2	40.4	56.8
Monoclonal B-cell Lymphocytosis	12.6	32.9	20.3	61.7	14	31.3	17.2	55.1	11.2	34.5	23.3	67.4
Monoclonal gammopathy of undetermined significance	27	72	45	62.5	29.3	63.9	34.6	54.2	24.8	79.5	54.7	68.8
Lymphoproliferative disorder not otherwise specified	6.6	16	9.4	58.9	7.9	19.5	11.6	59.3	5.3	12.6	7.4	58.3

Appendix A7 R Code for Calculating Total Prevalence

The data is available in STATA format. Calculations are performed in R version 3.0.1. The "foreign" R library is used to read STATA format data. Calculations of observed prevalence and estimates of total prevalence for all subtypes in this study are then calculated in R.

N.B. the words after "#" are comments on the code:

```
library(foreign)
d <- read.dta("hmrn.dta")
#We use the example of AML

##1 Prediction of incidences by spline regression
#1.1 HMRN population for males and females
M<-
c(107160,119103,124558,117089,107301,108539,128167,133
384,123962,113034,124644,99325,86445,75680,63721,50210
,47593)
# population of males in five years age group

F<-
c(104373,113668,120175,114423,110671,114524,134554,138
410,125696,113814,125792,99606,90097,83945,79433,70475
,101461)
# population of females in five years age group

#####################################################
# Population comes from census in the UK
# Numbers can be found in Chapter Three
```

```
#####################################################
```

```
age1<-
c(2,7,12,17,22,27,32,37,42,47,52,57,62,67,72,77,90)
```

use middle age for each age group

```
e<-
table(factor(d$cutage,levels=c(0,5,10,15,20,25,30,35,4
0,45,50,55,60,65,70,75,80)))
```

number of cases in every age group

```
i <- ((e/7)/M)*100000
```

incidence of every age group for males, change M to F for females calculation

```
#########################################################
```

Show annual number of cases

```
#########################################################
year<-c(1,2,3,4,5,6,7)
y<-c(table(d$year))

plot(year,y)
#########################################################
```

1.2 regression spline

```
library(splines)
es.sm <- lm(i ~ bs(age1, df=6))
summary(es.sm <- lm(i ~ bs(age1, df=6)))

Inc<-function(c,modsp=es.sm){
```

```
     predict(modsp, newdata=data.frame(age1=c))
     }

plot(Inc(2:82),type='l', ylab='Incidence (per
100,000)',xlab="Age", col=4, lwd=2)
points(age1, i,type="p", lwd=1, col=3, pch=19)

##################################################
```
#incidence cannot be below 0

#so we need to control that incidence(I)>=0

```
##################################################

t<-seq(0,100,length.ou=101)
I<-function(t){
            (Inc(t)>0)*Inc(t)
          }
```
avoid the incidence under 0 after estimation

2 survival function

```
library(splines)
library(survival)

Surv(d$time, d$status)
km<-survfit(Surv(d$time, d$status)~factor(d$agegrp))
```
#Kaplan Meier by age groups

```
plot(km,col=c(1:9), xlab="years", ylab
="Survival",from=0,to=100)
legend("bottomleft",pch,lty=1,
lwd=c(1,2,2,2,2,2,2,2,3),
```

```
    legend=c("0-10", "10-20", "20-30","30-40","40-
50","50-60","60-70","70-80","over 80"),

    col=c(1:9))

s <- survreg(Surv(d$time, d$status) ~
pspline(AgeDiagnosis,df=6), dist="weibull", data=d)

summary(s)
```

Function to return the scale parameter of the Weibull distribution from a fitted model

```
scale <- function(t){

       return(exp(predict(s,
        newdata=data.frame(AgeDiagnosis=t),type="lp")))

       }
```

Shape parameter from a fitted Weibull regression

```
p <- 1/s$scale
```

Survival function from a fitted Weibull regression

```
S<- function(duration,t)

  (exp(-(duration)/scale(t))^p)
```

3 Calculate N and R

3.1 read data from *.txt file: general mortality;

```
Mort<-read.table('PE.txt', header=T)
```

(1-mortality) data is saved in advance

3.2 calculate completeness index

```
N <- function(x, upper) {
    tmp=c()
       for (t in 0:upper) {
          tmp[t]=I(t)* survival(x+1-t, t) /
prod(mort$M[t:x])   }
sum(tmp)
     }
```

#change as "prod(mort$F[t:x])" for females

```
R <- function(x) if (x<8) 1 else
     1-N(x,x-7)/N(x,x)

r<-c()
x<-c(2,7,12,17,22,27,32,37,42,47,52,57,62,67,72,77,90)
n<-length(x)
for (i in 1:n ) r[i]<-R(x[i])
r[which(r==0)]<-1
r[is.na(r)]<-1
print(data.frame(x=x,R=r))
```

4 calculate prevalence

```
No<-
table(factor(d$agegrp3,levels=c(0,5,10,15,20,25,30,35,
40,45,50,55,60,65,70,75,80)))
No
```

show the number of observed prevalent cases

```
Nt<-No/r
Nt
```

```
# show the number of total prevalent cases
preobs<-sum(No)/sum(M)*100000
# calculate for observed prevalence for males
pretot<-sum(Nt)/sum(M)*100000
# calculate for total prevalence for males
# change as “sum(F)”for females
ratio<-preobs/pretot
preobs
# show observed prevalence (per 100,000)
pretot
# show total prevalence (per 100,1000)
ratio
# show the ratio of observed prevalence over total prevalence
```

Appendix A8 Abbreviations used in this thesis

Table A- 27 Abbreviations in this study

Abbreviation	The meaning in this study
AACR	Australasian Association of Cancer Registries
AIHW	Australian Institute of Health and Welfare
AIRTUM	The Italian Association of Cancer Registries
AL	Acute leukaemia
ALL	Acute lymphoblastic leukaemia
AML	Acute myeloid leukaemia
ANCR	The Association of the Nordic Cancer Registries
APML	Acute promyelocytic myeloid leukaemia
ASP	Age standardized proportion
CCR	Canadian Cancer Registry
CHILDPREV	Childhood Prevalence
CLL	Chronic lymphocytic leukaemia
CML	Chronic myelogenous leukaemia
CMML	Chronic myelomonocytic leukaemia
CTR	The Connecticut Tumour Registry
DCO	Death certificate only
DisMod	Disease model
DLBCL	Diffuse large B-cell lymphoma
ECSG	Epidemiology & Cancer Statistics Group
EEA	European Economic Area
EU	European Union
FAB	French, American, and British Cooperative Group
IACR	International Association of Cancer Registries
IARC	International Agency for Research on Cancer
ICD	The International Classification of Diseases
ICD-O-3	International Classification of Disease for Oncology, 3rd Edition
ILSG	International Lymphoma Study Group
IPM	Incidence, prevalence, and mortality
HILIS	HMDS Integrated Laboratory Information System
HIV	Human immunodeficiency virus
HL	Hodgkin lymphoma
HM	Haematological malignancy
HMDS	Haematological Malignancy Diagnostic Service
HMRN	Haematological Malignancy Research Network
LPDs NOS	Lymphoproliferative disorder not otherwise specified
MBL	Monoclonal B-cell Lymphocytosis
MCL	Mantle cell lymphoma
MENA	Middle East and Northern Africa
MDS	Myelodysplastic syndromes

Table A- 27 continued

Abbreviation	The meaning in this study
MDTs	Multi-disciplinary teams
MGUS	Monoclonal gammopathy of undetermined significance
MIAMOD	Mortality Incidence Analysis Model
MM	Multiple myeloma
MPN	Myeloproliferative neoplasms
MZL	Marginal zone lymphoma
NA	Not Available
NCCCR	North Carolina Central Cancer Registry
NCIN	National Cancer Intelligence Network
NHIS	National Health Interview Survey
NHL	Non-Hodgkin lymphoma
NHS	National Health Service
PIAMOD	Prevalence Incidence Analysis Model
REAL	Revised European- American Lymphoma
SEER	Surveillance Epidemiology and End Results
SSA	Sub-Saharan Africa
TRM	Transition rate method
UK	United Kingdom
US	United States of America
WHO	World Health Organization

References

Adami, H.O. et al., (1989). The prevalence of cancer in Sweden 1984. *Acta Oncologica (Stockholm, Sweden)*, 28(4), 463–470.

AIHW, (2012). Cancer survival and prevalence in Australia: period estimates from 1982 to 2010. [pdf]. Available at: http://www.aihw.gov.au/publication-detail/?id=10737422720 [Accessed July 29, 2013].

Alexanian, R., et al., (1968). Melphalan therapy for plasma cell myeloma. *Blood*, 31(1), 1-10.

Armitage, J.O., (2010). Early-stage Hodgkin's lymphoma. *The New England Journal of Medicine*, 363(7), 653–662.

Attal, M. et al., (2006). Maintenance therapy with thalidomide improves survival in patients with multiple myeloma. *Blood*, 108(10), 3289–3294.

Auvinen, A. et al., (2002). Lead-time in prostate cancer screening (Finland). *Cancer Causes & Control: CCC*, 13(3), 279–285.

Bagguley T. et al., (2012). Hematological malignancies & cancer registration in England (2004-2008). [pdf] NCIN. Available at: www.ncin.org.uk/view?rid=1725 [Accessed May 18, 2013].

Barlogie, B. et al., (2004). Treatment of multiple myeloma. *Blood*, 103(1), 20–32.

Becher, H., et al., (2009). Using Penalized Splines to Model Age-and Season-of-Birth-Dependent Effects of Childhood Mortality Risk Factors in Rural Burkina Faso. *Biometrical Journal*, 51(1), 110-122.

Bennett, J. M., et al., (1976). Proposals for the Classification of the Acute Leukaemias French-American-British (FAB) Co-operative Group. *British Journal of Haematology*, 33(4), 451-458.

Brameld, K.J. et al., (2002). Increasing 'active prevalence' of cancer in Western Australia and its implications for health services. *Australian and New Zealand Journal of Public Health*, 26(2), 164–169.

Bray, F., et al., (2013). Global estimates of cancer prevalence for 27 sites in the adult population in 2008. *International Journal of Cancer*, 132(5), 1133-1145.

Byrne, J., Kessler, L.G. and Devesa, S.S., (1992). The prevalence of cancer among adults in the United States: 1987. *Cancer*, 69(8), 2154–2159.

Cancer Research UK, (2011). Cancer incidence for all cancers combined. [Online]. Available at: http://info.cancerresearchuk.org/cancerstats/incidence/all-cancers-combined/ [Accessed November 12, 2011].

Cancer Research UK, (2013). Different types of non-Hodgkin lymphoma. [Online]. Available at: http://www.cancerresearchuk.org/cancer-help/type/non-hodgkins-lymphoma/about/types/the-most-common-types-of-non-hodgkins-lymphoma/ [Accessed April 12, 2013].

Capocaccia, R. and De Angelis, R., (1997). Estimating the completeness of prevalence based on cancer registry data. *Statistics in Medicine*, 16(4), 425–440.

Capocaccia, R. et al., (2002). Measuring cancer prevalence in Europe: the EUROPREVAL project. *Annals of Oncology: Official Journal of the European Society for Medical Oncology / ESMO*, 13(6), 831–839.

Carpenter, W.R. et al., (2011). Getting cancer prevalence right: using state cancer registry data to estimate cancer survivors. *Cancer Causes & Control : CCC*, 22(5), 765–773.

CCR, (2012). Canadian Cancer Registry. [Online]. Available at http://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=3207&Item_Id=1633&lang=en[Accessed July 29, 2013]

Census Dissemination Unit, (2001). Census Dissemination Unit. [Online]. Available at http://cdu.mimas.ac.uk/[Accessed October 13, 2011]

Cleves, M. et al., (2010). *An Introduction to Survival Analysis Using Stata, Third Edition* 3rd ed., Stata Press.

Colonna, M. et al., (2000). National cancer prevalence estimation in France. *International Journal of Cancer. Journal International du Cancer*, 87(2), 301–304.

Crocetti, E., et al. (2013). Cancer prevalence in United States, Nordic Countries, Italy, Australia, and France: an analysis of geographic variability. *British Journal of Cancer*. (109), 219-228

Cronin, K.A. et al., (2006). Additional common inputs for analyzing impact of adjuvant therapy and mammography on U.S. mortality. *Journal of the National Cancer Institute. Monographs*, (36), 26–29.

Curado,M.P. et al, (2007). *Cancer Incidence in Five Continents Vol. IX*. Lyon: IARC Scientific Publication No. 160.

Cutler, J. and Ederer, F. (1958). Maximum utilization of the life table in analyzing survival. *Journal of Chronic Diseases* 8, 699-712.

Davies, A. J., et al., (2007). Transformation of follicular lymphoma to diffuse large B‐cell lymphoma proceeds by distinct oncogenic mechanisms. British journal of haematology, 136(2), 286-293.

De Angelis, G. et al., (1994). MIAMOD: a computer package to estimate chronic disease morbidity using mortality and survival data. *Computer Methods and Programs in Biomedicine*, 44(2), 99–107.

De Angelis, R. et al., (2007). Cancer prevalence estimates in Italy from 1970 to 2010. *Tumori*, 93(4), 392–397.

DeVita, V. T., Serpick, A. A., & Carbone, P. P. (1970). Combination chemotherapy in the treatment of advanced Hodgkin's disease. *Annals of Internal Medicine*, 73(6), 881-895.

Ederer, F., Axtell, L.M. and Cuter, S.J., (1961). The relative survival rate: a statistical methodology. *National Cancer Institute Monograph*, 6, 101–121.

Ellison, L. F., and Wilkins, K. (2009). Cancer prevalence in the Canadian population. *Health Rep*, 20(1), 7-14.

Engholm G., et al., (2013). NORDCAN: Cancer Incidence, Mortality, Prevalence and Survival in the Nordic Countries, Version 5.3. Association of the Nordic Cancer Registries. Danish Cancer Society. [Online] Available from http://www.ancr.nu, [Accessed July 15, 2013]

Esna-Ashari, F., et al., (2012). Colorectal Cancer Prevalence According to Survival Data in Iran-2007. *Iranian Journal of Cancer Prevention*, 2(1), 15–18.

Estève, J., Benhamou, E. and Raymond, L., (1994). *Statistical Methods in Cancer Research. Volume IV. Descriptive Epidemiology.* IARC Scientific Publications, (128), 1–302.

Eurocare, (2011). MIAMOD and PIAMOD Software. [Online]. Available at: http://www.eurocare.it/MiamodPiamod/tabid/60/Default.aspx [Accessed August 3, 2012].

Feldman, A.R. et al., (1986). The prevalence of cancer. Estimates based on the Connecticut Tumor Registry. *The New England Journal of Medicine*, 315(22), 1394–1397.

Ferlay, J et al., (2010). Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. *International Journal of Cancer*, 127(12), 2893–2917.

Fermé, C., et al., (2007). Chemotherapy plus involved-field radiation in early-stage Hodgkin's disease. *New England Journal of Medicine*, 357(19), 1916-1927.

Fiorentino, F., et al. (2011). Modelling to estimate future trends in cancer prevalence. *Health Care Management Science*, 14(3), 262-266.

Forman, D. et al., (2003). Cancer prevalence in the UK: results from the EUROPREVAL study. *Annals of Oncology: Official Journal of the European Society for Medical Oncology / ESMO*, 14(4), 648–654.

Fritz, A., (2000). *International Classification of Diseases for Oncology: ICD-O, 3rd edn* 3rd ed., Geneva: World Health Organization.

Gail, M.H. et al., (1999). Two approaches for estimating disease prevalence from population-based registries of incidence and total mortality. *Biometrics*, 55(4), 1137–1144.

Gambacorti-Passerini, C. et al., (2011). Multicenter independent assessment of outcomes in chronic myeloid leukemia patients treated with imatinib. *Journal of the National Cancer Institute*, 103(7), 553–561.

Gatta, G., et al. (2011). Rare cancers are not so rare: The rare cancer burden in Europe. *European Journal of Cancer*, 47(17), 2493-2511.

Gigli, A., Simonetti, A. and Capocaccia, R., (2004). Validation of complete prevalence by age groups. Working Paper IRPPS, 1/2004

Gigli, A. et al., (2006). Estimating the variance of cancer prevalence from population-based registries. *Statistical Methods in Medical Research*, 15(3), 235–253.

Gjerstorff M. L. (2011). The Danish cancer registry. *Scandinavian journal of public health*, 39(7 suppl), 42-45.

GLOBOCAN, (2008). Estimated cancer Incidence, Mortality, Prevalence and Disability-adjusted life years (DALYs) Worldwide in 2008. [Online]. Available at: http://globocan.iarc.fr/ [Accessed July 29, 2013]

Golestan, B. et al., (2009). An estimation of the chronic rejection of kidney transplant using an eternal Weibull regression: a historical cohort study. *Archives of Iranian Medicine*, 12(4), 341–346.

Gras, C., Daurès, J.P. and Tretarre, B., (2006). Three approaches for estimating prevalence of cancer with reversibility. Application to colorectal cancer.

In M. Nikulin, D. Commenges, and C. Huber, eds. *Probability, Statistics and Modelling in Public Health*. Springer US, 169–186. [pdf] Available at: link.springer.com/content/pdf/10.1007%2F0-387-26023-4_12.pdf [Accessed January 4, 2013].

Guzzinati, S., et al. (2012). Cancer prevalence in Italy: an analysis of geographic variability. *Cancer Causes & Control*, 23(9), 1497-1510.

Haberland, J. et al., (2010). German cancer statistics 2004. *BMC Cancer*, 10, 52.

Hakama, M. et al., (1975). Incidence, mortality or prevalence as indicators of the cancer problem. *Cancer*, 36(6), 2227–2231.

Harris, N. L., et al. (2000 a). Lymphoma classification–from controversy to consensus: the REAL and WHO Classification of lymphoid neoplasms. *Annals of Oncology*, 11(suppl 1), S3-S10.

Harris, N.L. et al., (2000 b). The World Health Organization Classification of Hematological Malignancies Report of the Clinical Advisory Committee Meeting, Airlie House, Virginia, November 1997. *Modern Pathology*, 13(2), 193–207.

Hehlmann, R., Hochhaus, A. and Baccarani, M., (2007). Chronic myeloid leukaemia. *The Lancet*, 370(9584), 342–350.

Herrmann, C., et al. (2013). Cancer survivors in Switzerland: a rapidly growing population to care for. *BMC Cancer*, 13(1), 287.

Hewitt, M., Breen, N. and Devesa, S., (1999). Cancer Prevalence and Survivorship Issues: Analyses of the 1992 National Health Interview Survey. *Journal of the National Cancer Institute*, 91(17), 1480–1486.

HMDS, (2011). Haematological Malignancy Diagnostic Service. [Online]. Available at: http://www.hmds.info/ [Accessed August 3, 2012].

HMRN, (2011). Haematological Malignancy Research Network. [Online]. Available at: http://www.hmrn.org/ [Accessed August 3, 2012].

Hoffbrand, V., Moss, P. and Pettit, J., (2006). *Essential Haematology 5th ed*., Wiley-Blackwell.

Hoogenveen, R.T. and Gijsen, R., (2000). Dutch DisMod for Several Types of Cancer. [pdf]. Rijksinstituut voor Volksgezondheid en Milieu. [pdf] Available at: https://rivm.openrepository.com/rivm/bitstream/10029/.../260751004.pdf [Accessed June 8, 2012].

Horning, S. J., and Rosenberg, S. A., (1984). The natural history of initially untreated low-grade non-Hodgkin's lymphomas. *New England Journal of Medicine*, 311(23), 1471-1475.

Howard, M.R. and Hamilton, P.J., (2007). *Haematology: An Illustrated Colour Text 3rd ed*., London: Churchill Livingstone.

Hughes-Jones, N., Wickramasinghe, S.N. and Hatton, P.C., (2008). *Haematology 8th ed.*, Wiley-Blackwell.

IARC, (2012).Cancer Statistics (International Agency for Research on Cancer). [Online]. Available at: http://www-dep.iarc.fr/ [Accessed May 21, 2012].

IARC, (2013 a). Global initiative for cancer registry development in low- and midlle-income countries. [Online]. Available at: http://gicr.iarc.fr/en/whatwedo-where.php [Accessed June 8, 2013].

IARC, (2013 b). Population pyramid. [Online]. Available at: http://www-dep.iarc.fr/WHOdb/graph5_sel.asp [Accessed July 2, 2013].

Jensen OM. et al, (1991). *Cancer registration principles and methods.* Lyon: *Scientific Publication No. 95*.

Krogh, V. and Micheli, A., (1996). Measure of cancer prevalence with a computerized program: an example on larynx cancer. *Tumori*, 82(3), 287–290.

Kruijshaar, M.E., Barendregt, J.J. and Hoeymans, N., (2002). The use of models in the estimation of disease epidemiology. *Bulletin of the World Health Organization*, 80(8), 622–628.

Kumar, S.K. et al., (2008). Improved survival in multiple myeloma and the impact of novel therapies. *Blood*, 111(5), 2516–2520.

Kyle, R. A., et al.,(2010). Monoclonal gammopathy of undetermined significance (MGUS) and smoldering (asymptomatic) multiple myeloma: IMWG consensus perspectives risk factors for progression and guidelines for monitoring and management. Leukemia, 24(6), 1121-1127.

Landgren, O., et al., (2009). Monoclonal gammopathy of undetermined significance (MGUS) consistently precedes multiple myeloma: a prospective study. Blood, 113(22), 5412-5417.

Leon G. (2008). *Epidemiology, 4th Edition*. Philadelphia: Sauders Elsevier.

Levi, F. et al., (2002). Trends in mortality from Hodgkin's disease in western and eastern Europe. *British Journal of Cancer*, 87(3), 291–293.

Lennert K., (1978). *Malignant Lymphomas Other Than Hodgkin's Disease.* New York: Springer-Verlag.

Lossos, I. S., et al., (2002). Transformation of follicular lymphoma to diffuse large-cell lymphoma: alternative patterns with increased or decreased expression of c-myc and its regulated genes. Proceedings of the National Academy of Sciences, 99(13), 8886-8891.

Louchini, R. et al., (2006). Trends in cancer prevalence in Quebec. *Chronic Diseases in Canada*, 27(3), 110–119.

LSHTM, (2012). Tools for Cancer Survival Analysis | London School of Hygiene & Tropical Medicine. [Online]. Available at: http://www.lshtm.ac.uk/eph/ncde/cancersurvival/tools/ [Accessed December 4, 2012].

Ludwig, H. et al., (2010). Current Multiple Myeloma Treatment Strategies with Novel Agents: A European Perspective. *The Oncologist*, 15(1), 6–25.

Lutz, J.M. et al., (2003). Cancer prevalence in Central Europe: the EUROPREVAL Study. *Annals of oncology: official journal of the European Society for Medical Oncology / ESMO*, 14(2), 313–322.

Maddams, J., Utley, M., and Møller, H. (2012). Projections of cancer prevalence in the United Kingdom, 2010–2040. *British Journal of Cancer*.107, 1195-1020

Marcos-Gragera, R., et al. (2011). Survival of European patients diagnosed with lymphoid neoplasms in 2000–2002: results of the HAEMACARE project. *Haematologica*, 96(5), 720-728.

Mariotto, A. et al., (1999). Cancer prevalence in Italian regions with local cancer registries. *Tumori*, 85(5), 400–407.

Mariotto, A.B. et al., (2006). Projecting the number of patients with colorectal carcinoma by phases of care in the US: 2000-2020. *Cancer Causes & Control: CCC*, 17(10), 1215–1226.

Mariotto, A. B. et al., (2009). Long-term survivors of childhood cancers in the United States. *Cancer Epidemiology Biomarkers & Prevention*, 18(4), 1033-1040.

Mariotto, A. B., et al., (2011). Projections of the cost of cancer care in the United States: 2010–2020. *Journal of the National Cancer Institute*, 103(2), 117-128.

Marti, G. E., et al., (2005). Diagnostic criteria for monoclonal B‐cell lymphocytosis. British journal of haematology, 130(3), 325-332.

Marti, G.E., (2009). The changing definition of CLL. *Blood*, 113(18), 4130–4131.

Mauer, A.M. and Simone, J.V., (1976). The current status of the treatment of childhood acute lymphoblastic leukemia. *Cancer Treatment Reviews*, 3(1), 17–41.

Maynadié, M., et al. (2013). Survival of European patients diagnosed with myeloid malignancies: a HAEMACARE study. *Haematologica*, 98(2), 230-238.

Mehrabian, A.A. et al., (2010). Gastric Cancer Prevalence, According To Survival Data in Iran (National Study-2007). *Iranian J Public Health*, 39(3), 20–26.

Merrill, R.M. et al., (2000). Cancer prevalence estimates based on tumour registry data in the Surveillance, Epidemiology, and End Results (SEER) Program. *International Journal of Epidemiology*, 29(2), 197–207.

Micheli, A. et al., (1999). Cancer prevalence in Italian cancer registry areas: the ITAPREVAL study. ITAPREVAL Working Group. *Tumori*, 85(5), 309–369.

Micheli , A. et al., (2002 a). Cancer prevalence in European registry areas. *Annals of Oncology: Official Journal of the European Society for Medical Oncology / ESMO*, 13(6), 840–865.

Micheli, et al., (2002 b). Contrasts in cancer prevalence in Connecticut, Iowa, and Utah. *Cancer*, 95(2), 430–439.

Miguel, JF S, Creixenti, J.B. and Garcia-Sanz, R., (1999). Treatment of multiple myeloma. *Haematologica*, 84(1), 36–58.

Möller, T., et al., (2003). Cancer prevalence in Northern Europe: the EUROPREVAL study. *Annals of Oncology: Official Journal of the European Society for Medical Oncology / ESMO*, 14(6), 946–957.

Morton L M., et al., (2007). Proposed classification of lymphoid neoplasms for epidemiologic research from the Pathology Working Group of the International Lymphoma Epidemiology Consortium (InterLymph). *Blood*, 110(2): 695-708.

National Cancer Institute, (2012). SEER*Stat Software. [Online]. Available at: http://seer.cancer.gov/seerstat/ [Accessed August 3, 2012].

National Cancer Intelligence Network and Cancer Research UK, (2009). Cancer Incidence and Survival by Major Ethnic Group, England 2002-2006.

National Cancer Intelligence Network (NCIN), (2010). One, Five and Ten-year Cancer Prevalence, *National Cancer Intelligence Network*. [pdf]. NCIN Available at: www.ncin.org.uk/view?rid=76 [Accessed August 2, 2012].

NCIN, (2012). National Cancer Intelligence Network. [Online]. Available at: http://www.ncin.org.uk/home.aspx [Accessed November 28, 2012].

NHS, (2011). Cancer Networks | NCAT. [Online]. Available at: http://www.ncat.nhs.uk/what-is-ncat/cancer-networks [Accessed October 12, 2011].

NORDCAN, (2010). NORDCAN. [Online]. Available at: http://www-dep.iarc.fr/nordcan/English/frame.asp [Accessed October 12, 2011].

Office for National Statistics, (2001). Census: Aggregate data. UK Data Service Census Support. [Online]. Available at: http://casweb.mimas.ac.uk/ [Accessed August 3, 2012].

Office for National Statistics, (2008). Postcode directories. Office for National Statistics. [Online]. Available at: http://www.ons.gov.uk/ons/guide-method/geography/products/postcode-directories/index.html [Accessed August 3, 2012].

Office for National Statistics, (2012). Ethnicity and National Identity in England and Wales 2011. [Online]. Available at: http://www.ons.gov.uk/ons/rel/census/2011-census/key-statistics-for-local-authorities-in-england-and-wales/rpt-ethnicity.html [Accessed April 13, 2014].

Parker, S. L., et al., (1996). Cancer statistics, 1996. *CA: a cancer journal for clinicians*, 46(1), 5-27

Parkin, D M. et al., (2001). Estimating the world cancer burden: Globocan 2000. *International Journal of Cancer. Journal International du Cancer*, 94(2), 153–156.

Parkin, D M., (2006). The evolution of the population-based cancer registry. *Nature Reviews. Cancer*, 6(8), 603–612.

Pisani, P., Bray, F. and Parkin, D. M., (2002). Estimates of the world-wide prevalence of cancer for 25 sites in the adult population. *International Journal of Cancer*, 97(1), 72–81.

Polednak, A.P., (1997). Estimating the prevalence of cancer in the United States. *Cancer*, 80(1), 136–141.

Racine, J.S., (2011). A Primer on Regression Splines. [pdf]. Available at: cran.r-project.org/web/packages/crs/vignettes/spline_primer.pdf. [Accessed May 8, 2013]

Rappaport. H., (1966). Tumors of the hematopoietic system. *Atlas of Tumor Pathology*. Vol. Section III. Washington, Washington, DC: Armed Forces Institute of Pathology.

RARECARE, (2013). Surveillance of Rare Cancer in Europe. [Online]. Available at: http://www.rarecare.eu/default.asp [Accessed May 12, 2014]

Rawstron, A. C., et al., (2008). Monoclonal B-cell lymphocytosis and chronic lymphocytic leukemia. New England Journal of Medicine, 359(6), 575-583.

Roman, E. and Smith, A., (2011). Epidemiology of lymphomas. *Histopathology*, 58(1), 4–14.

Rosenberg S.A., et al., (1982). National Cancer Institutesponsored study of classification of non Hodgkin's lymphoma:summary and description of Working Formulation for clinicalusage. The non Hodgkin's lymphoma classification project. *Cancer*, 49(21), 12–35.

Pui, C. H., Campana, D., & Evans, W. E. (2001). Childhood acute lymphoblastic leukaemia–current status and future perspectives. The lancet oncology, 2(10), 597-607.

Pui, C. H., & Evans, W. E. (2006). Treatment of acute lymphoblastic leukemia. New England Journal of Medicine, 354(2), 166-178.

Salles, G.A., (2007). Clinical Features, Prognosis and Treatment of Follicular Lymphoma. *ASH Education Program Book*, 2007(1), 216–225.

Salomon, J. A, Gakidou, E. and Murray, C.J.L., (1999). Methods for modeling the HIV/AIDS epidemic in sub-Saharan Africa. [pdf]. Available at: www.who.int/healthinfo/paper03.pdf [Accessed May 28, 2012].

Salomon, J. A. and Murray, C.J., (2001). Modelling HIV/AIDS epidemics in sub-Saharan Africa using seroprevalence data from antenatal clinics. *Bulletin of the World Health Organization*, 79(7), pp.596–607.

Sant, M., et al. (2009). EUROCARE-4. Survival of cancer patients diagnosed in 1995–1999. Results and commentary. *European Journal of Cancer*, 45(6), 931-991.

Sant, M., et al. (2010). Incidence of hematologic malignancies in Europe by morphologic subtype: results of the HAEMACARE project. *Blood*, 116(19), 3724-3734.

Schrijvers, C.T.M. et al., (1994). Validation of Cancer Prevalence Data from a Postal Survey by Comparison with Cancer Registry Records. *American Journal of Epidemiology*, 139(4), 408–414.

SEER, (2012). Surveillance Epidemiology and End Results Program. [Online]. Available at: http://www.seer.cancer.gov/ [Accessed November 28, 2012].

Sehn, L. H., et al., (2005). Introduction of combined CHOP plus rituximab therapy dramatically improved outcome of diffuse large B-cell lymphoma in British Columbia. Journal of Clinical Oncology, 23(22), 5027-5033.

Shanafelt, T. D.,et al., (2010). Monoclonal B-cell lymphocytosis (MBL): biology, natural history and clinical management. Leukemia, 24(3), 512-520.

Shi, J., et al.,. (2004). Transformation of myelodysplastic syndromes into acute myeloid leukemias. Chinese medical journal, 117(7), 963-967.

Silva, I., (1999). *Cancer Epidemiology: Principles and Methods* 2nd Revised edition., Lyon: World Health Organization.

Simonetti, A. et al., (2008). Estimating complete prevalence of cancers diagnosed in childhood. *Statistics in Medicine*, 27(7), 990–1007.

Skjelbakken, T., Løchen, M.-L. and Dahl, I.M.S., (2002). Haematological malignancies in a general population, based on information collected from a population study, hospital records, and the Cancer Registry of Norway: the Tromsø Study. *European Journal of Haematology*, 69(2), 67–75.

Smith, A. et al., (2010). The Haematological Malignancy Research Network (HMRN): a new information strategy for population based epidemiology and health service research. *British Journal of Haematology*, 148(5), 739–753.

Stenbeck, M., Rosén, M. and Sparén, P., (1999). Causes of increasing cancer prevalence in Sweden. *Lancet*, 354(9184), 1093–1094.

Swerdlow, A., Silva, I.D.S. and Doll, R., (2001). *Cancer Incidence and Mortality in England and Wales: Trends and Risk Factors*, Oxford: Oxford University Press.

Swerdlow, A.J., (2003). Epidemiology of Hodgkin's disease and non-Hodgkin's lymphoma. *European Journal of Nuclear Medicine and Molecular Imaging*, 30(1), S3–S12.

Swerdlow, S.H. and Cancer International Agency for Research, (2008). *WHO Classification of Tumours of Haematopoietic and Lymphoid Tissues: International Agency for Research on Cancer* 4th ed., 2008., Lyon: IARC: WHO.

Tabata, N. et al., (2008). Partial cancer prevalence in Japan up to 2020: estimates based on incidence and survival data from population-based cancer registries. *Japanese Journal of Clinical Oncology*, 38(2), 146–157.

The Mount Sinai Hospital, (2012). Acute Lymphoblastic Leukemia Information. *The Mount Sinai Hospital*. [Online]. Available at: http://www.mountsinai.org/patient-care/health-library/diseases-and-conditions/acute-lymphoblastic-leukemia [Accessed December 18, 2012].

UKACR (2013). United Kingdom Association of Cancer Registries. [Online].Available at: http://www.ukacr.org/[Accessed December 20, 2013]

Verdecchia, A. et al., (1989). A method for the estimation of chronic disease morbidity and trends from mortality data. *Statistics in Medicine*, 8(2), 201–216.

Verdecchia, A et al., (2001). Incidence and prevalence of all cancerous diseases in Italy: trends and implications. *European Journal of Cancer (Oxford, England: 1990)*, 37(9), 1149–1157.

Verdecchia, A, De Angelis, G. and Capocaccia, R., (2002). Estimation and projections of cancer prevalence from cancer registry data. *Statistics in Medicine*, 21(22), pp.3511–3526.

Verdecchia, A et al., (2002). A comparative analysis of cancer prevalence in cancer registry areas of France, Italy and Spain. *Annals of Oncology: Official Journal of the European Society for Medical Oncology / ESMO*, 13(7), 1128–1139.

Verdecchia, A. et al., (2007). Methodology for estimation of cancer incidence, survival and prevalence in Italian regions. *Tumori*, 93(4), 337–344.

Visser, O., et al. (2012). Incidence, survival and prevalence of myeloid malignancies in Europe. *European Journal of Cancer*. 48, 3257-3266.

Wang, Z. Y., and Chen, Z. (2008). Acute promyelocytic leukemia: from highly fatal to highly curable. *Blood*. 111(5), 2505-2515.

Wang, A.-H. et al., (2010). Summary of 615 patients of chronic myeloid leukaemia in Shanghai from 2001 to 2006. *Journal of Experimental & Clinical Cancer Research: CR*, 29, 20.

WHO, (2008). *WHO Classification of Tumours of Haematopoietic and Lymphoid Tissues*. Lyon: International Agency for Research on Cancer.

WHO, (1994). International Statistical Classification of Disease and Related Health Problems, ICD -10. Vol. 3, Alphabetical index. Geneva: World Health Organization.

Wiggins, C.L. et al., (2010). Age disparity in the dissemination of imatinib for treating chronic myeloid leukemia. *The American Journal of Medicine*, .123(8), 764.e1–9.

Wobker, S., Yeh, W. and Carpenter, W., (2010). Focus on Survivorship: Refining Complete Prevalence Estimates Using Local Cancer Registry Data. *Cancer Epidemiology Biomarkers & Prevention*, 19(3), 897–897.

Youlden, D., Health, Q.G.-Q. and Baade, P.D., (2005). *Cancer Prevelence in Queensland 2002*. [pdf] Queensland Government - Queensland Health. Available at: www.health.qld.gov.au/hic/reports/cancer_prev.pdf [Accessed June 10, 2012].

Zanetti, R. et al., (1999). The prevalence of cancer: a review of the available data. *Tumori*, 85(5), 408–413.