Quantification of replication present in HIV reports and effect of patient movement between wards on MRSA acquisition

Wenwen Huo

Department of Mathematics and Statistics
University of Strathclyde

Degree of Doctor of Philosophy, 2014

Contents

List of Figures

List of Tables

Acknowledgements

First and foremost, I would like to thank my supervisors Dr David Greenhalgh and Professor Chris Robertson for supporting and helping me during the PhD course. I also thank Dr Kim Kavanagh for providing advice on the analysis of MRSA acquisition. My friends, especially Jiafeng, Wei, Hongrui, Alan and Chang, and my family including my parents and my husband have always supported me and I am really appreciate it. In addition, I would like to thank Statistics in Medicine for their suggestions and for accepting my paper.

Abstract

Outbreaks of widely spread infectious diseases, such as Human Immunodeficiency Virus (HIV), Severe Acute Respiratory Syndrome (SARS) and Swine flu (H1N1) and hospital acquired infections, such as Meticillin-resistant Staphylococcus Aureus (MRSA) and Clostridium Difficile, are serious health problems which have been tackled by the World Health Organization and international health protection agencies. Various statistical analyses have contributed a remarkable effect on providing scientific evidence on which to base political decisions and infection control strategies. In this project, we focused on two infectious diseases: HIV and MRSA and the research project is divided into two separate parts. One is the quantification of replication in HIV anonymous test reports and the other is the effect of patient movement between wards on the acquisition of MRSA.

The first research project is concerned with the analysis of an anonymous HIV test dataset. The data is collected as a set of birthdays and it is possible that there is repeated sampling of the same person. The aim is to quantify the amount of replication in the HIV data using a maximum likelihood technique and then give the confidence intervals for the estimated amount of replication using the bootstrap method.

The data were provided by the Public Health Laboratory Service (PHLS), Colindale, London in 1994, who were interested in a statistical method to estimate multiple counting that possibly existed in the database. The data consists of individual records of the number of AIDS cases diagnosed, with birthdates from 1901 to 1973. There were two datasets provided by the PHLS, one of which contained 1,134 records and was provided in 1991. The other dataset was provided in 1994 with the sample size 17,137. An estimate

of the true number of distinct individuals as well as the percentage of replication was obtained by programming the maximum likelihood calculation in the languages R and C. This technique is based upon evaluation of the probability that two records with the same birthdate represent two separate individuals as opposed to the same person reported twice.

The results for the 1991 dataset showed that there were five out of sixteen birth years (i.e. 31.25% of the observed records in the 1991 dataset) with replication in the true number of distinct individuals. In the results of the 1994 dataset, the majority of the birth years (57/73) recorded the correct number of distinct individuals in the observations.

The 95% confidence intervals for the estimated amount of replication were calculated by applying a parametric bootstrap method. The results show that the birth years in the 1991 dataset with non-zero estimated amount of replication (the birth years of 1931, 1934, 1935, 1943 and 1944) have comparatively wide 95% bootstrap confidence intervals, which implies higher uncertainty of the true amount of replication. A similar conclusion was obtained from the results of 95% bootstrap confidence intervals for the 1994 dataset. Comparing the results within the same birth years recorded in the 1991 dataset and the 1994 dataset, the data indicate that the confidence intervals for the 1994 dataset are mainly narrower than the corresponding ones in the 1991 dataset. The conclusion of this study illustrates the drawback of recording the HIV patients only with date of birth, which has now been improved by combining with ‘Soundex’ codes for the surname and gender.

The second part of the project aims to estimate the impact of patient movement within a hospital on the risk of MRSA acquisition by using data from the MRSA screening admission and discharge studies in Scotland which took place in two hospitals in 2010.

The data consist of an admission-only database (7,181 patients), a discharge-only database (2,432 patients) and a combined admission-discharge cohort (2,792 patients). The third database has complete information on MRSA status on admission, on discharge, as well as data on the wards the patient was in while in hospital. In order to understand

the effect of potential risk factors on MRSA acquisition, a multivariate logistic regression model was constructed to analyse the effects of the number of wards a patient was in on MRSA acquisition as well as other risk factors. Receiver Operating Characteristic (ROC) curves were plotted and the individual area under the curve (AUC) was also calculated for indicating the reliability and the accuracy of the prediction of the models.

Furthermore, we modelled the dynamic patient movement and assessed the effect of being in a ward with MRSA by imputing the unknown date of transfer, simulating the missing length of stay (along with the simulation envelope). The timelines of MRSA infection and carriage pressure in each ward of the two hospitals were then mapped for all patients in the three databases, imputing where necessary. Patient movement was measured as a volume indicator in terms of the frequency of ward to ward transfer and as cohabiting in the same ward. By using logistic regression within a bootstrap simulation, we estimated the odds ratio of acquisition of MRSA association with being in a ward with MRSA present, which was given by averaging the estimated effects from the fitted models, and generating the 95% confidence intervals.

The results indicate that the number of wards that patients had moved through and patients being in a ward with MRSA present do not affect the risk of acquiring MRSA significantly over and above the patient level risk factors such as age and the presence of open wounds or catheters. Some further work which can be done in an MRSA screening programme is suggested as an implementation study.

Chapter 1

Overview of Thesis

In this thesis, we describe modelling and imputation applications in the analysis of epidemiological data. Specifically, we want to understand how to make inference in the presence of unknown data using various methods. We briefly introduce two main techniques for accounting for the unknown data, namely the maximum likelihood method and imputation. For the material of this thesis, we focus on two widespread infectious diseases, which are Human Immunodeficiency Virus (HIV) and Methicillin-Resistant Staphylococcus Aureus (MRSA). The thesis consists of two parts, which are (i) estimation of the replication present in HIV reports, Chapters 2-4, and (ii) to estimate the impact of patient movement between wards within a hospital on MRSA acquisition, Chapters 5-7.

Unknown data can cause many statistical challenges and distort the inference about the population. In the dataset of HIV reports, there are individuals who have repeated positive HIV tests for reasons such as they do not believe the first HIV test result and want to check it, or they have moved area and the test has been repeated. Although the techniques for recording HIV infected individuals have now been improved, it is impossible to completely eliminate the replication problem. Inaccurate information on the number of unique HIV infections can affect the surveillance analysis, improvements on the cost-effective treatment of HIV infections and so on. Hence it is important to estimate the number of HIV infections reliably and to estimate the amount of replication.

In the first part of the thesis, our goal is to estimate the amount of replication in the anonymous HIV reporting provided by the Public Health Laboratory Service in London. The datasets presented the distribution of birthdates of HIV infected individuals and there were multiple records with the same birthdate. However, it is not known if they are the same person or not. Chapter 3 covers the methodology used to investigate the amount of replication. The method we develop in the first part is based upon the maximum likelihood technique. We also discuss the confidence interval for the estimated replication based upon the bootstrap method. The analysis is carried out using the statistical software package 'R' and programming language C. The language C is used since the algorithm is complicated and it considerably cuts the running time of the calculation. There are three chapters regarding to the first part of this thesis, namely (i) the introduction and literature review chapter including the background of HIV, the replication problem present in HIV datasets, the previous studies on replication work and the available data used for the following analysis (Chapter 2), (ii) the methodology to quantify the amount of replication (Chapter 3) and (iii) the results of the replication present in HIV reports (Chapter 4).

In the second part of this thesis, we use a different method to account for the unknown data - namely imputation. The main aim of the second part is to estimate the effect of patient movement within a hospital on MRSA acquisition. The patient movement can be characterised in two ways. One is measured as a volume indicator in terms of frequency of ward to ward transfer (i.e. the number of wards that a patient had stayed at during their hospital stay) and the other is measured as cohabitation did a patient stay in wards which are known to have MRSA present. Few studies have been published on MRSA acquisition in the general hospital population and there is limited information on the effect of patient movement on the risk of acquiring MRSA. The original intention was to study this as a prelude to estimation of parameters in a dynamic MRSA transmission model. However, some of the data that we need is missing namely the movement dates of patients between wards so we need to impute the missing data. Evidence on factors associated with the MRSA acquisition in the general hospital population is necessary. One especially important factor is whether being in a ward simultaneously in the presence of another patient with MRSA affects the chances of a patient acquiring MRSA. This is highly relevant for decision making on implementation of the Universal MRSA Screening Programme which was launched in 2007 in Scotland to prevent the transmission of MRSA in hospital.

The MRSA Screening Programme and literature on MRSA acquisition are reviewed in Chapter 5. We analyse the data taken from a one-year MRSA Screening Pathfinder Programme. In Chapter 6, we introduce the statistical techniques to investigate the association between risk factors and MRSA acquisition, taking into account missing data. It also covers the logistic regression models. However, the dates of movements are unavailable and collecting the data might be difficult and expensive. Hence we use imputation and simulation to make inference in the presence of missing data. Chapter 7 gives the details on the imputation and simulation for the unknown data. Then descriptive statistical analysis using tables and time line charts are used to summarise the MRSA infection pressure during the study period by ward. The acquisition of MRSA while in hospital can be linked to patient movement by the logistic regression method, which is also mentioned in Chapter 7. We use the statistical software package 'R' for the statistical analysis in the second part of the thesis since this is better for the manipulation of the relatively complex patient movement data.

Finally in Chapter 8, we briefly summarise the results of this thesis and discuss the limits and weaknesses of the study. In addition, we propose some further work that can be done with regard to both parts of the thesis.

Chapter 2

HIV introduction and literature review

In epidemiological research, modelling and imputation are common techniques. In this PhD thesis, different modelling and imputation methods are demonstrated and used for the analysis of two different infectious diseases, which are Human Immunodeficiency Virus (HIV) and meticllin-resistant *Staphylococcus aureus* (MRSA). Both infectious diseases are widely spread and are considered as serious public health problems all over the world. Disease intervention is an important mechanism to prevent the onset and development of diseases in populations [62]. Statistical analysis, which is based upon epidemiological data, can provide scientific evidence to help the governments on making decisions of disease intervention policies. Particularly, the prevalence of a disease, mortality and relationships between the occurrence of diseases and various descriptive characteristics of individuals in a population are commonly highlighted in the statistical analysis. However, it is common that some information is missing in the corresponding epidemiological dataset, causing difficulties in yielding valid results. Thus imputation and modelling are usually applied to estimate or predict the unknown data.

This PhD thesis is divided into two parts, which are (i) quantification of replication present in HIV reports and (ii) effect of patient movement between wards and MRSA acquisition. The main techniques of statistical analysis in both parts are modelling and imputation for estimating the unknown information. In the first three chapters, we will focus on the first part of the thesis, which is quantification of replication present in HIV reports. Then the analysis for the second part of this thesis will be demonstrated in the latter three chapters. Now the literature review of HIV background and the replication problem in the datasets of HIV reports will be introduced.

Sexual and reproductive health and HIV are major public health issues. These can cause severe impact on societal and economic well-being and thus there will be benefits from reliable information being available on the relevant quantities such as number of HIV infections and HIV incidence to improve public health. In the first part of this PhD thesis, we will focus on the replication problem present in HIV reports and study the methodology of quantifying the overcounting of HIV datasets. There will be three chapters for the first part of this thesis, which are a literature review, methodology to quantify the amount of replication and results of the replication present in HIV reports. In this literature review chapter, we will introduce the background of HIV, the replication problem of HIV, previous work and the available data which will be used for the further analysis of the HIV replication problem in the next two chapters in this PhD thesis.

2.1 Background.

First of all, in this section we will give a brief introduction on HIV, including the medical, biological and economic aspects.

HIV is a genus lentivirus (i.e. slowly replicating retrovirus) that causes Acquired Immune Deficiency Syndrome (AIDS) which is a life-threatening clinical condition in humans. AIDS was first clinically recognised in 1981 in the United States [11], which leads to progressive failure of the immune system that allows opportunistic infections and cancers to thrive. In 1983, two separate research groups led by R Gallo and L Montagnier respectively declared at the same time that a novel retrovirus which was successfully isolated from an AIDS patient may have been infecting AIDS patients and this virus was named HIV in 1986 [5], [9], [42]. Infection with HIV commonly occurs by the transfer of HIV contained in body fluids such as blood, semen, vaginal fluid, pre-ejaculate and breast milk as a media associated with sexual contact, injecting drug use, mother-to-child transmission and blood transfusion. On the other hand casual contact does not cause HIV infections. Especially, sexual intercourse is a major mode of transmission. The largest proportion of all HIV infections happen through unprotected sexual contact [26]. According to statistics from the Health Protection Agency, 95% of HIV colonisations and infections in the UK in 2010 were acquired by sexual contact [115]. The risk of HIV infection is higher for those who have multiple sex partners and unsafe sexual practices. Thus AIDS is also considered as a sexually transmitted disease (STD).

Globally, 34.0 million people were living with HIV at the end of 2011, where 2.5 million people were new HIV infections [127]. The report of the Joint United Nations Programme on HIV and AIDS (UNAIDS) in 2011 also showed that 1.8 million people died from AIDS-related death in 2010. The majority of HIV infected people are in Sub-Saharan Africa which continues to bear the brunt of the global epidemic. In the UK, an estimated 96,000 people were living with HIV, where 30,800 African born heterosexuals were living with HIV [124]. The prevalence of HIV infection has been declining in recent years in global efforts to address the AIDS epidemic. Compared to the number of new HIV infections across the world in 2001, there were 700,000 fewer in 2011 [127]. However, the treatment and prevention of HIV is still a primary issue in global public health and HIV infection in humans is considered as a pandemic disease by the World Health Organisation. Understanding of HIV pathogenesis can benefit the development of HIV treatments.

2.1.1 The virology of HIV.

The unique gene structure of HIV has been studied and two types of HIV have been characterized which are HIV-1 and HIV-2. The majority of HIV infection is caused by

HIV-1 globally, which is more virulent and infective [44]. HIV is typically responsible for long duration illnesses with a long incubation period [70]. A study reported that the mean of the incubation period is around ten years with a 95% confidence interval of 8.4 to 11.2 years [6], which indicated that the incubation period is also variable. Gigli et. al. [43] also pointed out that the effect of age in infection time enhances the uncertainty of the incubation period. For young people under 25, the incubation time shows rather large variability.

HIV infects vital cells in the human immune system such as CD4 T cells, macrophages, and dendritic cells [24], which protect the body against various bacteria, viruses and other germs. After HIV enters into the cells, it makes copies of itself by cellular transcription and in the meantime it attacks and kills the CD4 T cells, leading to a low level of CD4 T cells. Eventually, when the number of CD4 T cells declines to a critical level, the human cell-mediated immunity is lost and the body becomes progressively more susceptible to opportunistic infections. This process may take a long time, during which time it is asymptomatic. Most people usually experience a short, flu-like illness such as fever, rash and a severe sore throat two to six weeks after HIV infection, which is known as seroconversion illness [115]. There are two classic symptoms of AIDS, which are swelling of the lymph nodes of the neck and physical weakness [115].

2.1.2 The diagnosis and treatment of HIV.

HIV infection is identified either by the detection of HIV-specific antibodies in serum or plasma using antibody testing or by the presence of virus using polymerase chain reaction, p24 antigen testing [34]. Antibody testing based on blood and other body fluids is a common method used to detect HIV infection. In addition, the enzyme immunoassay screen test which is a highly sensitive and specific test and shortens the time from exposure to detection of HIV infection (i.e. the window period) to within two to three weeks, combined with the Western blot test, which is a confirmatory test are currently on the market for the diagnosis of HIV infection [88].

Although it is impossible to eradicate HIV yet, early pharmacological intervention gives the best chance at preserving the integrity of the immune system [92]. Since the mid 1990s there have been combinations of antiretroviral drugs which is highly active antiretroviral therapy (HAART) for HIV infection will delay the onset of AIDS and increase the lifespan of HIV infected individuals. Because of the access to HAART which reduces the viral load (i.e. a measure of the amount of HIV) in blood, AIDS-related mortality and morbidity has dropped dramatically, especially in affluent countries where the facilities and sources are relatively sufficient [79], [80], [90]. A report by Rochstroh et al. [101] also suggested that management of underlying hepatitis B and/or hepatitis C in patients with HIV infection under HAART is important in preventing morbidity and mortality. In addition, prevention strategies are suggested to reduce the risk of acquiring HIV, consisting of promotion in education and condom distribution to prevent HIV transmission. A routine antenatal HIV screening test for pregnant women is also recommended in order to reduce the risk of mother-to-child transmission by early interventions such as taking HAART and avoiding breastfeeding. In the UK, 96% of pregnant women accepted the routine antenatal HIV screening test in 2010 [3]. In 2011, the National Institute for Health and Clinical Excellence in the UK suggested a further implementation of HIV testing in the two high-risk groups which are men who have sex with men and black African communities [37].

The introduction of HIV testing improves the understanding of the high risk groups and identifies the number of HIV colonised patients who are likely to be AIDS patients in future, giving a better idea of the spread of the AIDS disease.

2.1.3 The economic cost of HIV infection.

The economic burden associated with prevention and treatment of HIV-related disease is a considerable problem of global concern. The treatments of HIV infection usually require a high financial cost either for individuals or for nations. In addition, HIV infections also affect economic growth by reducing the availability of human capital. In some heavily infected areas, the AIDS epidemic has led to an increased mortality which results in reduced productivity by a smaller skilled population and labour force [50]. Hutchinson et al. [58] reported that the cost of new HIV infections in the United States in 2002 was estimated at 36.4 billion dollars, including 6.7 billion dollars in direct medical costs and 29.7 billion in productivity losses. The UK has the fastest growing HIV epidemic in Europe and the cost of treatment and care increased from 104 million pounds in 1997 to 483 million pounds in 2006, which has risen 4.6 fold between 1997 and 2006 [117]. The projected annual cost for the treatments for people living with HIV in UK is between 721 million pounds and 758 million pounds by 2013 [74].

Due to the serious economic consequences of HIV and AIDS to both individuals and nations, it is imperative to have as accurate information as possible on the number of HIV infections so that improvements on cost-effective treatment and care strategies can be made [31].

As we discussed above, AIDS is an STD and especially homosexual activities between men are considered as having high risk in spreading HIV. The common routes of HIV infections also include injecting drug use. Hence there is a social stigma attached to a diagnosis of HIV or AIDS. In addition, the long and variable incubation period for HIV means that patients usually carry HIV asymptomatically for several years, which may cause individuals to be ignorant of their HIV infection and thus spread the HIV virus. It is difficult to obtain reliable estimates of the scale of the epidemic or data which contributes valuable information to surveillance analysis [26]. Therefore the replication problem in the available data of the number of HIV infections has attracted considerable research attention. In the next section, we will present the replication problem existing in the data of HIV infections.

2.2 The replication problem of HIV.

In the UK, the first death of AIDS was reported in 1982 [12] and then the reporting system was based on voluntary HIV reports in Scotland made to the Scottish Centre for Infection and Environmental Health (which is now Health Protection Scotland in Glasgow) and voluntary HIV reports in England and Wales made to the Public Health Laboratory Service (PHLS) (which is in Colindale, London) [51]. In this section, we will introduce the replication problem in the HIV data recording system in the 1990s.

Due to the known common routes of HIV infections including injecting drug use and unprotected promiscuity (especially homosexuality), HIV infection carries a stigma of personal irresponsibility or moral fault. The negative social implications imply that the confidentiality of reporting for patients is important. In many countries, laws establishing the confidentiality of AIDS information were established. In the UK, in order to ensure the confidentiality of AIDS reporting, patient names are not held on the databases, but a simple coding of the surname to a four-digit alphanumeric code (i.e. 'Soundex' code) is usually recorded instead [82].

In the early 1990's the PHLS were concerned about the accuracy of their database. Because of confidentiality, patient names were not recorded on the database but a short report accompanied an HIV positive diagnosis which generally contained the date of birth of the patient. There was no way to know whether or not two or more birth records with the same date of birth correspond to individuals multiply recorded in the database. Some individuals may have had more than one HIV test. There is a known example of an individual having had five HIV tests.

In the early 1990's whilst some reports had associated 'Soundex' codes the 'Soundex' codes were missing for a substantial proportion of the database. Thus the PHLS was interested to know whether statistical information on potential replication in the database was contained in the distribution of birth dates in the database. A cross-sectional sample of the database as it stood in 1991 was sent to Strathclyde University for statistical analysis.

In the period between 1991 and 1994 the PHLS made strenuous attempts to reduce the amount of replication in the database including eliminating known duplication and trying to ensure that as much as possible of the existing database had ‘Soundex’ codes. In 1994 the distribution of birth dates in the entire dataset was again sent to Strathclyde University for statistical analysis to see whether the level of replication had reduced.

There were multiple reasons why a HIV positive individual was being repeatedly recorded in the database which led to the replication problem. First, because of the serious lethal and incurable condition of HIV infection, an individual diagnosed as HIV positive might not believe the test result or might want to double check by being retested elsewhere. As a result, all the multiple test results for the same person were forwarded to the database and recorded. The second reason is that it was National Health Service (NHS) policy that an individual who reported to a new clinic or general practitioner as HIV positive was usually required to get another test before receiving HIV-related treatment. Additionally, individuals when taking HIV tests might sometimes use false names because of the social stigma associated with an AIDS diagnosis.

As it is impossible to completely eliminate replication in the anonymous reporting of the PHLS dataset in the 1990s, statistical methods can be used to estimate multiple counting in the dataset.

Recently, an advanced HIV surveillance system has been used including the new diagnoses system and the cross-sectional annual survey of prevalent HIV infections diagnoses (SOPHID) for collecting number of patients in a calendar year who attend for HIV-related care at an NHS site in England, Wales and Northern Ireland [77], [96]. Furthermore, the Health Protection Agency has developed a new database for collecting HIV reports recently (i.e. the HIV and AIDS Reporting System), which includes the data on site code where an individual received HIV-related treatments, the date that the patient was first diagnosed as HIV positive in the UK, the country of birth as well as ‘Soundex’ code and so on [30].

Considering the replication problems which was impossible to be completely eliminated in the anonymous reporting of the PHLS dataset in the 1990's, statistical methods can be used to estimate multiple counting in the database.

2.3 Previous work on the replication problem.

In this section, we focus on the previous work on the observed statistical methods associated with the problem of replicated birth dates.

Larsen estimated the number of individuals in a register from the number of distinct birth dates based on the classical occupancy theory, where the assumption was that the only information available was the birth date of each person [68]. In other words, based on the register which consisted of distinct birth dates records, the estimation of the true number of registered people can be obtained. The basic theory of the classical occupancy problem was addressed by Feller, which provided the probability of exactly m empty cells in a total of n cells where r balls were occupied randomly in these n cells [35]. Larsen defined n to be the number of consecutive days in a sequence of possible birth dates (for example 365 days); r was the true number of registered individuals born in this sequence of observed birth dates, which was aimed to be estimated in this problem; b was the observed number of distinct birth dates recorded in the register (i.e. occupied birth dates in n possible birth dates); and $m = n - b$ was the number of empty birth dates in the sequence of n possible birth dates. Thus r can be estimated using the approximate maximum likelihood method. Barabesi et al. [8] systematically introduced various approximations of the maximum likelihood estimate in the classical occupancy model. Larsen [68] proposed an approximate maximum likelihood estimate $\hat{r}_0$ for r on a basis of a Poisson asymptotic framework when n is large, which is

$$\hat{r}_0 = n \log(n/m),$$

with the corresponding approximate variance

$$V(\hat{r}_0) \approx ne^{-r/n}.$$

The numerical calculations for approximate maximum likelihood estimate showed that the point estimator $\hat{r}_0$ with the corresponding 95% confidence interval was quite near to the point estimator of r calculated from the exact maximum likelihood method with the corresponding 95% confidence interval when n was large.

Larsen also pointed out an alternative approach to estimate the true number of individuals in a birth date register, where the number of individuals was taken to be a random variable reflecting the stochastic nature of registration. Specifically, he supposed that the new individuals were registered with a certain intensity and the number of observed birth dates reflected both the nature of arrival of new individuals and the overlapping of their birth dates [68]. Then the estimated true number of individuals can be obtained by

$$\hat{r}_1 = n \log(n/m),$$

with the corresponding approximate variance

$$V(\hat{r}_1) \approx n(n-m)/m.$$

As an example he used the number of registrations of Chlamydia infections occurring in 1989 in a Danish country. Larsen applied his approach introduced above to estimate the true number of individuals in the register as well as the corresponding variance of the estimate. In addition, Larsen [68] and Song et al. [120] also applied a classical occupancy model to an United States national AIDS surveillance data which contained partial individual identifiers such as sex, birth date and 'Soundex' code but reported duplications to estimate the population size of AIDS cases. They proposed considering individuals as balls and various combinations of sex, birth date and 'Soundex' code as cells. However, the estimate of the true number of individuals proposed by Larsen is only suitable when $\hat{r} < n$ [8]. An assumption underlying the classical occupancy method is that the partial personal identifiers such as birth of date do not vary over time for each individual [120]. On the other hand, if the partial personal identifiers change or entry errors are made, the classical occupancy method would underestimate the number of true replications. For example, a woman's 'Soundex' code could be changed after marriage.

Moreover, the replication problem can also be considered as related to a record linkage problem [1], [84]. Record linkage is the process of determining that two or more records probably refer to the same individual [118]. For example, suppose that there are two files: file A and file B with records pertaining to individual cases [60]. Both files contain identifiers with the same information to be matched such as date of birth and sex and each of the files is assumed to contain no duplicate records. Further suppose that n is the number of records on file A and m is the number of records on file B, then there are $n \times m$ pairs of records in total required to be analysed in order to classify each pair as matched record pair or unmatched record pair. Consequently, both records from each matched pair are considered as essentially one record but each record from an unmatched pair is considered as essentially a distinct record.

There are two basic methods for record linkage, which are deterministic linkage and the more complex probabilistic linkage. Particularly, the deterministic linkage is the simplest record linkage, which generates links based on the number of individual identifiers such as date of birth and sex that match among the available dataset [102]. It is effected only when there is an exact match on all linking variables. Note that a linking variable is a single criterion (i.e. identifier) utilized to establish or partially establish record linkage [118]. Deterministic linkage is suitable when the records in the dataset have a common identifier or when there are several representative identifiers. For example, Muse et al. discussed the evaluation of the quality of anonymous record linkage with the New York State AIDS Registry and a hospital discharge file using multiple computer algorithm deterministic linkage [83]. In this study, the records from two population based files including hospital identification codes, dates of hospitalisation, sex and date of birth were linked using a deterministic procedure. Using the number of true links between the two files which were identified by using manual verification of additional information (not contained in the two files) such as the address of the patient and the phone number of the patient, the sensitivity (i.e. the proportion of the true links which are correctly identified by the computer algorithm) and positive predictive value (i.e. the percentage of computerised links which are true links verified manually) were calculated.

The probabilistic record linkage takes into account weights of each identifier based on its ability to correctly identify a match or a non-match and then evaluates the probability of two given records referring to the same individual. This approach is normally used when there is more than one linkage variable and it is suitable for information-poor situations [83]. The weight for each identifier can be estimated by a $\log_2$ of the odds ratio that the two records in this identifier refer to the same individual against those two records refer to different individuals. From a mathematical point of view, the weight for each identifier can be expressed as

$$\log_2(m/n),$$

where m is the probability that an identifier agrees given that the records being examined are a matched pair and n is the probability that an identifier agrees given that the records being examined are an unmatched pair. Note that m would be 1.0 in the case of perfect data, but this is rarely true in practice because, for example, records are prone to miscoding. Thus the estimation of the m probability is usually required. In particular, if prior knowledge of the datasets is available, a bootstrap method is proposed. By using the standard agreement rate for each identifier (i.e. the standard m known from other independent datasets) as an initial estimate, the m probability can be iteratively recalculated according to a bootstrap method [57], [131]. Howe et al. [57] and Arellano et al. [7] pointed out that commonly used identifiers demonstrate similar agreement rates across independent datasets. Hence a bootstrap method is appropriate to be used for estimating the probability parameters for common identifiers in a probabilistic record linkage. However this approach requires manually reviewing indeterminate record pairs [48]. Another common approach for the estimation of the probability parameter m is Naive Bayes based on a training dataset (i.e. a suitable amount of representative data) [85]. The Naive Bayes approach guaranteed the conditional independence (i.e. the density function for each comparison pair is different within the match or non-match classes) which is used as an assumption within the Naive Bayesian network and is useful by making the computation much more tractable [85], [135]. However, the training dataset can be expensive to be obtained and hence the Expectation-Maximisation (EM) algorithm can be applied to derive the m probability based on a presumed initial value [48]. Considering the estimation for the probability parameter n in the probabilistic record linkage, it is usually easy and straightforward. Take the month of birth as an example, the probability parameter n is $\frac{1}{12}$ when assuming that the month of birth is approximately uniformly distributed. However, the estimation approaches introduced above are also suitable for estimating the n probability.

In general, by comparing the total weight which combines all the weights for each identifier calculated above to the two threshold values, there are three different kinds of linkage defined for the records: (i) the 'definite links' with a total weight above the upper threshold; (ii) the 'non-links' with a total weight below the lower threshold and (iii) the 'possible links' with a total weight between two thresholds, which can be addressed by human review [57]. However, Grannis et al. [48] pointed out that a single true-link threshold can be established to avoid human review. In other words, a single threshold can be picked, above which a link is declared and below which a non-link is declared, so that human review can be removed. They demonstrated the calculation of a match likelihood score for each record pair in the probabilistic linkage model which summarised the component weights of each identifier and was used to compare with the single threshold to determine a link or a non-link. The algorithm of a match likelihood score was based on the EM algorithm for the estimation of probability parameters. This study also showed that the EM algorithm estimated linkage probability parameters (i.e. the corresponding m and n probabilities) with acceptable accuracy.

Generally speaking, the probabilistic record linkage method is useful for linking a new set of records which is added to the database with the old one, which can rank agreement between different matching linking variables and incorporate effects such as data transcription errors [51], [60]. Probabilistic linkage software has been developed, which utilizes a mathematical algorithm to determine whether two records should be linked or not based on the information in the datasets. A study addressed by Clark and Hahn [21] suggested that the probabilistic record linkage is more adaptable for general use compared to the deterministic record linkage, especially for linking large amounts of data. In addition, Elmagarid et al. [29] pointed out that probabilistic data linkage can be regarded as a Bayesian inference problem and they also described the Bayes decision rules based on a likelihood ratio with minimum error for the purpose of detecting the duplicate records with multiple identifiers in the dataset.

The efficiency of probabilistic record linkage can be measured by the positive predictive value which is the proportion of records linked by the algorithm that truly do match. A 'duplicate method' described by Blakely et al. [13] can be used to calculate the positive predictive value within the probabilistic linkage procedure when there is only one match for each record which is quite common in epidemiology (e.g. linking between a mortality file and a population file). They also pointed about that the 'duplicate method' is appropriate for the linkage using anonymous data since it does not require a validation subset with detailed personal information such as name and address. The positive predictive value estimated by the 'duplicate method' was proven to be robust to sensitivity analyses.

Considering the application of data linkage on the HIV database, Ades et al. [2] pointed out that data linkage in anonymous surveys can be used to investigate the local prevalence and incidence of the worldwide epidemic. They used an unlinked anonymous neonatal seroprevalence survey with electronic record linkage of data from child health computers (including mother's age, ethnic status and both parents' country of birth) and samples prior to anonymous HIV tagging and testing to assess the HIV prevalence in the UK. The electronically linked data were also sent to the Office for National Statistics. Recently, Rice et al. [96] established a cohort of HIV-diagnosed adults using deterministic record linkage on the data in the 1998 to 2007 SOPHID database, new diagnoses database and Office for National Statistics death records to assess the situation of attendance at HIV-related services for HIV diagnosed adults in England, Wales and Northern Ireland.

Apart from using record linkage approaches to estimate the HIV prevalence in the dataset, Goubar et al. [47] developed a Bayesian framework for synthesis of different sources of surveillance information, implemented through Markov chain Monte Carlo methods, in order to estimate HIV prevalence and proportion of HIV diagnosed people within the SOPHID database in England and Wales. They also pointed out that the data were found to be inconsistent but can be resolved by introducing 'bias adjustment' parameters. Moreover, Presanis et al. also applied the evidence synthesis approach based on the Bayesian framework on multiple data sources to create a transmission dynamic model and estimate the HIV prevalence among men who have sex with men in England and Wales [94].

In this thesis, we are interested in assessing the duplication in the PHLS dataset. In particular, only one linkage variable (i.e. the date of birth) is available for analysis. Thus the classical data linkage techniques are not appropriate here. We shall use maximum likelihood methods to estimate the percentage overcounting present in the PHLS dataset.

2.3.1 Previous work on the replication problem in the PHLS dataset.

The majority of the researches on the replication problem in the PHLS dataset were done by Greenhalgh, Doyle and Mortimer. In this subsection, we will briefly demonstrates the results obtained by Greenhalgh et al. that have been published.

Firstly, Doyle et al. [27] applied three statistical methods to detect whether there was a greater number of replication of individuals than expected by chance alone in the 1991 PHLS dataset of HIV diagnoses where only the birth dates were held. They proposed the theoretical distribution of the number of birth date replicates, which will be particularly demonstrated in the next chapter. By using a simple χ^2 test, they detected five out of eleven records with large sample sizes in the 1991 PHLS dataset having more replication than would have been expected by chance alone (at a 5% significance level). For the other five records with small sample sizes, an incomplete ranking scheme was applied, which revealed that one out of five records in the 1991 PHLS dataset have more replication than expected by chance alone. In addition, they also applied a ranking scheme of pairs to test whether there is replication in the 1991 PHLS dataset. In general, this study pointed out that the replication existed in the 1991 PHLS dataset of HIV diagnoses.

Then Greenhalgh and Doyle [51] developed the pairs test which is suitable for both small and large sample sizes for detecting replication in the 1991 PHLS dataset of HIV diagnoses and took into account the effect of seasonality of birth dates (i.e. birth dates are not randomly distributed). They showed that the effect of seasonality is negligible. Similarly, five out of 16 records in the 1991 PHLS dataset showed evidence of having more replication than expected by chance alone. However, those two studies did not quantify the amount of replication in the data which is worthwhile to be acknowledged when using this dataset. In addition, Greenhalgh et al. [52] also proposed a maximum likelihood method to estimate the probability of overcounting of individuals in a given record. However, the maximum likelihood method was only applied for the five records with small sample sizes in the 1991 PHLS dataset, where the conclusion that one out of five records in the dataset showed evidence of overcounting of individuals was the same as the results obtained by the partial ranking method which is suitable only for small sample sizes.

In our thesis, we focus on quantifying the amount of replicated individuals in the HIV diagnoses dataset based on the maximum likelihood method addressed by Greenhalgh et al. in 1999 [52] (which was introduced above) since previous work has already shown a considerable amount of replicated individuals. We will extend the application of the maximum likelihood method to the target datasets sent to us by the PHLS in 1991 and 1994. The datasets that we will use for the further analysis in this thesis will be demonstrated in the next section.

2.4 The available data.

The PHLS was interested in a statistical method to test whether individuals were being repeatedly counted in the database. The reported HIV positive individuals were divided according to their year of birth in the PHLS database. In 1991, the PHLS sent Strathclyde University a database of HIV diagnoses which only contained the distribution of birth dates of individuals. Particularly, for each birth year, the number of birth dates in that year for which there was at least one record in the database and the corresponding number of individuals observed in that year were recorded in the database. No information on 'Soundex' codes was included in the 1991 PHLS dataset. The information in the 1991 PHLS database which was sent to us for the replication analysis is displayed in Table A.1 in Appendix A. In the 1991 dataset, the birth year of the recorded individuals ranged from 1929 to 1944. For a given birth year, the number of individuals who were born in this year was included in the dataset and the record of the number of birth dates for those individuals was presented as a vector, consisting of the singletons, doubletons, tripletons and so on. For a given birth year, the distribution of the number of birth records can be expressed as

$$\boldsymbol{S} = (S_1, S_2, S_3, \cdots, S_n),$$

where S_1 denotes singleton birth records, S_2 denotes doubleton birth records, S_3 denotes tripleton birth records, up to S_n which denotes n-tuple birth records. A singleton represents a single birthdate (i.e. the birth records in the singletons had distinct birth dates). A doubleton represents that two birth records having the same birth date (i.e. each pair of birth records in the doubletons had the same birth dates). Similarly, an n-tuple is a birth date which appears in exactly n records in the dataset (i.e. n birth records in one n-tuple had the same birth dates). Therefore, the observed number of birth records in this given birth year can be calculated by $\sum_{i=1}^{n} iS_i$ which can also be considered as the observed number of individuals. Note that only non-zero birth year record tuples were recorded. All non-recorded birth year tuples are zeros. For example, for a given birth year of 1939 in the 1991 PHLS dataset, the total number of individuals reported as HIV positive was 99 and the record of number of birth dates can be expressed as (69, 13, 2) i.e. the $s_1 = 69$, $s_2 = 13$, $s_3 = 2$, $s_4 = s_5 = \cdots = 0$. Thus this represents that there are 69 birth records which have 69 distinct birth dates and 13 pairs of birth records where each pair of birth records had one distinct birth date and 2 tripletons where each triple means a birth date repeated three times. Hence the observed number of birth records in this birth year of 1939 is $69 + 2 \times 13 + 3 \times 2 = 101$. However it is difficult to tell from this information whether the multiple recording (such as doubletons, tripletons and so on) corresponds to one individual recorded repeatedly or multiple distinct individuals with the same birth date.

The 1994 dataset ranged from the 1901 birth year to the 1973 birth year, and has a larger population size compared to the 1991 dataset. The 1994 PHLS dataset of HIV diagnoses was also sent to us for the further analysis, which is displayed in Table A.2 in Appendix A. The same notation of birth dates records as used in the 1991 dataset was applied to the 1994 dataset.

In this thesis, the replication problem in both 1991 and 1994 PHLS datasets will be highlighted. Although the replication analysis cannot identify whether a particular pair of records is a pair of true replications or non-replications, it helps to estimate the magnitude of true number of distinct individuals reported as AIDS in a surveillance system [120]. As we mentioned in this literature chapter, the maximum likelihood method is the basic technique that will be introduced specifically in the next chapter. Moreover, we will also develop the bootstrap method for the purpose of generating the 95% confidence interval for the estimate of replicated records of birth dates. In Chapter 4, the results obtained by the methods introduced in Chapter 3 for both 1991 and 1994 datasets will be illustrated.

Chapter 3

Methodology to quantify the amount of replication

According to the study presented by Greenhalgh, Doyle and Mortimer [27], replication is present in the AIDS dataset sent from the PHLS in 1991. Our main target is that of developing an algorithm to estimate the amount of replication existing in the two datasets provided by the PHLS AIDS center. This chapter presents a general approach to iterative computation for the maximum likelihood estimate of the true number of distinct individuals as well as the estimated replication percentage based on the maximum likelihood technique. In general, the majority of the birth years in the dataset have several potential true numbers of distinct individuals, which can be derived from the observed sample. In order to obtain the maximum likelihood estimator of the true number of distinct individuals, we have to construct the likelihood function for each possible true number of distinct individuals and the true probability distribution of the number of HIV tests taken by an individual which is the probability of obtaining the observed replication vector given this true number of distinct individuals and probability distribution. For most of the birth years in the dataset, there is more than one likelihood function and therefore the maximum likelihood estimate can be obtained by maximising over all the results of the likelihood functions using the statistical software package R or the scientific programming language C.

For a given birth year, from observed replication vector $\boldsymbol{s} = (s_1, s_2, \cdots, s_n)$ with the corresponding sample size $r = \sum_{i=1}^{n} is_i$ we can derive a series of potential true numbers of distinct individuals r_j. Under each potential true individual record, the corresponding replication vectors can be generated based on the observed one so that a likelihood function can be constructed. We develop an iterative method to do the calculation.

3.1 The method of deriving the potential replication vectors.

Based on the observed replication vector $\boldsymbol{s} = (s_1, s_2, \cdots, s_n)$, the observed number of distinct individuals (i.e. the observed sample size) is $r_{obs} = \sum_{i=1}^{n} is_i$. However it is possible that one or even several persons were recorded repeatedly. One of the possibilities is that the records in each of the s_2 doubletons is actually the same individual and so are the records in each of the s_3 tripletons, $\cdots$, and s_n n-tuples, which leads to the minimum possible true number of distinct individuals $\bar{r}_{min} = \sum_{i=1}^{n} s_i$. Obviously, the maximum possible true number of distinct individuals is just the observed sample size $\bar{r}_{max} = r_{obs}$. Moreover, the rest of the possible true records are able to be calculated as $\bar{r}_i = \bar{r}_{max} - i$ where i is the number of repeated records that are assumed to exist in the observed sample and also $0 \leq i \leq \bar{r}_{max} - \bar{r}_{min}$. Specifically $i = 1$ means that there is one person recorded repeatedly and thus the sample size becomes one less than the observed sample size, i.e. $\bar{r}_1 = r_{obs} - 1$. Similarly $i = 2$ means there are two persons recorded twice or one person recorded three times and the corresponding sample size is $\bar{r}_2 = r_{obs} - 2$ and so on. Therefore, the potential numbers of true total distinct individual records is $\bar{r}_{max} - \bar{r}_{min} + 1$.

Take the birth year 1931 with the observed replication vector $(s_1, s_2, s_3) = (37, 6, 1)$ as an example. The observed sample size (i.e. the maximum possible true number of distinct individuals) is 52 ($37 + 2 \times 6 + 3 \times 1$) since there are 37 individuals having a different birth date throughout the year, 6 pairs of persons where each two have the same birth date and a trio of persons where the three persons have the same birth date. Also the minimum possible true number of distinct individuals can be calculated by considering the six pairs of persons as six distinct individuals and the trio of persons as a single individual that were recorded repeatedly, which means that $\bar{r}_{min} = 37 + 6 + 1 = 44$. According to the algorithm introduced above, the other possible true numbers of distinct individuals are $\bar{r}_1 = 51, \bar{r}_2 = 50, \cdots$, and $\bar{r}_7 = 45$ respectively. Therefore, there are 9 possible potential numbers of true distinct individual records.

With a given potential true number of distinct individuals that is derived above, we are interested in the corresponding replication vectors. An iterative method has been developed to generate the sets of the replication vectors based on the observed one. The following graph (Figure 3.1) demonstrates the idea of the algorithm clearly. Considering moving one tuple only one step every time from the right to the left, the new replication vectors are able to be derived. For the purpose of illustrating the methodology of deriving the potential replication vectors by the Figure 3.1 we assume that it is always true that at least one record exists in s_k ($1 \leq k \leq n$) i.e. $s_k \geq 1$ where $1 \leq k \leq n$. Otherwise we skip this step and move on to the next one. In other words, if $s_k = 0$ ($2 \leq k \leq n$) we do nothing about it and proceed to s_{k-1} to check whether it is larger than 0.

Firstly, based on the observed replication vector $(s_1, s_2, s_3, \cdots, s_{n-2}, s_{n-1}, s_n)$ with sample size r_{obs} we move one tuple from the last element s_n to the left element s_{n-1} which is only step from s_n, giving us the first new replication vector $(s_1, s_2, s_3, \cdots, s_{n-2}, s_{n-1} + 1, s_n - 1)$. Here we assume that $s_n \geq 1$. This new replication vector implies that there is one replication in an n-tuple, i.e. there is a person that was recorded twice. In other words, the true number of distinct individuals for the new replication vector becomes $r_{obs} - 1$. We denote this first new replication vector as $(t_1^{(1)}, t_2^{(1)}, t_3^{(1)}, \cdots, t_{n-2}^{(1)}, t_{n-1}^{(1)}, t_n^{(1)})$. Then returning to the observed replication vector we move one tuple from s_{n-1} ($s_{n-1} \geq 1$) one place to the left to s_{n-2}, getting the second new replication vector $(s_1, s_2, s_3, \cdots, s_{n-2}+$

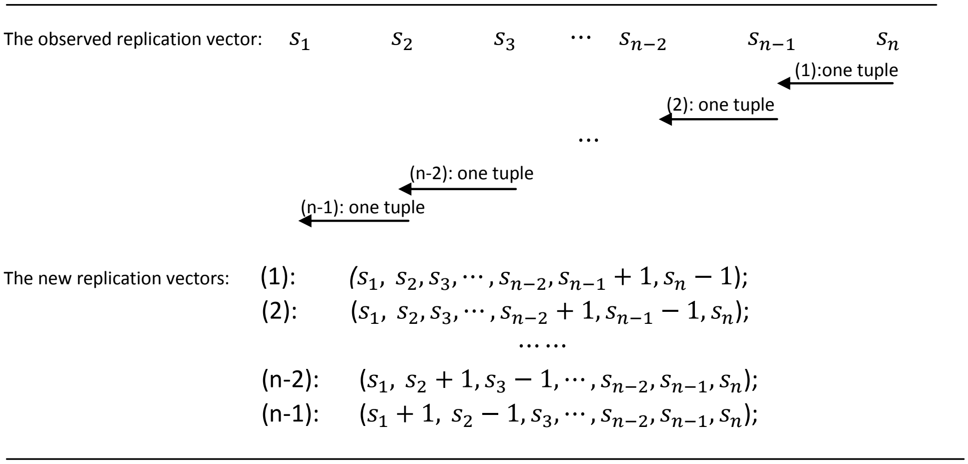

The observed replication vector: s_1 s_2 s_3 $\cdots$ s_{n-2} s_{n-1} s_n
(1):one tuple
(2): one tuple
$\cdots$
(n-2): one tuple
(n-1): one tuple
The new replication vectors: (1): $(s_1, s_2, s_3, \cdots, s_{n-2}, s_{n-1}+1, s_n-1)$;
(2): $(s_1, s_2, s_3, \cdots, s_{n-2}+1, s_{n-1}-1, s_n)$;
$\cdots\cdots$
(n-2): $(s_1, s_2+1, s_3-1, \cdots, s_{n-2}, s_{n-1}, s_n)$;
(n-1): $(s_1+1, s_2-1, s_3, \cdots, s_{n-2}, s_{n-1}, s_n)$;

Figure 3.1: The method of deriving the true potential replication vectors.

$1, s_{n-1}-1, s_n)$ denoting this as $(t_1^{(2)}, t_2^{(2)}, \cdots, t_{n-1}^{(2)}, t_n^{(2)})$. It also suggests that one of the $n-2$ persons who have the same birth date was recorded twice, leading to an overcount in $(n-1)$-tuples and an undercount in $(n-2)$-tuples. Similarly, another $n-3$ new replication vectors can be generated in the same way shown above. Note that the entire set of the new replication vectors obtained at this stage have the same sample size $\bar{r}_1 = r_{obs} - 1$ and they will be treated as the potential true ones. This procedure is illustrated diagrammatically by Figure 3.1.

In the second stage, we carry out the same procedure presented above with regard to each of the replication vectors we obtained before. For the replication vector $(t_1^{(1)}, t_2^{(1)}, t_3^{(1)}, \cdots, t_{n-2}^{(1)}, t_{n-1}^{(1)}, t_n^{(1)})$ with sample size $\bar{r}_1$ as a base replication vector, up to another $n-1$ new replication vectors can be derived by moving tuples one step to the left, from $t_k^{(1)}$ to $t_{k-1}^{(1)}$ where $1 < k \leq n$. Note that it might not be possible to obtain the full $n-1$ new replication vectors since it is possible that $t_k^{(1)} = 0$ was obtained in the first stage. In this procedure moving one individual to the left in replication vector $(t_1^{(1)}, t_2^{(1)}, t_3^{(1)}, \cdots, t_{n-2}^{(1)}, t_{n-1}^{(1)}, t_n^{(1)})$ corresponds to needing to add one additional repeated record to the proposed new true replication vector to obtain the observed replication vector. As a result, the proposed true number of distinct individuals decreases to $\bar{r}_2 = \bar{r}_1 - 1$. Applying the same procedure to the rest of the replication vectors obtained in the first stage, we can get up to $(n-1)^2$ new replication vectors in total treated as the potential true replication vectors with the sample size $\bar{r}_2$. However, it is possible that some of the new replication vectors derived in this stage appear several times. In other words, the same outcome is likely to be derived from different base replication vectors. In each stage, it has to be checked that whether the new replication vector already exists in the set of potential true replication vectors and if so the repeated one has to be eliminated. As a matter of fact, the total number of the potential replication vectors with sample size $\bar{r}_2$ is less than $(n-1)^2$. In order to find an upper bound for the amount of potential replication vectors in the second stage, we consider that there are $n-1$ base replication vectors obtained in the first stage and we also assume that all the elements t_i $(1 \leq i \leq n)$ in the vector are strictly positive integers. Comparing the first group of potential replication vectors derived from the first base replication vector $(s_1, s_2, \cdots, s_{n-1}+1, s_n-1)$ obtained in the first stage to the second group of replication vectors derived from the second base replication vector $(s_1, s_2, \cdots, s_{n-2}+1, s_{n-1}-1, s_n)$ obtained in the first stage we notice that there is one repeated replication vector existing in the second group which needs to be eliminated. Thus the amount of new replication vectors in the first group is still $n-1$ while the number of new replication vectors in the second group becomes $n-1-1$. Similarly, by checking the third group of the new potential replication vector we could find two repeated ones. Consequently, we can calculate the upper bound for the number of potential replication vectors in the second stage is $(n-1)+(n-1-1)+(n-1-2)+\cdots+(n-(n-1)) = \frac{(n-1)\times n}{2}$.

An iterative route is able to be constructed by repeating the procedure introduced above for each of the replication vectors we obtained before and so on. The calculation can be complicated for a large sample size, so we use the statistical software R. In general, except for the observed replication vector $(i=0)$ the upper bound for the amount of the potential true replication vectors with the corresponding sample size $\bar{r}_i$ is $\binom{n+i-2}{i}$ if $n \geq 2$. If $n=1$ (i.e. only singletons are observed) or $i=0$ (i.e. there is no replication) then there is at most one potential replication vector corresponding to the observed one. It can be detected from the pattern of deriving the potential replication vectors based on the observed one. The general conclusion is demonstrated in Theorem 3.1 which is proved below.

Generally speaking, considering an observed replication vector $(s_1, s_2, \cdots, s_n)$ ($n \geq 1$ and s_1, s_2, s_3, $\cdots$, s_n are all strictly positive integers) with observed sample size r_{obs}, we aim to find an upper bound for the number of potential replication vectors with sample size $\bar{r}_i = r_{obs} - i$ ($i \geq 0$). It is clear that if $n = 1$ there are only singletons contained in the observed replication vector, i.e. we must have all individuals distinct. Hence for $i = 0$ there is just one potential replication vector corresponding to the observed one (i.e. the observed one itself) and for $i \geq 1$ there are no potential replication vectors with true sample size $\bar{r}_i = r_{obs} - i$ corresponding to the observed one. Theorem 3.1 in relation to the calculation of an upper bound for the number of potential replication vectors can be proved, which are given as follows.

Theorem 3.1 *(i) If $n = 1$, $i = 0$ then there is just one potential replication vector corresponding to the observed one with $\bar{r}_i = r_{obs} - i$.*

(ii) If $n = 1$, $i \geq 1$ then there are no potential replication vectors corresponding to the observed one with $\bar{r}_i = r_{obs} - i$.

(iii) If $n \geq 2$ then there are at most $\binom{n+i-2}{i}$ potential replication vectors corresponding to the observed one with sample size $\bar{r}_i = r_{obs} - i$.

Proof. We shall prove the result by mathematical induction on the size n of the observed replication vector $(s_1, s_2, \cdots, s_n)$. The results for $n = 1$ have already been shown.

For $n = 2$, consider the observed replication vector is (s_1, s_2) with sample size r_{obs}. There is at most one potential true replication vector with $\bar{r}_i = r_{obs} - i$, namely

$$(s_1 + i, s_2 - i)$$

i.e. i true singletons are observed as doubletons. Note that we need $s_2 \geq i$ here for this true replication vector to be feasible, but even if $s_2 < i$, one is still an upper bound for the number of true replication vectors with $\bar{r}_i = r_{obs} - i$. Therefore the result demonstrated in Theorem 3.1 is true for $n = 2$.

As an induction hypothesis we assume that the result is true for $n \leq n_0$, where $n_0 \geq 2$. This implies that the upper bound for the number of potential replication vectors corresponding to an observed replication vector $(s_1, s_2, \cdots, s_{n_0})$ with the true sample size $\bar{r}_i = r_{obs} - i$ is $\binom{n_0+i-2}{i}$. Now suppose that $n = n_0 + 1 \geq 3$, and the observed replication vector is $(s_1, s_2, \cdots, s_{n_0+1})$. For any potential true replication vector corresponding to the observed one $(s_1, s_2, \cdots, s_{n_0+1})$ with the true sample size $\bar{r}_i = r_{obs} - i$, we consider the number of observed $(n_0 + 1)$-tuples, which is denoted by j, that are not actually true $(n_0 + 1)$-tuples in the potential true replication vector. It is obvious that j must be between 0 and i inclusive. The upper bound for the number of potential true replication vectors can be derived by considering all of the respective possible values for j.

If $j = i$ then this means that i observed $(n_0 + 1)$-tuples are actually true singletons, doubletons, $\cdots$, (n_0-1) or n_0-tuples. (Actually in this case all the observed (n_0+1)-tuples must be n_0-tuples since the true sample size for the potential true replication vector is $r_{obs} - i$, i.e. the i records in the $(n_0 + 1)$-tuples are only able to move one step to the left so that each of them gives an unit decrement in the observed sample size. However in order to develop an argument that will work for all j the general statements are given here.) Hence any potential replication vector could be written as

$$(s_1^{'}, s_2^{'}, s_3^{'}, \cdots, s_{n_0}^{'}, s_{n_0+1} - i).$$

Moreover, with regard to the corresponding sample size $\bar{r}_i = r_{obs} - i$ for the potential replication vector we have

$$\sum_{i=1}^{n_0} is_i^{'} + (n_0 + 1)(s_{n_0+1} - i) = r_{obs} - i.$$

Since

$$r_{obs} = \sum_{i=1}^{n_0} is_i + (n_0+1)s_{n_0+1},$$

it is straightforward that

$$\begin{aligned} \sum_{i=1}^{n_0} is_i^{'} + (n_0+1)(s_{n_0+1} - i) &= \sum_{i=1}^{n_0} is_i + (n_0+1)s_{n_0+1} - i. \\ \sum_{i=1}^{n_0} is_i^{'} &= \sum_{i=1}^{n_0} is_i + n_0 i. \end{aligned} \tag{3.1.1}$$

If $s_{n_0+1} < i$ then clearly there are no potential replication vectors corresponding to the observed one with $j = i$. On the other hand, suppose that $s_{n_0+1} \geq i$ and i observed (n_0+1)-tuples in the observed replication vector $(s_1, s_2, \cdots, s_{n_0+1})$ are true n_0-tuples giving the new replication vector

$$(s_1, s_2, \cdots, s_{n_0-1}, s_{n_0} + i, s_{n_0+1} - i).$$

This leads to another new replication problem where $n = n_0$ and the new observed replication vector is considered as

$$(s_1, s_2, \cdots, s_{n_0-1}, s_{n_0} + i) \tag{3.1.2}$$

with the observed sample size denoted by $r_{obs}^{(1)}$. In this new replication problem, we look for the number of potential true replication vectors $(s_1^{'}, s_2^{'}, s_3^{'}, \cdots, s_{n_0}^{'})$ corresponding to the observed one (3.1.2) with true number of distinct individuals $r_{obs}^{(1)} - i^{(1)}$ where all the elements in these potential replication vectors here are the same as the ones contained in the first n_0 elements of the potential replication vectors $(s_1^{'}, s_2^{'}, s_3^{'}, \cdots, s_{n_0}^{'}, s_{n_0+1} - i)$ in the previous replication problem. Here $r_{obs}^{(1)}$ is the number of individuals in the observed replication vector (3.1.2). Hence we can establish a map

$$(s_1^{'}, s_2^{'}, s_3^{'}, \cdots, s_{n_0}^{'}) \longrightarrow (s_1^{'}, s_2^{'}, s_3^{'}, \cdots, s_{n_0}^{'}, s_{n_0+1} - i)$$

which is a one-to-one correspondence between potential true replication vectors in the new problem and potential true replication vectors in the original problem with $j = i$ and furthermore the upper bound for the number of potential true replication vectors in the new problem is also an upper bound for the number of potential replication vectors in the original problem. According to the result in (3.1.1), we can derive that

$$\sum_{i=1}^{n_0} is_i^{'} = \sum_{i=1}^{n_0} is_i + n_0 i = r_{obs}^{(1)} - i^{(1)}.$$

Hence $i^{(1)} = 0$.

By the induction hypothesis, the upper bound for the number of potential replication vectors in the new problem $(n = n_0)$ with $i^{(1)} = 0$ is

$$\binom{n_0 - 2}{0} = 1.$$

Hence we can conclude that there is at most one potential true replication vector in our original problem $(n = n_0 + 1)$ with $j = i$.

Next we consider the case $j = i-1$ so that exactly $i-1$ of the observed (n_0+1)-tuples in the original problem (where the observed replication vector is $(s_1, s_2, \cdots, s_{n_0}, s_{n_0+1})$ with sample size r_{obs}) are actually singletons, doubletons, tripletons, $\cdots$, $n_0 - 1$ or n_0-tuples. Hence the potential true replication vectors with sample size $\bar{r}_i = r_{obs} - i$ can be expressed as

$$(s_1^{'}, s_2^{'}, s_3^{'}, \cdots, s_{n_0}^{'}, s_{n_0+1} - (i-1)).$$

Arguing as before, concerning the corresponding true sample size for the potential replication vectors we can deduce that

$$\begin{aligned}\sum_{i=1}^{n_0} is_i^{'} + (n_0+1)(s_{n_0+1} - (i-1)) = r_{obs} - i, \\ = \sum_{i=1}^{n_0} is_i + (n_0+1)s_{n_0+1} - i.\end{aligned}$$

Hence

$$\sum_{i=1}^{n_0} is_i^{'} = \sum_{i=1}^{n_0} is_i + n_0(i-1) - 1. \tag{3.1.3}$$

Now again suppose that all the $i-1$ observed (n_0+1)-tuples are actually n_0-tuples, which leads to another new replication problem to be considered where the new observed replication vector becomes

$$(s_1, s_2, \cdots, s_{n_0-1}, s_{n_0} + i - 1) \tag{3.1.4}$$

and the corresponding observed sample size is denoted by $r_{obs}^{(2)}$. Thus we focus on the number of potential true replication vectors $(s_1^{'}, s_2^{'}, s_3^{'}, \cdots, s_{n_0}^{'})$ with the true sample size $r_{obs}^{(2)} - i^{(2)}$, (whose elements here are also consistent with the ones in the potential true replication vectors $(s_1^{'}, s_2^{'}, s_3^{'}, \cdots, s_{n_0}^{'}, s_{n_0+1} - (i-1))$ in the original problem), corresponding to the new observed replication vector (3.1.4). For $s_{n_0+1} \geq i-1$ the map

$$(s_1^{'}, s_2^{'}, s_3^{'}, \cdots, s_{n_0}^{'}) \longrightarrow (s_1^{'}, s_2^{'}, s_3^{'}, \cdots, s_{n_0}^{'}, s_{n_0+1} - (i-1))$$

establishes a one-to-one correspondence between potential true replication vectors in the new problem and potential replication true replication vectors in the original problem with $j = i-1$. Note that for $s_{n_0+1} < i-1$ in the original problem, there are no potential replication vectors with $j = i-1$. Therefore in any case the number of potential true replication vectors in the new problem is an upper bound for the number of potential replication vectors in the original problem with $j = i-1$. Similarly as before, according to the result in (3.1.3) it can be inferred that

$$\begin{aligned}\sum_{i=1}^{n_0} is_i^{'} &= \sum_{i=1}^{n_0} is_i + n_0(i-1) - 1 \\ &= r_{obs}^{(2)} - i^{(2)}.\end{aligned}$$

Hence we deduce that $i^{(2)} = 1$ since $r_{obs}^{(2)} = \sum_{i=1}^{n_0-1} is_i + n_0(s_{n_0} + i - 1) = \sum_{i=1}^{n_0} is_i + n_0(i-1)$.

By the induction hypothesis, it is known that the number of potential replication vectors in the new problem with $i^{(2)} = 1$ is at most

$$\binom{n_0+1-2}{1} = \binom{n_0-1}{1}.$$

Hence we can conclude that there are at most $\binom{n_0-1}{1}$ potential replication vectors in our original problem with $j = i - 1$.

In general consider $j = p$ where $0 \leq p \leq i$ which means that exactly p of the observed (n_0+1)-tuples in the original observed replication vector $(s_1, s_2, s_3, \cdots, s_{n_0}, s_{n_0+1})$ must actually be singletons, doubletons, tripletons, $\cdots$ or n_0-tuples. The corresponding potential true replication vectors can be obtained as

$$(s_1^{'}, s_2^{'}, s_3^{'}, \cdots, s_{n_0}^{'}, s_{n_0+1} - p)$$

with the same true number of distinct individuals $r_{obs} - i$. Applying the same argument as before we deduce that

$$\begin{aligned}\sum_{i=1}^{n_0} is_i^{'} + (n_0+1)(s_{n_0+1} - p) &= r_{obs} - i \\ &= \sum_{i=1}^{n_0} is_i + (n_0+1)s_{n_0+1} - i,\end{aligned}$$

which gives

$$\sum_{i=1}^{n_0} is_i^{'} = \sum_{i=1}^{n_0} is_i + n_0 p - (i - p). \qquad (3.1.5)$$

Now consider a new replication problem where the observed replication vector is

$$(s_1, s_2, \cdots, s_{n_0-1}, s_{n_0} + p) \qquad (3.1.6)$$

with the sample size $r_{obs}^{'}$. In this new replication problem we look for potential true replication vectors $(s_1^{'}, s_2^{'}, s_3^{'}, \cdots, s_{n_0}^{'})$ corresponding to the observed replication vector

(3.1.6) with true number of individuals $r_{obs}^{'} - i^{'}$. Based on the result in (3.1.5),

$$\begin{aligned}\sum_{i=1}^{n_0} is_i^{'} &= \sum_{i=1}^{n_0} is_i + n_0 p - (i-p) \\ &= r_{obs}^{'} - i^{'}.\end{aligned}$$

Thus we have $i^{'} = i - p$, i.e. the true sample size for the potential replication vectors in this new replication problem is $r_{obs}^{'} - i^{'} = r_{obs}^{'} - (i-p)$.

For $s_{n_0+1} \geq i - p$, the map $(s_1^{'}, s_2^{'}, s_3^{'}, \cdots, s_{n_0}^{'}) \longrightarrow (s_1^{'}, s_2^{'}, s_3^{'}, \cdots, s_{n_0}^{'}, s_{n_0+1} - (i-p))$ establishes a one-to-one correspondence between potential replication vectors in the new problem and potential replication vectors in the original problem with $j = i - p$. By the induction hypothesis an upper bound for the number of potential replication vectors in the new problem with the corresponding sample size $r_{obs}^{'} - (i-p)$ is

$$\binom{n_0 + i - p - 2}{i - p}.$$

Hence an upper bound for the total number of potential replication vectors over all values of j $(0 \leq j \leq i)$ is

$$\binom{n_0 - 2}{0} + \binom{n_0 - 1}{1} + \binom{n_0}{2} + \cdots + \binom{n_0 + i - 2}{i} = \binom{n_0 + i + 1 - 2}{i},$$

which can be proved by Lemma 3.2. This completes the proof of Theorem 3.1. □

Lemma 3.2 *For* $i \geq 0$, $\binom{n_0-2}{0} + \binom{n_0-1}{1} + \binom{n_0}{2} + \cdots + \binom{n_0+i-2}{i} = \binom{n_0+i+1-2}{i}$.

Proof. Using mathematical induction on i, for $i = 0$, it is obvious that

$$\binom{n_0 - 2}{0} = \binom{n_0 - 1}{0} = 1.$$

Hence the result is true. For $i = 1$, we also have the true result that

$$\binom{n_0 - 2}{0} + \binom{n_0 - 1}{1} = 1 + n_0 - 1 = n_0 = \binom{n_0}{1}.$$

Assume that it is true for i_0, i.e.

$$\binom{n_0 - 2}{0} + \binom{n_0 - 1}{1} + \binom{n_0}{2} + \cdots + \binom{n_0 + i_0 - 2}{i_0} = \binom{n_0 + i_0 + 1 - 2}{i_0}.$$

Then for $i_0 + 1$, we have

$$\begin{aligned}
&\binom{n_0 - 2}{0} + \binom{n_0 - 1}{1} + \binom{n_0}{2} + \cdots + \binom{n_0 + (i_0 + 1) - 2}{i_0 + 1}, \\
&= \binom{n_0 - 2}{0} + \binom{n_0 - 1}{1} + \binom{n_0}{2} + \cdots + \binom{n_0 + i_0 - 2}{i_0} + \binom{n_0 + i_0 + 1 - 2}{i_0 + 1}, \\
&= \binom{n_0 + i_0 + 1 - 2}{i_0} + \binom{n_0 + i_0 + 1 - 2}{i_0 + 1}, \text{ (using the induction hypothesis here)} \\
&= \binom{n_0 + i_0 - 1}{i_0} + \binom{n_0 + i_0 - 1}{i_0 + 1}, \\
&= \frac{(n_0 + i_0 - 1)!}{i_0!(n_0 - 1)!} + \frac{(n_0 + i_0 - 1)!}{(i_0 + 1)!(n_0 - 2)!}, \\
&= \frac{(n_0 + i_0 - 1)!}{(i_0 + 1)!(n_0 - 1)!}(i_0 + 1 + n_0 - 1), \\
&= \frac{(n_0 + i_0)!}{(i_0 + 1)!(n_0 - 1)!}, \\
&= \binom{n_0 + i_0}{i_0 + 1}, \\
&= \binom{(n_0 + 1) + (i_0 + 1) - 2}{i_0 + 1}.
\end{aligned}$$

Therefore the result is true for $i_0 + 1$. In conclusion, Lemma 3.2 follows by induction. □

For example consider the observed replication vector $(s_1, s_2, \cdots, s_8)$. We can generate the upper bounds for the number of potential replication vectors at each stage. Clearly according to the method of generating the potential replication vectors we introduced above, the upper bound for the number of potential replication vectors at the first stage is 7 (which is also equal to $\binom{8+1-2}{1}$ for $i = 1$). Taking the situation that repeated replication vectors possibly occur at each stage into account, the upper bound for the amount of potential replication vectors at the second stage ($i = 2$) equals $7 \times 1 + 6 \times 1 + \cdots + 1 \times 1 = \frac{(7+1)\times 7}{2} = 28$ after getting rid of the repeat vectors in the set of new potential replication vectors. It can also be proved by Theorem 3.1 where for $i = 2$, $\binom{8+2-2}{2} = \binom{8}{2} = 28$. With regard to the general formula for the upper bound of the potential replication vectors at the second stage we mentioned before $\frac{(n-1)\times n}{2}$, the result in this example is consistent with this since $n = 8$ here. Recall that the result was derived according to $\frac{(n-1)\times n}{2}$ by considering the number of potential true replication vectors which can be obtained from the stage above ($i = 1$) after taking each base replication vector, considering moving one tuple from each element one stage to the left and then eliminating duplicate replication vectors. Based on a similar procedure, we are able to obtain the upper bounds for the number of potential replication vectors at the third, fourth stage etc. which are as follows:

$$\begin{aligned}
Stage\ 3 :& 7 \times 1 + 6 \times 2 + 5 \times 3 + 4 \times 4 + 3 \times 5 + 2 \times 6 + 1 \times 7, \\
&= 7 \times \frac{1!}{1!0!} + 6 \times \frac{2!}{1!1!} + 5 \times \frac{3!}{1!2!} + 4 \times \frac{4!}{1!3!} + \cdots + 1 \times \frac{7!}{1!6!} = 84, \\
Stage\ 4 :& 7 \times 1 + 6(2 + 1) + 5(3 + 2 + 1) + 4(4 + 3 + 2 + 1) + \cdots \\
&+ 1(7 + 6 + 5 + 4 + 3 + 2 + 1), \\
&= 7 \times \frac{2!}{2!0!} + 6 \times \frac{3!}{2!1!} + 5 \times \frac{4!}{2!2!} + 4 \times \frac{5!}{2!3!} + \cdots + 1 \times \frac{8!}{2!6!} = 210, \\
Stage\ 5 :& 7 \times \frac{3!}{3!0!} + 6 \times \frac{4!}{1!3!} + 5 \times \frac{5!}{2!3!} + 4 \times \frac{6!}{3!3!} + \cdots + 1 \times \frac{9!}{3!6!} \\
&\cdots\cdots
\end{aligned}$$

Therefore we can deduce that an upper bound for the amount of potential replication vectors with the true sample size $\bar{r}_i$ ($i \geq 2$) at the ith stage is $\sum_{j=1}^{7}(8 - j)\frac{(i-3+j)!}{(i-2)!(j-1)!} = \binom{8+i-2}{i}$.

Proof. As we introduced before, the potential replication vectors are derived by moving one tuple only one step every time from the right to the left. We denote a_p as one tuple moving from $(p+1)$-tuples to p-tuples. Here $1 \leq p \leq 7$ since $n = 8$. For example, a_1 means one tuple moving from a doubleton to a singleton. Therefore two a_p's (denoted as a_{p_1}, a_{p_2}) mean that there are two tuples moving one space from the right to the left, leading to the true sample size for the potential replication vectors equal to $\bar{r}_i$ (i.e. the ith stage). Thus we can obtain all the possible movement patterns of the tuples at each stage after removing all the repeated ones, which will give a set of corresponding potential replication vectors. For example, at the second stage ($i = 2$), we can generate the movement patterns of the tuples which are expressed by a_{p_1}, a_{p_2} and this can be used to derive the potential replication vectors based upon the observed one (shown in Table 3.1). Note that the movement pattern a_1, a_2 is equivalent to the movement pattern a_2, a_1, which gives the same replication vector at the second stage. In other words, the order of the movements does not affect the derived new replication vectors.

Table 3.1: Movement patterns of tuples in the second stage.

Movement patterns of tuples							Number of distinct possible patterns
a_1, a_1	a_1, a_2	a_1, a_3	a_1, a_4	a_1, a_5	a_1, a_6	a_1, a_7	7
a_2, a_2	a_2, a_3	a_2, a_4	a_2, a_5	a_2, a_6	a_2, a_7		6
a_3, a_3	a_3, a_4	a_3, a_5	a_3, a_6	a_3, a_7			5
a_4, a_4	a_4, a_5	a_4, a_6	a_4, a_7				4
a_5, a_5	a_5, a_6	a_5, a_7					3
a_6, a_6	a_6, a_7						2
a_7, a_7							1

Hence the upper bound of the amount of potential replication vectors with the true sample size $\bar{r}_2$ at the second stage is $7 + 6 + 5 + \cdots + 1$. Based on the movement patterns for the potential replication vectors at the second stage, we are able to derive the corresponding movement patterns at the third stage (i.e. $i = 3$). For example, based on the movement pattern a_1, a_1 in Table 3.1, there are seven possibilities of a new tuple moving one place from the right to the left, which give seven potential replication vectors with the true sample size $\bar{r}_3$. The corresponding movement patterns are (i) a_1, a_1, a_1; (ii) a_1, a_1, a_2; (iii) a_1, a_1, a_3; (iv) a_1, a_1, a_4; (v) a_1, a_1, a_5; (vi) a_1, a_1, a_6; (vii) a_1, a_1, a_7. According to the movement pattern a_1, a_2 at the second stage, there are six new distinct possibilities of a new tuple moving one place from the right to the left, after considering removing the same movement pattern a_1, a_2, a_1 as we derived above. Similarly, the upper bound for the number of distinct potential replication vectors with the true sample size $\bar{r}_3$ at the third stage is

$$
\begin{aligned}
&(7+6+5+\cdots+1)+(6+5+\cdots+1)+(5+4+\cdots+1)+(4+3+2+1)+\cdots+1,\\
=&7+6\times 2+5\times 3+4\times 4+3\times 5+2\times 6+1\times 7,\\
=&7\times\frac{1!}{1!0!}+6\times\frac{2!}{1!1!}+5\times\frac{3!}{1!2!}+4\times\frac{4!}{1!3!}+\cdots+1\times\frac{7!}{1!6!},\\
=&\sum_{j=1}^{7}(8-j)\frac{j!}{1!(j-1)!},\\
=&\binom{8+1}{3},
\end{aligned}
$$

as stated previously.

According to the procedure of generating the movement patterns of the tuples we used above at the i_0'th stage, we have an upper bound from the (i_0-1)'th stage of

$$
\begin{aligned}
&(7+6+5+\cdots+1)+(6+5+\cdots+1)\times\frac{(i_0-2)!}{(i_0-3)!1!}+(5+4+\cdots+1)\times\frac{(i_0-1)!}{(i_0-3)!2!}+\\
&(4+3+2+1)\times\frac{i_0!}{(i_0-3)!3!}+(3+2+1)\times\frac{(i_0+1)!}{(i_0-3)!4!}+(2+1)\times\frac{(i_0+2)!}{(i_0-3)!5!}+\\
&1\times\frac{(i_0+3)!}{(i_0-3)!6!},\\
=&7+6\times\left(1+\frac{(i_0-2)!}{(i_0-3)!1!}\right)+5\times\left(1+\frac{(i_0-2)!}{(i_0-3)!1!}+\frac{(i_0-1)!}{(i_0-3)!2!}\right)+\\
&4\times\left(1+\frac{(i_0-2)!}{(i_0-3)!1!}+\frac{(i_0-1)!}{(i_0-3)!2!}+\frac{i_0!}{(i_0-3)!3!}\right)+\\
&3\times\left(1+\frac{(i_0-2)!}{(i_0-3)!1!}+\frac{(i_0-1)!}{(i_0-3)!2!}+\frac{i_0!}{(i_0-3)!3!}+\frac{(i_0+1)!}{(i_0-3)!4!}\right)+\\
&2\times\left(1+\frac{(i_0-2)!}{(i_0-3)!1!}+\frac{(i_0-1)!}{(i_0-3)!2!}+\frac{i_0!}{(i_0-3)!3!}+\frac{(i_0+1)!}{(i_0-3)!4!}+\frac{(i_0+2)!}{(i_0-3)!5!}\right)+\\
&1\times\left(1+\frac{(i_0-2)!}{(i_0-3)!1!}+\frac{(i_0-1)!}{(i_0-3)!2!}+\frac{i_0!}{(i_0-3)!3!}+\frac{(i_0+1)!}{(i_0-3)!4!}+\frac{(i_0+2)!}{(i_0-3)!5!}+\frac{(i_0+3)!}{(i_0-3)!6!}\right),
\end{aligned}
$$

$$
\begin{aligned}
=&7+6\times\left(\binom{i_0-3}{i_0-3}+\binom{i_0-2}{i_0-3}\right)+5\times\left(\binom{i_0-3}{i_0-3}+\binom{i_0-2}{i_0-3}+\binom{i_0-1}{i_0-3}\right)+\\
&4\times\left(\binom{i_0-3}{i_0-3}+\binom{i_0-2}{i_0-3}+\binom{i_0-1}{i_0-3}+\binom{i_0}{i_0-3}\right)+\\
&3\times\left(\binom{i_0-3}{i_0-3}+\binom{i_0-2}{i_0-3}+\binom{i_0-1}{i_0-3}+\binom{i_0}{i_0-3}+\binom{i_0+1}{i_0-3}\right)+\\
&2\times\left(\binom{i_0-3}{i_0-3}+\binom{i_0-2}{i_0-3}+\binom{i_0-1}{i_0-3}+\binom{i_0}{i_0-3}+\binom{i_0+1}{i_0-3}+\binom{i_0+2}{i_0-3}\right)+\\
&1\times\left(\binom{i_0-3}{i_0-3}+\binom{i_0-2}{i_0-3}+\binom{i_0-1}{i_0-3}+\binom{i_0}{i_0-3}+\binom{i_0+1}{i_0-3}+\binom{i_0+2}{i_0-3}+\binom{i_0+3}{i_0-3}\right),\\
=&7+6\times\binom{i_0-1}{i_0-2}+5\times\binom{i_0}{i_0-2}+4\times\binom{i_0+1}{i_0-2}+3\times\binom{i_0+2}{i_0-2}+2\times\binom{i_0+3}{i_0-2}+\\
&1\times\binom{i_0+4}{i_0-2},\\
=&7\times\frac{(i_0-2)!}{(i_0-2)!0!}+6\times\frac{(i_0-1)!}{(i_0-2)!1!}+5\times\frac{i_0!}{(i_0-2)!2!}+4\times\frac{(i_0+1)!}{(i_0-2)!3!}+\\
&3\times\frac{(i_0+2)!}{(i_0-2)!4!}+2\times\frac{(i_0+3)!}{(i_0-2)!5!}+1\times\frac{(i_0+4)!}{(i_0-2)!6!},\\
=&\sum_{j=1}^{7}(8-j)\frac{(i_0-3+j)!}{(i_0-2)!(j-1)!}.
\end{aligned}
$$

However

$$\sum_{j=1}^{7}(8-j)\frac{(i_0-3+j)!}{(i_0-2)!(j-1)!}=7\binom{i_0-2}{0}+6\binom{i_0-1}{1}+5\binom{i_0}{2}+\cdots+\binom{i_0+4}{6}.$$

But note that by Lemma 3.2,

$$\binom{i_0-2}{0}+\binom{i_0-1}{1}+\cdots+\binom{i_0+4}{6}=\binom{i_0+5}{6},$$

$$\binom{i_0-2}{0}+\binom{i_0-1}{1}+\cdots+\binom{i_0+3}{5}=\binom{i_0+4}{5},$$

$\cdots$

$$\binom{i_0-2}{0}+\binom{i_0-1}{1}=\binom{i_0}{1},$$

$$\binom{i_0-2}{0} = \binom{i_0-1}{0}.$$

Hence

$$\sum_{j=1}^{7}(8-j)\frac{(i_0-3+j)!}{(i_0-2)!(j-1)!} = \binom{i_0+5}{6} + \binom{i_0+4}{5} + \binom{i_0+3}{4} + \cdots + \binom{i_0-1}{0},$$
$$= \binom{i_0+6}{6},$$

which is the same upper bound as derived previously. □

This upper bound is also valid for $i = 1$, where there are at most seven replication vectors, and $i = 0$, where there is at most one, replication vector.

To illustrate the method of deriving the potential replication vectors from an given observed one, we take the replication vector (37, 6, 1) with the birth year 1931 and observed sample size 52 as an example. Figure 3.2 presents the results of an algorithm for deriving the potential replication vectors based on the observed one.

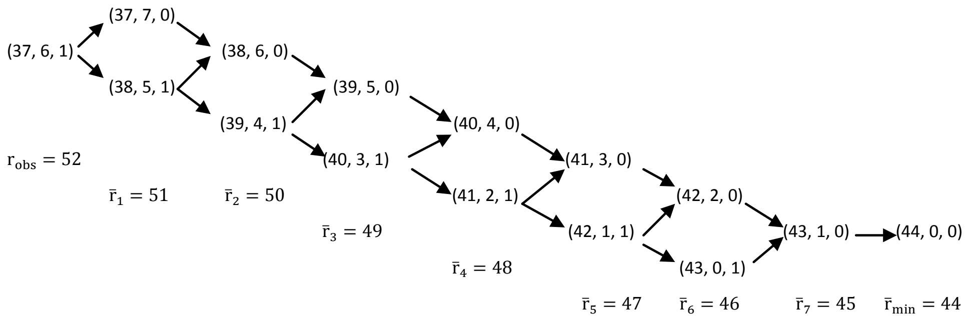


Figure 3.2: The results of potential replication vectors for the observed vector (37,6,1) for the birth year 1931.

Assuming that there is one person recorded twice in the birth year 1931, we consider moving one tuple from the tripletons to the doubletons giving us the potential replication vector (37, 7, 0) with the sample size 51. That is the tripleton actually means two unique individuals instead of three distinct persons having the same birth date. Another outcome with the postulated true sample size 51 is (38, 5, 1) which can be obtained by moving one tuple from the doubletons to the singletons. It indicates that exactly one of the six doubletons is a single individual that was doubly counted. Based on those two potential replication vectors, we can then derive new replication vectors by applying the same algorithm to them. Given that there are two replication vectors present at the current stage, three new replication vectors can be obtained: (38, 6, 0), (38, 6, 0) and (39, 4, 1) where the first result is derived from the potential replication vector (37, 7, 0) and the other two are derived based on the replication vector (36, 5, 1). We notice that the first and the second outcomes are exactly the same although they are derived from different vectors. Therefore, we have to eliminate one of them and the set of the potential replication vectors with the true sample size 50 consists of two vectors: (38, 6, 0) and (39, 4, 1). By means of repeating the same procedure and getting rid of the replication vector that already exists in the set of the potential replication vectors at each stage, all the potential replication vectors can be generated as shown in Figure 3.2.

To complete the calculation for deriving all the potential replication vectors with the corresponding true sample size based on the observed one, we mainly use ‘for’ loops in R and at each stage if the new replication vector is the same as one of the potential replication vectors it has to be eliminated.

As the set of potential replication vectors is obtained associated with the corresponding true number of distinct individuals, the probabilities of the replication vectors are able to be deduced so that the likelihood functions for each potential true sample size can be constructed.

3.2 The construction of the likelihood functions.

The main target of this project is to accurately estimate the amount of replication present in the dataset which will be produced by the maximum likelihood method. We suppose that individuals in a given birth year are independently and randomly sampled. Moreover, the individuals are chosen without replacement from the population consisting of all people born in that birth year. In order to calculate the likelihood function of a replication vector given the corresponding true number of distinct individuals, the probability of occurrence for all the potential replication vectors has to be estimated. Therefore, the probability distribution of the replication vectors should be known at first.

According to Theorem 1 of Greenhalgh, Doyle and Mortimer [52], the probability distribution of the replication vector can be expressed as follows:

$$P(S_1 = s_1, S_2 = s_2, \cdots, S_n = s_n) = \frac{d!}{s_1!s_2!\cdots s_n!(d-t)!}\frac{r!}{(1!)^{s_1}(2!)^{s_2}\cdots(n!)^{s_n}}\frac{1}{d^r} \quad (3.2.1)$$

where $t = s_1 + s_2 + \cdots + s_n$ is the total observed number of tuples in a given year; d is the number of days throughout a year which is chosen as 365 in this project and r is the sample size (i.e. $r = \sum_{i=1}^{n}(is_i)$). As a matter of fact, it is the probability of the individuals having the same birth date which was recoded as the given true replication vector. By applying this theorem, the probability of obtaining a particular replication vector can be calculated for each of the potential replication vectors including the observed one. Using the example we took above, the observed replication vector (37, 6, 1) with the corresponding true number of distinct individuals (i.e. the observed sample size) 52 leads to the probability of obtaining this replication vector being

$$\frac{365!}{37!6!1!(365-37-6-1)!}\frac{52!}{(1!)^{37}(2!)^{6}(3!)^{1}}\frac{1}{365^{52}} = 0.004515.$$

When using R to do the calculation for the distribution of the replication vector, the large factorial argument (for example 365) usually causes difficulties in obtaining the results. Hence we take the logarithm on both sides of the distribution formula, which gives us the log-probability

$$\begin{aligned}\log P&(S_1 = s_1, S_2 = s_2, \cdots, S_n = s_n)\\ &= \log(d!) + \log(r!) - (\log(s_1!) + \log(s_2!) + \cdots + \log(s_n!) + \log((d-t)!)\\ &+ s_2 \log(2!) + s_3 \log(3!) + \cdots + s_n \log(n!) + r \log(d)).\end{aligned} \tag{3.2.2}$$

Once we get the log-probability of a certain replication vector, the corresponding probability can be calculated as $exp(\log P(S_1 = s_1, S_2 = s_2, \cdots, S_n = s_n))$. Then we focus on the probability of the occurrence of the potential replication vectors, including the observed one. A general approach has been developed for calculating it by considering the number of tuples moving from large tuples on the right to the small tuples on the left.

We have already described how to calculate the potential true replication vectors $\boldsymbol{T} = (T_1, T_2, \cdots, T_n)$. Given the observed replication vector $\boldsymbol{s} = (s_1, s_2, \cdots, s_{n-1}, s_n)$ and one of the potential replication vectors $\boldsymbol{t} = (t_1, t_2, \cdots, t_{n-1}, t_n)$, we can define a unique nonnegative integer vector $x = (x_1, x_2, \cdots, x_{n-2}, x_{n-1})$ giving the number of tuples which have been moved from the different positions on the right of the observed replication vector to the left to get the potential true replication vector $\boldsymbol{t}$. Specifically, x_i can be uniquely determined as follows:

$$\begin{cases} x_1 = t_1 - s_1, \\ x_2 = t_2 - s_2 + x_1, \\ x_3 = t_3 - s_3 + x_2, \\ \qquad \cdots \\ x_{n-2} = t_{n-2} - s_{n-2} + x_{n-3}, \\ x_{n-1} = t_{n-1} - s_{n-1} + x_{n-2}, \\ x_{n-1} = s_n - t_n. \end{cases} \tag{3.2.3}$$

Clearly, x_1 is the difference between the new singletons t_1 and the observed singletons s_1, indicating the number of tuples that have been moved from doubletons S_2, tripletons

S_3, $\cdots$ and n-tuples S_n to the original singletons S_1 which gives rise to the new singletons T_1. Similarly, x_2 denotes the number of tuples moved from S_3, S_4, $\cdots$ and S_n to both doubletons S_2 and singletons S_1. Likewise x_3 represents the number of tuples moved from S_4, S_5, $\cdots$ and S_n to the S_1, S_2 and S_3 categories. Since it is impossible that any tuples will be moved into the last element of the vector S_n, x_{n-1} represents the number of tuples moved out of the n-tuples. As the total number of tuples, which is the total number of distinct birth dates in the sample, must remain constant we must have

$$s_1 + s_2 + \cdots + s_n = t_1 + t_2 + \cdots + t_n$$

and given the first $n-2$ equations in (3.2.3) the last two equations in (3.2.3) are equivalent. However, we are not only interested in the number of tuples moving from the right to the left but also in the particular number of the tuples ending up in a p-tuple t_p that come from a certain element s_q (here $q > p$).

Suppose that the non-negative integer $x_{i,j}$ is defined as the number of tuples moving from the observed replication vector element s_{j+i} to the proposed true replication vector t_i, for $1 \leq i \leq n$, $1 \leq j \leq n-i$. For example, $x_{1,1}$ means the number of the singletons t_1 that come from the doubletons s_2 and $x_{2,2}$ means the number of tuples moving from the four-tuples s_4 to the doubletons t_2 in the potential true replication vector. Figure 3.3 demonstrates the idea of the algorithm clearly.

The next stage is to derive the equations expressing x_1, x_2, $\cdots$, x_{n-1} in terms of the $x_{i,j}$. Remember that x_1 is the number of tuples that have been moved from doubletons s_2, tripletons s_3 $\cdots$ and n-tuples s_n to the potential true singletons t_1. Therefore, the equation expressing x_1 in terms of $x_{1,j}$ $(1 \leq j \leq n-1)$ is $\sum_{j=1}^{n-1} x_{1,j} = x_1$ (see Figure 3.3). Similarly x_2 is the total number of the doubletons t_2 coming from the right-hand side s_3, s_4, $\cdots$, s_n is $\sum_{j=1}^{n-2} x_{2,j}$, plus the number of the singletons t_1 coming from s_3, s_4, $\cdots$, s_n,

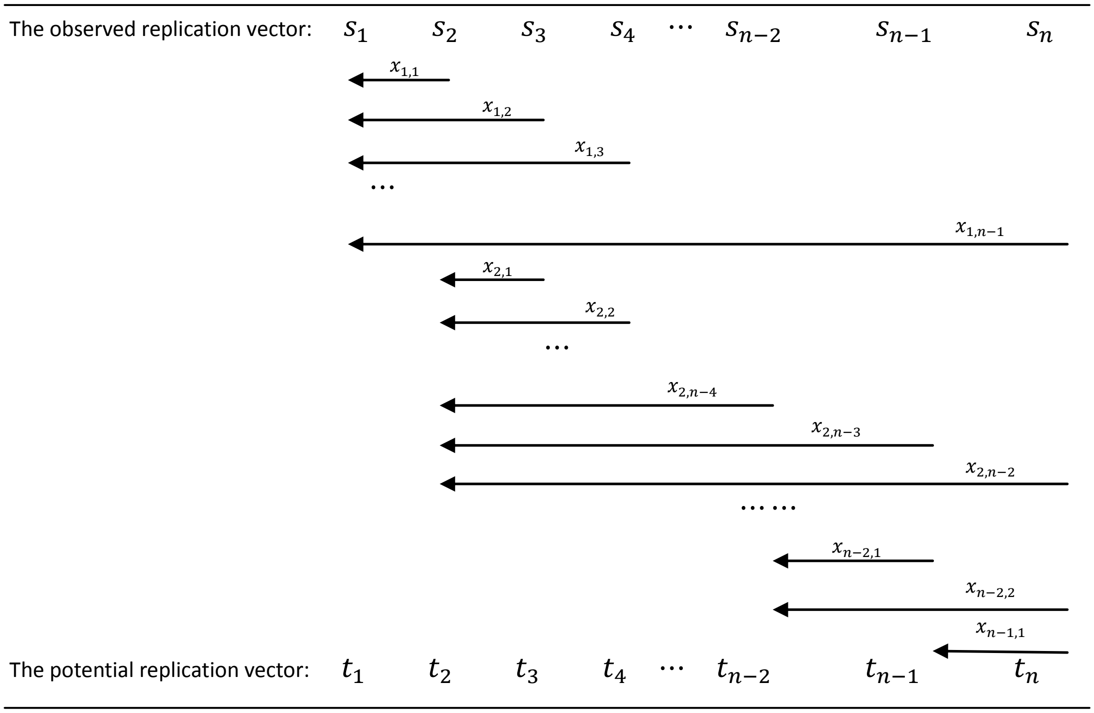


Figure 3.3: The moving pattern of the observed replication vector.

$\sum_{j=2}^{n-1} x_{1,j}$. As a result, we deduce that

$$x_2 = \sum_{j=1}^{n-2} x_{2,j} + \sum_{j=2}^{n-1} x_{1,j}.$$

In general, the set of the equations expressing x_1, x_2, $\cdots$, x_{n-1} in terms of $x_{i,j}$'s are as follows:

$$\begin{cases} x_1 = \sum_{j=1}^{n-1} x_{1,j}, \\ x_2 = \sum_{j=1}^{n-2} x_{2,j} + \sum_{j=2}^{n-1} x_{1,j}, \\ x_3 = \sum_{j=1}^{n-3} x_{3,j} + \sum_{j=2}^{n-2} x_{2,j} + \sum_{j=3}^{n-1} x_{1,j}, \\ x_4 = \sum_{j=1}^{n-4} x_{4,j} + \sum_{j=2}^{n-3} x_{3,j} + \sum_{j=3}^{n-2} x_{2,j} + \sum_{j=4}^{n-1} x_{1,j}, \\ \qquad \cdots\cdots \\ x_{n-2} = \sum_{j=1}^{2} x_{n-2,j} + \sum_{j=2}^{3} x_{n-3,j} + \cdots + \sum_{j=n-3}^{n-2} x_{2,j} + \sum_{j=n-2}^{n-1} x_{1,j}, \\ x_{n-1} = x_{n-1,1} + x_{n-2,2} + \cdots + x_{2,n-2} + x_{1,n-1}. \end{cases} \tag{3.2.4}$$

It is obvious that given x_1, x_2, $\cdots$, x_{n-1} the solution of equations (3.2.4) for the set $x_{i,j}$ is unlikely to be unique. For each possible solution set $x_{i,j}$, the corresponding probability of obtaining the given potential true replication vector can be calculated. Note that for a feasible set of $x_{i,j}$ we must always have

$$\begin{cases} t_1 \geq x_1 = x_{1,1} + x_{1,2} + x_{1,3} + \cdots + x_{1,n-1}, \\ t_2 \geq x_{2,1} + x_{2,2} + \cdots + x_{2,n-2}, \\ t_3 \geq x_{3,1} + x_{3,2} + \cdots + x_{3,n-3}, \\ \qquad \cdots \\ t_{n-2} \geq x_{n-2,1} + x_{n-2,2}, \\ t_{n-1} \geq x_{n-1,1}. \end{cases} \tag{3.2.5}$$

The first of these inequalities is obvious from equation (3.2.3). It also follows from looking at Figure 3.3 as for any feasible set of $x_{i,j}$ the final number of singletons in the proposed true replication vector (t_1) must be at least the sum of all the tuples from $\boldsymbol{s}$ that have moved to t_1 (i.e. $x_{1,1}$ from s_2, $x_{1,2}$ from s_3, $x_{1,3}$ from s_4, $\cdots$, $x_{1,n-1}$ from s_n). Thus

$$t_1 \geq x_{1,1} + x_{1,2} + x_{1,3} + \cdots + x_{1,n-1}.$$

Similarly considering the final number of doubletons in the proposed true replication vector, Figure 3.3 shows that

$$t_2 \geq x_{2,1} + x_{2,2} + x_{2,3} + \cdots + x_{2,n-2}.$$

The remaining inequalities in (3.2.5) follow similarly. Note also that by considering the number of k-tuples in the observed replication vector $\boldsymbol{s}$ we see that for $k = 2, 3, \cdots, r$,

$$s_k \geq \sum_{l=1}^{k-1} x_{l,k-l}. \tag{3.2.6}$$

Considering the general observed replication vector $\boldsymbol{s} = (s_1, s_2, \cdots, s_n)$, we assume that the probability distribution for a randomly chosen individual having a given number of positive HIV tests recorded in the dataset is defined as an unknown probability vector $\boldsymbol{p} = (p_1, p_2, \cdots, p_n)$. In particular, p_i $(1 \leq i \leq n)$ is the probability that an individual has had exactly i positive HIV tests. It is certain that $0 \leq p_i \leq 1$ and $\sum_{i=1}^{n} p_i = 1$. Based on a given observed replication vector $\boldsymbol{s} = (s_1, s_2, \cdots, s_n)$ and one of the derived potential replication vectors $\boldsymbol{t} = (t_1, t_2, \cdots, t_n)$ associated with the non-negative values $x_{i,j}$, the probability that for $j = 1, 2, \cdots n - i$ exactly $x_{i,j}$ i-tuples (in $\boldsymbol{t}$) are observed as $(i + j)$-tuples (in $\boldsymbol{s}$) and everyone else has had exactly one positive HIV test is

$$\frac{t_i!}{(t_i - \sum_{j=1}^{n-i} x_{i,j})! \prod_{j=1}^{n-i} x_{i,j}!} f_{i,0}^{t_i - \sum_{j=1}^{n-i} x_{i,j}} \prod_{j=1}^{n-i} f_{i,j}^{x_{i,j}}.$$

Here $f_{i,j}$ is the probability of the replication needed so that a true i-tuple is observed as an $(i + j)$-tuple in $\boldsymbol{s}$. Note that $t_i \geq \sum_{j=1}^{n-i} x_{i,j}$ by inequalities (3.2.5). For example consider the singletons (t_1) and suppose that $x_{1,3} = x_{1,4} = x_{1,5} = \cdots = x_{1,n-1} = 0$ but $x_{1,1}$ and $x_{1,2}$ are non-zero. Then the probability that exactly $x_{1,1}$ out of t_1 single individuals have had two positive HIV tests and $x_{1,2}$ distinct individuals in the singletons are also overcounted as a tripleton (i.e. $x_{1,2}$ distinct individuals have had exactly three positive HIV tests) is

$$\begin{aligned}\frac{t_1!}{x_{1,1}!(t_1-x_{1,1})!}\frac{(t_1-x_{1,1})!}{x_{1,2}!(t_1-x_{1,1}-x_{1,2})!}f_{1,0}^{t_1-x_{1,1}-x_{1,2}}f_{1,1}^{x_{1,1}}f_{1,2}^{x_{1,2}}\\ =\frac{t_1!}{x_{1,1}!x_{1,2}!(t_1-x_{1,1}-x_{1,2})!}f_{1,0}^{t_1-x_{1,1}-x_{1,2}}f_{1,1}^{x_{1,1}}f_{1,2}^{x_{1,2}}.\end{aligned}$$

Here $f_{1,0}$ is the probability that a proposed true singleton been tested once (i.e. $f_{1,0} = p_1$) and $f_{1,1}$, $f_{1,2}$ are the probabilities that a true singleton is treated by mistake as a doubleton (i.e. $f_{1,1} = p_2$) and a tripleton respectively (i.e. $f_{1,2} = p_3$). As for $f_{2,0}$ which is the probability that a true doubleton corresponds to two single persons having the same birth date and both of them had exactly one positive HIV test, the formula can be written as $f_{2,0} = p_1 \times p_1$. Concerning the definition of $f_{2,1}$ (the probability that a observed tripleton is actually a true doubleton), it represents that exactly one of the two single persons in the doubleton has had a positive HIV test twice. Thus $f_{2,1} = p_1p_2 + p_2p_1 = 2p_1p_2$. $f_{2,2}$ is the probability that a true doubleton is observed as a four-tuple which means that either one of the two distinct individuals in the doubleton took an HIV positive test three times or both two persons in the doubleton took an HIV test exactly twice each. Hence, it leads to the formula of $f_{2,2}$ that is $f_{2,2} = p_1p_3 + p_3p_1 + p_2p_2 = 2p_1p_3 + p_2^2$. Similar arguments show that

$$\begin{aligned}&f_{1,0} = p_1, f_{1,1} = p_2, f_{1,2} = p_3, f_{1,3} = p_4,\\ &f_{2,0} = p_1^2, f_{2,1} = (p_1p_2 + p_2p_1), f_{2,2} = (p_1p_3 + p_3p_1 + p_2p_2),\\ &f_{3,0} = p_1^3, f_{3,1} = (p_1p_1p_2 + p_1p_2p_1 + p_2p_1p_1),\end{aligned}$$

and

$$f_{4,0} = p_1^4.$$

Note that $f_{i,j} = \sum_{\boldsymbol{\xi}} p_{\xi_1}p_{\xi_2}\cdots p_{\xi_i}$ where the sum is over all $\boldsymbol{\xi} = (\xi_1, \xi_2, \cdots, \xi_i)$ such that $\xi_1+\xi_2+\xi_3+\cdots+\xi_i = i+j$. So $f_{i,j}$ is the sum of all products of i p's whose subscripts sum to $i+j$.

In practice we can use the following lemma to calculate the $f_{i,j}$ by the mathematical induction.

Lemma 3.3 *For $i \geq 2$,*

$$f_{i,j} = p_1 f_{i-1,j} + p_2 f_{i-1,j-1} + p_3 f_{i-1,j-2} + \cdots + p_{j+1} f_{i-1,0}.$$

Proof. The result is clearly true for $i = 2$. Assume that it is true for $i-1$ then consider

$$f_{i,j} = \sum_{\boldsymbol{\xi}} p_{\xi_1} p_{\xi_2} \cdots p_{\xi_i}$$

where the sum is over all $\boldsymbol{\xi} = (\xi_1, \xi_2, \cdots, \xi_i)$ such that $\xi_1+\xi_2+\xi_3+\cdots+\xi_i = i+j$. Thus

$$\begin{aligned} f_{i,j} = p_1 & \sum_{\boldsymbol{\xi}_1 \in \Omega_1} p_{\xi_{1,1}} p_{\xi_{1,2}} \cdots p_{\xi_{1,i-1}} \\ + p_2 & \sum_{\boldsymbol{\xi}_2 \in \Omega_2} p_{\xi_{2,1}} p_{\xi_{2,2}} \cdots p_{\xi_{2,i-1}} \\ + p_3 & \sum_{\boldsymbol{\xi}_3 \in \Omega_3} p_{\xi_{3,1}} p_{\xi_{3,2}} \cdots p_{\xi_{3,i-1}} \\ + \cdots & \\ + p_{j+1} & \sum_{\boldsymbol{\xi}_{j+1} \in \Omega_{j+1}} p_{\xi_{j+1,1}} p_{\xi_{j+1,2}} \cdots p_{\xi_{j+1,i-1}} \end{aligned}$$

where $\Omega_1 = \{\boldsymbol{\xi}_1 = (\xi_{1,1}, \xi_{1,2}, \cdots, \xi_{1,i-1})$ such that $\xi_{1,1}+\xi_{1,2}+\cdots+\xi_{1,i-1} = i+j-1\}$, the second sum is over all $\Omega_2 = \{\boldsymbol{\xi}_2 = (\xi_{2,1}, \xi_{2,2}, \cdots, \xi_{2,i-1})$ such that $\xi_{2,1}+\xi_{2,2}+\cdots+\xi_{2,i-1} = i+j-2\}$, the third sum is over all $\Omega_3 = \{\boldsymbol{\xi}_3 = (\xi_{3,1}, \xi_{3,2}, \cdots, \xi_{3,i-1})$ such that $\xi_{3,1}+\xi_{3,2}+\cdots+\xi_{3,i-1} = i+j-3\}$, $\cdots$ and the last one over $\Omega_{j+1} = \{\boldsymbol{\xi}_{j+1} = (\xi_{j+1,1}, \xi_{j+1,2}, \cdots, \xi_{j+1,i-1})$ such that $\xi_{j+1,1}+\xi_{j+1,2}+\cdots+\xi_{j+1,i-1} = i-1\}$. Hence we can get

$$f_{i,j} = p_1 f_{i-1,j} + p_2 f_{i-1,j-1} + p_3 f_{i-1,j-2} + \cdots + p_{j+1} f_{i-1,0}.$$

The result of Lemma 3.3 follows. □

According to these quantities, the likelihood function for a true replication vector $\boldsymbol{t}$ given the true number of distinct individuals $\bar{r}$ and the unknown probability distribution $\boldsymbol{p}$ can be constructed.

$$
\begin{aligned}
L(t|\bar{r}, \boldsymbol{p}) = \sum_{\boldsymbol{x}} & \frac{t_1!}{x_{1,1}!x_{1,2}! \cdots x_{1,n-1}!(t_1 - x_{1,1} - x_{1,2} - \cdots - x_{1,n-1})!} \\
& \times \frac{t_2!}{x_{2,1}!x_{2,2}! \cdots x_{2,n-2}!(t_2 - x_{2,1} - x_{2,2} - \cdots - x_{2,n-2})!} \\
& \times \cdots \times \frac{t_n!}{x_{n-1,1}!(t_n - x_{n-1,1})!} \\
& \times f_{1,0}^{t_1 - x_{1,1} - x_{1,2} - \cdots - x_{1,n-1}} f_{1,1}^{x_{1,1}} f_{1,2}^{x_{1,2}} \cdots f_{1,n-1}^{x_{1,n-1}} \\
& \times f_{2,0}^{t_2 - x_{2,1} - x_{2,2} - \cdots - x_{2,n-2}} f_{2,1}^{x_{2,1}} f_{2,2}^{x_{2,2}} \cdots f_{2,n-2}^{x_{2,n-2}} \\
& \times f_{3,0}^{t_3 - x_{3,1} - x_{3,2} - \cdots - x_{3,n-3}} f_{3,1}^{x_{3,1}} f_{3,2}^{x_{3,2}} \cdots f_{3,n-3}^{x_{3,n-3}} \\
& \times \cdots \times f_{n-1,0}^{t_{n-1} - x_{n-1,1}} f_{n-1,1}^{x_{n-1,1}} \times f_{n,0}^{t_n} \times P
\end{aligned}
\tag{3.2.7}
$$

where

$$
\begin{aligned}
f_{1,j} &= p_{j+1}, \qquad 0 \leq j \leq n-1, \\
f_{i,j} &= p_1 f_{i-1,j} + p_2 f_{i-1,j-1} + p_3 f_{i-1,j-2} + \cdots + p_{j+1} f_{i-1,0},
\end{aligned}
$$

and

$$
P = Pr(S_1 = t_1, S_2 = t_2, \cdots, S_n = t_n).
$$

Note that given the observed replication vector $\boldsymbol{s}$ and the proposed replication vector $\boldsymbol{t}$ we can calculate the integer vector $(x_1, x_2, \cdots, x_{n-2}, x_{n-1})$ from equations (3.2.3). Then in equation (3.2.7) the sum is taken over the set of values $x_{i,j}$ where $1 \leq i \leq n-1$, $1 \leq j \leq n-i$. The $x_{i,j}$'s can be calculated from equations (3.2.4) constrained by the inequalities

(3.2.6). Note that $\boldsymbol{t}$ then automatically satisfies the equations (3.2.5). Furthermore, remember that $x_{i,j}$ is the number of tuples moving from $(i+j)$-tuples to i-tuples and $f_{i,j}$ is the probability of the replication needed so that a true i-tuple is observed as an $(i+j)$-tuple. Hence assuming that individuals taking the HIV tests are independent, it is obvious that $f_{i,j}^{x_{i,j}}$ is the probability of the replication needed so that $x_{i,j}$ true i-tuples are observed as $(i+j)$-tuples. Note that for the t_i postulated true i-tuples, $t_i - \sum_{j=1}^{n-i} x_{i,j}$ of them must correspond to i-tuples in the observed replication vector $\mathbf{s}$ which have no repeated records, i.e. each of these i-tuples consists of i distinct individuals who have had exactly one HIV test. There are $i \times (t_i - \sum_{j=1}^{n-i} x_{i,j})$ of these individuals who had exactly one HIV test in total. Thus, $f_{i,0}^{t_i - \sum_{j=1}^{n-i} x_{i,j}}$ is the probability that these $i \times (t_i - \sum_{j=1}^{n-i} x_{i,j})$ individuals have each had exactly one HIV test (since $f_{i,0} = p_1^i$, clearly $f_{i,0}^{t_i - \sum_{j=1}^{n-i} x_{i,j}} = p_1^{i\times(t_i - \sum_{j=1}^{n-i} x_{i,j})}$). From (3.2.7) we can see that the likelihood function can be expressed as a sum of terms where each term is a product of powers p_i for $1 \leq i \leq n$ and the sum of the powers of p_i is always equal to the true sample size $\bar{r}$. In other words, we can assume that after simplification the likelihood function $L(t|\bar{r}, \boldsymbol{p})$ is a sum of terms such as

$$Cp_1^{\tau_1} p_2^{\tau_2} \cdots p_k^{\tau_k}$$

where C is a constant, $\boldsymbol{p} = (p_1, p_2, \cdots, p_n)$ and $1 \leq k \leq n$ and it is always true that $\sum_{i=1}^{k} \tau_i = \bar{r}$.

It should be pointed out that the likelihood function (3.2.7) is also suitable for the observed replication vector $\boldsymbol{s} = (s_1, s_2, \cdots, s_n)$. Because there is no movement of the tuples, the values of $x_{i,j}$ are all zeros, which makes the likelihood function for $\boldsymbol{s}$ become

$$L(\boldsymbol{s}|r_{obs}, \boldsymbol{p}) = f_{1,0}^{s_1} f_{2,0}^{s_2} \cdots f_{n,0}^{s_n} P = p_1^{\sum_{i=1}^{n} s_i} P = p_1^{r_{obs}} Pr(S_1 = s_1, \cdots, S_n = s_n). \tag{3.2.8}$$

As we mentioned before, it is likely that a given postulated true sample size may correspond to more than one potential true replication vector. Since the objective is to estimate the amount of replication accurately in a given birth year, the likelihood functions should include the probabilities of obtaining all the potential replication vectors for each of the true postulated sample sizes. From a mathematical point of view, the likelihood function for a given postulated true sample size $\bar{r}$ associated with the corresponding potential true replication vectors $\boldsymbol{t}^{(1)}, \boldsymbol{t}^{(2)}, \cdots, \boldsymbol{t}^{(k)}$ is

$$L(\boldsymbol{t}_{birthyear}|\bar{r}, \boldsymbol{p}) = L(\boldsymbol{t}^{(1)}|\bar{r}, \boldsymbol{p}) + L(\boldsymbol{t}^{(2)}|\bar{r}, \boldsymbol{p}) + \cdots + L(\boldsymbol{t}^{(k)}|\bar{r}, \boldsymbol{p}).$$

3.3 An example of calculating the likelihood function.

Take the observed replication vector (28, 1, 1) with the birth year 1925 in the 1994 dataset as an example to construct the likelihood functions. Firstly, we calculate the potential true number of distinct individuals as well as the observed sample size.

$$r_{obs} = 28 + 2 \times 1 + 3 \times 1 = 33; \bar{r}_{min} = 28 + 1 + 1 = 30.$$

Hence there are four potential true number of distinct individual records. In other words, except for the observed sample size 33 and minimum potential sample size 30 there are two other potential true sample sizes which are $\bar{r}_1 = r_{obs} - 1 = 32$ and $\bar{r}_2 = r_{obs} - 2 = 31$. Next, the corresponding potential true replication vectors should be derived. Based on the original data (28, 1, 1), there are two outcomes with the postulated sample size 32. If the tripleton is actually a doubleton indicating one replication exists in the observed sample, the new replication vector becomes (28, 2, 0). The other result is (29, 0, 1) showing that one of the twenty-nine distinct individuals in the singleton is observed as a doubleton. After getting rid of the repeated replication vectors derived from (28, 2, 0) and (29, 0, 1), there is only one potential replication vector with sample size 31 which is (29, 1, 0). Finally, it is straightforward to obtain the last potential replication vector for the minimal possible true sample size 30 (i.e. moving the doubleton to the singleton, giving the vector (30, 0, 0)). The results are shown clearly in table 3.2:

Table 3.2: The result for the data in the birth year 1931.

Potential Sample Size $\bar{r}_i$	Potential Replication Vectors
33	$(28, 1, 1)$
32	$(28, 2, 0); (29, 0, 1)$
31	$(29, 1, 0)$
30	$(30, 0, 0)$

Next we are going to calculate the probabilities of the potential replication vectors respectively:

$$\begin{aligned}
P(\boldsymbol{S} = (28, 1, 1)) &= exp(\log(365!) + \log(33!) - (\log(28!) + \log((365 - 28 - 1 - 1)!) \\
&\quad + \log(2!) + \log(3!) + 33 * \log(365))) = exp(-4.245127) = 0.01433391; \\
P(\boldsymbol{S} = (28, 2, 0)) &= exp(-1.436272) = 0.2378128; \\
P(\boldsymbol{S} = (29, 0, 1)) &= exp(-4.515885) = 0.01093392; \\
P(\boldsymbol{S} = (29, 1, 0)) &= exp(-0.9831117) = 0.3741451; \\
P(\boldsymbol{S} = (30, 0, 0)) &= exp(-1.225252) = 0.2936838.
\end{aligned}$$

For each potential true sample size, a likelihood function containing the probability of getting all the potential replication vectors with that sample size given the unknown parameter $\boldsymbol{p}$ is able to be derived. Define the non-negative integer vectors $\boldsymbol{x} = (x_1, x_2)$ and $\tilde{\boldsymbol{x}} = (x_{1,1}, x_{1,2}, x_{2,1})$. The equation sets with regard to $\boldsymbol{x}$ and $\tilde{\boldsymbol{x}}$ can be derived according to the methods introduced in Section 3.2. For the replication vector (28, 2, 0) denoted as $t^{(1)}$, we can build the equations of x_i and $x_{i,j}$ $(1 \leq i \leq 2, 1 \leq j \leq 2)$ as follows:

$$\begin{cases}
x_1^{(1)} = t_1^{(1)} - s_1 = 28 - 28, \\
x_2^{(1)} = t_2^{(1)} - s_2 + x_1^{(1)} = 2 - 1 + x_1^{(1)}, \\
x_{1,1}^{(1)} + x_{1,2}^{(1)} = x_1^{(1)}, \\
x_{1,2}^{(1)} + x_{2,1}^{(1)} = x_2^{(1)}.
\end{cases}$$

The results can be obtained as

$$\begin{cases} x_1^{(1)} = 0, \\ x_2^{(1)} = 1, \\ x_{1,1}^{(1)} = x_{1,2}^{(1)} = 0; x_{2,1}^{(1)} = 1. \end{cases}$$

Due to the non-negativity of the integers $x_{1,1}^{(1)}$, $x_{1,2}^{(1)}$ and $x_{2,1}^{(1)}$, there is only one set of solutions valid for the equation set. That is one doubleton in the potential true replication vector (28, 2, 0) is observed as a tripleton in the original database. Consequently, the likelihood function for the potential replication vector (28, 2, 0) given the true sample size $\bar{r}$ and unknown parameter $\boldsymbol{p}$ is

$$\begin{aligned} L(\boldsymbol{t}^{(1)} = (28, 2, 0)|\bar{r} = 32, \boldsymbol{p}) &= \frac{28!}{0!0!(28-0)!}\frac{2!}{1!(2-1)!}p_1^{28}(p_1^2)^{(2-1)}(2p_1p_2)(p_1^3)^0 Pr(\boldsymbol{S} = \boldsymbol{t}^{(2)}) \\ &= 4p_1^{31}p_2 0.2378128 = 0.9512512p_1^{31}p_2. \end{aligned}$$

Specifically, the coefficient $\frac{28!}{0!0!(28-0)!}\frac{2!}{1!(2-1)!}$ indicates the possible number of ways that one doubleton can be observed as a tripleton. The probability of the repeated HIV tests needed so that a true doubleton is observed as a tripleton is $f_{2,1}^1 = (2p_1p_2)^1$. Thus, the corresponding probability of one repeated HIV test for an individual whose birth date is in a doubleton (i.e. there are 31 persons having exactly one positive HIV test and one person in a doubleton being tested twice) is $\frac{28!}{0!0!(28-0)!}\frac{2!}{1!(2-1)!}2p_1^{31}p_2$. Moreover, it is obvious that sum of the powers of p_1, p_2 and p_3 equals the true sample size of this potential replication vector (here $31 + 1 + 0 = \bar{r} = 32$).

For another potential replication vector $\boldsymbol{t}^{(2)} = (29, 0, 1)$ with the same true sample size 32, we use the same method to generate the likelihood function.

$$\begin{cases} x_1^{(2)} = t_1^{(2)} - s_1 = 29 - 28, \\ x_2^{(2)} = t_2^{(2)} - s_2 + x_1^{(2)} = 0 - 1 + x_1^{(2)}, \\ x_{1,1}^{(2)} + x_{1,2}^{(2)} = x_1^{(2)}, \\ x_{1,2}^{(2)} + x_{2,1}^{(2)} = x_2^{(2)}. \end{cases}$$

It gives the results that

$$\begin{cases} x_1^{(2)} = 1, \\ x_2^{(2)} = 0, \\ x_{1,1}^{(2)} = 1; x_{1,2}^{(1)} = x_{2,1}^{(1)} = 0. \end{cases}$$

As a result, the likelihood function for the potential replication vector $\boldsymbol{t}^{(2)}$ is

$$\begin{aligned} &L(\boldsymbol{t}^{(2)} = (29,0,1)|\bar{r} = 32, \boldsymbol{p}) \\ =&\frac{29!}{1!0!(29-1)!}\frac{0!}{0!(0-0)!} \times p_1^{(29-1)} p_2 (p_1^2)^0 (2p_1p_2)^0 (p_1^3)^1 Pr(\boldsymbol{S} = t^{(2)}), \\ =&29p_1^{31}p_2 0.01093392 = 0.3170837 p_1^{31} p_2. \end{aligned}$$

Now we are able to generate the likelihood function of obtaining the potential replication vectors with the true sample size $\bar{r} = 32$ which is

$$\begin{aligned} L(\boldsymbol{t}_{1931}^{(2)}|\bar{r} = 32, \boldsymbol{p}) &= L(\boldsymbol{t}^{(1)}|\bar{r} = 32, \boldsymbol{p}) + L(\boldsymbol{t}^{(2)}|\bar{r} = 32, \boldsymbol{p}), \\ &= 0.9512512 p_1^{31} p_2 + 0.3170837 p_1^{31} p_2 = 1.268335 p_1^{31} p_2. \end{aligned}$$

Clearly, the sum of the powers of p_1 and p_2 is 32 which is exactly the same as the true sample size $\bar{r}$.

Similarly, we can obtain the likelihood function for the true sample size 31 by applying the same procedure as before. Table 3.2 shows that only one potential replication vector (29,1,0), was derived with the true sample size 31. Define the non-negative integers $\dot{x}_1$, $\dot{x}_2$ and $\dot{x}_{1,1}$, $\dot{x}_{1,2}$, $\dot{x}_{2,1}$. We have

$$\begin{cases} \dot{x}_1 = 29 - 28, \\ \dot{x}_2 = 1 - 1 + \dot{x}_1, \\ \dot{x}_{1,1} + \dot{x}_{1,2} = \dot{x}_1, \\ \dot{x}_{1,2} + \dot{x}_{2,1} = \dot{x}_2, \end{cases}$$

with the two possible sets of answers

$$\begin{cases} \dot{x}_1 = 1, \\ \dot{x}_2 = 1, \\ \dot{x}_{1,1} = 1, \dot{x}_{1,2} = 0, \dot{x}_{2,1} = 1, \end{cases}$$

or

$$\begin{cases} \dot{x}_1 = 1, \\ \dot{x}_2 = 1, \\ \dot{x}_{1,1} = 0, \dot{x}_{1,2} = 1, \dot{x}_{2,1} = 0. \end{cases}$$

Therefore, for the replication vector (29,1,0), the likelihood function is

$$\begin{aligned} L(\boldsymbol{t}^{(3)}_{1931}|\bar{r} = 31, \boldsymbol{p}) &= \Bigg[\frac{29!}{1!0!(29-1-0)!}\frac{1!}{1!(1-1)!}p_1^{(29-1-0)}p_2(p_1^2)^{1-1}(2p_1p_2)^1 \\ &+ \frac{29!}{0!1!(29-0-1)!}\frac{1!}{0!(1-0)!}p_1^{(29-0-1)}p_2^0p_3(p_1^2)^{1-0}(2p_1p_2)^0\Bigg] \\ &\times Pr(\boldsymbol{S} = (29,1,0)), \\ &= [29 \times 2p_1^{(28+1)}p_2^{(1+1)} + 29p_1^{(28+2)}p_3] \times 0.3741451, \\ &= 21.7004158p_1^{29}p_2^2 + 10.8502079p_1^{30}p_3. \end{aligned}$$

Considering the sums of the powers of the parameters p_i, we detect that in both terms the sums are equal to 31 which is the postulated true sample size here. Specifically in the first term, the power of p_1 is 29 and the power of p_2 is 2 giving a total of 31. Similarly in the second term of the likelihood function, the powers of p_1 and p_3 are 30 and 1 respectively whose sum is 31 as well.

In order to get the likelihood function for the minimal possible true sample size $\bar{r} = 30$ with the corresponding potential replication vector (30,0,0), the non-negative integers defined as $\ddot{x}_1$, $\ddot{x}_2$ and $\ddot{x}_{1,1}$, $\ddot{x}_{1,2}$, $\ddot{x}_{2,1}$ have to be calculated. According to the equation sets (3.2.3) and (3.2.4), we have

$$\begin{cases} \ddot{x}_1 = 30 - 28, \\ \ddot{x}_2 = 0 - 1 + \ddot{x}_1, \\ \ddot{x}_{1,1} + \ddot{x}_{1,2} = \ddot{x}_1, \\ \ddot{x}_{1,2} + \ddot{x}_{2,1} = \ddot{x}_2, \end{cases}$$

we can get

$$\begin{cases} \ddot{x}_1 = 2, \\ \ddot{x}_2 = 1, \\ \ddot{x}_{1,1} = \ddot{x}_{1,2} = 1, \dot{x}_{2,1} = 0, \end{cases} \tag{3.3.1}$$

or

$$\begin{cases} \ddot{x}_1 = 2, \\ \ddot{x}_2 = 1, \\ \ddot{x}_{1,1} = 2, \ddot{x}_{1,2} = 0, \dot{x}_{2,1} = 1. \end{cases} \tag{3.3.2}$$

However, the answer set (3.3.2) does not satisfy the condition that $s_k \geq \sum_{l=1}^{k-1} x_{l,k-l}$ (see (3.2.6)) Therefore it should be deleted so the corresponding likelihood function is

$$\begin{aligned} L(\boldsymbol{t}_{1931}^{(4)}|\bar{r} = 30, \boldsymbol{p}) = & \frac{30!}{1!1!(30-1-1)!} \frac{0!}{0!(0-0)!} p_1^{(30-1-1)} p_2 p_3 Pr(\boldsymbol{S} = (30,0,0)), \\ & = 30 \times 29 p_1^{28} p_2 p_3 \times 0.2936838 = 255.504906 p_1^{28} p_2 p_3. \end{aligned}$$

As for the observed replication vector (28,1,1) with the sample size 33, it is straightforward to get the likelihood function based on the formula (3.2.8) which is $L(\boldsymbol{t}_{1931}^{(1)}|\bar{r} = 33, \boldsymbol{p}) = p_1^{33} Pr(\boldsymbol{S} = (28,1,1)) = 0.01433391 p_1^{33}$.

In these simple examples it was straightforward to calculate the likelihood function by hand. However in real data the number of observed birth records and individuals in a birth year becomes very large. Also the number of distinct records which correspond to the same birth date may be as high as eleven. Consequently the combinations of the possibilities for non-negative integers x_i and $x_{i,j}$ that must be taken into account during the construction of the likelihood function will increase dramatically. It is then impossible to calculate the likelihood function by hand so we have devised a computer algorithm written in R (and C) that can calculate the likelihood functions based on the method introduced in Section 3.2.

3.4 Maximum likelihood estimates.

Based on the likelihood function of the replication vectors given the true number of distinct individuals $\bar{r}$ and unknown parameter vector $\boldsymbol{p}$, we aim to calculate the maximum likelihood estimate $\hat{\boldsymbol{p}}$ which is the estimated probability vector that an individual has a given number of HIV tests. Furthermore, the corresponding true sample size $\bar{r}$ denoted by $\hat{r}$ is considered as the maximum likelihood estimate for the true number of distinct individuals. As we know, with a given birth year there is usually more than one likelihood function corresponding to different possible true sample sizes. Since each likelihood function with the corresponding postulated true sample size $\bar{r}$ leads to its own maximum likelihood estimate, the overall maximum likelihood estimate for the given birth year should be chosen as the one making the corresponding likelihood function be the largest among all the maximum values of different likelihood functions given the relative possible true sample sizes. From a mathematical point of view, the maximum likelihood estimate for a given birth year can be presented as $MLE = (\hat{\boldsymbol{p}}, \hat{r})$ such that the corresponding likelihood function $L_{MLE} = \max\{L_i(\hat{\boldsymbol{p}}_i, \bar{r}_i), 0 \leq i \leq \bar{r}_{max} - \bar{r}_{min}\}$, where L_i is the likelihood function for the derived true sample size $\bar{r}_i$ ($\bar{r}_0$ is the observed sample size) and $\hat{\boldsymbol{p}}_i$ is the maximum likelihood estimate derived from L_i.

Commonly, the likelihood function is a nonlinear polynomial function with the constraint of the known parameter $\boldsymbol{p}$ that $\sum_{j=1}^{n} p_j = 1$ and $0 \leq p_j \leq 1$. Lemma 3.4 which is taken from a study addressed by Greenhalgh, Doyle and Mortimer [52] explains a method to calculate the maximum likelihood function in a comparatively simple circumstance.

Lemma 3.4 *Suppose that $m \geq 1$ and k_1, k_2, $\cdots$, k_m are strictly positive real numbers. Then $p_1^{k_1} p_2^{k_2} \cdots p_m^{k_m}$ is maximised over $p_i \geq 0, 1 \leq i \leq m, p_1 + p_2 + \cdots + p_m = 1$ at $\hat{p}_i = \frac{k_i}{\sum_{j=1}^m k_j}$ when its value is $\frac{k_1^{k_1} k_2^{k_2} \cdots k_m^{k_m}}{(\sum_{j=1}^m k_j)^{(\sum_{j=1}^m k_j)}}$.*

For example, based on the likelihood functions we have obtained in Section 3.3 (shown in Table 3.3) the maximum likelihood estimate for the birth year 1925 can be calculated by applying Lemma 3.4.

Table 3.3: Likelihood functions for the birth year 1925.

True Sample Size $\bar{r}_i$	Likelihood Function L_i
33	$0.01433391 p_1^{33}$
32	$1.268335 p_1^{31} p_2$
31	$21.7004158 p_1^{29} p_2^2 + 10.8502079 p_1^{30} p_3$
30	$255.504906 p_1^{28} p_2 p_3$

Case (i). When $\bar{r}_0 = 33$ and the likelihood function is $L_0 = 0.01433391 p_1^{33}$, the maximum likelihood estimate $\hat{\boldsymbol{p}}_0$ is

$$p_1 = 1, p_2 = 0, p_3 = 0$$

i.e. $\hat{\boldsymbol{p}}_0 = (1, 0, 0)$ and the likelihood function with the observed sample size 33 becomes

$$L_0 = 0.01433391 \times 1^{33} = 0.01433391.$$

Case (ii). Based on Lemma 3.4, for $\bar{r}_1 = 32$ the likelihood function $L_1 = 1.268335 p_1^{31} p_2$ can be maximised at

$$p_1 = \frac{31}{31+1} = \frac{31}{32} = 0.96875, p_2 = \frac{1}{31+1} = \frac{1}{32} = 0.03125.$$

Equivalently, the maximum likelihood estimate $\hat{\boldsymbol{p}}_1 = (0.96875, 0.03125, 0)$ with $\hat{r}_1 = 32$ gives the maximum value for L_1 which is 0.014813 (=$1.268335 \times 0.96875^{31} \times 0.03125$).

Case (iii). With $\bar{r}_2 = 31$, the likelihood function is $L_2 = 21.7004158 p_1^{29} p_2^2 + 10.8502079 p_1^{30} p_3$. Since $p_1 + p_2 + p_3 = 1$, clearly p_2 can be expressed in terms of

p_1 and p_3 ($p_2 = 1 - p_1 - p_3$). Thus the likelihood function in this case becomes $L_2 = 21.7004158 p_1^{29}(1-p_1-p_3)^2 + 10.8502079 p_1^{30} p_3$ with the constraints of $0 \leq p_1 + p_3 \leq 1$ and $p_1, p_3 \in [0,1]$. In order to maximise L_2 subject to $0 \leq p_1 + p_3 \leq 1$, we define a Lagrangian $\mathsf{L} = L_2 - \lambda(p_1 + p_3 - 1)$ where complementary slackness gives $\lambda(p_1+p_3-1) = 0$ and $\lambda \geq 0$. By differentiating the Lagrangian with respect to p_1 and p_3 respectively, we can get

$$\begin{aligned}\mathsf{L}_{p_1} &= 21.7004158 * 29 * p_1^{28}(1-p_1-p_3)^2 - 21.7004158 * 2 * p_1^{29}(1-p_1-p_3) \\ &\quad + 10.8502079 * 30 * p_1^{29} p_3 - \lambda,\end{aligned}$$

and

$$\mathsf{L}_{p_3} = -21.7004158 * 2 * p_1^{29}(1-p_1-p_3) + 10.8502079 * p_1^{30} - \lambda.$$

Then we can neatly capture the results by writing

$$\mathsf{L}_{p_1} = 0, \quad \mathsf{L}_{p_3} = 0 \quad \text{and} \quad \lambda(p_1 + p_3 - 1) = 0.$$

From the equality $\lambda(p_1 + p_3 - 1) = 0$ it is clear that either (i) $p_1 + p_3 = 1$ or (ii) $\lambda = 0$. In case (i) if $p_1 + p_3 = 1$ then the solution satisfying these constraints and inequalities is $\lambda = 10.8502079 \times (30/31)^{30}$, $p_1 = 30/31$ and $p_3 = 1/31$. In this solution, the constraint is binding because $\lambda \neq 0$, and so the constraint $p_1 + p_3 = 1$. By using the bordered Hessian method (since the constraint is binding), the solution gives a local maximum. In case (ii) if $\lambda = 0$ and $p_1 = 0$ then the corresponding likelihood function L_2 becomes zero which is not a maximum. Hence $p_1 = 0$ is not part of a maximum likelihood estimate. If $\lambda = 0$ and $p_1 \neq 0$ then the solution of the equalities becomes $p_1 = 0.8602$ and $p_3 = -0.0753$ which is not a feasible solution. Hence, the maximum likelihood estimate for L_2 subject to the constraint $p_1 + p_2 + p_3 = 1$ ($0 \leq p_1, p_2, p_3 \leq 1$) is $\hat{\boldsymbol{p}}_2 = (0.96774, 0, 0.03226)$.

Case (iv). Similarly, the likelihood function with the proposed true sample size $\bar{r}_3 = 30$

has a maximum value of 0.04112 at

$$\begin{aligned} p_1 &= \frac{28}{28+1+1} = 0.93333, \\ p_2 &= \frac{1}{28+1+1} = 0.03333, \\ p_3 &= 1 - p_1 - p_2 = \frac{1}{28+1+1} = 0.03334. \end{aligned}$$

The results are summarised in Table 3.4 shown below.

Table 3.4: Maximum likelihood estimates for the birth year 1925.

True Sample Size $\bar{r}_i$	Maximum Likelihood Estimates $\hat{\boldsymbol{p}}_i$	Maximum Values of Likelihood Function L_i
33	(1, 0, 0)	0.014334
32	(0.96875, 0.03125, 0)	0.014813
31	(0.96774, 0, 0.03226)	0.13088
30	(0.93333, 0.03333, 0.03334)	0.04112

Comparing the maximum values of the likelihood functions in the four cases (see Table 3.4, we find that the likelihood function for the true sample size 31 has the largest maximum value, which implies that the maximum likelihood estimate for the birth year 1925 is $\hat{\boldsymbol{p}} = (0.96774, 0, 0.03226)$ with the estimated true sample size 31. In other words, we can conclude that for the birth year 1925 it is most likely that there were thirty-one individuals one of whom was recorded three times due to having had three HIV tests.

However, in practice the likelihood function could be much more complicated due to the large number of observations. Under these circumstances, Lemma 3.4 is no longer suitable for maximising the likelihood function. As a matter of fact, we can obtain the maximum likelihood estimates $\hat{r}$ and $\hat{\boldsymbol{p}}$ by using the constrained nonlinear optimisation package 'alabama' in R written by Varadhan and Grothendieck [129]. The principle of the package 'alabama' is the Augmented Lagrangian and Adaptive Barrier Minimisation Algorithm for optimising smooth nonlinear objective functions with constraints. The optimisation procedure begins with a starting point that is defined in a feasible region and a barrier is added to enforce the constraints by specifying the parameters. The algorithm proceeds by minimising the objective function over all the values of $\hat{r}$ and $\hat{\boldsymbol{p}}$. During the application of the 'alabama' package, a simple mathematical transformation of the likelihood function is used to increase the precision of the maximum calculation because of two reasons. One of the reasons is that by default the package will minimise a constrained nonlinear function. However we aim to maxmise the likelihood function subject to constraints, so we can do this by minimising the negative of this function. The second problem is that the likelihood function is usually quite small in practice which may cause the problem of obtaining accurate answers from the calculation in R due to numerical approximation difficulties. Therefore, we aim to minimise a new function when applying 'alabama' to do the analysis of the real dataset, which is $\tilde{L} = -log(L)$ where L is the original likelihood function we obtained by using the procedure introduced in Section 3.2. The results obtained from the calculations of the 'alabama' package are extremely close to the theoretical ones. For example, with the birth year 1925 the estimated parameter $\hat{\boldsymbol{p}}$ in R is (0.96777, 0, 0.03223) with $\hat{r} = 31$ and the likelihood function 0.13088. Compared with the theoretical estimates, we can conclude that the R estimation procedure is very accurate.

3.5 The calculation of the amount of replication.

Once the maximum likelihood estimates are obtained, we are able to calculate the amount of replication presented in the HIV dataset so that statistical inferences can be made on the amount of the replication of individuals. The amount of replication denoted as $r\hat{e}p$ is presented as a percentage. It can be estimated as

$$\frac{r - \hat{r}}{\hat{r}} \tag{3.5.1}$$

where r is the observed sample size and $\hat{r}$ is the estimated true sample size. Therefore, the estimated amount of replication is guaranteed to be non-negative since the observed sample size r is the maximum potential true sample size and $\hat{r} \leq r$ is always true.

Take the replication vector with the birth year 1925 as an example, the estimated true sample size is 31 with the estimated associated probability vector (0.96777, 0, 0.03223) according to the R calculation in Section 3.4. Hence based on the observed number of distinct individuals 33 the estimated amount of replication for the birth year 1925 is

$$r\hat{e}p_{1925} = \frac{33-31}{31} = \frac{2}{31} = 6.45\%.$$

Furthermore, there is an alternative method to calculate the amount of replication. According to the definition of the probability parameter p_i (i.e. p_i is the probability that an individual has had exactly i positive HIV tests), it is clear that this individual has had i HIV tests and $i-1$ of these are repeat tests. Hence the contribution to the overall amount of replication in the dataset from the ith value p_i of the probability vector $\boldsymbol{p}$ is $(i-1)p_i$. For example, p_2 is the probability that an individual had two HIV tests. It illustrates that there is one replication recorded in the dataset since an individual was actually considered as two distinct ones due to the repeated report. Hence the amount of replication from the second component p_2 of the probability vector $\boldsymbol{p}$ is $(2-1)p_2$ expressed as a percentage. Similarly, the definition of p_3 which is the probability that an individual had HIV test three times means that one individual was reported as three distinct persons. Thus there are two overcounted records in three postulated persons giving that the amount of replication from the third component p_3 of the probability vector $\boldsymbol{p}$ is $(3-1)p_3$. Therefore, the amount of the replication for a birth year is

$$(2-1)p_2 + (3-1)p_3 + (4-1)p_4 + \cdots + (n-1)p_n = \sum_{i=2}^{n}(i-1)p_i.$$

Here n denotes the last non-zero element of $\boldsymbol{p}$ (so n is guaranteed to be not larger than the highest number of the repeated birth dates which individuals were having in the observed replication vector).

For the birth year 1925, we can obtain the amount of replication by using this alternative method, which is shown as follows:

$$r\hat{e}p_{1925}^{alternative} = (2-1) \times 0 + (3-1) \times 0.03223 = 6.446\%.$$

Clearly, it is quite close to the result we obtained by using the previous method to estimate the amount of replication.

3.6 The parametric bootstrap method.

From statistical point of view, a point estimate of the amount of replication is of limited use. Hence we calculate the 95% confidence intervals for the estimated amount of replication by applying the parametric bootstrap method. Based on the estimated probability distribution for the replication, we can generate a large number of samples to do the simulation so that a confidence interval is able to be generated. The simulation of the bootstrap method is used as a measurement of the estimated amount of replication proposed before.

For a given birth year the sample size estimate $\hat{r}$ and corresponding replication probability vector estimate $\hat{\boldsymbol{p}}$ have been obtained by using the algorithm introduced in Section 3.4, then we can generate a large number of random bootstrap samples of $\hat{r}$ individuals with each individual having probability distribution of number of records $\hat{\boldsymbol{p}}$. Specifically, based on the replication probability vector estimate $\hat{\boldsymbol{p}} = (\hat{p}_1, \hat{p}_2, \hat{p}_3, \cdots, \hat{p}_n)$ we generate a set of random values d_i $(1 \leq i \leq \hat{r})$ for the ith of the $\hat{r}$ distinct individuals in our bootstrap sample. Each d_i takes a value between 1 and n inclusive according to the probability distribution $\hat{\boldsymbol{p}}$. In other words, $d_i = 1$ with probability $\hat{p}_1$, $d_i = 2$ with probability $\hat{p}_2$, $d_i = 3$ with probability $\hat{p}_3$, $\cdots$ and $d_i = n$ with probability $\hat{p}_n$. d_i corresponds to the number of times that the ith individual out of $\hat{r}$ distinct individuals in our bootstrap sample has had a HIV test. This is done by generating a corresponding random variable U_i with value u_i from $\mathbf{U}(0,1)$, a uniform distribution on $[0,1]$. The U_i are independent for each individual and each bootstrap sample.

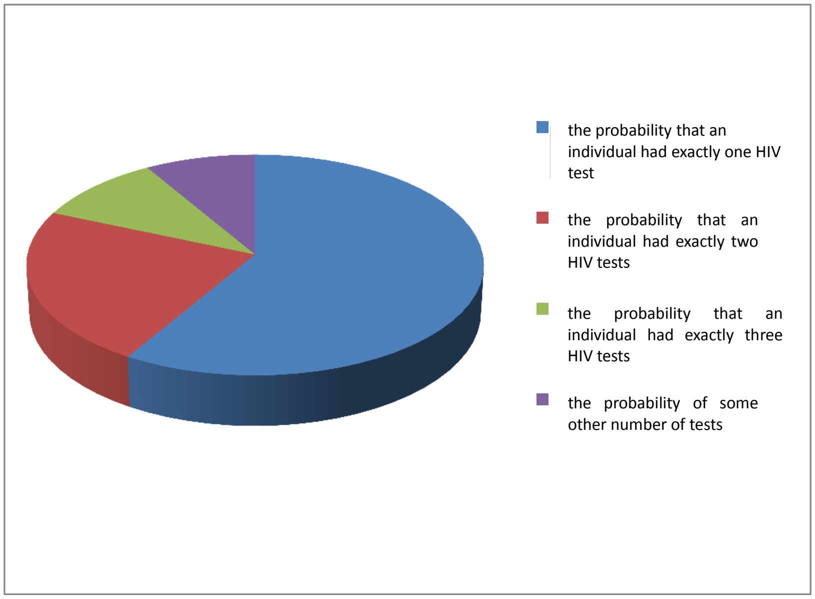
Then

$$
\begin{aligned}
&(i)\text{ If } 0 \le u_i \le \hat{p}_1,\ \text{ then } d_i = 1.\\
&(ii)\text{ If } \hat{p}_1 < u_i \le \hat{p}_1 + \hat{p}_2,\ \text{ then } d_i = 2.\\
&(iii)\text{ If } \hat{p}_1 + \hat{p}_2 < u_i \le \hat{p}_1 + \hat{p}_2 + \hat{p}_3,\ \text{ then } d_i = 3.\\
&(iv)\text{ If } \hat{p}_1 + \hat{p}_2 + \hat{p}_3 < u_i \le \hat{p}_1 + \hat{p}_2 + \hat{p}_3 + \hat{p}_4,\ \text{ then } d_i = 4.\\
&\qquad \cdots\cdots\\
&(n)\text{ If } \hat{p}_1 + \hat{p}_2 + \cdots + \hat{p}_{n-1} < u_i \le \hat{p}_1 + \hat{p}_2 + \cdots + \hat{p}_{n-1} + \hat{p}_n,\ \text{ then } d_i = n.
\end{aligned}
\tag{3.6.1}
$$

If the random value u_i is less than or equal to $\hat{p}_1$, then the ith individual in the bootstrap sample has had exactly one HIV test. We treat it as a singleton (i.e. $d_i = 1$). If u_i is chosen greater than $\hat{p}_1$ and less than or equal to $\hat{p}_1 + \hat{p}_2$, this implies that the ith individual in the bootstrap sample has had exactly two HIV tests. Equivalently d_i is considered to be a doubleton which is denoted as $d_i = 2$. Similarly, if the random value u_i is greater than $\hat{p}_1 + \hat{p}_2$ and less than or equal to $\hat{p}_1 + \hat{p}_2 + \hat{p}_3$ it shows that the ith individual has had exactly three HIV tests (where the random value d_i is treated as a tripleton, i.e. $d_i = 3$) and so on.

This can be illustrated clearly by the pie chart shown in Figure 3.4. Suppose that the whole area of the pie chart is one. Then the blue area (angle at the centre $360 \times \hat{p}_1\,^\circ$) corresponds to the probability that ith individual in the bootstrap sample has had exactly one HIV test ($0 \le u_i \le \hat{p}_1$). The red area (angle at the centre $360 \times \hat{p}_2\,^\circ$) corresponds to the probability that ith individual in the bootstrap sample has had exactly two HIV tests ($\hat{p}_1 < u_i \le \hat{p}_1 + \hat{p}_2$). The green area (angle at the centre $360 \times \hat{p}_3\,^\circ$) corresponds to the probability that the ith individual in the bootstrap sample has had exactly three HIV tests ($\hat{p}_1 + \hat{p}_2 < u_i \le \hat{p}_1 + \hat{p}_2 + \hat{p}_3$). The purple area (angle at the centre $360 \times (1 - \hat{p}_1 - \hat{p}_2 - \hat{p}_3)\,^\circ$) corresponds to the probability that the ith individual in the bootstrap sample has had

Figure 3.4: The probability distribution of the number of HIV tests of an individual in the bootstrap sample.



some other number of HIV tests $(\hat{p}_1 + \hat{p}_2 + \hat{p}_3 < u_i \leq 1)$.

For example, given the birth year 1925 with the estimated probability vector (0.96777, 0, 0.03223) and estimated sample size 31, we simulate 31 random samples from the uniform distribution $\boldsymbol{U}(0, 1)$ using computer software R which gives the results as follows:

$u_1 = 0.30403607, u_2 = 0.09584217, u_3 = 0.11918258, u_4 = 0.46913120, u_5 = 0.96680645,$

$u_6 = 0.31099250, u_7 = 0.12851647, u_8 = 0.36017595, u_9 = 0.18566244, u_{10} = 0.58873800,$

$u_{11} = 0.38195380, u_{12} = 0.84997603, u_{13} = 0.68574188, u_{14} = 0.14658660, u_{15} = 0.87304620,$

$u_{16} = 0.92822543, u_{17} = 0.86302061, u_{18} = 0.98687901, u_{19} = 0.05818043, u_{20} = 0.48908429,$

$u_{21} = 0.39215760, u_{22} = 0.33279110, u_{23} = 0.79075260, u_{24} = 0.12705525, u_{25} = 0.86916528,$

$u_{26} = 0.23403299, u_{27} = 0.16330792, u_{28} = 0.37997295, u_{29} = 0.27511547, u_{30} = 0.28733935,$

$u_{31} = 0.43947281.$

To assign the random values d_i for $1 \leq i \leq 31$, we define that

(i) If $0 \leq u_i \leq 0.96777$, then $d_i = 1$.

(ii) If $0.96777 < u_i \leq 0.96777 + 0$, then $d_i = 2$.

(iii) If $0.96777 < u_i \leq 1$, then $d_i = 3$.

Hence we have

$$d_1 = 1, d_2 = 1, d_3 = 1, d_4 = 1, d_5 = 1, d_6 = 1, d_7 = 1, d_8 = 1, d_9 = 1, d_{10} = 1,$$
$$d_{11} = 1, d_{12} = 1, d_{13} = 1, d_{14} = 1, d_{15} = 1, d_{16} = 1, d_{17} = 1, d_{18} = 3, d_{19} = 1, d_{20} = 1,$$
$$d_{21} = 1, d_{22} = 1, d_{23} = 1, d_{24} = 1, d_{25} = 1, d_{26} = 1, d_{27} = 1, d_{28} = 1, d_{29} = 1, d_{30} = 1,$$
$$d_{31} = 1.$$

From the simulation results, it is clear that there are 30 individuals with exactly one HIV positive test report and one individual with exactly three HIV test reports since d_{18} is a tripleton.

In order to give a better illustration for the method of generating bootstrap samples, we are also going to use a more general case as an example. That is for the above example of the birth year 1925 we build an artificial MLE probability vector $\hat{\boldsymbol{p}}$ with non-zero elements instead, which is assumed as $\hat{\boldsymbol{p}} = (0.95, 0.03, 0.02)$ here. Thus the criteria which are:

(i) If $0 \leq u_i \leq 0.95$, then $d_i = 1$,

(ii) If $0.95 < u_i \leq 0.95 + 0.03$ (i.e. $0.95 < u_i \leq 0.98$), then $d_i = 2$,

(iii) If $0.95 + 0.03 < u_i \leq 1$ (i.e. $0.98 < u_i \leq 1$), then $d_i = 3$, should be followed. Consequently based on the u_i's we obtained before, the values of d_i's become

$$\begin{aligned}
&d_1 = 1, d_2 = 1, d_3 = 1, d_4 = 1, d_5 = 2, d_6 = 1, d_7 = 1, d_8 = 1, d_9 = 1, d_{10} = 1,\\
&d_{11} = 1, d_{12} = 1, d_{13} = 1, d_{14} = 1, d_{15} = 1, d_{16} = 1, d_{17} = 1, d_{18} = 3, d_{19} = 1, d_{20} = 1,\\
&d_{21} = 1, d_{22} = 1, d_{23} = 1, d_{24} = 1, d_{25} = 1, d_{26} = 1, d_{27} = 1, d_{28} = 1, d_{29} = 1, d_{30} = 1,\\
&d_{31} = 1.
\end{aligned} \tag{3.6.2}$$

In this case, the simulation results shows that there are 29 individuals having had exactly one HIV test, one individual with two HIV test reports and the other individual with three HIV test reports.

In the next stage, birth dates of individuals are required to be randomly assigned to dates throughout the year in order to generate the simulated replication vectors. For the ith individual the d_i value gives the number of records corresponding to that individual. We now also need to simulate the birth date of that individual within the given birth year. Since we assume that there are 365 days in a year, the birth date can be expressed as the kth day out of 365 ($1 \leq k \leq 365$) which corresponds to a specific date. Then $\hat{r}$ integers B_1, B_2, $\cdots$, $B_{\hat{r}}$ are randomly selected with replacement from the set

$$\{\Omega : 1, 2, 3, \cdots, 365\}$$

which can be done directly in R. Next we combine the birth dates and number of distinct records of individuals in the bootstrap sample to get the observed replication vector for each bootstrap sample. Note that it is possible that some of these integers B_1, B_2, $\cdots$, $B_{\hat{r}}$ could be the same. Hence when we combine the birth dates and number of recorded HIV tests of individuals in the bootstrap sample we must take this into account. For example if a birth date b occurs exactly three times in a bootstrap sample as individual 1 who has had exactly one HIV test (i.e. $B_1 = b$, $d_1 = 1$), individual 10 who has had exactly three HIV tests ($B_{10} = b$, $d_{10} = 3$) and individual 26 who has had exactly two HIV tests ($B_{26} = b$, $d_{26} = 2$) then in the corresponding observed replication vector the birth date b gives rise to a six-tuple. In the first above example with the given birth year 1925, based on the simulation we have done before, 31 random numbers (d_i) are obtained from which thirty records are treated as singletons and the other one is treated as a tripleton. We choose 31 integers as birth date at random from 1 to 365 and assign them to B_i respectively. The birth dates which are chosen randomly with replacement are shown as follows:

$$B_1 = 32, B_2 = 148, B_3 = 100, B_4 = 245, B_5 = 215, B_6 = 193, B_7 = 191, B_8 = 28, B_9 = 62,$$
$$B_{10} = 166, B_{11} = 158, B_{12} = 223, B_{13} = 71, B_{14} = 285, B_{15} = 324, B_{16} = 326, B_{17} = 276,$$
$$B_{18} = 63, B_{19} = 360, B_{20} = 100, B_{21} = 189, B_{22} = 143, B_{23} = 48, B_{24} = 126, B_{25} = 316,$$
$$B_{26} = 344, B_{27} = 254, B_{28} = 301, B_{29} = 202, B_{30} = 139, B_{31} = 59.$$

$B_1 = 32$ means that it is the 32nd date in a year and $B_2 = 148$ represents the 148th date in a year, etc. Since the birth dates are randomly selected, we can simply assign B_i to d_i such that each individual corresponds to a particular birth date. As we mentioned before, it is possible that the same birth date is chosen repeatedly due to the sampling with replacement. In this example, the 100th date was chosen twice ($B_3 = 100$ and $B_{20} = 100$) at random. Consequently, two distinct individuals ($d_3 = 1$ and $d_{20} = 1$) who are assigned the birth dates B_3 and B_{20} respectively had the same birth date, leading to a doubleton in the new replication vector. Besides, the rest of the 28 singletons ($d_i = 1$, where $1 \leq i \leq 31$) are assigned different birth dates giving 28 singletons in the new replication vector. The three individuals in the tripleton ($d_{18} = 3$) had another distinct birth date which is the 301st day of the year ($B_{28} = 301$). Thus we can get a new bootstrap replication vector (28, 1, 1). On the other hand, if we consider the second example with the artificial MLE probability vector $\hat{\boldsymbol{p}} = (0.95, 0.03, 0.02)$ in the 1925 birth year there are two doubletons in the bootstrap sample since according to the explanation above the same birth dates from the simulation of B_i's give rive to one doubleton in the observed replication vector of the bootstrap sample and also we already had one individual with two HIV tests ($d_5 = 2$) before in the set of the number of times that an individual took HIV tests (see 3.6.2) which is a doubleton as well. Further the same birth dates B_3 and B_{20} make the number of the singletons decrease from 29 to 27. Therefore in this case the new observed replication vector in the bootstrap sample is (27, 2, 1).

Another example we intend to give is that considering a different set of birth dates B_i which has been selected for the example we used above for the birth year 1925 with the artificial MLE probability vector $\hat{\boldsymbol{p}} = (0.95, 0.03, 0.02)$. The B_i's are enumerated as follows:

$$B_1 = 32, B_2 = 148, B_3 = 100, B_4 = 245, B_5 = 100, B_6 = 215, B_7 = 191, B_8 = 28, B_9 = 62,$$
$$B_{10} = 166, B_{11} = 158, B_{12} = 223, B_{13} = 71, B_{14} = 285, B_{15} = 324, B_{16} = 326, B_{17} = 276,$$
$$B_{18} = 63, B_{19} = 360, B_{20} = 100, B_{21} = 189, B_{22} = 143, B_{23} = 48, B_{24} = 126, B_{25} = 316,$$
$$B_{26} = 344, B_{27} = 254, B_{28} = 301, B_{29} = 202, B_{30} = 139, B_{31} = 59.$$

Therefore, there are three same birth dates (B_3, B_5 and B_{20}) assigned to the third, fifth and twentieth individuals out of the 31 persons respectively. Then we combine the number of records of individuals with those three same birth dates together. Note that both the third and twentieth individuals have had exactly one HIV test (since $d_3 = d_{20} = 1$) respectively while the fifth individual is treated as a doubleton who has had the HIV tests twice (since $d_5 = 2$) according to the results in (3.6.2). Thus, one doubleton and two singletons (i.e four records) have the same birth dates, which leads to a four-tuple in the observed bootstrap replication vector. The eighteenth individual has a unique birth date but has had exactly three HIV tests giving rise to a tripleton in the observed replication vector. The other 27 distinct individuals, each of whom had an unique birth date, give rise to 27 singletons in the observed replication vector. As a result, the bootstrap replication vector becomes (27,0,1,1).

By regarding the new bootstrap replication vector as the observed one, the estimated true sample size associated with the corresponding HIV test probability vector can be obtained by applying the maximum likelihood estimation method. Consequently, the estimated amount of replication is able to be generated as well. According to the calculation in Section 3.3, the estimated true number of distinct individuals is 31 for the bootstrap replication vector (28, 1, 1) which is regarded as an observed one here in the first example. Compared with the observed sample size 33 in this example, the estimated amount of replication is 6.45%.

In order to generate the 95% confidence interval for the estimated amount of replication, we normally generate 100 samples of 'observed bootstrap' replication vectors with each of the simulated vectors calculated by the method shown above. Once we obtain the bootstrap replication vector samples, the corresponding estimated amount of replication for each bootstrap replication vector can be obtained based on the maximum likelihood method. In other words, using the bootstrap method we can generate 100 estimated sample sizes $\hat{r}_j$ with the corresponding probability vectors $\hat{\boldsymbol{p}}_j$ (where $1 \leq j \leq 100$), giving the 100 estimated amounts of replication based on the observed sample size in the bootstrap samples $r_{obs,j}$. From the mathematical point of view, the estimated amount of replication for the new bootstrap replication vector can be presented as $\frac{r_{obs,j}-\hat{r}_j}{\hat{r}_j}$ according to the definition given in Section 3.5. Recall that we proposed another method of estimating the amount of replication in Section 3.5 as $\sum_{j=2}^{n}(j-1)\hat{p}_j$, which can also be used for bootstrap replication. In this thesis, we use the former method to estimate the amount of bootstrap replication since as we demonstrated in Section 3.5, both methods give virtually the same answer. Thus we can obtain the 95% confidence interval for the estimated amount of replication by using R to get the 2.5% and 97.5% sample quartiles. Particularly, as the distribution of the amount of replication in the bootstrap samples is not necessarily symmetric, we use a technique that involves finding the quantiles from a reversed empirical bootstrap distribution [59]. Suppose that η denotes the true percentage replication in our sample, For each bootstrap sample we calculate $\eta^* - \hat{\eta}$ where η^* is the estimated percentage replication in the bootstrap sample and $\hat{\eta}$ is the estimated percentage replication. From these we find the empirical values δ^L and δ^U such that 2.5% of the adjusted observations lie below δ^L and 2.5% lie above δ^U. Hence we deduce the 95% bootstrap confidence interval for the true percentage replication η as

$$(\hat{\eta} - \delta^U, \hat{\eta} - \delta^L).$$

According to the algorithm for constructing the 95% confidence interval based on the parametric bootstrap method, a computer program written in R was devised which can efficiently generate the results by running the program. This program was comprehensively verified using detailed output from a large number of runs.

3.7 The validation of the method.

In order to validate the algorithms of the estimated amount of replication written in R and C, we compare the theoretical results with the computational ones by constructing a simple artificial replication vector due to the computability of the maximum likelihood function based on Lemma 3.4. Here we use the artificial replication vector (0, 0, 1) for illustration. According to the likelihood estimation method we have introduced before, the theoretical answer could be obtained. Considering the replication that possibly exists in this observed replication vector, there are three possible true sample sizes can be derived from the observed one which are 1, 2 and 3 with the corresponding replication vectors (1, 0, 0), (0, 1, 0) and (0, 0, 1) respectively. Thus the probabilities of the postulated replication vectors can be calculated, which are

$$\begin{aligned}
P(\boldsymbol{s} = (0,0,1), r = 3) &= \frac{365!}{0!0!1!(365-1)!}\frac{3!}{(1!)^0(2!)^0(3!)^1}\frac{1}{365^3} = 7.506099 \times 10^{-6}, \\
P(\boldsymbol{s} = (0,1,0), r = 2) &= \frac{365!}{0!1!0!(365-1)!}\frac{2!}{(1!)^0(2!)^1(3!)^0}\frac{1}{365^2} = 2.739726 \times 10^{-3}, \\
P(\boldsymbol{s} = (1,0,0), r = 1) &= \frac{365!}{1!0!0!(365-1)!}\frac{1!}{(1!)^1(2!)^0(3!)^0}\frac{1}{365^1} = 1
\end{aligned}$$

respectively with the likelihood functions of obtaining the potential replication vector given the corresponding true sample size following as:

$$
\begin{aligned}
&L(\boldsymbol{s} = (0,0,1)|\bar{r} = 3, \boldsymbol{p}) = 7.506099 \times 10^{-6} p_1^3,\\
&L(\boldsymbol{s} = (0,1,0)|\bar{r} = 2, \boldsymbol{p}) = 2 \times 2.739726 \times 10^{-3} p_1 p_2 = 5.479452 \times 10^{-3} p_1 p_2,\\
&L(\boldsymbol{s} = (1,0,0)|\bar{r} = 1, \boldsymbol{p}) = p_3.
\end{aligned}
$$

By applying Lemma 3.4, the maximum likelihood estimate $\hat{\boldsymbol{\theta}} = (\hat{r}, \hat{\boldsymbol{p}})$ for each individual likelihood function is straightforward to be obtained. The results are shown in the table below (Table 3.5) compared with the computational results obtained by R.

Table 3.5: The theoretical and computational maximum likelihood estimates for the artificial replication vector (0,0,1).

	$\hat{r}$	$\hat{\boldsymbol{p}}$	$L(\hat{r}, \hat{\boldsymbol{p}})$
Theoretical results	3	$(1, 0, 0)$	7.506099×10^{-6}
	2	$(0.5, 0.5, 0)$	1.369863×10^{-3}
	1	$(0, 0, 1)$	1
Computational results	3	$(1, 0, 0)$	7.505834×10^{-6}
	2	$(0.5009, 0.4991, 0)$	0.001369858
	1	$(0, 0, 1)$	1

Table 3.5 shows that the computational results are approximately equal to the theoretical ones which verifies to some extent the precision of the algorithm written in R. This is a simple illustration to check the accuracy of the R algorithm. We ran many similar checks and in each case the numerical results of the R program were very close to the theoretical results. Some examples of the maximum likelihood estimates for simple likelihood functions obtained from the theoretical and numerical algorithm (using Lemma 3.4 and software R respectively) are demonstrated in the Table 3.6, which shows that the maximisation routine in R works satisfactorily. It should be pointed out that the result of parameter estimated by using the optimisation routine in R is usually considered as zero when it is small enough (less than 10^{-14}).

Furthermore, the validation of the parametric bootstrap method can also be

Table 3.6: The theoretical and computational maximisation results for some example functions.

Likelihood Function		Theoretical Results	Computational Results
$p_1^{28}p_2^3p_3p_4$	$\hat{\boldsymbol{p}}$	(0.84848,0.09091, 0.03030,0.03031)	(0.84851,0.09087, 0.03026,0.03036)
	$L(\hat{\boldsymbol{p}})$	6.931522×10^{-9}	6.931502×10^{-9}
$p_1^{35}p_2^4p_3^2p_5$	$\hat{\boldsymbol{p}}$	(0.83333,0.09524, 0.04762,0,0.02381)	(0.83333,0.09522, 0.04760,0,0.02385)
	$L(\hat{\boldsymbol{p}})$	7.519887×10^{-12}	7.519887×10^{-12}
$p_1^{40}p_2^5p_4^2$	$\hat{\boldsymbol{p}}$	(0.85106,0.10638, 0,0.04255)	(0.85107,0.10634, 0,0.04259)
	$L(\hat{\boldsymbol{p}})$	3.89682×10^{-11}	3.896801×10^{-11}

performed by constructing simple artificial replication vectors with given known amounts of replication and generating the estimated amount of replication with associated bootstrap 95% confidence intervals. By means of the comparison between the estimated amount of replication with the 95% bootstrap confidence interval and the true ones, we are able to assess the performance of the bootstrap method. We take the simple artificial replication vector (9,1) with the given known true sample size 10 as the test example. In other words we assume that there are nine individuals with exactly one recorded HIV test and one individual with exactly two recorded HIV tests. When we apply the maximum likelihood estimation method demonstrated above, the estimated true sample size turns out to be 10 with corresponding estimated probability vector (0.9, 0.1). Therefore there is only one replicated record in the observed sample size 11. In other words the estimated amount of replication for the observed replication vector (9, 1) can be calculated as $\frac{1}{10}$ which is exactly the same as the given known true one. Based on the parametric bootstrap method outlined above, we can get the 95% bootstrap confidence interval for the estimated amount of replication which is $(0, 31.75\%)$. Clearly, the true amount of replication 10% lies in the bootstrap confidence interval.

However, this validation method has some disadvantages. The major issue is that the amount of replication that we arbitrarily chose as the given known true amount of replication existing in the observed replication vector could possibly rarely happen in reality. For example suppose that we constructed the artificial replication vector as (9,8) (i.e we observe nine birth dates with exactly one HIV test and exactly eight birth dates with exactly two HIV tests) and selected one as the given known true amount of replication. In other words there were really ten birth dates corresponding to just one individual and seven birth dates corresponding to two individuals. That is one of the ten individuals with a singleton birth date has actually had two HIV tests. However usually in this situation if we have a relatively small number of birth dates, true doubleton birth dates will be relatively rare compared with singleton birth dates. Hence for this replication vector (9,8) with a high number of doubletons it is very likely that more than one doubleton arises from an individual with a single birth date having had two HIV tests. Thus the example which we have constructed is unlikely to happen in practice. This may lead to a potential problem that when we construct the bootstrap samples using the parametric bootstrap method the observed bootstrap replication vectors are likely to be quite different from the one which we started with. Consequently it is likely that the corresponding 95% confidence interval for the estimated amount of replication excludes the true amount of replication in the original sample (constructed artificially), which is the potential problem here.

Nonetheless this does not necessarily mean that our algorithm of the bootstrap method is invalid. Instead the key issue is rather the selection of a test artificial replication vector which was unlikely to arise in practice. In order to overcome the drawback of the previous validation method, an alternative approach has been developed for the validation. We first assign the number of HIV tests that each individual has had. Then we allocate the birth dates at random to each individual in the artificially constructed sample. By doing this repeatedly at random, a set of observed replication vectors can be given. For each of these observed replication vectors, we can use the maximum likelihood method to derive the estimated true number of individuals and associated probability vector. Then a 95% bootstrap confidence interval for the amount of replication in that observed replication can be constructed as described in Section 3.6. In addition, the empirically simulated probability distribution of the observed replication vectors can be constructed and hence we can test whether the true amount of replication lies in the bootstrap confidence interval (it should be 95% of the time).

Using the same artificial replication vector (9, 1) as an example, we assume that nine individuals have had exactly one HIV test respectively and one individual has had two HIV tests. Hence it is obvious that there were actually ten distinct individuals giving rise to the true amount of replication which is $\frac{1}{10}$. In order to construct the bootstrap samples, first we choose 10 birth dates at random throughout a year and also assign them to each of the 10 individuals, one of whom has had two HIV tests. A set of 10 random numbers selected as the birth date from 365 days (denoted as B_i where $1 \leq i \leq 10$) were

$$B_1 = 323, B_2 = 171, B_3 = 193, B_4 = 188, B_5 = 333,$$
$$B_6 = 317, B_7 = 261, B_8 = 191, B_9 = 15, B_{10} = 80$$

which are distinct in this case.

Table 3.7: The pattern of birth dates assignment.

Individual	I_1	I_2	I_3	I_4	I_5	I_6	I_7	I_8	I_9	I_{10}
HIV tests	1	1	1	1	1	1	1	1	1	2
Birthdate	323	171	193	188	333	317	261	191	15	80

Suppose that the first nine individuals I_1, I_2, $\cdots$, I_9 are singletons and the tenth one I_{10} is a doubleton (i.e. one birth date with two HIV tests). Table 3.7 demonstrates the pattern of the birth dates B_i being allocated to the individuals. It implies that the ten individuals have had different respective birth dates, giving the new replication vector (9,1) associated with the estimated true sample size 10. Based on the observed number of individuals which is 11, the estimated amount of replication becomes 0.1 for the new replication vector constructed here.

Afterwards, we repeat the same procedure to generate additional 'observed' replication vectors. One set of the results we obtained for the random selection of the birth dates is shown as follows:

$$B_1 = 225, B_2 = 59, B_3 = 240, B_4 = 212, B_5 = 110,$$
$$B_6 = 252, B_7 = 32, B_8 = 8, B_9 = 8, B_{10} = 80.$$

Obviously, there are two same birth dates ($B_8 = B_9 = 8$) in the 10 random numbers which are assigned to the eighth and ninth individuals (both of whom have had exactly one HIV test) respectively and the other eight birth dates are assigned to the rest of the persons. In other words the two singletons (I_8, I_9) in the original sample had the same birth date. By combining the birth dates and number of recorded HIV tests of individuals together, this gives 7 singletons and 2 doubletons (i.e. the observed 'bootstrap' vector is (7,2)).

Consider another case for the selection of birth dates which is

$$B_1 = 342, B_2 = 365, B_3 = 256, B_4 = 84, B_5 = 273,$$
$$B_6 = 252, B_7 = 14, B_8 = 259, B_9 = 35, B_{10} = 252.$$

Both the sixth and tenth individual had the 252th date in a year as a birth date while the other eight individuals had a distinct birth date. That is the individual (I_{10}) who has had two HIV tests and the individual (I_6) who has had exactly one HIV test had the same birth date. Hence there was one birth date with exactly three corresponding HIV tests and the 'observed' replication vector in this case is (8,0,1).

By applying the maximum likelihood method, we can generate the estimated true sample sizes for the observed bootstrap replication vectors (7,2) and (8,0,1) which are both 9. Hence, the amounts of replication for both observed replication vectors in the bootstrap samples are estimated as $\frac{2}{9}$.

Generally speaking, we run the procedure introduced above repetitively to generate 10,000 'observed' replication vectors. By combining the same 'observed' replication vectors in the samples together, we are able to calculate the corresponding frequencies for the contribution to the probability distribution of 'observed' replication vectors. Specifically, the simulation results for this example (i.e. nine individuals have taken exactly one HIV test and one individual has taken exactly two HIV tests with given known true amount of replication $\frac{1}{10}$) are shown in Table 3.8.

Table 3.8: The 'observed' replication vectors associated with the frequencies and estimated sample sizes for the validation example.

'Observed' Bootstrap Replication Vector	Estimated Sample Size	Estimated Probability Vector	Frequency
(9,1)	10	(0.9,0.1)	8,779
(8,0,1)	9	(0.88887,0,0.11113)	238
(7,2)	9	(0.77776,0.22224)	937
(6,1,1)	8	(0.75,0.12499,0.12501)	21
(5,3)	8	(0.62382,0.37618)	25

It illustrates that the 'observed' replication vector (9,1) occurs 8,779 times out of 10,000 in the 'observed' samples which takes the highest frequency among the five simulated 'observed' replication vectors, followed by (7,2) which arises 937 times. Clearly, the last two 'observed' replication vectors (6,1,1) and (5,3) in the simulated samples get the comparatively lower frequencies (21 out of 10,000 and 25 out of 10,000 respectively). For each 'observed' replication vector, we also calculated the estimated sample size and the associated corresponding probability vector by using the maximum likelihood method (shown in Table 3.8) so that the relative 95% bootstrap confidence interval for the amount of replication can be generated. Table 3.9 shows the results.

Table 3.9: The 95% confidence intervals for the 'observed' bootstrap replication vectors.

'Observed' Replication Vector	Frequency	95% CI
(9,1)	87.7%	(0,31.75%)
(8,0,1)	2.38%	(0,66.67%)
(7,2)	9.37%	(0,55.56%)
(6,1,1)	0.21%	(0,80.63%)
(5,3)	0.25%	(0,75.0%)

Considering to control the computing time for simulating the observed birth record replication vector, we chose a small true number of individuals (10) as an example for the validation method. The assumed small true number of individuals result in relatively wide confidence intervals. Nonetheless the algorithm of generating the estimated true sample sizes as well as the corresponding probability vector performs satisfactorily. In the 87.8% of cases the sample size is estimated correctly and in another 11.7% of cases the sample size was estimated just one out. In 87.8% of cases where the sample size was estimated correctly the probability vector was also estimated correctly. Note that the simulations to calculate the observed bootstrap replication vector were conditional on there being exactly one individual in the dataset who had had exactly two reported HIV tests and that the simulations to calculate the 95% bootstrap confidence intervals were conditional on both the estimated sample size and the number of reported HIV tests that an individual had had following the estimated probability distribution.

3.8 Conclusion.

In this chapter, we discussed the method of generating the potential replication vectors and constructing the likelihood function given the true sample size and unknown HIV test distribution parameter $\boldsymbol{p}$. Based on the maximum likelihood method, two methods of the calculation for the amount of replication were illustrated. These methods gave very similar results. Besides, we mentioned the package 'alabama' in software R which is used performing the optimisation routine to obtain the maximum likelihood estimate. Due to the limitation of the point estimation, we introduced the parametric bootstrap method so that the 95% confidence intervals can be derived. Finally we did the validation of the methods including checking the accuracy of the algorithm written in R and assessing the performance of the bootstrap method constructing the artificial replication vector with the given known true amount of replication.

Although the algorithm written as the program in R has been comprehensively checked by the validation method and it is proved that the program works well for all the dataset, the times R takes to run for the large observed replication vector is currently much too long due to the large amount of for-loops contained in the program. We must use different software for the large dataset to conquer the problem. Preliminary investigation showed that the program will run much faster in C (roughly 2,000 times faster than the running speed of R). Therefore, we use C associated with the NAG library to rewrite the program greatly reducing the running time.

An alternative possible approach is an approximate method which allows individuals to have at most two HIV tests as most of the dataset provides comparatively large probabilities of an individual having one or two HIV tests and the other probabilities (the ones of an individual having had more than two HIV tests) being very small. Consequently, we treat the probabilities of an individual having more than three HIV positive tests to be zero (i.e. $p_3 = p_4 = \cdots = p_{11} = 0$). The method gives an approximate answer in the case that the probabilities of an individual having a large number of (three or more) HIV tests are very small but the computation program runs more quickly especially in R. However for the cases that the theoretical answer actually gives slightly larger probabilities of an individual having more than three HIV tests, the approximate method failed to generate correct answers, so we did not pursue this further.

This concludes our description of the theoretical methods that we used to obtain the maximum likelihood estimates for the amount of replication in the PHLS HIV test datasets and associated bootstrap confidence intervals. In the next chapter we shall give our results of the analysis of these datasets.

Chapter 4

Results of the replication present in HIV reports

4.1 Introduction.

Previously in Chapter 3 we have introduced the methods for estimating the amount of replication as well as constructing the 95% bootstrap confidence interval based on the maximum likelihood technique and parametric bootstrap method. We were given two datasets by the PHLS AIDS Center (the 1991 dataset and the 1994 dataset), containing numbers of repeated birthdates for individuals whose HIV positive tests had been reported. In this chapter we aim to present our results of the analysis of the amount of replication for both HIV datasets with the associated 95% bootstrap confidence intervals by applying the different programs written in R and C respectively.

For the 1991 dataset where the observed records show a fairly small sample size within each birth year (the maximum observed sample size in the 1991 dataset was 176 in the birth year 1944 with the highest number of repeated birthdates for individuals being 6), we are able to use the program written in R to calculate the amounts of replication within each birth year recorded in the 1991 dataset since the R program runs efficiently and accurately for the cases with relatively small observed sample sizes. The program basically has three parts. Given the observed replication vector $\boldsymbol{S}$ along with the observed sample size r_{obs}, the first part of the program calculated all the potential true numbers of distinct individuals $\{\bar{r}_i\}$ and for each of the possible true sample sizes $\bar{r}_i$ the corresponding potential replication vectors $\{\boldsymbol{T}_i\}$ also can be derived by running the first program. It is obvious that for a given observed replication vector within a birth year, the upper bound of the number of potential replication vectors increases significantly as the corresponding true potential sample size decreases, especially for the cases with a high number of repeated birthdates for individuals. In other words, for a given observed replication vector $(s_1, s_2, \cdots, s_n)$ which has the observed sample size r_{obs}, the upper bound of the number of potential replication vectors having the same derived potential true sample size $\bar{r}_i$ is the largest when $\bar{r}_i$ is the smallest in the set of potential true sample sizes (i.e. $\bar{r}_i = s_1 + s_2 + \cdots + s_n$). This may not only lead to the difficulty of constructing the likelihood function for a set of potential replication vectors given the derived potential true sample size $\bar{r}$ but also cause the problem of taking an extremely long running time in R since the likelihood function becomes very complicated for the cases with large observed sample sizes. In Chapter 3, we managed to find the upper bound for the total number of replication vectors successfully, which was used to define the size of the potential replication vectors in R. As we know, the new set of potential replication vectors with the same true sample size $\bar{r}_i$, which were derived at each stage using a straightforward procedure, probably contained repeated vectors due to the procedure we used to generate the potential replication vectors. Hence, when deriving the potential replication vectors in the first part of program, there is a checking procedure at each stage which runs every time that a new replication vector is generated. That is for a given true sample size $\bar{r}_i$ we compared the new derived vector with the ones in the set of potential replication vectors which were obtained before in this stage and if the new derived vector is the same as one that already existed it would be eliminated immediately. As a result, we were able to get the distinct potential replication vectors with the given possible true sample sizes.

Based on both the potential sample size $\bar{r}_i$ and the corresponding replication vector

$\boldsymbol{T}_i$ derived from the first part of the program, we can move forward to the second part which is used to calculate the probabilities for each potential replication vector obtained before. Theoretically, we should apply the Theorem 1 of Greenhalgh, Doyle and Mortimer [52] who proposed a mathematical formula to calculate the probability distribution of the replication vector (see (3.2.1) in Chapter 3). Hence the distribution of each potential replication vector $\boldsymbol{T}_i$ can be generated. However, in order to overcome the difficulty of evaluating the probabilities of obtaining a given potential replication vector by using the formula of Theorem 1 directly which was mentioned in Chapter 3, we used the logarithm of the probability formula (see (3.2.2) in Chapter 3) and applied this one instead of the formula in Theorem 1 in the second part of the R program to calculate the probabilities for all potential replication vectors.

Additionally, we initially created a look-up table in the form of a vector $\boldsymbol{v}$ which consists of the values of logarithm of the factorial of non-negative integers (i.e. $\log(k!)$, where $k \geq 0$ is a integer) so that the running time of the second part of program can be considerably reduced which is one of the main purposes when we are programming the R code. In other words, the vector $\boldsymbol{v}$ was set up before applying the formula of the logarithm of the probability, where the values $\{\log(c_i!), 0 \leq c_i \leq 365\}$ were assigned to the elements of the vector $\boldsymbol{v}$ in order (i.e. $v_i = \log(c_i!) = \sum_{j=1}^{c_i} \log(j)$). In that way, when running the program to calculate the probabilities for the given potential replication vectors, the created vector $\boldsymbol{v}$ can be called directly to substitute into the logarithm probability formula (refer to (3.2.2) in Chapter 3) so that the program would be more efficient.

The third part of the program aims to construct the likelihood function and then obtain the maximum likelihood estimate for the true number of distinct individuals within each birth year. According to the method introduced in Chapter 3 and based on the quantities calculated from the first two parts of the program, the third part of the program generated the likelihood function for each potential true sample size $\bar{r}_i$ by summing up the probabilities of all the true potential replication vectors $\boldsymbol{T}_i$ corresponding to the same given true sample size $\bar{r}_i$ weighted by the probability factors which are the probabilities of the number of positive HIV tests that individuals in the true potential replication vectors had, which leads to the overcounting of the number of distinct individuals and gives rise to the observed replication vector within the given birth year. As illustrated in the previous chapter, for a potential replication vector, there probably are different combinations of probability factors, which are treated as the coefficients in the formula of the likelihood function (see 3.2.7). We used the nested for-loops to generate all the possible factors and then for each given true sample size we produced the likelihood function which possibly involves a large number of the probabilities of obtaining the potential replication vectors. However the large number of loop statements (say 15 nested for-loops) in an R program causes the running time to be extremely long. In the 1991 dataset provided by the PHLS centre, the observed sample sizes were quite small which means that the likelihood functions for a given true sample size are, relatively speaking, easy to calculate. Hence, using the R program to construct the likelihood function for the 1991 dataset is sensible.

We constructed a likelihood function for each of the derived potential true number of distinct individuals. Generally speaking, several likelihood functions can be constructed as user-defined functions in R for a given birth year since there is usually more than one potential true sample size. Then for each fixed true potential sample size $\bar{r}_i$ within a birth year the program maximises the likelihood function with respect to the probability parameters $\boldsymbol{p}$ which are the probabilities of the number of times an individual has had a positive HIV test. A package in R called 'alabama' was used to maximise the likelihood function over $\boldsymbol{p}$ under constraints since the parameters p_i lie between zero and one and also follow the condition that $\sum_{i=1}^{n} p_i = 1$, where n is the dimension of the observed replication vector.

4.2 The application of the R package 'alabama'.

It is obvious that the likelihood function of the potential replication vectors given the true sample size is basically nonlinear. For the purpose of optimising a constrained nonlinear function, the package ‘alabama’ is suitable to be applied in R. In principle, it is based on the augmented Lagrangian adaptive barrier minimisation algorithm for optimising smooth nonlinear objective functions with linear or nonlinear equality and inequality constraints. As we mentioned in Section 3.4 of Chapter 3, the transformed new log-likelihood function $\tilde{L} = -log(L)$ is used as the objective function to be optimised in the ‘alabama’ package so that the minimum value of the negative log-likelihood function $\tilde{L}$ can be obtained which is equivalently the maximum value of the original likelihood function L constructed before in the third part of the R program. With regard to the parameters p_i $(i = 1, 2, \cdots, n)$ in the log-likelihood function $\tilde{L}$ for the replication vector $(S_1, S_2, \cdots, S_n)$, the constraints for the parameters are that $\sum_{i=1}^{n} p_i = 1$ and $0 \leq p_i \leq 1$. The equality constraint can be treated as $p_n = 1 - \sum_{i=1}^{n-1} p_i$ which is applied in the log-likelihood function $\tilde{L}$, making the optimisation problem a nonlinear minimisation with linear inequality constraints. In other words, we aim to minimise the nonlinear objective function $\tilde{L} = f(p_1, p_2, \cdots, p_{n-1})$ with the constraints that $0 \leq p_i \leq 1$ $(i = 1, 2, \cdots, n-1)$ and $\sum_{i=1}^{n-1} p_i \leq 1$. Hence, the feasible region for the parameters p_i is

$$\begin{cases} p_i \geq 0, \qquad i = 1, 2, \cdots, n-1, \\ 1 - \sum_{i=1}^{n-1} p_i \geq 0 \end{cases} \tag{4.2.1}$$

where $p_i \leq 1$ $(i = 1, 2, \cdots, n-1)$ can be guaranteed from the the second inequality in (4.2.1) and thus we have in total n constraints in the minimisation algorithm. To start the algorithm of optimisation in R, an arbitrary and feasible initial point should be chosen in the interior of the feasible region. From this initial point the algorithm proceeds to a neighbouring point, then a point near this neighbouring point and so on. This procedure continues so that a sequence of solutions can be obtained. Eventually it gives the optimum solution. However it is not allowed to choose the boundary of the feasible region as the initial point.

The program uses the augmented Lagrangian algorithm. The procedure consists of an inner loop augmented by a sequence of outer iterations. At each outer iteration a minimisation problem with constraints is approximately solved where Lagrange multipliers and penalty parameters which are augmented to the Lagrangian are updated in the master routine. A logarithmic barrier is added to enforce the constraints followed by the unconstrained optimisation algorithm in the inner loop. The barrier function is chosen so that the objective function should decrease at each outer iteration. We can specify the tolerance for convergence of outer iterations of the barrier and augmented Lagrangian algorithm as well as the maximum number of outer iterations to improve the precision of the algorithm. Referring to the inner loop of the algorithm, general-purpose optimisation based on the BFGS (i.e. the Broyden-Fletcher-Goldfarb-Shanno method which is a quasi-Newton method for solving nonlinear optimisation problems without constraints) is applied.

The ‘alabama’ package provides the estimated parameters $\hat{p}_j$ for $1 \leq j \leq n-1$ with the minimum value of the negative log-likelihood function $\tilde{L}(\hat{\boldsymbol{p}})$ for each potential true sample size $\bar{r}_i$ of a given observed replication vector $(s_1, s_2, \cdots, s_n)$. With the comparison of the estimated log-likelihood functions $\tilde{L}(\hat{\boldsymbol{p}})$ among all potential true sample sizes, the minimum one corresponds to the maximum likelihood estimate $\hat{\boldsymbol{p}}$ for the given observed replication vector. From a mathematical point of view, given the set of potential true sample sizes $\{\bar{r}_i\}$ (assuming $0 \leq i \leq k$) for an observed replication vector $(s_1, s_2, \cdots, s_n)$, we managed to minimise the negative log-likelihood function $\tilde{L}_i$ for each potential true sample size $\bar{r}_i$ respectively giving the estimated parameters $\hat{p}_{i,j}$ $(1 \leq j \leq n-1)$. Then the maximum likelihood estimate of p_j's for the observed replication vector $(s_1, s_2, \cdots, s_n)$ are the ones giving

$$min\{\tilde{L}_i(\hat{p}_{i,j}), 0 \leq i \leq k \text{ and } 1 \leq j \leq n-1\}$$

in the way of comparing all the minimum values of $\tilde{L}$ provided by the R package. Thus the corresponding potential true sample size $\bar{r}_i$ is considered as the estimated true sample size. Based on those quantities, we are able to obtain the maximum likelihood estimate of the probabilities that an individual had a certain number of positive HIV tests given $\hat{p}_n = 1 - \sum_{j=1}^{n-1} \hat{p}_j$, along with the maximum value of the likelihood function L which equals $exp(-\tilde{L}(\hat{p}))$ and the estimated amount of replication that can be calculated in a straightforward manner from formula (3.5.1) in Chapter 3.

4.2.1 The amount of replication results for the 1991 dataset.

By applying the R program introduced before, we are able to estimate the true number of distinct individuals for every birth year in the 1991 dataset so that the corresponding amount of replication can be derived. The results of the maximum likelihood estimation for the true sample size in the 1991 dataset are presented in Table 4.1.

From the results in Table 4.1, we can see that there are five out of sixteen birth years (i.e. 31.25% of the birth years in the 1991 dataset) having some estimated replication in the true number of distinct individuals, which are the 1931, 1934, 1935, 1943 and 1944 birth years respectively. For example, it is estimated that within the birth year 1943 there are 102 distinct individuals instead of the observed 113 records, where 90.21% out of 102 persons have had exactly one positive HIV test while 9.29% of them have had exactly two HIV tests and the number of individuals having had three HIV tests respectively is estimated to be 0.5% of the 102 distinct individuals. Similarly, the estimated true number of distinct individuals for the birth year 1934 is 36 with the probability that an individual has had exactly one positive HIV test is 0.6112 and the probability that an individual has had exactly two positive HIV tests is 0.3888. In the 1934 birth year, the probability that an individual has had more than three HIV tests can be considered as zero since only singletons and doubletons (i.e. the number of records having a distinct birthdate and the number of birthdates where there are exactly two records corresponding to that birthdate) have been recorded. Based on the observed sample size 50, we estimated that 14 out of 36 distinct individuals were overcounted by being recorded twice. The corresponding maximum value of the likelihood function demonstrates that the true sample size is 36

Table 4.1: The results of likelihood estimation for the 1991 dataset.

Year of Birth	Observed replication vector	Observed sample size r	Estimated parameters $\hat{\boldsymbol{p}}$	Maximum likelihood function $L(\hat{\boldsymbol{p}})$	Estimated sample size $\hat{r}$
1929	(26,1)	28	(1,0)	0.3864	28
1930	(23,1)	25	(1,0)	0.3794	25
1931	(19,2,1)	26	(0.8636,0.0909,0.0454)	0.0561	22
1932	(23,2)	27	(1,0)	0.1704	27
1933	(38,3)	44	(1,0)	0.2109	44
1934	(22,14)	50	(0.6112,0.3888)	0.0227	36
1935	(40,5,0,1)	54	(0.9804,0,0,0.0196)	0.0288	51
1936	(48,2)	52	(1,0)	0.1806	52
1937	(57,4,1)	68	(1,0,0)	0.0459	68
1938	(66,6)	78	(1,0)	0.1044	78
1939	(67,13,2)	99	(1,0,0)	0.0100	99
1940	(71,8)	87	(1,0)	0.0833	87
1941	(63,10)	83	(1,0)	0.0560	83
1942	(86,13,4)	124	(1,0,0)	0.0071	124
1943	(69,17,2,1)	113	(0.9021,0.0929,0.005)	0.0016	102
1944	(104,24,6,0,0,1)	176	(0.9941,0,0,0,0,0.0059)	0.0017	171

distinct individuals and that each individual has had either exactly one or exactly two positive HIV tests with probabilities 0.6112 and 0.3888 respectively. Additionally the probability of obtaining the true replication vector is 0.0227.

According to the quantities in Table 4.1, we are able to obtain the amount of replication by the two methods described in Chapter 3. The results are illustrated in Table 4.2.

Table 4.2: The results of amount of replication for the 1991 dataset.

Year of Birth	Amount of replication using $\frac{r-\hat{r}}{\hat{r}}$	Amount of replication using $\sum_{i=2}^{n}(i-1)\hat{p}_i$
1929	0	0
1930	0	0
1931	0.1818	0.1817
1932	0	0
1933	0	0
1934	0.3889	0.3888
1935	0.0588	0.0588
1936	0	0
1937	0	0
1938	0	0
1939	0	0
1940	0	0
1941	0	0
1942	0	0
1943	0.1078	0.1029
1944	0.02092	0.02095

This shows that the birth year 1934 has the highest amount of the replication among the records of the 1991 dataset which is 38.89%, followed by the amount of replication in the birth year 1931 where 18.17% of the 22 distinct individuals have been recorded repeatedly. Additionally, the birth years 1935 and 1944 show relatively lower amounts of replication which are 5.88% and 2.095% respectively. For the observed records in birth years 1931, 1934, 1935, 1943 and 1944, the results in Table 4.2 show no replication occurring in the observed records.

Comparing the results of the amount of replication calculated in different two ways, we can conclude that both algorithms give similar results. Hence it is sensible to use either formula for the calculation of the amount of replication.

To briefly summarise the evidence points to a significant amount of replication in the 1991 dataset. 31.25% of the birth years showed evidence of some replication. There were 37 records out of 1,097 in total estimated to be replicated in the 1991 dataset, which is equivalent to 3.37%. The estimated amount of replication looked highly skewed. One birth year had around 40% estimated replication but most had no estimated replication. Overall, the estimated amount of replication present in 1991 dataset was 37 records out of 1,097. i.e. the percentage of replication was estimated as 3.37%. Our results confirmed the results of Greenhalgh, Doyle and Mortimer [51], [27] in that there appears to be some replication present in this dataset. Moreover we have additionally quantified the amount of replication present.

4.3 The parametric bootstrap method in R.

We obtained the point estimates for the amount of replication in the dataset. Whilst point estimates are sometimes useful, their use is limited since they can be mean values with high associated variabilities. Confidence intervals are a standard way to quantify the variability in the amount of replication in our estimates. In this section, we shall apply the parametric bootstrap method to generate the 95% confidence intervals for the estimated amount of replication within each birth year so that the reliability of the estimated amount of replication can be indicated. As we discussed in Chapter 3, based on the estimated true sample size $\hat{r}$ for a given birth year and the estimated probability distribution for the number of HIV tests taken by an individual, we simulate the number of HIV tests taken by each of the estimated $\hat{r}$ individuals in the dataset using the R software (i.e. we choose $\hat{r}$ independent values from the estimated probability distribution using a uniform random variable in the R software) and then we categorise these random values into the corresponding tuples by following the criteria (3.6.1) in Chapter 3. Then we assigned the birthdates, which were sampled throughout a year (365 days) with replacement by applying the R base package ‘sample’, into each of the randomly simulated tuples in sequence. After combining the tuples (i.e. the values randomly simulated from the replication probability distribution and treated as the tuples) with the same birthdates, we generate an observed bootstrap replication vector. Hence the set of observed bootstrap samples of replication vectors can be derived by applying the same procedure repeatedly.

For every observed bootstrap replication vector denoted by $\boldsymbol{C}_j$ (where $j = 1, 2, \cdots, 100$ and $\boldsymbol{C}_j$ is the jth bootstrap sample in the set of observed bootstrap replication vectors), we apply the maximum likelihood method to estimate the corresponding true number of distinct individuals $\hat{r}_j$. This can be done by individually calling the R program of maximum likelihood estimation which we wrote before and applying it separately to each of the observed replication vectors. Consequently, we are able to calculate the amount of replication based on the jth observed bootstrap sample size $r_{obs,j}$ for the given observed bootstrap replication vector $\boldsymbol{C}_j$ according to the formula

$$\frac{r_{obs,j} - \hat{r}_j}{\hat{r}_j}.$$

Here $\hat{r}_j$ is the estimated true sample size from the output of the maximum likelihood estimation algorithm applied to the jth observed bootstrap replication vector and $r_{obs,j}$ is the corresponding observed sample size for the jth bootstrap replication vector. An alternative approach to calculate the amount of replication for each observed bootstrap replication vector is $\sum_{k=2}^{n}(k-1)\hat{p}_{k,j}$, where $\hat{p}_{k,j}$ ($k = 2, 3, \cdots, n$) is the estimated probability parameter of an individual having had exactly k positive HIV tests for the observed bootstrap replication vector $\boldsymbol{C}_j$. The results in Table 4.2 show clearly that the estimated amounts of replication are quite similar by using these two methods, which implies that both approaches are reasonable to calculate the amount of replication for the observed bootstrap samples. In this thesis, we use

$$\frac{r_{obs,j} - \hat{r}_j}{\hat{r}_j}$$

as the formula of estimated amount of replication for the observed bootstrap samples so that the 95% bootstrap confidence intervals for the amount of replication can be constructed. After estimating the amount of replication for all the observed bootstrap samples individually, the bootstrap confidence interval within the birth year can be generated by calculating the adjusted 2.5% and 97.5% quantiles of the estimated amounts of replication for the observed bootstrap samples. Here the bootstrap percentile method was introduced in Section 3.6 in Chapter 3. Additionally, we can also construct the 99% bootstrap confidence intervals for the estimated amount of replication by applying the same procedure.

4.3.1 The 95% and 99% confidence intervals for the estimated amount of replication in the 1991 dataset.

According to the approach outlined above, we obtained the 95% bootstrap confidence intervals as well as 99% bootstrap confidence intervals for every birth year in the 1991 dataset which are shown in the following table (Table 4.3):

The results in Table 4.3 show that the birth years in the 1991 dataset with non-zero estimated amount of replication (the birth year of 1931, 1934, 1935 and 1943) have comparatively wide 95% and 99% bootstrap confidence intervals, which implies higher uncertainty of the true amount of replication compared to the other cases with zero estimated amount of replication in different birth years, whereas the birth year 1944 from which the estimated amount of replication is also positive and the corresponding probability parameter contains quite a large value of p_1 has a small 95% (or 99%) confidence interval. It is believed that the wide confidence intervals are caused by the low value of the probability $\hat{p}_1$ (which is the probability that an individual took exactly one postive HIV test) and relatively small estimated true number of distinct individuals $\hat{r}$ in the dataset. However, although for the birth year 1944 there is a positive estimated amount of replication (2.092%), the large estimated true sample size (171) and the

Table 4.3: Amount of replication with confidence intervals for the 1991 dataset.

Year of Birth	Observed replication vector	Estimated amount of replication	95% Confidence Interval	99% Confidence Interval
1929	(26,1)	0	(0,12.00%)	(0,16%)
1930	(23,1)	0	(0,8.69%)	(0,13.64%)
1931	(19,2,1)	18.18%	(0,46.38%)	(0,52.10%)
1932	(23,2)	0	(0,12.50%)	(0,12.55%)
1933	(38,3)	0	(0,7.32%)	(0,12.85%)
1934	(22,14)	38.89%	(28.57%,61.76%)	(25.66%,96.23%)
1935	(40,5,0,1)	5.88%	(0,23.02%)	(0,29.62%)
1936	(48,2)	0	(0,4.00%)	(0,10.64%)
1937	(57,4,1)	0	(0,6.25%)	(0,7.95%)
1938	(66,6)	0	(0,6.94%)	(0,9.89%)
1939	(67,13,2)	0	(0,8.23%)	(0,8.80%)
1940	(71,8)	0	(0,10.21%)	(0,13.02%)
1941	(63,10)	0	(0,5.06%)	(0,7.82%)
1942	(86,13,4)	0	(0,5.08%)	(0,6.91%)
1943	(69,17,2,1)	10.78%	(0,20.44%)	(0,22.82%)
1944	(104,24,6,0,0,1)	2.092%	(0,10.06%)	(0,12.50%)

relatively high estimated probability of an individual having had exactly one HIV test make the corresponding 95% bootstrap confidence interval quite narrow, which implies the increased precision of the estimation of replication amount compared to the other four birth years with non-zero estimated amount of replication. Similarly, the 99% bootstrap confidence interval for the amount of replication within the 1944 birth year is narrower compared to the other four birth years 1931, 1934, 1935 and 1943, which can also be explained by the large estimated true sample size $\hat{r}$ and high estimated probability of exactly one HIV test taken by an individual $\hat{p}_1$.

Moreover, it is clear that the estimated amount of replication for each birth year in the 1991 dataset, which was obtained above (outlined in Table 4.2), lies in both the corresponding 95% bootstrap confidence interval and the 99% bootstrap confidence interval respectively. For those birth years in the 1991 dataset which had a zero estimated amount of replication, zero was also the lower bound of these 95% bootstrap confidence intervals. Although the distribution of the bootstrap estimates of the amount of replication for those birth years shows a little bit of skewness, a perfectly symmetric distribution would not be realistic since the amount of replication is constrained to be non-negative. In general, we conclude that the bootstrap percentile method is valid.

From Table 4.3, we can see that we are 95% confident the true amount of replication within the birth year 1929 lies between 0 and 12% and the corresponding 99% confidence interval is between 0 and 16% which obviously has a higher upper bound compared to the 95% confidence interval. Similarly, in the birth year 1930, we are 95% confident that the true amount of replication lies between 0 and 8.69% while there is 95% confidence that the true number of replication for the birth year 1934 lies between 28.57% and 61.57% which are quite high and demonstrates that there are a large true number of individuals who took HIV tests repeatedly. This can be explained by the fairly relatively small probability of an individual having exactly one HIV test ($\hat{p}_1 = 0.6112$) and also the small estimated true sample size ($\hat{r} = 36$) for the birth year 1934. As to the relative 99% confidence intervals for both the birth year 1930 and 1934, we are 99% sure that the true amount of replication lies in the intervals (0, 13.64%) and (25.69%, 96.23%) respectively. In the birth year 1936, we are 95% confident that the true percentage of amount of replication is in the interval of 0 to 4%, which is the most narrow one. It demonstrates a high precision of the distinct individual records within the birth year 1936 among all the observations by comparing the 95% confidence intervals for each birth year in the 1991 dataset. However the corresponding 99% confidence interval with the birth year 1936 becomes (0, 10.64%) which is significantly wider compared to the 95% confidence interval.

In general, although the 99% bootstrap confidence interval for each birth year is normally wider than the corresponding 95% bootstrap confidence interval as expected it reflects relatively similar results in confidence intervals with different significance levels. The exception to this is just one birth year (1934) with quite small estimated probability of an individual taking exactly one HIV test showing a remarkably increased upper bound in the 99% confidence interval.

4.4 The results for the 1994 dataset.

By applying the programs we described above, the estimated amount of replication with the associated 95% bootstrap confidence interval for the 1994 dataset can also be obtained. However, the observed replication vectors recorded in the 1994 dataset had in general larger sample sizes compared to the ones in the 1991 dataset which we analysed in the previous sections. Also the number of repeated birthdates could be large in the observed replication vectors for the 1991 dataset. For example, in the birth year 1962 the observed sample size was 929 with the highest record of repeated birthdates eleven for the birth year 1960. This causes difficulties in obtaining the results by running the R programs written above due to the extremely long running time. Therefore, we applied the programs written in the C language instead since it is much faster and much more efficient. A preliminary study showed that the running time of the C program is round 2,000 times faster than that of the corresponding R program to achieve the same target.

Basically, the idea of programming in C is the same as we introduced above for R. The first part of the program in C is used to derive all the potential replication vectors along with the corresponding potential true sample sizes. Then the second part of the program was built to generate the probability distribution for each potential replication vector derived before. The third part of program is used to construct the likelihood function and also to obtain the maximum likelihood estimate for the true sample sizes within each birth year so that the estimated amount of replication can be calculated afterward. In the C language, we applied the optimisation package 'e04ucc' in the NAG library for the purpose of maximising the likelihood functions.

4.4.1 The introduction of the optimisation program in the C language.

Generally speaking, we aim to optimise a nonlinear likelihood function with constraints on the parameters p_i ($i = 1, 2, \cdots, n$). The optimisation package nag_opt_nlp (e04ucc) is suitable to be applied here. It is designed to minimise an arbitrary smooth function subject to constraints which allows to include simple bounds on the variables (here the parameters p_i $(i = 1, 2, \cdots, n-1)$ are required to be greater than zero and less than one), linear constraints (specifically $p_n = 1 - \sum_{i=1}^{n-1} p_i$ in this project which also guarantees a constraint that $0 \leq p_n \leq 1$) and smooth nonlinear constraints (in our case we do not have any nonlinear constraint for the parameters p_i in the likelihood function so the number of nonlinear constraints are set to be zero here) using a sequential quadratic programming (SQP) method. It is obvious that the negative log-likelihood function $\tilde{L}$ which has also been applied in our optimisation program in R illustrated previously is considered as the objective function to be optimised in C by applying e04ucc. Mathematically, e04ucc solves the nonlinear programming problem which can be stated as follows:

$$\begin{aligned} &\underset{x \in \mathbb{R}^n}{\text{minimise}} \quad && \tilde{L(\boldsymbol{p})} = -log(L(\boldsymbol{p})) \\ &\text{subject to} \quad && 0 \leq \left\{ \begin{matrix} \boldsymbol{p} \\ \\ A_L \boldsymbol{p} \end{matrix} \right\} \leq 1. \end{aligned}$$

Here the vector $\boldsymbol{p}$ consists of the parameters p_i $(i = 1, 2, \cdots, n-1)$ and p_n is substituted by $1 - \sum_{i=1}^{n-1} p_i$ in the objective function $\tilde{L}(\boldsymbol{p})$, A_L is a 1 by $n-1$ constant matrix, here $(1, 1, \cdots, 1)$ (i.e. the coefficient matrix for the linear constraints) since there is only one linear constraint $0 \leq \sum_{i=1}^{n-1} p_i \leq 1$. Clearly, the objective function and the constraint functions are smooth (i.e. at least twice-continuously differentiable) which guarantees the assumptions of the SQP method.

nag_opt_nlp is based on the same algorithm as used in subroutine Nonlinear Programming and Systems Optimisation Laboratory (NPSOL) described Gill et al. [45], which is a software package that performs numerical optimisation. It solves nonlinear constrained problems using the sequential quadratic programming algorithm. Note that the upper and lower bounds specified for all the variables and for the constraint are always zero and one respectively. Hence we define the upper bounds as an n-dimensional vector $\boldsymbol{U}$ with all the elements being zeros. Similarly, the lower bounds vector $\boldsymbol{L}$ is also n-dimensional with all the elements being ones. The initial values for the parameters should be provided in the beginning which requires the variable bounds and linear constraint to be satisfied. Also the unspecified first partial derivatives of the objective function are approximated by finite differences.

The intermediate and final results can be obtained and printed out by default. The final results consist of the values of the parameters at the final iteration which are controlled according to the feasibility tolerances specified by the optimal parameters so that the values of the optimised parameters are expected to lie no more than the feasibility tolerances outside the upper or lower bounds, and the value of the Lagrange multiplier for the associated bound constraint which demonstrates the optimisation of the parameters based on the state of the variable that are also listed in the final results (e.g. if the parameter is optimal, the multiplier should be non-negative if the corresponding state is that the variable is on its lower bound, and non-positive if the state is that the variable is on the upper bound), and the numerical results of the likelihood function. The level of printed output can also be controlled by the optional parameters.

4.4.2 The estimated true number of distinct individuals for the birth years in the 1994 dataset.

The 1994 dataset contains the observed replication records from 1901 to 1973, where only non-zero birth year record tuples were recorded. In other words, all the non-recorded birth year tuples were zero in the dataset. For example, within the birth year 1916 the observed replication vector is (2,1) which that means $s_1 = 2, s_2 = 1$, and $s_i = 0$ $(i \geq 3)$. Based on the programs written in C, we are able to generate the estimated true sample sizes for each birth year in the 1994 dataset, followed by the associated amount of replication. The results of the likelihood estimation for the true number of distinct individuals within a given birth year are demonstrated in the following table (Table 4.4). Note that in Table

4.4, column four, we show only non-zero estimated values of $\hat{p}$. Those elements of $\hat{p}$ which are not shown are estimated as zero.

As shown in Table 4.4, we can see that in general the number of distinct individuals who had positive HIV tests increases from birth year 1901 to birth year the beginning of the 1960s while after that there is a gradual declining reduction in the number of HIV positive records. Table 4.4 shows that the figure peaked in the birth year of 1962. These individuals would have been thirty two years old in 1994. It is plausible that these individuals were most sexually active in the years in which HIV has been widespread. The results of maximum likelihood estimation implied that there was some replication due to the repeated HIV tests taken by individuals in sixteen out of seventy three birth years in the 1994 dataset. For example, within the birth year 1916 the estimated true number of distinct individuals is three rather than four which is the observed number of distinct records in the dataset. The corresponding estimated probability parameters indicate that 66.67% of individuals have had exactly one HIV test while the other 33.33% have had exactly two HIV tests leading to the estimated replication in the 1916 birth year. The probability that an individual has had more than three HIV tests in the birth year 1916 was estimated as zero since there is no evidence of more than three records with the same birthdate.

Similarly, for the birth year 1925 with the observed number of distinct individuals 33, the maximum likelihood estimate for the true number of distinct individuals turns out to be 31 and the corresponding probabilities that an individual has had exactly one, two and three HIV tests are estimated to be 0.9678, 0 and 0.0322 respectively. In other words, no individuals have exactly two HIV tests and 3.22% of individuals have exactly three HIV tests and are thus recorded as at least tripletons whereas the true amount of replication is actually less. With respect to the observed replication vector (28, 1, 1) for the birth year 1925, according to the results of maximum likelihood estimation it is clear that one of the 31 single individuals was observed as a tripleton (i.e. the observed tripleton record is in fact a single individual who took exactly three positive HIV tests) and all other birth

Table 4.4: The results of maximum likelihood estimation for the 1994 dataset.

Year of birth	Observed replication vector	Observed sample size r	Estimated parameters $\hat{\boldsymbol{p}}$	Maximum likelihood function $L(\hat{\boldsymbol{p}})$	Estimated sample size $\hat{r}$
1901	(0)	0	-	0	0
1902	(0)	0	-	0	0
1903	(1)	1	1	1	1
1904	(0)	0	-	0	0
1905	(2)	2	1	0.9973	2
1906	(0)	0	-	0	0
1907	(0)	0	-	0	0
1908	(1)	1	1	1	1
1909	(0)	0	-	0	0
1910	(0)	0	-	0	0
1911	(2)	2	1	0.9973	2
1912	(4)	4	1	0.9836	4
1913	(5)	5	1	0.9729	5
1914	(10)	10	1	0.8831	10
1915	(5)	5	1	0.9729	5
1916	(2,1)	4	(0.6667,0.3333)	0.4408	3
1917	(5,1)	7	(0.8334,0.1666)	0.3856	6
1918	(6)	6	1	0.9595	6

Table 4.4 - continued from previous page

Year of birth	Observed replication vector	Observed sample size r	Estimated parameters $\hat{\boldsymbol{p}}$	Maximum likelihood function $L(\hat{\boldsymbol{p}})$	Estimated sample size $\hat{r}$
1919	(8,1)	10	(0.8889,0.1111)	0.3529	9
1920	(6)	6	1	0.9595	6
1921	(3)	3	1	0.9918	3
1922	(7,2)	11	(0.7778,0.2222)	0.2771	9
1923	(13)	13	1	0.8056	13
1924	(17,1)	19	(1,0)	0.3060	19
1925	(28,1,1)	33	(0.9678,0,0.0322)	0.1309	31
1926	(17,0,1)	20	(0.9444,0,0.0556)	0.2472	18
1927	(22,1)	24	(1,0)	0.3726	24
1928	(26,2)	30	(1,0)	0.2132	30
1929	(39,1)	41	(1,0)	0.2443	41
1930	(35,4)	43	(1,0)	0.1045	43
1931	(37,6,1)	52	(0.8936,0.1064,0)	0.1852	47
1932	(51,4)	59	(1,0)	0.1802	59
1933	(68,3)	74	(1,0)	0.0388	74
1934	(60,8,2)	82	(0.9744,0,0.0257)	0.0161	78
1935	(59,8,1)	78	(1,0,0)	0.0400	78
1936	(69,10,2)	95	(1,0,0)	0.0203	95

Table 4.4 - continued from previous page

Year of birth	Observed replication vector	Observed sample size r	Estimated parameters $\hat{\boldsymbol{p}}$	Maximum likelihood function $L(\hat{\boldsymbol{p}})$	Estimated sample size $\hat{r}$
1937	(86,16)	118	(1,0)	0.0229	118
1938	(96,12,3)	129	(1,0,0)	0.0110	129
1939	(94,25,4)	156	(1,0,0)	0.0086	156
1940	(106,17,1)	143	(1,0,0)	0.0120	143
1941	(105,19,2)	149	(1,0,0)	0.0188	149
1942	(101,43,7,1)	212	(1,0,0,0)	4.32×10^{-4}	212
1943	(115,28,9,1)	202	(1,0,0,0)	0.0021	202
1944	(127,49,14,2,1)	280	(1,0,0,0,0)	7.317×10^{-4}	280
1945	(118,55,11,2,2)	279	(1,0,0,0,0)	5.500×10^{-5}	279
1946	(130,56,18,6)	320	(1,0,0,0)	4.359×10^{-4}	320
1947	(133,69,35,6,1,1)	411	(1,0,0,0,0)	2.955×10^{-5}	411
1948	(128,66,27,9,3)	392	(1,0,0,0)	5.230×10^{-5}	392
1949	(150,66,29,11,1)	418	(1,0,0,0)	2.134×10^{-5}	418
1950	(131,68,36,10,3)	430	(1,0,0,0,0)	4.185×10^{-5}	430
1951	(137,78,33,6,3,1,1)	444	(0.9977,0,0,0,0, 2.2775×10^{-3},0, 2.2500×10^{-5})	5.2351×10^{-6}	439

Table 4.4 - continued from previous page

Year of birth	Observed replication vector	Observed sample size r	Estimated parameters $\hat{\boldsymbol{p}}$	Maximum likelihood function $L(\hat{\boldsymbol{p}})$	Estimated sample size $\hat{r}$
1952	(131,91,41,13,3,2)	515	(1,0,0,0,0,0)	2.2389×10^{-5}	515
1953	(133,75,35,11,6,1,0,1,1)	485	(0.9936,0, 2.1320×10^{-3},0,0,0,0, 4.2638×10^{-3},4.200×10^{-6})	2.7198×10^{-7}	469
1954	(110,78,53,28,7,2,1)	591	(1,0,0,0,0,0,0)	1.2953×10^{-7}	591
1955	(88,104,57,24,11,1)	624	(1,0,0,0,0,0)	7.3892×10^{-8}	624
1956	(130,92,61,16,8,3,3,1)	648	(0.9937,0,0,0, 6.3000×10^{-3},0,0,0)	2.0853×10^{-9}	632
1957	(103,99,57,30,15,2,4,1,1)	724	(0.9972,0,0,0,0, 2.8000×10^{-3},0,0,0)	7.2675×10^{-9}	714
1958	(84,107,62,40,15,5,3)	770	(1, 0, 0, 0, 0, 0, 0)	2.7892×10^{-7}	770
1959	(86,91,75,37,18,4,5,1)	798	(1,0,0,0,0,0,0,0)	3.6988×10^{-8}	798
1960	(87,92,71,37,25,8,7,2,1,0,1)	890	(0.9977,0,0, 1.1405×10^{-3},0,0,0,0,0,0, 1.1595×10^{-3})	3.2041×10^{-11}	877
1961	(79,96,72,48,19,9,2,2)	858	(1,0,0,0,0,0,0,0)	2.0853×10^{-7}	858

Table 4.4 - continued from previous page

Year of birth	Observed replication vector	Observed sample size r	Estimated parameters $\hat{\boldsymbol{p}}$	Maximum likelihood function $L(\hat{\boldsymbol{p}})$	Estimated sample size $\hat{r}$
1962	(68,107,55,47,29,13,4,3,1,1)	929	(0.9978,0,0,0,0,0, 2.1810×10^{-3},0,0,0, 1.9000×10^{-5})	7.9214×10^{-12}	917
1963	(82,80,75,47,25,12,4)	880	(1,0,0,0,0,0,0)	3.3591×10^{-8}	880
1964	(80,87,76,42,18,12,5)	856	(1,0,0,0,0,0,0,0,0)	5.1125×10^{-9}	856
1965	(107,92,71,28,9,7)	703	(1,0,0,0,0,0)	4.7599×10^{-7}	703
1966	(109,96,52,22,15,2,1)	639	(1,0,0,0,0,0,0)	6.9163×10^{-7}	639
1967	(114,88,38,16,8)	508	(1,0,0,0,0)	2.479×10^{-6}	508
1968	(123,68,24,11,1)	380	(1,0,0,0,0)	3.833×10^{-5}	380
1969	(112,63,12,2,1,0,1)	294	(0.9965,0,0,0,0,0, 3.4722×10^{-3}, 2.7800×10^{-5})	7.9311×10^{-6}	288
1970	(107,41,9,0,1)	221	(0.9954,0,0,0,0.0046)	2.3290×10^{-4}	217
1971	(92,12,3)	125	(1,0,0)	0.01358	125
1972	(62,7)	76	(1,0)	0.1115	76
1973	(31,2)	35	(1,0)	0.2653	35

records correspond to unique individuals. Moreover, given the estimated true sample size as well as the estimated probability parameters, we can also obtain the corresponding maximum likelihood function 0.1309.

For the birth year 1962, the estimated probability that an individual has had exactly one HIV test is 0.9978 which is quite close to one and we also obtained that the small estimated probabilities that an individual has had seven and eleven HIV tests are 2.181×10^{-3} and 1.9×10^{-5} respectively. This reveals that there is a high probability that an individual has had exactly one HIV test whereas a singleton individual is recorded repeatedly as a seven-tuple with probability 2.181×10^{-3} and also it extremely rarely happens that a single individual is recorded repeatedly as an eleven-tuple since the corresponding probability is trivial (1.9×10^{-5}). Apart from this, the probabilities that an individual had exactly two, three, four, five, six, eight, nine or ten HIV tests in the 1962 observed birth year are all estimated to be zero, which means that no distinct individuals are recorded repeatedly as doubletons, tripletons, four-tuples, five-tuples, six-tuples, eight-tuples, nine-tuple and ten-tuples. Hence we can derive the corresponding estimated true number of distinct individuals which is 917 and the maximum likelihood function becomes fairly small (7.9214×10^{-12}).

On the other hand, for the birth year 1952, the observed sample size 515 is considered as the true one with the estimated probability replication vector (1, 0, 0, 0, 0, 0) which clearly illustrates that no one is estimated to have taken two or more HIV positive tests in 1952. In other words, it is believed that all the 515 individuals are distinct.

Table 4.4 presents that the majority of the birth years in the 1994 dataset recorded the correct number of distinct individuals in the observations, i.e. except for the birth years 1916, 1917, 1919, 1922, 1925, 1926, 1931, 1934, 1951, 1953, 1956, 1957, 1960, 1962, 1969 and 1970, the observed sample sizes within the birth years from 1901 to 1973 of the 1994 dataset are considered as the true number of distinct individuals with the probability that an individual took exactly one HIV test being equal to one. In general, the amount of replication present in the 1994 dataset was 100 records out of a sample of 17,272 which is the actual total number of individuals. It gives us a replication of 0.58% of the total number of distinct individuals present. However for the birth years containing replication, the estimated probability that an individual took exactly one HIV test is always significantly large and the estimated probability that an individual took more than two HIV tests is very small, especially for the birth years with large observed sample sizes. Furthermore, it is clear that the maximum value of the likelihood function given the corresponding estimated probability parameters and true number of distinct individuals becomes extremely small for the large observed sample sizes.

4.4.3 The estimated amount of replication for the birth years in the 1994 dataset.

Based on the quantities we have obtained before, the amount of replication for each birth year can also be calculated by using the formula (3.5.1) in Chapter 3 which are demonstrated in the following table (Table 4.5). Note that in Chapter 3 we introduced two methods to calculate the amount of replication and the results show that basically both methods give the same estimates of true amount of replication. In this thesis we are consistent with one of the methods to generate the amount of replication for birth years where the formula is expressed as

$$\frac{r_{obs} - \hat{r}}{\hat{r}}.$$

Here $\hat{r}$ is the estimate true sample size and r_{obs} is the observed sample size for a given birth year.

Table 4.5: The amount of replication for the birth years in the 1994 dataset.

Year of birth	Observed sample size	Estimated amount of replication (%)
1901	0	-
1902	0	-
1903	1	0
1904	0	-
1905	2	0
1906	0	-
1907	0	-
1908	1	0
1909	0	-
1910	0	-
1911	2	0
1912	4	0
1913	5	0
1914	10	0
1915	5	0
1916	4	33.33%
1917	7	16.67%
1918	6	0
1919	10	11.11%
1920	6	0
1921	3	0
1922	11	22.22%
1923	13	0
1924	19	0

Continued on next page

Table 4.5 – continued from previous page

Year of birth	Observed sample size	Estimated amount of replication (%)
1925	33	6.45%
1926	20	11.11%
1927	24	0
1928	30	0
1929	41	0
1930	43	0
1931	52	10.64%
1932	59	0
1933	74	0
1934	82	5.13%
1935	78	0
1936	95	0
1937	118	0
1938	129	0
1939	156	0
1940	143	0
1941	149	0
1942	212	0
1943	202	0
1944	280	0
1945	279	0
1946	320	0
1947	411	0
1948	392	0
1949	418	0

Continued on next page

Table 4.5 – continued from previous page

Year of birth	Observed sample size	Estimated amount of replication (%)
1950	430	0
1951	444	1.14%
1952	515	0
1953	485	3.41%
1954	591	0
1955	624	0
1956	648	2.53%
1957	724	1.40%
1958	770	0
1959	798	0
1960	890	1.48%
1961	858	0
1962	929	1.31%
1963	880	0
1964	856	0
1965	703	0
1966	639	0
1967	508	0
1968	380	0
1969	294	2.08%
1970	221	1.84%
1971	125	0
1972	76	0
1973	35	0

Table 4.5 shows clearly that the birth years with positive replication consist of 21.92% among all the birth year records in the 1994 dataset. As we can see from Table 4.4, for the birth year 1916 where there were four observed birth year records the maximum likelihood estimation indicates that there are in fact only three distinct individuals where one of them was recorded twice, giving the estimated amount of replication to be 33.33%. Similarly from Table 4.4, we can see that for the birth year 1925, there is exactly one individual in the estimated true sample size 31 recorded exactly three times leading to the observed number of birth year records being 33. Thus there are 31 distinct individuals and two of the birth year records correspond to repeat recordings (i.e. the estimated amount of replication for the birth year 1925 is 6.45%). For the birth year 1962 it was estimated that there were 917 true distinct individuals and the estimated amount of replication was 1.33%. Generally speaking, the positive estimated amounts of replication for the birth years 1916, 1917, 1919, 1922, 1925, 1926, 1931, 1934, 1951, 1953, 1956, 1957, 1960, 1962, 1969 and 1970 show quite small proportions, especially for those birth years with large sample sizes. On the other hand, the estimated amount of replication within the birth years which had the same estimated true sample size as the observed sample size such as 1952, 1967 etc. are estimated as zero. It makes sense that for the birth years only consisting of the singletons there is always no replication being estimated. Moreover, the figures in Table 4.5 show that when replication is estimated to be present in a birth year there is a downward trend in the amount of replication throughout the birth years. One possible explanation for this could be that the earlier birth years in the dataset correspond to older individuals who have been sexually active for longer and are likely to have had more HIV tests than the younger individuals in the later birth years in the dataset.

Compared to the replication results for the 1991 dataset illustrated in Table 4.2, it is obvious that the estimated amount of replication shows a significant decrease. Focusing on the birth years from 1929 to 1944, there are two birth years (1931 and 1934) in the 1994 dataset revealing replication estimated to be 10.64% and 5.13% respectively while in the 1991 dataset, five out of 16 birth years recorded repeated individuals. i.e. the number of birth years with positive replication is found to be half of the one found in the 1991 dataset. Also the corresponding positive estimated amounts of replication in the 1991 dataset are considerably larger than those in the 1994 dataset. This can be attributed to the fact that the PHLS took considerable trouble to clean the dataset and eliminate duplicate reports between 1991 and 1994. Another reason for the result of less estimated amount of replication in 1994 dataset compared to the 1991 dataset is that new cases identified between 1991 and 1994 were consequently more likely to correspond to unique individuals as the PHLS were now alert to the problem of potential replication in the database and could consequently improve the methods of identifying repeated individuals between 1991 and 1994.

4.4.4 The parametric bootstrap confidence intervals for the amount of replication in the 1994 dataset.

By applying the same method illustrated in the previous Section 4.3 to the 1994 dataset, we are able to construct the corresponding 95% bootstrap confidence intervals for the amount of replication so that the variability of the estimated amount of replication we have obtained before can be quantified.

Based on the random variables of size $\hat{r}$ (where $\hat{r}$ is the estimated sample size for a given birth year using the maximum likelihood method) simulated by the 'unifRand()' package in the C library, which presents the number of HIV tests taken by the estimated $\hat{r}$ individuals, and the birthdates which are randomly sampled throughout a year with replacement by using 'rand()' in the C library, the bootstrap replication vectors which are treated as observed bootstrap replication vectors can be generated by combining the tupletons assigned with the same birthdate. For each observed bootstrap replication vector, the maximum likelihood method programmed in the C language is applied to estimate the corresponding true sample size $\hat{r}_{bootstrap}$ and the associated estimated probability vector $\hat{\boldsymbol{p}}_{bootstrap}$ of taking a certain number of HIV tests. According to the formula of estimating the amount of replication mentioned in the previous chapter, the amount of replication for each bootstrap replication vector can be estimated by

$$\frac{r_{obs,bootstrap} - \hat{r}_{bootstrap}}{\hat{r}_{bootstrap}}$$

based on the quantities before. Here $r_{obs,bootstrap}$ is the observed sample size for the given bootstrap replication vector. Therefore, both the 95% bootstrap confidence interval and the 99% bootstrap confidence interval for the amount of replication can be constructed based on the bootstrap samples of estimated amount of replication. It is possible that the distribution of the estimated amount of replication for the bootstrap replication vectors shows slight skewness, since the estimated amount of replication is zero in the majority of cases, giving the mean of the bootstrap samples of amount of replication equal to zero whereas there usually are a few positive results of estimated amount of replication arising when estimating the true sample size for observed bootstrap samples. Thus we use the adjusted quantiles method introduced in Section 3.6, which involves the empirical distribution of the bootstrap samples in the previous chapter, to calculate the 95% and the 99% bootstrap confidence intervals. The results for the 1994 dataset are illustrated in the following table (Table 4.6).

Table 4.6: The 95% and 99% bootstrap confidence intervals for the amount of replication in the 1994 dataset.

Year of birth	95% Confidence Interval	99% Confidence Interval
1901	-	-
1902	-	-
1903	(0%,0%)	(0%,0%)
1904	(0%,0%)	(0%,0%)
1905	(0%,0%)	(0%,0%)
1906	(0%,0%)	(0%,0%)

Continued on next page

Table 4.6 – continued from previous page

Year of birth	95% Confidence Interval	99% Confidence Interval
1907	(0%,0%)	(0%,0%)
1908	(0%,0%)	(0%,0%)
1909	(0%,0%)	(0%,0%)
1910	(0%,0%)	(0%,0%)
1911	(0%,0%)	(0%,0%)
1912	(0%,0%)	(0%,0%)
1913	(0%,0%)	(0%,0%)
1914	(0%,0%)	(0%,0%)
1915	(0%,0%)	(0%,0%)
1916	(0,66.67%)	(0,66.67%)
1917	(0,50.00%)	(0,60.07%)
1918	(0%,0%)	(0%,0%)
1919	(0,25.00%)	(0,33.33%)
1920	(0%,0%)	(0%,0%)
1921	(0%,0%)	(0%,0%)
1922	(0,55.56%)	(0,55.67%)
1923	(0%,0%)	(0%,0%)
1924	(0,11.76%)	(0,11.76%)
1925	(0,19.35%)	(0,19.37%)
1926	(0,33.33%)	(0,37.54%)
1927	(0,17.29%)	(0,17.29%)
1928	(0,7.14%)	(0,7.14%)
1929	(0,10.81%)	(0,10.81%)
1930	(0,10.47%)	(0,13.19%)
1931	(0,28.26%)	(0,30.27%)

Continued on next page

Table 4.6 – continued from previous page

Year of birth	95% Confidence Interval	99% Confidence Interval
1932	(0,7.27%)	(0,7.27%)
1933	(0,7.35%)	(0,10.45%)
1934	(0,17.81%)	(0,20.00%)
1935	(0,9.95%)	(0,11.43%)
1936	(0,6.74%)	(0,7.97%)
1937	(0,4.48%)	(0,7.30%)
1938	(0,2.38%)	(0,6.64%)
1939	(0,4,73%)	(0,7.32%)
1940	(0,9.60%)	(0,11.73%)
1941	(0,4.20%)	(0,5.12%)
1942	(0,2.91%)	(0,4.56%)
1943	(0,3.64%)	(0,5.07%)
1944	(0,2.19%)	(0,2.98%)
1945	(0,5.20%)	(0,9.93%)
1946	(0,2.18%)	(0.3.82%)
1947	(0,1.46%)	(0.2.25%)
1948	(0,1.82%)	(0,2.91%)
1949	(0,1.56%)	(0,2.82%)
1950	(0,1.86%)	(0,3.72%)
1951	(0,3.54%)	(0,5.92%)
1952	(0,1.55%)	(0,3.88%)
1953	(0,4.84%)	(0,6.18%)
1954	(0,1.69%)	(0,3.38%)
1955	(0,1.28%)	(0,2.56%)
1956	(0,4.04%)	(0,4.09%)

Continued on next page

Table 4.6 – continued from previous page

Year of birth	95% Confidence Interval	99% Confidence Interval
1957	(0,2.17%)	(0,4.20%)
1958	(0,1.17%)	(0,2.59%)
1959	(0,0.75%)	(0,2.05%)
1960	(0,2.43%)	(0,3.54%)
1961	(0,0.70%)	(0,2.26%)
1962	(0,1.91%)	(0,3.27%)
1963	(0,1.48%)	(0,2.27%)
1964	(0,1.52%)	(0,2.69%)
1965	(0,1.85%)	(0,3.27%)
1966	(0,2.22%)	(0,2.41%)
1967	(0,1.52%)	(0,3.74%)
1968	(0,2.12%)	(0,4.34%)
1969	(0,4.00%)	(0,9.56%)
1970	(0,7.96%)	(0,11.32%)
1971	(0,5.04%)	(0,5.17%)
1972	(0,5.56%)	(0,8.57%)
1973	(0,9.38%)	(0,9.41%)

From Table 4.6, it is clear that for the birth years with large sample sizes the corresponding 95% confidence intervals for the estimated amount of replication are much narrower. Based on the results in Table 4.4, we can see that from the birth year 1937 (except the last two years 1972 and 1973), the estimated true sample size becomes fairly large. Considering the corresponding 95% bootstrap confidence intervals, Table 4.6 demonstrates that the confidence intervals are relatively narrower compared to the ones with small sample sizes in the birth years before 1936. Moreover, the narrow confidence intervals imply high accuracy of the estimation of the amount of replication. For example, within the birth year 1938 associated with the estimated sample size 129 and the estimated amount of replication 0, the corresponding confidence interval presents that we are 95% sure that the true amount of replication lies between 0 and 2.38%. It is obvious that the maximum likelihood estimate of the amount of replication for the birth year 1938 is included in this 95% confidence interval and moreover it is necessarily the lower bound of the confidence interval, which is sensible since the amount of replication is always non-negative. For the birth year 1966, there is 95% confidence that the true amount of replication lies between 0 and 2.22%, which means that from a statistical point of view the true amount of individuals in birth year 1966 who had repeated HIV tests is less than 15 among all the 639 individuals recorded in the 1994 dataset. The narrow 95% confidence interval for the birth year 1966 also implies low variability. Similarly, for the birth year 1959 the corresponding 95% confidence interval shows that the true amount of replication lies between 0 and 0.75% with the probability 0.95. Especially, the birth years containing only singletons (from 1901 to 1915, 1918, 1920, 1921 and 1923) have the 95% confidence intervals (0%,0%) which have the upper and lower bounds are both zero. This is because there are no individuals sharing the same birthdate (i.e no one had more than one HIV test) in the observed replication vector, giving that the probability of an individual having had exactly one HIV test is one. Moreover, the small observed sample sizes in those birth years also means that it is very likely to simulate distinct birth dates in the bootstrap samples. Hence the simulated bootstrap replication vectors were the same as the observed one in the original dataset, which gives no replication.

On the other hand, we are 95% confident that the true amount of replication within the birth year 1917 lies between 0 and 50.00%, which shows large variability (i.e poor accuracy) because of the fairly small estimated sample size. For the birth year 1916, the results in Table 4.6 show that we are 95% confident that the true amount of replication lies between 0 and 66.67% which clearly contains the estimated amount of replication 33.33%. It is believed that the wide confidence interval is caused by the small estimated sample size as well as the associated relatively low probability of an individual taking exactly one HIV test. Comparing the 95% confidence interval corresponding to the birth year 1960 and the one corresponding to the birth year 1922, although the estimated amount of replication for both cases are both non-zero the birth year 1960 with the larger estimated sample sizes shows a considerably narrower confidence interval. In general the birth years with the larger estimated sample sizes have narrower confidence intervals.

As for the 99% confidence intervals although the width of the majority of 99% confidence intervals are wider than corresponding 95% confidence intervals, which is sensible as the significance level decreases, the upper bound of the individual 99% confidence intervals are quite close to those of the corresponding 95% confidence intervals. For example within the birth year 1933, the upper bound of the 95% confidence interval is 7.35% while the upper bound of the corresponding 99% confidence interval is 10.45% which is fairly close. Moreover, the same conclusions as for the 95% confidence intervals which we obtained above can also be drawn from the results of the 99% confidence intervals. That is the cases with large estimated true sample size and high probability of an individual having had exactly one HIV test have narrower confidence interval (which also means lower variability of the true amount of replication for a given birth year) compared to those with small sample sizes and lower probability of an individual taking exactly one HIV test.

Generally speaking, we can see from Table 4.6 that the accuracy of the estimation of the amount of replication increases with the birth year due to the growing true number of distinct individuals.

Comparing the results within the same birth year recorded in the 1991 dataset and the 1994 dataset (shown in Table 4.3 and Table 4.6 respectively), the data indicates that the confidence intervals for the 1994 dataset are mainly narrower than the corresponding ones in the 1991 dataset.

4.5 Discussion.

In this chapter we briefly introduced the computer programs written in both the statistical software R and the C language, including the packages for optimisation in R and C respectively. Within a given birth year, the programs were used to derive all the potential replication vectors, where for each potential replication vector the corresponding probability of obtaining this potential replication vector can also be calculated by using the program. Eventually the maximum likelihood estimate for the true number of distinct individuals for the given birth year can be obtained based on the optimisation packages in the numerical library of either software R or the C language. Despite the fact that the program written in R has a problem of a long running time for the birth years with extremely large sample sizes, it gives highly accurate estimation for the true number of distinct individuals. With regard to those birth years with relatively large observed sample sizes, we used the program written in the C language instead which applied the same programming techniques as in R so that the running time could be cut down. This was in accordance with the preliminary investigation about the comparison of running times between R and C. For the data of observed replication vectors associated with the corresponding observed sample size given by the PHLS in 1991 and 1994, we obtained the results consisting of the estimated true number of distinct individuals, estimated probabilities of an individual having a certain number of HIV tests and the corresponding maximum value of the likelihood function. Hence according to these quantities, the estimated amount of replication was calculated.

The results show that generally the amount of replication declines as the birth year increases in both datasets. Comparing the results for the 1991 dataset (shown in Table 4.2) and the 1994 dataset (shown in Table 4.5), we found that there were 31.25% of birth years in the 1991 dataset presenting replication whereas only 21.92% of the birth years in the 1994 dataset showed replication. We also estimated that the replication present in the 1991 dataset was 3.37% and it was 0.58% for the 1994 dataset. Moreover, for those birth years which appeared in both datasets the amount of replication for birth years in the 1994 dataset appears smaller than in the 1991 dataset as expected, which indicates improved precision. The precision of the 1994 dataset has been improved both by eliminating existing repeated records which were in the 1991 dataset and adopting a more stringent approach to eliminate replication in new records entering the database between 1991 and 1994. It is believed that in the more recent years the establishment of the surname Soundex code used in recording the data provides better identification of duplicate reporting of the same individual. We also found that the years where replication were estimated to be present by the method used here were the same as the ones identified by the matching pairs method [51]. The same conclusion reinforces our confidence in the results.

We also constructed the 95% confidence intervals for the amount of replication using the parametric bootstrap method. This quantifies the variability of the estimation of the amount of replication. The greater proportion of the birth years with large estimated sample sizes and relatively high estimated probability of an individual having had exactly one HIV test in the 1994 dataset corresponds to narrower 95% and 99% bootstrap confidence intervals, compared to the ones in 1991 dataset. This concludes our statistical analysis of the amount of replication in the anonymous PHLS HIV test data.

In the next chapter we shall move on to discussing a new problem related to a infectious disease of MRSA. We shall discuss statistical modelling and imputation techniques for accessing the effect of patient movements within a hospital associated with acquiring MRSA while in hospital.

Chapter 5

MRSA introduction and literature review

Public health, especially hospital-acquired infectious disease, is a global problem of concern. In the second part of our thesis, we focus on one particular hospital-acquired infection caused by methicillin-resistant *Staphylococcus aureus* (MRSA). The infection of MRSA is difficult to be treated in humans and the increase in the number of cases of MRSA has triggered the attention of governments all over the world. A large number of studies on MRSA have been published and several scientific prevention and control strategies have also been proposed. In Scotland, universal MRSA screening for preventing the wide spread of MRSA infection in hospitals has been implemented [107]. The main aim of the second part of this thesis is to analyse the effects of patient movement and exposure to MRSA in hospital on acquisition of MRSA using the data from an one-year MRSA screening pilot study [111].

There will be three chapters (Chapter 5-7) in this study. In this chapter which is the literature review, four main topics are covered, namely (i) the medical, biological and economic background of MRSA, (ii) the published studies on MRSA, (iii) the introduction of the MRSA Screening Programme launched in Scotland and (iv) the method of collecting data within the MRSA Screening Programme that we will use for further analysis in this

Phd thesis. The aim of this literature review is to look at the evidence for the effect of patient movement on the risk of acquiring MRSA. Currently cohorting of MRSA patients in a small ward or isolation of MRSA patients in a single bedded ward is one of the strategies for reducing transmission. The effect of patient movement has not previously been investigated using these data.

Risk factor analysis is an important technique to identify and understand relevant factors which affect the behaviour and risk of acquiring a disease. In Chapter 6, we will apply risk factor analysis to identify potential risk factors which are associated with the risk of acquiring MRSA in hospital. Specifically, we focus on assessing the effect of patient movement, which is measured in our data by the number of wards a patient stayed in, on MRSA acquisition using the logistic regression method.

In Chapter 7, we aim to construct the variables on MRSA exposure based upon the patient movement throughout the hospital and analyse the effects of those exposure variables on the risk of MRSA acquisition using bootstrapped logistic regression. The pattern of patient movement within a hospital can be mapped based on the dates of admission, dates of discharge, dates of transfer between wards in the same hospital, length of stay in hospital and information on the wards that patients had been to. However there were missing data on dates of admission, dates of discharge and dates of transfer to another wards in the same hospital. Thus a multiple imputation procedure is developed to try and overcome the effects of the missing data. On the other hand, if the complete data on dates of admission, dates of discharge and dates of transfer to other wards in the same hospital are available, the analysis for the effects of exposure to MRSA variables in Chapter 7 would have been much better and much more informative. In addition, because of the poor quality of the data in this study, the dynamic modelling for transmission of MRSA in hospital cannot be implemented here.

A thread running through this PhD thesis is the modelling and imputation of unknown quantities using different statistical methods. Recall that in the first part of this thesis, we investigated the replication problem in the PHLS HIV datasets and imputed the amount of replication present in those HIV reports based on the maximum likelihood technique, while in the second part, the imputation of unknown movement dates between wards is used to construct the variables of exposure to MRSA and analyse the effects of exposure to MRSA variables associated with MRSA acquisition.

5.1 Background.

S. aureus is the most common cause of hospital-acquired infection [86], which can be a severe detriment to the welfare of patients, leading to patient morbidity and mortality and places a large burden on health-care resources [46], [49], [71], [93]. *S. aureus* involving MRSA, and methicillin-susceptible *Staphyloccus aureus* (MSSA), is a Gram positive bacterium able to cause localised skin infections, cellulitis, pneumonias and bacteraemias [65], [81]. *S. aureus* causes 25% to 35% of endocarditis cases [40]. In the 1960s, MRSA was identified from clinical specimens obtained from hospitalized patients [61]. MRSA is reported as the most frequently isolated organism in skin and soft tissue Healthcare Associated Infections (HAIs) [28] and it also causes bone, joint and surgical HAIs [72]. MRSA infections, which are resistant to antibiotic methicillin, have a higher in-hospital mortality than MSSA infections which are sensitive to methicillin [97]. The majority of patients habouring nosocomial pathogens such as MRSA typically carry asymptomatically, with overt infections developing in only a proportion of patients [22].

The spread of MRSA has generated much attention over the world. Although the incidence of MRSA has declined recently in several European countries, infection with MRSA remains a major cause of morbidity and mortality in patients admitted to hospital, particularly those in intensive care units (ICUs) [100]. Findings of the questionnaire survey undertaken in English ICUs in 2000, for the investigation of MRSA prevalence and variation in infection control policy across ICUs in England, showed that one in six patients in English ICUs were colonised, infected, or both [53]. The proportion of MRSA isolates in ICUs in the United States was 59.5% in 2003 according to the results from a large surveillance study [122]. Generally speaking, ICUs are considered as 'high risk' wards where patients are more likely to have MRSA compared to patients admitted into general wards in hospital. In addition, a number of investigations of MRSA prevalence in general hospitals have been carried out recently. During 1999 to 2001, the European antimicrobial resistance surveillance survey showed a wide variation in MRSA rates across Europe. For example 37% of blood isolates in UK were MRSA positive but only 3% of those taken in the Netherlands, Sweden, Denmark and Iceland were MRSA positive [17]. In England, the number of MRSA infections has risen since 2002 and peaked in 2005-2006 [104] whilst the number of cases for MRSA infection reduced by 30% from 2007 to 2010 according to a report by UK National Health Service (NHS) [38]. MRSA still continues to be a threat in public health nowadays and consequently infection prevention and control measures are important health protection priorities.

5.1.1 The treatments and pharmacology of MRSA.

In addition to studying the epidemiology of MRSA, the microbiological mechanism of MRSA and effective treatments for MRSA have also been developed. MRSA is resistant to multiple antibiotics such as tetracyclines, which have been widely used in humans as broad-spectrum antibiotics, aminoglycosides, macrolides, lincosamides and others [32], [75], [132]. A study by Trzcinski and Cooper [126] analysed the mechanisms of resistance of MRSA to tetracyclines, which showed that diverse genotypes of MRSA isolates determined the corresponding resistances to different antibiotics. In addition, this study established a method (i.e. a double disc diffusion method) which was the phenotypic identification of resistance to tetracyclines in MRSA. In the 1970s and 1980s, small outbreaks of infection caused by epidemic MRSA strains (EMRSA-1 and EMRSA-3) occurred in the UK and before the mid-1990s MRSA was typically resistant to tetracyclines. Several years later, a new epidemic strain of MRSA (emergence of EMRSA-15 and EMRSA-16) became endemic in most British hospitals [121]. This strain was believed to be susceptible to tetracycline and a decrease in resistance to tetracycline has been observed [126]. Based on the known property of a certain antibiotic restriction of MRSA, an effective treatment recommended by Cunha [23] was using the new drug of daptomycin, combined with other drugs such as linezolid and vancomycin if it was necessary. Recently, a decolonisation therapy which uses topical antimicrobials such as chlorhexidine and intranasal mupirocin has been applied [100]. In Scotland, a five day standard decolonisation course has been implemented, which consists of mupirocin nasal treatment (three times daily) in conjunction with an antiseptic body wash [130].

5.1.2 The economic cost of MRSA.

The incidence of patients infected or colonised with MRSA is a considerable socio-economic burden in the UK. HAIs are estimated to cost the UK NHS one billion pounds per year and a fifth of HAIs are caused by MRSA [89]. The mortality and morbidity of MRSA influence the cost of healthcare and resource utilisation, leading to the impact of the economic burden of MRSA. MRSA infection is associated with an increase in the length of stay in hospital and the costs of hospitalisation such as drugs, professional staff, medical records, admission services, isolation and so on. A survey undertaken in United States hospitals from 1998 to 2003 showed that the total economic burden of *S. aureus* increased significantly [119]. In 2003, the estimate for total economic burden of *S. aureus* was approximately $14.5 million for all inpatient stays, where the majority of the costs were associated with patient surgical stays [87]. A news report by ‘The Telegraph’ newspaper in June 2008 indicated that every MRSA infection case cost the UK NHS an extra £9,000 [25]. Therefore, cost-effectiveness prevention and infection control measures are pressingly needed and a few studies have been focused on modelling cost-effectiveness intervention strategies. In the next subsection, we will introduce some common prevention and infection control measures that have been implemented in hospitals recently.

5.1.3 The recent common infection control measures and prevention of MRSA.

MRSA is thought to be mainly spread through patient-to-patient transmission, which may be mediated by contacts from transiently-colonised health-care workers [22]. Infection with MRSA can increase the length of hospital stay, risk of death, and treatment cost for an inpatient [100]. Patients may be colonised by MRSA asymptomatically and this increases the risk of developing a clinical MRSA infection and is a source of cross-infection [100]. Thus the infection control measures, and prevention of MRSA, have become the priority for concern for the governments. National guidelines for preventing the spread of MRSA recommend contact precautions such as (i) hand hygiene, (ii) wearing of disposable gloves, aprons and gowns by health-care workers, and (iii) isolation (i.e. the placement of patients in single rooms, or in cohort wards) which may interrupt cross-infection through physical or behavioural barriers. Moreover, decolonisation treatment is also used for eliminating or suppressing MRSA by topical antimicrobials such as chlorhexidine and intranasal mupirocin [100].

In the UK the 'Clean your hands' campaign was launched and has been ongoing since September 2004 [104]. This was supported by the Department of Health (DH) and the National Patient Safety Agency and was first introduced to NHS hospitals in England and Wales. The aim of the 'Clean your hands' campaign was to ensure healthcare staff in NHS hospitals perform hand hygiene correctly at the right time in the right place to prevent cross-infection caused by healthcare workers. The campaign involved the provision of alcohol hand rub at the bedside of every patient, distribution of posters reminding healthcare workers to clean their hands, regular audit and feedback of compliance and the provision of materials empowering patients to remind healthcare workers to clean their hands.

In addition, early and effective detection of colonised (infected) patients by screening patients for MRSA colonisation or infection on admission allows timely implementation of targeted infection control measures to prevent transmission or infection. A recent review pointed out that many screening methods exist such as conventional culture, chromogenic agars and polymerase chain reaction tests [100]. In order to guarantee timely identification and intervention to reduce the risk of infection to both colonised and non-colonised patients, a sensitive, accurate and economic screening method is required.

In order to prevent the spread of MRSA and decrease the risk of cross-infections of MRSA in hospital, multiple infection prevention and control measures have been implemented in the last five years. These include a national hand hygiene campaign, dissemination of infection prevention and control guidance and implementation of care bundles [108]. However, the control strategies vary from hospital to hospital [14], [18], [55], [63]. One of the reasons is that only a few of the possible combinations of interventions have been examined in clinical trials, which are required to establish well-designed infection control strategies. An MRSA Screening Programme has been implemented since 2007 in Scottish NHS hospitals and continues to be improved with the aim of establishing an efficient and economic prevention strategy. The Scottish Government Health Directorate commissioned NHS Quality Improvement Scotland (NHSQIS) to develop a Health Technology Assessment (HTA) on the clinical effectiveness and cost effectiveness of MRSA screening [111].

The spread of MRSA infections has been of considerable concern to governments and clinical and academic epidemiologists. Recently, a large number of studies have been published on MRSA. In the next section, we will briefly review the published works on MRSA acquisition in hospitals.

5.2 Published studies on MRSA acquisition in hospitals.

Most research has focused on the investigation of MRSA acquisition in presumed ‘high risk’ wards such as ICUs, renal specialties, and cardiological specialties [110]. The rate of MRSA acquisition in general wards was 1.7%-3.2% per stay [36], [63], [98], whilst the MRSA acquisition rate in an ICU was much higher (17%) [114]. The prevalence of MRSA colonisation or infection on admission to ICUs in the UK was higher than that of many other countries [19], [64]. Published rates of MRSA colonised patients on admission showed 6.8% for an Australian ICU [76], 6.9% among 14 French ICUs [73] and 10% for an English ICU [123]. Due to the relatively high prevalence of MRSA in ICUs, the majority of studies which have been published concentrate on analysing MRSA acquisition in ICUs. These studies involve testing the proposed infection control measures and intervention strategies in ICUs and constructing dynamic transmission models of MRSA in ICUs.

5.2.1 Published studies on MRSA in ICUs.

Isolation of MRSA positive patients usually is considered as a common infection control method to reduce the spread, and its benefit above other contact precautions, such as hand hygiene, and wearing disposable gloves and aprons, was investigated. In particular, the effectiveness of isolating MRSA positive patients in ICUs to prevent transmission of MRSA was assessed by Cepeda et al. [18] based on a prospective one year two-centre study. Findings in this study suggested that isolation of ICU patients who were colonised or infected with MRSA into single rooms or cohorted bays does not reduce cross-infection, over and above the use of standard precautions, in an environment where MRSA is endemic. Therefore, reduction in the number of bed moves was recommended in ICUs, where MRSA is endemic, thus allowing better resource use in ICUs and minimising the risks from both the transfer and isolation itself. The value of source isolation during a confirmed point source outbreak of MRSA involving a single strain was not addressed by Cepeda et al. This is discussed by Christensen et al., who focused on a reported outbreak in Norwegian hospitals and showed that contact precautions proved to be sufficient to prevent transmission of MRSA [20]. The isolation of MRSA positive patients may not be directly associated with the interruption of MRSA spread when the other contact precautions have been taken into account. However, isolation was still recommended as a part of comprehensive control measures in general hospitals since it was possible that the major benefit for isolation came predominantly from the skilled practices of nursing staff such as relatively high compliance with contact precautions rather than the reduction of airborne transmission which is not a major factor in MRSA spread [33]. Further work can be undertaken within the MRSA Screening Pathfinder Project to estimate the effect of isolation of MRSA positive patients when the other contact precautions have been taken into account.

Robotham et al. [100] proposed cost-effective MRSA control strategies in ICUs according to the economic evaluation based on a dynamic transmission model in England and Wales. This study indicated that a strategy of universal topical decolonisation was optimal in the short term and that combining universal screening on admission using polymerase chain reaction with targeted decolonisation was likely to represent good value for money. Compared to the studies on the MRSA Screening Pathfinder Project, this study took parameter uncertainty such as the distribution of length of stay into account although it limited the focus to ICUs.

Routine time series data can usually be collected and may be analysed for a number of purposes. For example, Batra et al. [10] reported that a chlorhexidine-based surface antiseptic protocol can interrupt transmission of MRSA in ICUs except for MRSA strains carrying *qacA/B* genes which may be unaffected or potentially spread more rapidly, using segmented regression models on interrupted time series data collected from ICUs in St. Thomas' Hospital site of London. On the other hand, one of the main topics for MRSA acquisition from an academic perspective is modelling the dynamic MRSA transmission process which is also based on time series data. However, the published studies of MRSA transmission models mostly described the situation in ICUs.

Since the epidemic transmission process of MRSA is partially observed and nosocomial pathogens are typically carried asymptomatically, the acquisition times are difficult to observe and the evaluation of the important epidemiological parameters such as transmission rate (i.e. the rate of transmission to each susceptible patient from each colonised or infected patient) is also complicated. Understanding the route of transmission can be important for the design of optimal infection control strategies. Several rudimentary studies have modelled the underlying transmission process and estimated the transmission parameter in ICUs using various methods, which took the communicable nature of MRSA (i.e. patient-to-patient transmission) into account, and a number of transmission models for MRSA have been proposed. As pointed out by Bonten et al. [15], significant fluctuations in the incidence and prevalence of colonisation and infection is possible in small populations for example an ICU and it is suggested that a stochastic model is analysed.

Recently stochastic modelling has been applied to studies of transmission of hospital pathogens especially MRSA in ICUs. Pelupessy et al. [91] proposed a Markov model for routine hospital surveillance data in ICUs to estimate the transmission rate using maximum likelihood techniques. However, this Markov model assumed that a sequence of surveillance swabs can detect carriage with certainty, which is almost impossible in practice. In addition, a model based on a different stochastic modelling approach was proposed by Cooper et al. [22], which was a susceptible-infection-susceptible hidden Markov model (SIS HMM) to explicitly describe the unobserved epidemic process of transmission. This was the first application of a hidden Markov model to analyse the epidemic data. Some available extensive priori information for important parameters in this proposed structured HMM can provide lower and upper bounds for the parameters, which can overcome the problems with collinearity. However, there were some shortcomings in this study. For example, such a hidden Markov model may be appropriate for a single ward or unit but for a larger hospital population the application is limited as the algorithm becomes slow, numerical problems may occur and the assumption that all patients are equivalent may not be valid in practice. Markov chain Monte Carlo (MCMC) methods are recently well-designed techniques, which can be applied to infer colonisation times for partially observed infectious diseases. This approach can be used to fit a HMM when the state space is large (i.e. the number of beds is large), so that the numerical difficulties can be overcome.

Forrester et al. [39] developed methodology to estimate the transmission rate parameters of a transmissible nosocomial pathogen which allowed for imperfect sensitivity of swab and conform to the susceptible-colonised-removed paradigm based on reversible jump Markov chain Monte Carlo (RJMCMC) methods within a Bayesian framework. This model was applied to MRSA data of the ICUs of the Princess Alexandra Hospital in Australia. Findings in this study showed that the transmission rate from isolated patients was lower than from non-isolated patients.

The quantitative effects of interventions can also be analysed by applying similar methods for the purpose of improving the design of effective infection control measures in ICUs. For example, the efficacy of isolation measures in reducing transmission of MRSA using routine surveillance data from ICUs in Boston, Massachusetts, US was assessed by Kypraios et al. [67] based on a stochastic model using MCMC within a Bayesian framework. This study indicated that nares (i.e. nose) surveillance identifies a large majority of carriers and the effectiveness of barrier precautions showed an overall benefit but this benefit is inconsistent within different types of ICUs.

However, the application of the proposed stochastic models are limited to a small population and their value for a large hospital population is not clear yet. In addition, all these studies did not account for variations in host risk factors for acquisition (such as as comorbidities, age, severity of illness, wounds and so on). In our study, we aim to investigate potential risk factors for MRSA acquisition and this may indicate suitable parameters to use in a stochastic dynamic transmission model.

5.2.2 Published studies on MRSA in general hospitals.

To our knowledge, only a relatively small number of studies currently published focused on MRSA acquisition in a hospital population. The majority of these suggested that MRSA acquisition in hospital was associated with the healthcare workers or the contaminated environment especially the overload of work for healthcare staff and insufficient nursing practice. A number of other studies in general hospitals identified the risk factors associated with the risk of MRSA acquisition. For example, a one year study undertaken in an acute hospital in Hong Kong indicated that having age above 60, being in a residential care home for the elderly, prolonged hospitalisation and being in a residential care home for the elderly with patients with open wounds were identified as the risk factors for hospital-acquired MRSA [69]. The prevalence of MRSA infection regarding mortality and morbidity as well as the strain types of MRSA have also been investigated in general hospitals. Melzer et al. [78] undertook a study at Guy's and St. Thomas' Hospitals in South London from 1995-2000 to compare the incidence of mortality as well as to compare the rates of infection directly attributable to MRSA and MSSA. The findings of this study showed that there was a higher statistical proportion of death due to MRSA infection, compared with MSSA infection, but no significant difference was found between rates of disseminated MRSA and MSSA infection. A large-scale study took place in one primary care trust (PCT) in the Leeds Teaching Hospitals NHS Trust. This study determined the molecular epidemiology of MRSA colonising a large sample of elderly residents of care homes in the Leeds PCT over a four-year period [56].

To date, there is limited research focusing on the association between patient movement such as number of wards where the patient resided per hospital stay, and the risk of MRSA acquisition. The paper written by Velzen et al. [130] within the MRSA Screening Pathfinder Project in Scotland suggested that the number of wards was not a significant risk factor associated with MRSA acquisition, though a detailed analysis was not presented. This study estimated the proportion of patients who acquired MRSA in hospital and identified the main risk factors associated with MRSA acquisition as age above 64 years, self-reported renal failure and self-reported presence of open wounds. It also suggested that cross-transmission remained as an important issue in hospital. However, the association between the number of wards with other risk factors and the trend of MRSA acquisition associated with number of wards were not systematically analysed in this study. In the next chapter of this thesis, we will focus on not only the association between number of wards a patient resides in during their hospital stay and MRSA acquisition but also the association between the number of wards and the other risk factors.

5.3 MRSA Screening Pathfinder Project in Scotland.

As we mentioned in Section 5.1, a MRSA Screening Programme was launched in Scotland in order to help prevent the spread of MRSA in hospitals. Also a Health Technology Assessment, published by NHSQIS in 2007, was developed in Scotland to assess the clinical effectiveness and cost effectiveness of MRSA screening.

This report indicated that universal screening of patients on admission to hospital and efficient isolation of those MRSA colonised patients will reduce the prevalence of MRSA colonisation or infection within the patient population. It also recommended that laboratory nasal screening of all in-patient admissions using chromogenic agar was likely to be the most clinically and cost effective strategy for MRSA screening in NHS Scotland [99]. A one-year MRSA Screening Pathfinder Project was established in NHS Scotland commissioned by the Scottish Government Healthcare Associated Infection Task Force. The aims of this project were to (i) investigate the proposed model, (ii) test HTA findings and (iii) examine the feasibility and implications of the proposed screening strategies for health boards, in order to provide scientific evidence for further national policy decision making [110], [111]. This study protocol (AREC reference number 09/MRE00/50, R&D reference NRS09BA01) was approved by the Scotland A

Research Ethics Committee, Edinburgh in June 2009 [106]. A large intervention study was undertaken in two Scottish acute hospitals for the Pathfinder Project [105]. The Pathfinder study has addressed many organisational issues in healthcare regarding the underlying assumptions within the HTA model on MRSA screening [111].

5.3.1 Previous conclusions for the MRSA Screening Pathfinder Project.

Initially the HTA did not recommend a Clinical Risk Assessment (CRA) which is a questionnaire to identify patients at higher risk of MRSA carriage, involving demographics of patients and risk factors for MRSA colonisation such as medical history of MRSA colonisation or infection, having wounds or ulcers, on admission within the MRSA Screening Programme due to the cost-ineffectiveness. However, the findings of the MRSA Screening Pathfinder indicated that the use of CRA could successfully identify 80.7% of colonised patients and is economically efficient [110]. Moreover, a simple CRA involving three questions: (i) Has the patient any previous history of MRSA colonisation or MRSA infection at any time in the past? (ii) has the patient been admitted from somewhere other than their own home? (iii) does the patient have a wound, ulcer or implanted medical device which was present before admission to hospital?, was shown to be adequately effective to identify MRSA carriers. Thus future implementation of CRA for all admissions was recommended.

A literature survey showed that *S. aureus* colonises the nasal cavity of about 30% of the healthy population [113]. Direct nasal swab screening combined with culture on chromogenic agar, which was also recommended by HTA, has been the routine methodology for detecting MRSA carriage in Scotland and in many other countries [107]. However, the efficacy of universal nasal swabbing for MRSA was only 66% of the 'gold standard' diagnoses detected within the MRSA Screening Pathfinder [107], [112]. The 'gold standard' combined results from nasal, axillary, throat and perineal swabs plus swabs from wound or indwelling medical device sites, with broth culture on chromogenic agar and nutrient broth enrichment and sub culture on chromogenic agar. The report for the MRSA Screening Pathfinder by NHS Scotland recommended using nasal plus perineal swabbing for detecting MRSA carriage (for patients who felt that it was difficult or unacceptable to do perineal swabbing, throat swabbing was used instead), which gave 82.2% of 'gold standard' diagnoses detected, combined with the CRA for all admissions for pre-emptive management (i.e. isolation or decolonisation for patients at high risk of MRSA colonisation when awaiting laboratory confirmation) [107].

According to further analysis from the MRSA Screening Pathfinder project, screening taking the form of CRA and laboratory testing for MRSA colonisation is highly clinically effective and cost-effective as a first stage screening process. The Pathfinder study also indicated that universal MRSA screening involving applying CRA for all admissions to identified high-risk patients as a first line screening tool and using swabbing and culture to those identified high-risk patients as the second line screening tool may be associated with a reduction in MRSA prevalence and infection incidence. Furthermore, it is highly acceptable to patients and the public [111]. However the debate on optimal MRSA screening continues regarding (i) the clinical and cost effectiveness, (ii) the feasibility and potential for rollout of the MRSA screening programme, and (iii) the acceptability of MRSA screening for all acute in-patient admissions [111].

Evidence was also found in the MRSA Screening Pathfinder Programme that 3.9% of patients were MRSA colonised on admission and 2.9% of patients were MRSA colonised on discharge and 0.8% of patients were MRSA positive on admission but MRSA negative on discharge whilst the majority of patients remained MRSA negative both on admission and on discharge [106], [109]. There were cases of patients who were not MRSA positive on admission and acquired MRSA during hospital stay, 1.3% of all patients who were screened both on admission and discharge in the MRSA Screening Programme acquired MRSA whilst in hospital [106]. A large Swiss study also reported half of the patients who developed an infection during hospital stay were not MRSA positive on admission [54]. Both studies suggested that a number of patients acquired MRSA through cross-transmission (including transmission within the ward and transmission outside the ward such as underlying transmission by health-care workers) whilst in hospital. The evidence suggests that there is MRSA acquisition through cross-transmission for patients who are MRSA negative on admission to hospital.

However, few studies have been published on MRSA acquisition in the general hospital population and there is limited information on the effect of colonisation pressure on the risk of acquiring MRSA [133]. Evidence on its potential effects on MRSA acquisition in the general hospital population is highly relevant for decision making about further guidance of MRSA presentation and control strategy and implementation of universal screening for MRSA.

5.3.2 The aims of this study.

Our study is further work on analysing the effect of potential risk factors, especially patient movement, on the risk of MRSA acquisition within the MRSA Screening Pathfinder Project. The primary objective for our study is to assess the role of patient movement within a hospital, which can be treated as an indicator of potential for cross-transmission within the ward, on the propensity of patients to acquire MRSA in hospital. Specifically, we aim to investigate three main questions:

(i) Is there any evidence that the probability of acquisition of MRSA in hospitals is greater among patients who are transferred among wards within the hospital compared to patients who remain in the one ward throughout their stay?

(ii) is there any evidence that the probability of acquisition of MRSA in hospitals is greater among patients who are transferred to wards within the hospital where there are known patients who are colonised or infected with MRSA compared to patients who are in wards with no known MRSA colonisation or infection?

(*iii*) is there any evidence that the probability of acquisition of MRSA in hospitals is affected by length of stay only among those patients who spend some of their time in hospital in wards where MRSA is known to be present?

If the patient movement is not significantly related to the risk of MRSA acquisition, then this may suggest that underlying transmission may be by the route of health-care workers or by the existence of improper disinfection for the beds used by MRSA colonised or infected patients. The hand hygiene compliance may be a considerable issue in general hospitals, where improvement is recommended. Moreover, it is possible that hand hygiene was less carefully adhered to for patients not known to be MRSA positive while a delay between collection of screening cultures and results becoming available is always present in practice. Thus MRSA colonised patients might be a potential source of spread during this time unless carriage had been reported previously. Therefore additional intervention policies on controlling MRSA will be recommended.

The second objective in our study is to identify the potential risk factors for MRSA acquisition to improve pre-emptive management of these high risk patients. The previous report within the MRSA Screening Pathfinder Project identified three risk factors for acquisition of MRSA, which were: (i) age above 64, (ii) self reported renal failure, and (iii) self reported presence of open wounds [106]. However, in this study the number of wards which gives information about patient movements were not fully included for the analysis. In our study, the association between the risk of MRSA acquisition and the potential risk factors including the number of wards will be reworked in Chapter 6 with the addition of an in depth evaluation of the role of the number of wards a patient is resident in.

In summary, we briefly introduced the medical, biological and economic background of MRSA and reviewed some published works on MRSA acquisition in this chapter. Generally speaking, infection with MRSA is difficult to treat in humans. MRSA causes severe morbidity and mortality, which leads to a large impact and economic burden. The spread of MRSA is a focus of global public health concern. The risk of acquiring MRSA is especially high for patients admitted to ICUs in hospitals. A large number of published studies were focused on MRSA acquisition in ICUs. Among these were studies on dynamic modelling of MRSA transmission and the assessments on MRSA control strategies in ICUs. In addition, studies on MRSA acquisition in general wards in hospitals were also illustrated in Section 5.2.2.

In order to reduce the incidence of MRSA-related mortality and the rate of MRSA infections, the most effective treatments and prevention and control strategies, such as hand hygiene, isolation and decolonisation, have been proposed. In Scotland, a universal MRSA Screening Programme was launched in general hospitals in 2007, so that patients admitted in hospital are routinely screened for MRSA. Simultaneously, the assessments of the clinical effectiveness and cost effectiveness of MRSA Screening Programme were developed.

In this chapter, the one-year MRSA Screening Pathfinder Project for investigating the proposed MRSA screening was also described. In addition, most of the main findings within the MRSA Screening Pathfinder Project were also reviewed. The primary objective of the second part of this thesis is to investigate the effects of patient movements and exposure to MRSA in hospitals on MRSA acquisition based upon the data from the MRSA Screening Pathfinder Project. The method of collecting the data which will be used for the analysis in this second part of the thesis will be described in the next section.

5.4 Data collection for the MRSA Screening Pathfinder Project.

We will assess the role of patient movement on MRSA acquisition within the MRSA Screening Pathfinder Project in Scotland. As the data collected from this project will be used for the analysis, we describe the data collection in this section.

Universal screening for MRSA took place in two acute care hospitals in two NHS Boards from February to August 2010 (seven months) [106]. Dedicated administrative staff were employed to assist in data collection whilst the clinical tasks were done by ward staff, additional ancillary staff, dedicated nursing staff and dedicated screeners depending on the workload [107]. Patients who were willing to take part in the MRSA Screening Programme were required to sign a consent form. For those patients, the CRA which is a questionnaire described in Section 5.3.1, asking whether patients had received antibiotics in the year before admission and whether they suffered from specified co-morbidity (diabetes, renal failure, chronic obstructive pulmonary disease), open wounds, sores, or ulcers, was administered on admission. Patient admission information on the patient data form was collected from the hospital Patient Administration System (PAS) and nursing notes. CRA and consent form were scanned into a holding database using TELEform® scanning software [107].

All consenting patients were swabbed on admission at four body sites: anterior nares, perineum, axillae and throat, which was undertaken within 48 hours of admission. Note that swabs also were taken from wounds and devices if applicable. Discharge screening was also undertaken within 24 hours before discharge. The samples were taken by trained nurses and screening assistants and then were sent to the laboratory. Those samples taken from one patient were inoculated onto Oxoid's Brilliance MRSA Agar medium. Then they were pooled and inoculated into Oxoid selective manitol enrichment broth and incubated at 37°C for 18-24 hours before being plated on Oxoid Brilliance MRSA agar [107]. The MRSA colonised results were confirmed by Vitek 2 AST/ID testing and genotyped in an MRSA reference laboratory to allow identification of the acquisition of new MRSA strains in hospital. Two molecular methods were applied to identify the MRSA genotype: the pulse-field gel electrophoresis method and multilocus Variable-Number Tandem-Repeat analysis.

A patient was considered MRSA positive if one or more of the corresponding individual swab samples enriched in broth were MRSA positive. Patients who were

MRSA positive on admission were isolated or cohorted immediately. The decolonisation treatment (i.e. the standardised intervention protocol consisting of mupirocin nasal treatment three times a day for five days in conjuction with five days of use of antiseptic wash) was applied to all MRSA positive patients [106]. After the decolonisation treatment, a re-test for MRSA was required, followed by a second decolonisation course if applicable. However, the results of the re-test for MRSA were not included in our analysis. Patients who were discharged before finishing decolonisation were advised to finish the full decolonisation course after discharge.

Briefly, in our study, data on demographics and risk factors for MRSA acquisition were collected for the analysis: patient identity number, gender, age, admission/discharge specialty, length of stay, number of wards, number of days in isolation facilities, date of admission/discharge, unit admitted to (high risk unit such as General Surgery ward and Orthopaedics Elective ward or low risk such as General Medicine ward and Medical Receiving ward), co-morbidity and being on decolonisation treatment at discharge, negative/positive results on admission and discharge and date of swab. Note that information on co-morbidity was obtained from the CRA which was completed on admission. The collected data which were visually and automatically validated were imported into a Structured-Query-Language database at Health Protection Scotland.

The data which will be used in our study consists of (i) an admission only database (7,181 patients) where the information for patients on discharge is incomplete (i.e. date of discharge and MRSA measure on discharge are missing), (ii) a discharge only database (2,432 patients) where the information for patients on admission is incomplete (i.e. date of admission and MRSA measure on admission are missing) and (iii) a combined admission-discharge cohort (2,792 patients). The latter database has complete information on MRSA status on admission, on discharge, as well as data on the wards the patient was in while in hospital. The admission only and discharge only databases also include the information on wards the patient was in while in hospital. These different databases are because of the way patients gave consent. Patients had to consent separately for admission and discharge data collection. For patients who consented to take part at admission, only data collected at admission was available and date of discharge was not collected. Similarly for patients who did not consent at admission but did consent on discharge their date of admission was not available to the researchers.

In this study, we are particularly interested in estimating the impact of patient movement within general hospitals on the risk of MRSA acquisition using the data from the Universal MRSA Screening Program in Scotland. In Chapter 6, we aim to understand the effect of potential risk factors on MRSA acquisition (i.e. question (i) in Section 5.3.2) using the admission-discharge cohort only. In Chapter 7 the dynamics of patient movement will be modelled using all three databases. Then the effect of being in a ward in which there is a patient or patients with MRSA (i.e. questions (ii) and (iii) in Section 5.3.2) will be assessed based on multiple imputation using the admission-discharge cohort. As the data on the dates of transfer to another ward in the same hospital were missing in all three databases and dates of admission, and discharge were missing from the discharge only and admission only cohorts respectively, the pattern of patient movement and the variables of exposure to MRSA cannot be constructed directly. Hence multiple imputation must be used. The analysis in Chapter 7 for the effects of exposure to MRSA associated with MRSA acquisition would be more informative if all these dates were available.

Chapter 6

MRSA Screening Pathfinder Programme: Risk factors for acquisition for MRSA in the hospital

6.1 Introduction.

Methicillin-resistant Staphylococcus Aureus (MRSA) is a common health problem for concern across the UK [16]. In order to reduce the transmission of MRSA within healthcare settings in the UK, a policy of Universal MRSA Screening Program was implemented for all the elective patients admitted into hospital [125]. However, the effectiveness of universal MRSA screening for patients on admission is controversial [95].

One of the report of the Universal MRSA Screening Program Pilot Study published by National Health Service for Scotland (NHS Scotland) identified three risk factors associated with MRSA acquisition. These were age over 64, self reported renal failure and self reported open wounds [106]. However, the assessment of the effect of patient movement (i.e. the number of wards that patients stayed in per hospital stay) on MRSA acquisition was not analysed in detail in the previous NHS Scotland report.

In this chapter, the effect of the number of wards that patients had moved through while in hospital on the acquisition of MRSA is analysed as a primary objective, assuming that the admission screening was reasonably sensitive to detect MRSA on admission [107]. There are four main aims in this chapter: (i) investigate whether the risk of MRSA acquisition is higher among patients who were in a relatively large number of wards; (ii) investigate the association between the number of wards and other risk factors which are identified as having significant effects on MRSA acquisition and are possible confounding factors; (iii) investigate the trend in risk of MRSA acquisition associated with increasing levels of number of wards and other related risk factors; (iv) derive an updated multivariable logistic regression model for MRSA acquisition.

The data used in this chapter come from the multicentre prospective cohort study within the Universal MRSA Screening Program. The MRSA screening results were taken from the patients both on admission and at discharge in two acute care hospitals in two NHS boards. One of the hospitals is a large district general hospital called Crosshouse hospital, which is situated in Ayrshire and Arran NHS board (i.e. south-west Scotland) and contains 590 beds. The other one is Aberdeen Royal Infirmary which is a large teaching hospital within NHS Grampian and contains 893 beds. The patients in the elective orthopaedic ward of the Aberdeen Royal Infirmary, which is located in the adjacent Woodend hospital, were also included in the study population of the Aberdeen Royal Infirmary [106].

We focus on assessing the odds ratios of acquiring MRSA among patients in the admission-discharge cohort who were MRSA negative on admission. All patients aged 16 and older who were admitted to any ward in the two acute hospitals, and also stayed at least one night in hospital, were eligible for inclusion in the study. Patients who had not been screened on admission or at discharge were excluded from the acquisition analysis. Therefore in this chapter, we are concerned about the admission-discharge cohort, which has complete information on MRSA status on admission, at discharge, as well as data on the wards the patient was in while in hospital. The details of the data collection were reported in the previous chapter.

6.2 Methods.

The main aim of this chapter is to analyse the effect of the number of wards a patient resided in while in hospital as a potential risk factor on MRSA acquisition. In this section, we will introduce the definition of MRSA acquisition cases, the potential risk factors involved in the analysis and the statistical methods.

6.2.1 Case definitions for MRSA acquisition.

A patient was considered MRSA positive i.e. colonised with MRSA if the lab results for any of the swabs on either admission or discharge showed MRSA positive. On the contrary, a patient was considered MRSA negative i.e. not colonised with MRSA, if all of the swabs on admission or discharge were tested to be MRSA negative. A patient was considered to acquire MRSA during the stay in hospital i.e. a patient was colonised by a new strain of MRSA if one of the following three cases was met [106]:

- The patient was MRSA negative on admission and MRSA positive on discharge.
- The patient was MRSA positive on both admission and discharge but acquired a new strain of MRSA during hospital stay (as shown by genotyping).
- The patient was MRSA negative on both admission and discharge, but developed an MRSA infection during the hospital stay.

In this study, we find that there were 34 patients who were MRSA negative on admission and MRSA positive on discharge in the admission-discharge cohort but no patient developed an MRSA infection or acquired a new strain of MRSA during the hospital stay in this cohort. Note that the patients who died during the period of this study are excluded.

6.2.2 Potential risk factors.

The putative risk factors for MRSA acquisition on the basis of a plausible prior hypothesis consisting of reported association with MRSA acquisition were available for the entire admission-discharge cohort of the study. Overall, there are 12 potential risk factors: gender, age, discharge specialty, length of stay, the situation of isolation (i.e. whether the patient had been isolated during hospital stay), patient movement through the hospital (i.e. the number of wards that the patient had been in during hospital stay), being on decolonisation treatment at discharge, and a further five potential risk factors relating to co-morbidity which include diabetes, Chronic Obstructive Pulmonary Disease (COPD), open wounds or ulcers, renal failure and antibiotic use during the past year. As introduced in the previous chapter, the data on co-morbidity were collected from an admission risk assessment questionnaire administered on admission using a standardised data form.

6.2.3 Statistical analysis.

Before analysing the potential risk factors associated with MRSA acquisition, we cleaned the records of ward code and number of wards in the dataset. Specially, we modified any mistyping of ward codes in Aberdeen Royal Infirmary within Grampian NHS board by comparing the observed records of ward code in the dataset with the corresponding true ward codes in the hospital. Hence, the number of wards that a patient had moved through were also amended according to the modified ward codes in the dataset (i.e. totalling the non-empty modified records of ward codes). For example, in the Aberdeen Royal Infirmary dataset, a patient was in Ward 10, I1, 10. The ward code record ‘I1’ can be considered as a computer identified mistake, which should be ‘11’ instead. After reasonably correcting the ward codes, we adjust the total number of wards by adding up the number of ward codes recorded for an individual patient. In this case, a patient was in Wards 10, 11 and then moved back to 10 this is recorded as 3 wards.

We used two modelling approaches to estimate the odds ratios of the risk factors associated with the risk of acquiring MRSA while in hospital, which are categorical logistic regression and linear logistic regression. In epidemiological researches, it is common to use the categorical variables in the analysis, allowing easy interpretation and presentation of results but this represents a loss of information by comparison with using the numerical values of the measured or counted variables. In order to understand the associations between the risk factors, such as the associations between the number of wards and the age of patient, the length of stay, we also investigated the patterns of the pairwise categorical risk factors in tables.

Firstly, we carried out the categorical analysis for MRSA acquisition using a binary logistic regression model in R software to estimate the odds ratios for the potential risk factors. As MRSA acquisition in the hospital is rare, the odds ratio which is obtained from the logistic regression can also be interpreted as a relative risk. Univariate analysis was used to detect the potential risk factors which have the significant effects on MRSA acquisition. The number of wards is always retained as this is the principal variable in this analysis. Other variables with low p-values ($p < 0.10$) were considered for inclusion in the multivariable logistic regression model. Variables with multiple levels (i.e. more than three categories) were tested for the corresponding overall p-values with the Wald test. For those significant risk factors associated with MRSA acquisition, the mutual association between pairwise risk factors was assessed using a χ^2 test.

In order to investigate a multiplicative interaction between the risk factors and the potential confounding effects associated with the risk of MRSA acquisition, a stratified analysis was applied and the stratified-specific odds ratios were calculated for assessing the effects of the risk factors after controlling for the possible confounding effects.

Since the sample sizes for some of the strata can be slightly small, exact logistic regression was applied to estimate the stratum-specific odds ratio and the corresponding p-value, which were compared with the results obtained from the binary logistic regression. The exact logistic regression is based on a Markov Chain Monte Carlo approach, which uses uniform sampling within the Markov chain for generating the Monte Carlo samples, so that the parameters for the exposure variables within the logistic model can be simulated. We use the R library of ‘elrm’ package to do the analysis.

The consistency of the stratum-specific odds ratios across strata refers to the absence of multiplicative interaction. So we used the Woolf method to test the null hypothesis which is that all the stratum-specific odds ratios are the same (i.e. $OR_1 = OR_2 = \cdots = OR_i$ where i is the number of strata) against the alternative hypothesis which is that at least one of the stratum-specific odds ratios differs from another one using a 5% significance level. In this study, we also used an alternative approach for investigating the multiplicative interaction, which is the likelihood ratio test based upon the nested logistic regression models to identify the significance of multiplicative association. Here the Bonferroni method was used to adjust the p-value for the multiple interaction testing.

We assessed the possible confounding effects between the potential risk factors using stratified analysis. In the multivariable analysis, a confounding effect causes a distortion of the underlying association between the exposure of interest and MRSA acquisition. In order to investigate the association between the exposure of interest and the risk of MRSA acquisition by controlling for the possible confounding effect, the Cochran-Mantel-Haenszel (CMH) test was employed under the stratified-specific analysis using a 5% significance level. This assumes that there is no interaction. The CMH test is powerful since it specifies the alternative hypothesis and excludes the presence of interaction. Specifically, the data are stratified into a series of strata, each of which contains individuals that share common values of all the relevant confounders. For each strata, the data on MRSA acquisition and exposure of interest are displayed as a two-by-two table and thus the association between MRSA acquisition and the exposure of interest can be measured by the stratum-specific odds ratio. In this test, we assumed that there is no interaction effect. By applying the CMH test on the stratified data, the null hypothesis which states that the MRSA acquisition and the exposure of interest are independent, controlling for the possible confounding effects (i.e. the stratum-specific odds ratios $OR_1 = OR_2 = \cdots = OR_i = 1$) against the alternative hypothesis which is that the $OR_1 = OR_2 = \cdots = OR_i \neq 1$ can be tested.

Under the assumption that there is no striking interaction associated with the strata variables, the adjusted odds ratio for MRSA acquisition associated with the individual exposure of interest can also be estimated by using the Mantel-Haenszel (MH) method which provides the average odds ratio estimates across strata [62]. Comparing the adjusted odds ratio with the corresponding crude odds ratio means that we are able to investigate the significance of any potential confounding bias.

Furthermore, we were interested in investigating the causal link between the exposure of interest and MRSA acquisition by using the trend test with 5% significance level. A significant result suggests a linear trend in the risk of MRSA acquisition as the exposure of interest increases or decreases. Moreover, we also used ordered variables in the logistic regression to investigate possible nonlinear trends for the potential risk factors.

Backward variable selection for all potential risk factors was applied to construct the multivariable logistic regression using a 5% significance level.

However, the logistic regression based upon the categorical variables may have disadvantages of an inevitable loss of information and power. Using categorisation in the analysis causes underestimation of the extent of the variation in risk [103]. Hence we also used a linear logistic regression model based upon the numeric variables. A generalised additive model was also used based upon penalised regression smoothers for the purpose of investigating any nonlinear association between MRSA acquisition and potential quantitative risk factors. This model allows a rather flexible specification of the dependence of the MRSA acquisition and numeric exposure variables [136]. A final numeric multivariable numeric logistic model was constructed based upon backward selection.

In the end, the categorical logistic model was compared with the corresponding numeric logistic model. We divided the population in the admission-discharge cohort into two groups. One is a control group, including two thirds of the population which were selected at random and the other one is a test group, which has one third of the population and was also randomly selected. Then we fitted the categorical logistic model to the control group and applied this model to the data of the test group, yielding predicted responses. The sensitivity and specificity of the fitted categorical model was calculated by comparing the predicted responses to the observed ones. In this way, a Receiver Operating Characteristic (ROC) curve was plotted for the model. The accuracy of the model was then determined from the area under the curve (AUC). Similarly, we fitted the linear logistic model to the data of the control group and predicted the responses of MRSA acquisition in the test group. A ROC curve was also plotted based upon the corresponding sensitivity and specificity. Thus the AUC was obtained to investigate the accuracy of the model. Eventually, a more reliable model can be detected by comparing the AUCs. The greater the AUC the better the model is for prediction. In this chapter, we used a bootstrap test to compare the AUCs from the categorical and linear models statistically. This test investigated the null hypothesis that there is no difference between those two AUCs.

6.3 Univariate risk factor analysis.

In total, 2,724 patients in the admission-discharge cohort were included in the MRSA acquisition analysis. The measurements for the MRSA results on admission and on discharge show that there were 34 patients (1.25%), who were MRSA negative on admission but MRSA positive on discharge. In this section, we aim to not only assess the association between the number of wards and MRSA acquisition but also to identify the other significant risk factors associated with MRSA acquisition.

6.3.1 The univariate analysis of the categorised variable of number of wards associated with MRSA acquisition.

The histogram of the number of wards (shown in Figure 6.1) shows that the number of patients decreases exponentially as the number of wards that the patients had moved through increases. The majority of patients stayed only in one ward while in hospital. The frequency of patients who had stayed in more than three wards is low.

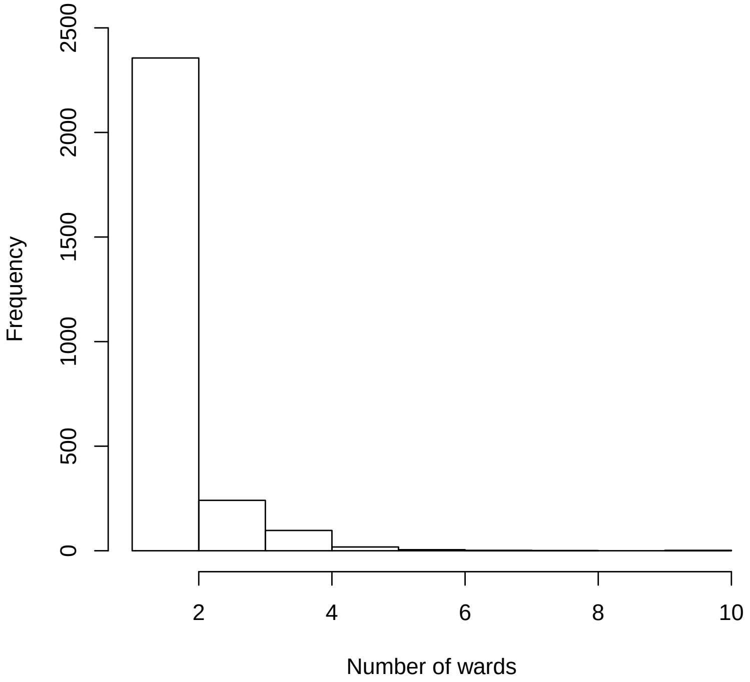


Figure 6.1: The barchart of number of wards patients had stayed in during their stay in hospital.

In order to understand the trends associated with the number of wards, we categorise the variable into three levels: one ward, two wards and three or more wards for application of the univariate analysis as well as the multivariable risk factor analysis in latter sections.

Table 6.1 illustrates the number of patients who acquired MRSA while in hospital for each category of number of wards. It shows that the percentage of the patients acquiring MRSA increases as the number of wards increases from one ward to three or more wards.

Table 6.1: Table of MRSA acquisition for patients with different number of wards.

Number of wards	MRSA acquisition		
	0	1	Total
1 ward	1,518 (99.09%)	14 (0.91%)	1532
2 wards	812 (98.54%)	12 (1.46%)	824
≥3 wards	358 (97.81%)	8 (2.19%)	366
Total	2,686	34	

In addition, for each category of the number of wards, the vast majority of the patients did not acquire MRSA while in hospital.

Table 6.2: Univariate risk analysis for the categorised number of wards.

Risk factor	Categories	OR	p-value	95% CI	Combined p-value (Wald test)
Number of wards	1 ward	1			
	2 wards	1.60	0.23	(0.74,3.48)	
	≥ 3 wards	2.42	0.05	(1.00,5.82)	0.13

Applying the univariate risk analysis shows that a patient who moved between more wards has an increased risk of acquiring MRSA while in hospital in Table 6.2. Specifically, a patient who had moved through three or more wards has about two and a half times as high a risk of acquiring MRSA compared to a patient who remained in one ward during their stay in hospital. The odds ratio for MRSA acquisition for the patients staying in two wards is 1.60 in comparison with the patients staying in one ward with p-value 0.23 which indicates that the risk of MRSA acquisition for the patients staying in two wards is not significantly different from the risk of MRSA acquisition for the patients staying only in one ward in hospital. The overall effect of the categorised number of wards is not strongly associated with MRSA acquisition due to the slightly large p-value (0.13 using the Wald test). However, the trend test with p-value=0.039 suggests that the risk of MRSA acquisition increases as the number of wards increases.

6.3.2 The univariate analysis of other categorised variables associated with MRSA acquisition.

The results of the other potential risk factors for MRSA acquisition based on the univariate logistic regression are shown in Table 6.3.

Table 6.3: Univariate risk factor analysis for MRSA acquisition ($N = 2,724$).

Variables	Categories	OR	p-value	95% CI	Combined p-value (Wald test)
Gender	Male	1			na
	Female	1.18	0.75	(0.57,2.20)	na
Age (years)	≤ 49	1			
	$50 - 64$	2.35	0.30	(0.47,11.68)	
	$65 - 79$	4.83	0.04	(1.10,21.18)	
	≥ 80	9.99	0.003	(2.20,45.34)	0.003
Discharge specialty	Medicine	1			
	A&E	1.52	0.70	(0.19,12.13)	
	Cardiology	0.53	0.41	(0.11,2.42)	
	Care of the elderly	2.42	0.26	(0.52,11.37)	
	Oncology	1.06	0.95	(0.13,8.46)	
	Orthopedics	0.41	0.11	(0.14,1.22)	
	Nephrology	1.73	0.48	(0.37,8.06)	
	Surgery	0.75	0.51	(0.32,1.77)	0.46
Length of stay	1 night	1			
	2-3 nights	0.61	0.52	(0.14,2.73)	
	4-7 nights	1.92	0.31	(0.55,6.73)	
	≥ 8 nights	2.34	0.19	(0.66,8.26)	0.088
Patient has been isolated	No	1			
	Yes	0.51	0.53	(0.06,4.16)	na
Co-morbidity: diabetes	No	1			
	Yes	0.85	0.76	(0.30,2.43)	na
Co-morbidity: COPD	No	1			
	Yes	1.80	0.22	(0.69,4.70)	na
Co-morbidity: wounds/ulcers	No	1			
	Yes	3.04	0.01	(1.31,7.08)	na
Co-morbidity: renal failure	No	1			
	Yes	4.58	0.006	(1.57,13.33)	na
Antibiotic use in year prior to discharge	No	1			
	Yes	1.62	0.17	(0.81,3.25)	na
Decolonisation treatment on discharge	No	1			
	Yes	4.04	0.18	(0.53,31.05)	na

There are four out of eleven potential risk factors showing potentially significant association with MRSA acquisition in hospital (i.e. age, length of stay, open wounds or ulcers and renal failure) with the corresponding p-values not larger than the significance level 0.1. The results in Table 6.3 imply that age has a strong association with MRSA acquisition due to a low overall p-value (i.e $p = 0.003$ using a Wald test). Clearly, elderly patients have a higher risk of MRSA acquisition compared to younger patients. The crude odds ratios increase as age group increases. Patients aged over 64 years old have significantly higher risk of acquiring MRSA compared to the patients aged 49 or under with the corresponding p-values less than 0.05. For example, patients 80 years old and over were around 10 times as likely to acquire MRSA during their stay in hospital compared to the younger patients aged 49 or under. On the other hand for the patients between 50-64 years old, the risk of acquiring MRSA is 2.35 times as high as the patients aged 49 years or under. From a statistical point of view, the risk of MRSA acquisition for the patients aged 50-64 years old is not significantly different from the risk of MRSA acquisition for the patients aged 49 years old or under due to the relatively high p-value.

The risk of MRSA acquisition changes as the length of stay changes and the corresponding overall p-value=0.088 from the Wald test. The unadjusted odds ratios for the length of stay indicate that patients staying for over 8 nights have 2.34 times as high a risk of MRSA acquisition compared to the patients staying only for one night. On the other hand, the risk of acquiring MRSA for the patients staying for two to three nights is 0.61 times the risk of MRSA acquisition for the patients staying for one night, though the confidence interval is wide.

We can see from Table 6.3 that the patients with self-reported open wounds or ulcers were almost three times more likely to acquire MRSA than the patients with intact skin ($p = 0.01$). Furthermore, the risk of MRSA acquisition for a patient with renal failure is more than four times that of a patient without renal failure. Thus, the potential risk factor of renal failure is also considered as a highly significant risk factor associated with MRSA acquisition (with p-value=0.009).

6.4 Association between potential risk factors.

In this section, we aim to investigate all the pairwise associations between the number of wards and the important risk factors from Table 6.3 which are age, length of stay, open wounds or ulcers and renal failure. The results are illustrated by two-way tables in Table 6.4.

Table 6.4: Two-way tables of association between potential risk factors.[1]

		Age				p-value (χ^2-test)
		≤ 49	50-64	65-79	≥ 80	
Number of wards	1 ward	398 (64.61%)	475 (60.13%)	513 (53.05%)	146 (41.83%)	< 0.001
	2 wards	162 (26.30%)	207 (26.20%)	315 (32.57%)	140 (40.12%)	
	≥ 3 wards	56 (9.09%)	108 (13.67%)	139 (14.38%)	63 (18.05%)	
Length of stay	1 night	127 (20.62%)	104 (13.16%)	101 (10.42%)	31 (8.88%)	< 0.001
	2-3 nights	229 (37.17%)	252 (31.90%)	247 (25.49%)	65 (18.62%)	
	4-7 nights	165 (26.79%)	260 (32.91%)	344 (35.50%)	119 (34.10%)	
	≥ 8 nights	95 (15.42%)	174 (22.03%)	277 (28.59%)	134 (38.40%)	
Wounds/ ulcers	Yes	70 (11.76%)	53 (6.94%)	66 (7.03%)	22 (6.55%)	0.002
	No	525(88.24%)	711 (93.06%)	873 (92.97%)	314 (93.45%)	
Renal failure	Yes	14 (2.36%)	21 (2.78%)	27 (2.90%)	15 (4.52%)	0.300
	No	580 (97.64%)	734 (97.22%)	903 (97.10%)	317 (95.48%)	

Continued on the next page

[1] Note that the percentage in the table is the percentage of total in the column.

Table 6.4-continued from previous page

		Length of stay				p-value (χ^2-test)
		1 night	2-3 nights	4-7 nights	≥ 8 nights	
Number of wards	1 ward	317 (87.57%)	540 (68.09%)	461 (51.97%)	214 (31.47%)	< 0.001
	2 wards	44 (12.15%)	237 (29.89%)	315 (35.51%)	228 (33.51%)	
	≥ 3 wards	1 (0.28%)	16 (2.02%)	111 (12.52%)	238 (35.00%)	
Wounds/ ulcers	Yes	30 (8.57%)	53 (6.94%)	67 (7.78%)	61 (9.26%)	0.425
	No	320 (91.43%)	711 (93.06%)	794 (92.22%)	598 (90.74%)	
Renal failure	Yes	8 (2.30%)	21 (2.76%)	22 (2.58%)	26 (4.01%)	0.314
	No	340 (97.70%)	739 (97.24%)	832 (97.42%)	623 (95.99%)	
		Number of wards				
		1 ward	2 wards	≥ 3 wards		
Wounds/ ulcers	Yes	131 (8.82%)	47 (5.92%)	33 (9.30%)		0.033
	No	1,354 (91.18%)	747 (94.08%)	322 (90.70%)		

Continued on the next page

Table 6.4-continued from previous page

		Number of wards				p-value (χ^2-test)
		1 ward	2 wards	$\geq$ 3 wards		
Renal failure	Yes	36 (2.43%)	22 (2.81%)	19 (5.44%)		0.011
	No	1,444 (97.57%)	760 (97.19%)	330 (94.56%)		
		Wounds/ulcers				
		Yes	No			
Renal failure	Yes	17 (8.10%)	193 (2.51%)			< 0.001
	No	193 (91.90%)	2,333 (97.49%)			

The two-way tables of the potential risk factors (in Table 6.4) show that age is strongly associated with length of stay, number of wards and open wounds or ulcers whereas the association between age and renal failure is not significant. It is clear that the elderly patients visited a larger number of wards in hospital. The percentage of the patients 80 years or over moving through three or more wards (18.05%) is almost double the percentage for the patients 49 years or under who had moved through three or more wards (9.09%). Thus the elderly patients tend to move through a relatively large number of wards. A similar conclusion is reached using length of stay. 38.40% of patients aged over 80 are in hospital for eight days or more, compared to 15.42% of those under 50. Considering the association between open wound or ulcers and age of patients, the results in Table 6.4 demonstrate that a higher percentage of patients aged 49 years or under have open wounds or ulcers compared to the patients 50 years or over. Although compared to the young patients, elderly patients have a slightly higher proportion who have renal failure on admission to hospital, there is no statistical association between age and renal failure due to the high p-value of the χ^2-test.

As expected, the results show that the length of stay is highly associated with the number of wards. To be more specific, a patient who had stayed in hospital for a long time had usually moved through a relatively large number of wards. It is clear in the two-way table that the percentage of the patients staying for eight nights or more and moving through three or more wards is significantly higher than the percentage of the patients staying for only one night and moving through three or more wards. Note that one patient is reported as being in three wards yet spending only one night in hospital. This is possible but very unusual. On the other hand, with respect to both co-morbidity risk factors: open wounds or ulcers and renal failure, the tests have high p-values which means that from a statistical point of view there is no significant association between length of stay and open wounds or ulcers or renal failure.

The association between the number of wards and the risk factors of open wounds or ulcers and renal failure have small p-values in Table 6.4, which implies that the number of wards is strongly associated with open wounds or ulcers and it is also significantly associated with renal failure. The majority of the patients did not have renal failure when they were admitted into the hospitals, but the percentage of patients who had renal failure increases as the number of wards increases. Although the proportion of patients with open wounds or ulcers does not show a straightforward increasing trend as the number of wards that the patients had moved through increases, the p-value for the χ^2-test gives evidence that the number of wards is associated with open wounds or ulcers.

The analysis also indicates that there is convincing evidence of a high association between renal failure and open wounds or ulcers due to the corresponding low p-value. Generally speaking, a patient with renal failure has a high probability of also having open wounds or ulcers since the percentage of patients with a wound or ulcer who also have renal failure (8.10%) is much higher than the percentage of patients without open wounds having renal failure (2.51%).

In this section, we analysed the pairwise association between the potential risk factors. The strong associations between some of the risk factors suggest potential confounding issues in the multivariable analysis. In the next section, we will investigate this problem.

6.5 Analysis of potential confounding effects and interactions.

There is a strong association between the age of patients and MRSA acquisition (based on the results of the univariate risk factor analysis in Section 6.3) and high correlations between age and the number of wards, length of stay and open wounds or ulcers respectively according to the small *p*-values (<0.05) in the two-way tables (shown in Table 6.4). Consequently we investigate age as a potential confounder in the multivariable risk factor analysis.

In this section, we go through the steps of the analysis of the possible confounding effect of age associated with the number of wards on MRSA acquisition as an example. The details of the analysis for the other risk factors are shown in Appendix B.

Table 6.5: Stratified risk analysis of number of wards by different age groups.

		MRSA acquisition			
Age	No. of wards	No	Yes	OR	*p*-value
≥80 years old	1 ward	143 (97.95%)	3 (2.05%)	0.24	0.058
	2 ward	137 (97.86%)	3 (2.14%)	0.25	0.067
	≥3 wards	58 (92.06%)	5 (7.94%)	1	
65-79 years old	1 ward	504 (98.25%)	9 (1.75%)	2.46	0.394
	2 ward	310 (98.41%)	5 (1.59%)	2.23	0.467
	≥3 wards	138 (99.28%)	1 (0.72%)	1	
50-64 years old	1 ward	473 (99.58%)	2 (0.42%)	0.45	0.519
	2 ward	204 (98.55%)	3 (1.45%)	1.57	0.696
	≥3 wards	107 (99.07%)	1 (0.93%)	1	
≤49 years old	1 ward	398 (100%)	0 (0%)	0	0.994
	2 ward	161 (99.38%)	1 (0.62%)	0.34	0.450
	≥3 wards	55 (98.21%)	1 (1.79%)	1	

To investigate the possible confounding effect of age, we stratify the association between MRSA acquisition and the number of wards according to the age groups and assess the stratified-specific associations. However a small number of patients in a stratum will make the interpretation difficult. In Table 6.5, there is no patient recorded in the dataset who was 49 years old or under and who had stayed in one ward and acquired MRSA. Hence we use three or more wards as the baseline in each age group stratum to estimate the stratified-specific odds ratios for MRSA acquisition. The results show that the elderly patients (80 years or over) are more likely to acquire MRSA if they moved through a large number of wards. The strata-specific odds ratio for one ward is 0.24 with p-value 0.058 where three or more wards is treated as a baseline in the stratum of age of 80 years old or older and the stratum-specific odds ratio for two wards in the stratum of age of 80 years old or older is 0.25 with p-value 0.067. The p-value (0.08) using the Wald test, which exceeds the significance level 0.05, implies that the effect of the number of wards among the patients 80 years old or over is not significantly associated with MRSA acquisition.

For the patients aged between 65 and 79, the number of wards is not associated with MRSA acquisition since the p-value of the Wald test is high (0.7). The stratum-specific odds ratio for two wards is 2.23 with p-value 0.467 while the stratum-specific odds ratio for one ward is 2.46 with the p-value 0.467, both of which imply that although the risk of MRSA acquisition for the patients aged 65-79 who stayed in two wards or less is more than two times as high as the risk of MRSA acquisition for patients aged 65-79 staying in three wards or more, there is no statistically significant difference between the effect of patients staying in two or less wards and the effect of patients staying three or more wards associated with MRSA acquisition.

For the patients whose age is 50-64 years, the combined p-value 0.40 (Wald test) indicates that the effect of number of wards is not significant on MRSA acquisition.

Since there is no record of a patient aged 49 years old or under who stayed only in one ward and acquired MRSA in hospital in the dataset, the stratum-specific odds ratio for MRSA acquisition associated with one ward in the stratum of age of 49 or under has no meaningful assessment.

In Table 6.5, we can see that the application of the stratified analysis makes the sample size in each stratum of age become relatively small. In particular, the number of patients who acquired MRSA in hospital within each stratum of age is quite small. In this situation, the exact logistic regression is capable of giving more reliable estimates of the stratum-specific odds ratios as well as the corresponding p-values. The results obtained from exact logistic regression under the stratified analysis are shown in Table 6.6.

Table 6.6: Stratified risk analysis of number of wards by different age groups using exact logistic regression.

Age	No. of wards	MRSA acquisition		$\mathrm{OR}_{\mathrm{LogXact}}$	p-value (LogXact)
		No	Yes		
≥80 years old	1 ward	143 (97.95%)	3 (2.05%)	0.25	0.052
	2 ward	137 (97.86%)	3 (2.14%)	0.26	0.099
	≥3 wards	58 (92.06%)	5 (7.94%)	1	
65-79 years old	1 ward	504 (98.25%)	9 (1.75%)	2.51	0.329
	2 ward	310 (98.41%)	5 (1.59%)	2.27	0.240
	≥3 wards	138 (99.28%)	1 (0.72%)	1	
50-64 years old	1 ward	473 (99.58%)	2 (0.42%)	0.47	0.484
	2 ward	204 (98.55%)	3 (1.45%)	1.58	1.00
	≥3 wards	107 (99.07%)	1 (0.93%)	1	
≤49 years old	1 ward	398 (100%)	0 (0%)	0.14	0.12
	2 ward	161 (99.38%)	1 (0.62%)	0.32	0.463
	≥3 wards	55 (98.21%)	1 (1.79%)	1	

Compared to the results in Table 6.5, the stratum-specific odds ratios obtained by exact logistic regression are approximately the same as the stratum-specific odds ratios obtained by the logistic regression. Generally speaking, the exact logistic regression can generate more reliable estimates of stratum-specific odds ratios in stratified analysis but the conclusions are not different compared to the ones obtained by the logistic regression.

Now we investigate the multiplicative interactions between age and number of wards associated with MRSA acquisition based upon the stratified analysis shown in Table 6.5. In order to use Woolf's method to detect the homogeneity of the stratum-specified odds ratio across the strata, we dichotomise the number of wards into two groups which are one ward and two or more wards. The high p-value (0.176) suggests that the stratum-specific odds ratios for MRSA acquisition associated with number of wards are consistent across the strata, i.e. age does not modify the effect of number of wards on MRSA acquisition. This conclusion validates the assumption of the CMH test which will be used later. An alternative test, which is the likelihood ratio test, for investigating the multiplicative interactions will also be used in Section 6.6.1.

We use the CMH method to assess the independence between the number of wards and MRSA acquisition, controlling for the potential confounding effect of age. Then based on the MH method we are able to estimate the adjusted odds ratio for MRSA acquisition across the strata of age. The results are shown in Table 6.7, where the category of one ward is treated as the baseline in the analysis.

Table 6.7: The estimation of average odds ratio for the number of wards, stratified by age.

	p-value (CMH method)	OR_{MH}	95% CI
2 wards vs. 1 ward	0.43	1.36	(0.63,2.94)
$\geq$3 wards vs. 1 ward	0.16	1.89	(0.78,4.62)

The results in Table 6.7 show that the stratum-specific odds ratios for MRSA acquisition in patients staying in two wards relative to the patients staying in one ward are equal to one in a consistent manner across the different age groups due to the high p-value which is 0.43. By controlling for the possible confounding effect of age, the risk of MRSA acquisition for the patients staying in two wards is 1.36 times as high as the risk of MRSA acquisition for the patients staying in one ward. Compared to the crude odds ratio in the univariate analysis which is 1.60 shown in Table 6.2, this adjusted odds ratio decreases by 15%. The adjusted odds ratio for MRSA acquisition associated with three or more wards in the comparison with one ward indicates that the risk of MRSA acquisition for patients staying in three or more wards is 1.89 times as high as the risk of MRSA acquisition for the patients staying in one ward but the corresponding p-value 0.16 indicates that the effect is not statistically related to MRSA acquisition. Compared to the corresponding crude odds ratio shown in Table 6.2 (2.42), the adjusted odds ratio decreases by 21.9%. This indicates that there is a confounding effect of age associated with number of wards, influencing MRSA acquisition.

In this subsection, we analysed the multiplicative interaction between age and the number of wards associated with MRSA acquisition. We also investigated the potential confounding effect of age associated with the number of wards on MRSA acquisition. The results showed that there is no multiplicative interaction between age and the number of wards, and the confounder, age, is significantly associated with the number of wards and also influences MRSA acquisition. The analysis of the other potential confounding variables, such as number of wards, length of stay, open wounds or ulcers and renal failure, and their possible interaction effects are demonstrated in Appendix B.

Briefly, the results in Appendix B show that there is no multiplicative interaction between the pairwise potential risk factors of age, number of wards, length of stay, open wounds or ulcers and renal failure. Those risk factors can be considered as potential confounders since they are potentially causally related to MRSA acquisition. Although age is not likely to be causal, it is associated with an impaired immune system among older patients. Similarly for renal failure. Length of stay and number of wards both have positive correlation with the risk of acquiring MRSA. Open wounds or ulcers might also be causal. The analysis of the potential confounders implies that age is an actual confounder which has impact on the effect of length of stay and open wounds or ulcers. Furthermore, the number of wards also has a confounding effect associated with age and the length of stay. The length of stay has a confounding effect associated with age and the number of wards. The risk factors of open wounds or ulcers and renal failure are mutually confounded.

6.5.1 The analysis of the trend test for the risk of MRSA acquisition adjusting for confounding factors.

In the previous subsection, we did not develop a detailed understanding of how the risk of MRSA acquisition changes over the exposure levels. In epidemiological studies, it is common that an exposure variable has a natural ordering. In this subsection, we aim to investigate whether the risk of MRSA acquisition increases or decreases as the level of exposure of interest increases.

In order to detect a trend in the risk of MRSA acquisition with respect to the categorical risk factors which have natural orders, we use the trend test under the assumption that there is no interaction between age and the number of wards. The trend test is adjusted by taking the possible confounding effects of another variable into account. The previous subsection shows that there is no interaction, which indicates that if the trend in risk exists, it is considered to be consistent across all strata. In this subsection, we use the investigation of an increase or a decrease in the risk of MRSA acquisition as the number of wards increases as an example. Obviously, in order to avoid the confounding effect of age associated with the number of wards, the trend test is based on the stratified analysis. We assign the values of 0, 1 and 2 to the ordered categories of the qualitative variable of number of wards (i.e. '1 ward', '2 wards' and '$\geq$ 3 wards'). According to the quantities in Table 6.5, the trend test based upon the logistic regression with age as a categorical variable and number of wards using 0, 1, 2 yields a p-value of 0.13, indicating that there is no evidence that the risk of MRSA acquisition has an increasing risk as the number of wards increases adjusted for age.

The details of the trend tests for the other potential exposure variables, which take the corresponding confounding effects into account, are demonstrated in Appendix B. The results show that the risk of MRSA acquisition increases as the level of age increases.

6.6 Multivariable logistic regression model.

Based on the results of the previous univariate analysis, we now aim to construct a multivariable model to describe the pattern of how the risk of MRSA acquisition changes when accommodating all the potential risk factors. Firstly, we investigate the potential interactions between potential risk factors for inclusion using the likelihood ratio test and compare the results with the ones obtained above by the Woolf method. Then two multivariable logistic models, one of which involves the categorical exposure variables and the other involves the numeric variables respectively, are constructed to relate the risk factors to MRSA acquisition. From mathematical point of view, the multivariable logistic regression model can be expressed as $\log(\frac{p}{1-p}) = a + \mathbf{bx}$, where p denotes the risk of acquiring MRSA in hospital and $\mathbf{x}$ denotes the risk factors related to MRSA acquisition. In this section, we focus on the analysis of the categorical multivariable model, in which the association between risk factors and MRSA acquisition is easy and straightforward to interpret.

6.6.1 Testing for interactions using likelihood ratio test.

We use the likelihood ratio test to test all the plausible interactions between the potential risk factors based upon the multivariable logistic model. The Bonferroni method is used to adjust the p-value for multiple testing. Equivalently, in this study, the adjusted significance level becomes $\frac{0.05}{10} = 0.005$ since there are five potential risk factors involved in the multivariable analysis and thus there are ten possible interaction terms. The results in Table 6.8 show that there is no pairwise interaction between the potential risk factors though the interaction between age and length of stay has a small p-value.

For example, the likelihood ratio test for the interaction between the length of stay and age (i.e. the product of length of stay and age) gives a p-value (0.013) which is larger than the adjusted significance level 0.005, indicating that the combined effect of age and length of stay does not have a significant association with MRSA acquisition in

Table 6.8: The likelihood ratio tests for plausible interactions.

Interaction	p-value
age×length of stay	0.013
age×number of wards	0.414
age×wounds/ulcers	0.438
age×renal failure	0.385
length of stay×number of wards	0.546
length of stay×wounds/ulcers	0.837
length of stay×renal failure	0.962
number of wards×wounds/ulcers	0.974
number of wards×renal failure	0.637
renal failure×wounds/ulcers	0.996

the multivariable logistic model, with Bonferroni adjustment. Similarly, the p-values of the likelihood ratio tests for the interactions between the number of wards and age, the length of stay, open wounds or ulcers and renal failure are 0.414, 0.546, 0.837 and 0.962 respectively, meaning that none of the p-values for the interactions reach the adjusted significance level (0.005). Hence this illustrates that there is no significant effect of the interactions between the number of wards and other potential risk factors associated with the MRSA acquisition. The conclusion of non-significant interactions is the same as the result we obtained by the Woolf method in the previous sections.

6.6.2 Multivariable analysis for the categorical model.

Now we construct the categorical multivariable model, taking all the significant categorical exposure variables into account. According to the analysis of the univariate logistic regression illustrated in Table 6.3 in Section 6.3, five potential risk factors with significant results are considered to be included into the multivariable logistic model, which are age, length of stay, number of wards, open wounds or ulcers and renal failure. Note that the number of wards is included in the multivariable analysis since one of the main objectives of the analysis presented in this chapter is to assess the association between number of wards and MRSA acquisition adjusting for the other risk effects.

Table 6.9: Multivariable analysis of risk factors for MRSA acquisition.

Risk factor	Categories	Adjusted OR[2]	p-value	95% CI	Combined p-value (Wald test)
Age	≤49 years	1			0.013
	50-64 years	2.30	0.31	(0.46,11.51)	
	65-79 years	4.47	0.05	(1.01,19.88)	
	≥ 80 years	8.39	0.01	(1.80,39.18)	
Length of stay	1 night	1			0.47
	2-3 nights	0.58	0.48	(0.13,2.63)	
	4-7 nights	1.41	0.61	(0.38,5.22)	
	≥ 8 nights	1.34	0.68	(0.34,5.27)	
Number of wards	1 ward	1			0.75
	2 wards	1.26	0.58	(0.56,2.86)	
	≥3 wards	1.44	0.47	(0.53,3.87)	
Wounds/ulcers	No	1			
	Yes	2.89	0.02	(1.18,7.10)	
Renal failure	No	1			
	Yes	2.98	0.06	(0.94,9.47)	

As we can see from Table 6.9, there are three risk factors: length of stay, number of wards and renal failure becoming nonsignificant in the multivariable logistic model. This may be caused by the confounding effects. Taking the effect of length of stay, number of wards, open wounds or ulcers and renal failure into account, the adjusted odds ratio for MRSA acquisition increases significantly as age increases from 49 or under to 80 years or over. The adjusted effect of age remains significant (since the adjusted p-value< 0.05).

[2] the odds ratio by taking the other risk factors into account

However, the adjusted odds ratio for the patients aged 80 years old or over declined to 8.39 whereas the crude odds ratio for the patients aged 80 years old or over is 9.99. This can be explained by the confounding effect of length of stay which was demonstrated in Section 6.5. Compared to the results in the univariate analysis (shown in Table 6.3), the adjusted odds ratios for the length of stay change dramatically as well. Specifically, the adjusted odds ratio for the patients staying for eight nights or over becomes much lower than the corresponding crude odds ratio in the univariate analysis. Moreover, the adjusted p-value of the Wald test for the length of stay increases from 0.088 in the univariate analysis to 0.47 in the multivariable analysis, implying that the positive effect of the length of stay on MRSA acquisition disappears with the inclusion of the effects of the other risk factors, principally age. The adjusted risk of MRSA acquisition increases when the number of wards that the patients stayed in increases. The corresponding adjusted p-value for number of wards in the multivariable logistic model means that the adjusted effect of number of wards is non-significant. Compared to the unadjusted odds ratios for MRSA acquisition associated with the number of wards in Table 6.2, the adjusted one decreases significantly. This indicates that some of the other exposure variables affect the association between number of wards and MRSA acquisition. Similarly, the adjusted p-values (< 0.05) for open wounds or ulcers and renal failures indicate that there remain strong associations between open wounds or ulcers/renal failure and MRSA acquisition adjusting for the effects of other risk factors. Taking all putative risk factors into account, the effect of renal failure on the acquisition of MRSA becomes weaker. This can be seen from the significant decreases in the adjusted odds ratio (2.98) compared to the unadjusted odds ratio for renal failure (4.58 shown in Table 6.3).

Then we derive a new model where length of stay is excluded since this variable is not significant in the multivariable model. Number of wards is retained at this time as it is the primary research question in this chapter.

Comparing the adjusted odds ratios for the remaining risk factors in the new fitted model in Table 6.10 to the corresponding adjusted odds ratios in the full model in Tables

Table 6.10: Multivariable analysis of nested model without length of stay.

Risk factor	Categories	Adjusted OR	p-value	95% CI	Combined p-value (Wald test)
Age	≤49 years	1			
	50-64 years	2.43	0.28	(0.49,12.61)	
	65-79 years	4.93	0.03	(1.12,21.77)	
	≥ 80 years	9.50	0.004	(2.05,43.95)	0.0069
Number of wards	1 ward	1			
	2 wards	1.39	0.41	(0.63,3.08)	
	≥3 wards	1.82	0.19	(0.74,4.51)	0.41
Wounds/ulcers	No	1			
	Yes	3.04	0.01	(1.25,7.38)	
Renal failure	No	1			
	Yes	2.85	0.07	(0.90,8.97)	

6.9, we can see that both adjusted odds ratios for age and number of wards in the new model become larger when length of stay is excluded. This indicates that there is a possible confounding effect of length of stay on age and the number of wards (the details of the analysis are illustrated in Appendix B). However based on the Wald test, the adjusted effect of the number of wards in this multivariable model is not significant. We perform the χ^2 difference test to investigate the hypothesis of whether the coefficient of number of wards is zero in this model and it produces a high p-value (0.415) which indicates that it is reasonable to exclude the number of wards from this model.

Hence we generate a second new model which involves three only risk factors: age, open wounds or ulcers and renal failure. The results are presented in Table 6.11, which shows that the risk of MRSA acquisition increases significantly as age increases, by adjusting for the effects of open wounds or ulcers and renal failure. The risk of acquiring MRSA is more than five times higher for a patient above 64, compared to a patient who was younger than 50 years old. For a patient aged 80 years old or over, the risk of MRSA acquisition is more than ten times higher compared to the patients aged 49 years or under. In addition, patients with open wounds or ulcers are 3.02 times as likely to acquire MRSA compared to the patients without open wounds or ulcers and patients with renal failure have 3.17 times as high risk of MRSA acquisition compared to the patients without renal failure. Note that this is the same model as the one in the paper written by Velzen et al. [130].

Table 6.11: Multivariable analysis of nested model without length of stay and number of wards.

Risk factor	Categories	Adjusted OR	p-value	95% CI	Combined p-value (Wald test)
Age	$\leq$49 years	1			
	50-64 years	2.53	0.26	(0.51,12.63)	
	65-79 years	5.16	0.03	(1.17,22.76)	
	$\geq$ 80 years	10.57	0.002	(2.31,48.39)	0.0034
Wounds/ulcers	No	1			
	Yes	3.02	0.01	(1.25,7.26)	
Renal failure	No	1			
	Yes	3.17	0.04	(1.04,9.71)	

In this section, we constructed a categorical multivariable model to describe the pattern of MRSA acquisition related to the potential risk factors. Only age, open wounds or ulcers and renal failure remain significantly associated with MRSA acquisition. Age is a primary risk factor related to the risk of MRSA acquisition. Taking the other exposure variables into account, the older a patient is the higher the risk of acquiring MRSA while in hospital. There are no significant multivariable interactions between the potential risk factors based on the Bonferroni method. Now we are interested in investigating whether there is nonlinearity of the categorical variables associated with MRSA acquisition in the multivariable analysis, which will be assessed in the next section.

6.7 The analysis of trend of risk factors.

In order to investigate the nonlinear trend of the potential categorised risk factors, we use ordered variables in the logistic regression. Section 6.7.1 has the results from the analysis for the logistic regression model which involves linear, quadratic and cubic terms of the ordered exposure variables. Although the categorical logistic regression is easy and straightforward to interpret the relationship between the potential risk factors and the risk of MRSA acquisition, this may lead to an inevitable loss of information and power. Therefore, we use the numeric variables in the multivariable model to enhance the power of the analysis (i.e. increase the degrees of freedom in the full model) [41]. In addition, we will investigate the potential nonlinearity of the numeric variables related to the risk of MRSA acquisition, using the generalised additive models. The results of the investigation for the generalised additive model will be presented in Section 6.7.2.

6.7.1 The analysis of the logistic regression with ordered factors.

Table 6.12 has the univariate p-values of significance tests for the ordered risk factors. Generally speaking, the large p-values for the quadratic and cubic terms in the model of MRSA acquisition depending on the age illustrate that neither of the quadratic or cubic terms of age require to be included in the model. The p-value for the linear trend of the effect of age associated with MRSA acquisition is 0.0013, revealing that there is strong evidence that the linear trend of age is significant in the model of MRSA acquisition with regard to the effect of age.

Similarly, the linear trend of the length of stay is significant for the risk of MRSA acquisition but there is no evidence that quadratic or cubic terms of length of stay would be included in the model with respect to MRSA acquisition.

Moreover, by investigating the nonlinear trend for the number of wards in the model of MRSA acquisition, similar conclusions can be drawn. That is the linear trend of the number of wards as a p-value of 0.048 which is close to 0.05, whereas the quadratic term has a large p-value. Note that in Table 6.12, the results of nonlinear trend tests for open wounds or ulcers and renal failure are excluded since open wounds or ulcers and renal failure are binary variables ([0,1]).

The estimates from the multivariable logistic model including all the potential categorised risk factors which are ordered are presented in Table 6.13. There is convincing

Table 6.12: Univariate ordered risk factor analysis.

Risk factor	Tested trend	Coefficient	Standard error	p-value
Age (categorised)	Linear	1.705	0.529	0.0013
	Quadratic	-0.063	0.456	0.888
	Cubic	0.032	0.369	0.935
Length of stay (categorised)	Linear	0.827	0.450	0.067
	Quadratic	0.347	0.430	0.419
	Cubic	-0.582	0.408	0.153
Number of wards (categorised)	Linear	0.626	0.316	0.048
	Quadratic	-0.024	0.299	0.937

evidence that age shows a linear association with MRSA acquisition and the corresponding p-value is 0.0035. The p-values (>0.1) of the linear trend for both length of stay and number of wards suggest that neither are associated with MRSA acquisition adjusting for age, open wounds or ulcers and renal failure. There is not evidence that nonlinear terms in age are required and the linear logistic regression is adequate to model MRSA acquisition.

Table 6.13: Multivariable ordered risk factor analysis.

Risk factor	Tested trend	Coefficient	Standard error	p-value
Age	Linear	1.576	0.540	0.0035
Length of stay	Linear	0.395	0.495	0.425
Number of wards	Linear	0.256	0.358	0.474
Wounds/ulcers	Linear	1.063	0.456	0.02
Renal failure	Linear	1.092	0.590	0.06

In this subsection, we investigated the nonlinearity of the categorical variables associated with the risk of MRSA acquisition and the results showed that the effect of categorical variables with higher orders is not significant on the risk of MRSA acquisition. As we mentioned before, the numeric logistic regression model will increase the power of the analysis compared to the categorical model. Thus in the next subsection, we will investigate the nonlinearity of the numeric variables associated with the risk of MRSA acquisition and then construct the corresponding multivariable logistic model including the numeric exposure variables.

6.7.2 The analysis for the generalised additive model.

In the previous section, nonlinearity in terms of ordered categorical variables and quadratic or cubic effects was investigated. In order to investigate the nonlinearity of the numeric risk factors in more detail, generalised additive models were applied in this section. From mathematical point of view, the generalised additive model can be expressed as $\log(E(Y)) = \beta + \mathbf{f}(\mathbf{X})$, where Y is the response (i.e. MRSA acquisition), $\mathbf{f}$ is the smoothed functions and $\mathbf{X}$ is the risk factors related to the response.

Using the numeric age rather than the categorised age for each patient as a risk factor in the generalised additive model, the plot is displayed in Figure 6.2, where the solid line represents the estimated effects and the dashed lines are 95% confidence limits which is estimated by the Bayesian approach (i.e. strictly Bayesian credible intervals).

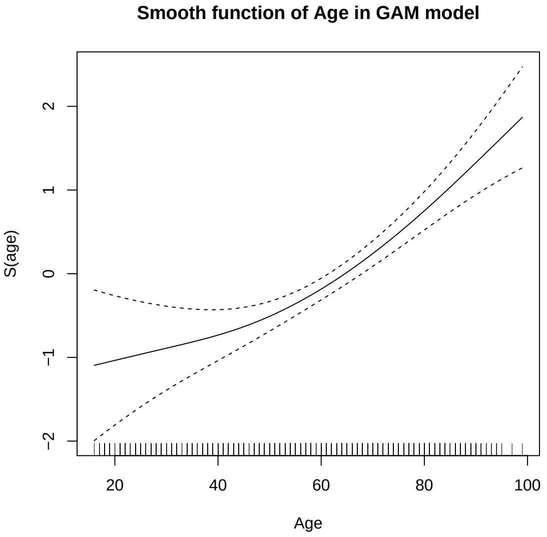


Figure 6.2: The smooth function of age in the generalised additive model.

Clearly, the effect of age is estimated as a smoothed curve with the corresponding effective degrees of freedom of 1.70 which reflects the flexibility of the fitted model and is determined by the smoothing parameter (i.e. the degree of smoothness) for the penalized regression spline of age. The 95% Bayesian confidence limits are wide on the tails since the number of relatively young and old patients is small. The penalised age is slightly curved.

Furthermore, we also investigated the adjusted effects of age, number of wards that the patients had moved through and the length of stay by applying a multivariable generalised additive model, where those three risk factors are all treated as numeric variables. The plots of the corresponding smooth functions of the three risk factors are presented in Figure 6.3.

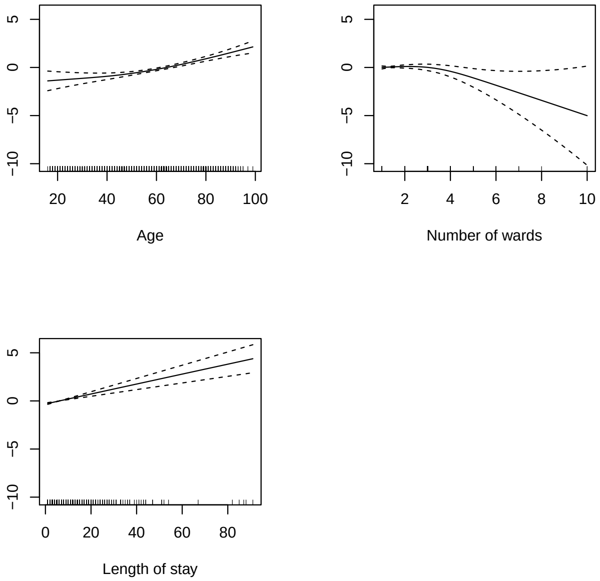


Figure 6.3: The smooth functions of age, number of wards, length of stay in the full multivariable generalised additive model, with associated 95% confidence limits.

According to the estimated results of the multivariable generalised additive model involving the five potential risk factors of age (as a numeric variable), number of wards (as a numeric variable), length of stay (as a numeric variable), open wounds or ulcers and renal failure, the adjusted effect of age on the risk of MRSA acquisition is estimated as a smooth curve with the effective degrees of freedom of 1.59. The adjusted effect of length of stay is estimated as a straight line related to the risk of MRSA acquisition, corresponding to one degree of freedom. In other words, length of stay shows a linear effect on MRSA acquisition, controlling for the effects of age, number of wards, open wounds or ulcers and renal failure. On the other hand, the adjusted effect of number of wards is estimated as a smooth curve, corresponding to 1.89 effective degrees of freedom, which indicates that the linearity of number of wards might not be adequate to model the risk of MRSA acquisition. The 95% confidence interval for the number of wards becomes wider when the number of wards increases. This is because the number of the patients who had stayed in a large number of wards is relatively small. The same phenomenon of the wide confidence interval in the right tail can be found in the plot of the length of stay due to the similar reason that the number of the patients who had stayed for a long period is small.

Then we use the likelihood ratio tests to investigate if the smoothed terms of age and the number of wards are necessary. For example, we compare the multivariable model including the penalized regression of the number of wards to the model including a linear term of the number of wards and a high p-value (0.326) for the χ^2 test means that there is no need to include a smoothed term of the number of wards in the multivariable model. Similarly, the test for the penalized regression of age has a p-value 0.21 for the nonlinear terms which suggests that there is no significant nonlinear effect of age on the risk of MRSA acquisition.

In this section, we used the logistic regression model with the ordered exposure variables to investigate the nonlinearity of the categorical variables and the results showed that the effects of the categorical variables of age, open wounds or ulcers and renal failure are significant on the risk of MRSA acquisition. Since using the categorical variables in the multivariable logistic model weaken the power of the analysis, we are interested in investigating the multivariable model including the numeric variables instead. The generalised additive model was used to investigate the nonlinearity of the numeric exposure variables in this section. We can conclude that a linear model is adequate to interpret the relationship between the potential risk factors and the risk of MRSA acquisition. Thus in the next section, we will construct a linear model including the numeric variables and assess the effects of those variables on the risk of MRSA acquisition.

6.8 The multivariable logistic analysis with the numeric risk factors.

Now we build the model which includes the linear terms of the numeric variables of age, number of wards and length of stay. According to the results in the previous section, a linear model is adequate to clearly demonstrate the effects of age, number of wards and length of stay associated with the risk of MRSA acquisition (shown in Table 6.14).

Table 6.14: Multivariable model with all numeric risk factors.

Risk factor	Adjusted OR	p-value	95% CI
Age (numeric)	1.05	0.00022	(1.02,1.08)
Number of wards (numeric)	0.87	0.507	(0.59,1.30)
Length of stay (numeric)	1.05	0.0062	(1.01,1.08)
Wounds/ulcers	2.49	0.054	(0.98,6.32)
Renal failure	2.51	0.128	(0.77,8.20)

Table 6.14 shows that age as a numeric variable is linearly related with MRSA acquisition and the risk of MRSA acquisition increases by a factor of 1.05 as the age of patients increases by one, controlling for the potential risk factors of number of wards, length of stay, open wounds or ulcers and renal failure. The adjusted odds ratio of the length of stay shows that the risk of MRSA acquisition increases by a factor of 1.05 as the length of stay increases by one. And the corresponding small p-value (0.0062) implies that the effect of the numeric length of stay is significant on the risk of MRSA acquisition. On the other hand, the adjusted effect of number of wards is not significant on the risk of MRSA acquisition due to the relatively high p-value (0.507).

In both the multivariable full model involving all the categorised ordered risk factors (shown in Table 6.13 in Section 6.7.1) with the corresponding multivariable full model with the numeric ones (which is shown in Table 6.14), the adjusted effect of age is significantly associated with MRSA acquisition. Furthermore there is no significant association with the number of wards adjusted for the other risk factors. The numeric multivariable full model shows significant effect of length of stay on the risk of MRSA acquisition whereas the adjusted effect of the length of stay is not significant in the corresponding categorical multivariable full model. One of the reasons is that the categorical variables lose some variability and conceal a certain information about the details of the data since this model assumes that the relation between MRSA acquisition and length of stay is constant within each category (i.e. any change in effect within a category will be lost) [41]. Generally speaking, it is adequate to use linearity of the risk factors to model the risk of MRSA acquisition. One of the disadvantages of the numeric multivariable logistic regression model is that the interpretation of the effects of the risk factors associated with MRSA acquisition is not quite as simple as it is for the categorical models.

By applying the backward selection using the 5% significant level, the results for the final numeric multivariable model which involves three risk factors of age, length of stay and open wounds or ulcers are shown in Table 6.15.

Table 6.15: Multivariable analysis of risk factors for MRSA acquisition involving numeric age and length of stay.

Risk factor	Categories	Adjusted OR	p-value	95% CI	Combined p-value (Wald test)
Age (numeric)		1.06	0.00015	(1.03,1.08)	
Length of stay		1.04	0.001	(1.02,1.06)	
Wounds/ulcers	No	1			
	Yes	3.07	0.012	(1.08,7.39)	

In Table 6.15, we can see that the effects of age, length of stay and open wounds or ulcers are all strongly associated with MRSA acquisition. Particularly, the risk of MRSA acquisition increases by a factor of 1.06 as each increment of age, after controlling for the effects of length of stay and open wounds or ulcers. The adjusted odds ratio for length of stay also demonstrates that the risk of MRSA acquisition increases by a factor of 1.04 as each increment of length of stay. For the patients with open wounds or ulcers, the risk of MRSA acquisition is 3.07 times as high as the risk of MRSA acquisition for the patients without open wounds or ulcers.

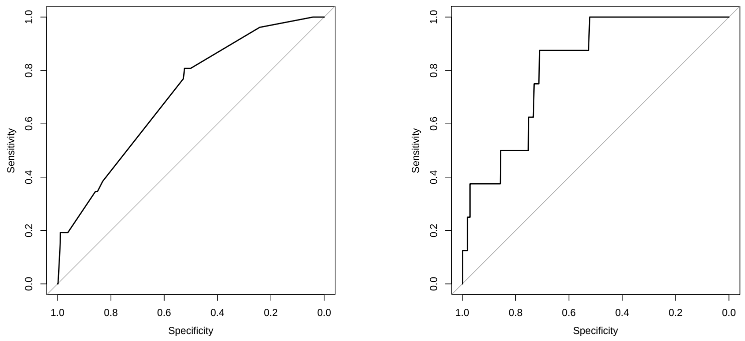
In this section, we constructed a numeric multivariable logistic regression model which included three numeric exposure variables of age, length of stay and open wounds or ulcers. The results showed that the risk of MRSA acquisition increases as the age and/or length of stay increases. The patients with open wounds or ulcers have higher risk of acquiring MRSA. Recall that the categorical multivariable model showed that three out of five potential risk factors, which are age, open wounds or ulcers and renal failure, are significant on the risk of MRSA acquisition. Now we are interested in assessing that which model is more reliable and predictive.

6.8.1 The comparison of the numeric model and the categorical model.

In this subsection, we compare the final categorical model shown in Table 6.11 and the final numeric model shown in Table 6.15. The sensitivity and specificity of each model is established using the model predictions and observations. Then the ROC curve can be plotted for each model, with the corresponding AUC calculated (shown in Figure 6.4). The AUC for the numeric model which is 0.812 is slightly greater than the AUC for the categorical model which is 0.788. However, comparing those two AUCs based upon a bootstrap test, the p-value (0.198) indicates that the AUC for the numeric model is not statistically greater than the AUC for the categorical model. This means that there is no convincing evidence that the numeric model is more reliable and predictive than the categorical one.

6.9 Conclusion.

In this chapter, we investigated the effects of the potential risk factors associated with MRSA acquisition in the admission-discharge cohort ($N = 2,724$) where 1.26% of the patients acquired MRSA while in hospital. The plausible interactions as well as the



(a) ROC curve for categorical model on the test cohort $N = 906$. The AUC is 0.788.

(b) ROC curve for numeric model on the test cohort $N = 906$. The AUC is 0.812.

Figure 6.4: The ROC curves for categorical and numeric models with the corresponding AUCs.

possible confounders were analysed. As a result, three categorical risk factors for acquiring MRSA are identified: age group, self reported open wounds or ulcers and self reported renal failure. Age group is strongly related to MRSA acquisition which is considered as the predominant risk factor. Elderly patients (65 years old or older) are more likely to acquire MRSA compared with the younger patients aged 49 years old or under. Patients with open wounds or ulcers are more likely to acquire MRSA than the patients without open wounds or ulcers when admitted into the hospital. Similarly, patients with renal failure are 4.58 times as likely to acquire MRSA in comparison with the patients without renal failure.

Generally speaking, the number of wards when treated as a grouped variable is not significantly associated with MRSA acquisition in the univariate analysis. However, the risk of MRSA acquisition is slightly higher for patients staying in three or more wards compared to the patients staying in one ward. By taking the effects of other risk factors into account, the grouped number of wards does not have a significant effect on MRSA acquisition.

The number of wards has strong associations with other four risk factors involving age, length of stay, open wounds and renal failure. Age as a dominant risk factor associated with MRSA acquisition is a confounder with respect to the other risk factors of length of stay, number of wards and open wounds or ulcers. The risk factors of open wounds or ulcers and renal failure are mutually confounding. By applying Woolf's test to investigate the consistency of the odds ratios across strata, the results showed that there was no multiplicative interaction among the pairwise potential risk factors. In addition, there is no obvious trend in the risk of MRSA acquisition as the number of wards increases.

Generalised additive regression models were also applied to generate the estimated effects of the risk factors which were modelled as a smooth line by penalized regression spline. The conclusion drawn from this analysis is that there are linear trends of exposure variables in the risk of MRSA acquisition. Hence, we used the linear numeric variables in constructing the logistic model for the risk of MRSA acquisition. In this model length of stay is associated with MRSA acquisition in contrast to the use of the corresponding categorical variable which may be easier to interpret. This illustrates the loss of the information in the data by the use of categorisation.

We constructed linear numeric multivariable model which included two numeric variables of age, length of stay and a binary variable, open wounds or ulcers. The results showed that the risk of MRSA acquisition increases as the age of a patient increases. The long length of stay may also increase the risk of MRSA acquisition and the patients with open wounds or ulcers have higher risk of acquiring MRSA. We compared this numeric model with the categorical multivariable model obtained before, the results showed that from a statistical point of view there is no obvious evidence that one of those two models is more reliable or predictive though the linear model had a higher area under the ROC curve.

Although the screening program is important to identify MRSA positive patients, the findings in this chapter provide evidence that the cross-transmission of MRSA still takes place in Scottish hospitals and hence implementing contact precaution and infection control in the hospital is also important to prevent the cross-transmission. This study can be considered to be representative of the general Scottish in-patient population but selection bias towards relatively healthier patients exists, which may lead to underestimation of the proportion of patients acquiring MRSA. For example, patients who died in hospital cannot be included in the admission-discharge cohort as they could not consent on discharge.

This study has some limitations. A prospective cohort study was undertaken to help determine risk factors for acquiring MRSA in hospital since the longitudinal observations were collected in a defined period. In this chapter, only the data in the admission-discharge cohort were used for the analysis since we needed both the MRSA measurements on admission and on discharge. This may lead to an underestimation of the proportion of patients acquiring MRSA. In this chapter, the multiple testing was addressed by adjusting the significance level using the Bonferroni method. A large number of multiple null hypotheses, for example, in the test of the potential pairwise interactions, means that this procedure will have little power.

Since there were only a small number of patients acquiring MRSA while in hospital (only 34 patients) compared to the study population, data have little power to identify which model is a better fit to the data and the power of the study is low.

Patient movement can be characterised by both the frequency of the movement and also by cohabiting where a patient is in a ward at the same time as there are MRSA patients in the same ward. In the next chapter, we focus on modelling the dynamic patient movements and assess the effect of being in a ward with other MRSA patients on the MRSA acquisition. Specifically, two main questions are assessed in the next chapter: (i) is there any evidence that the probability of acquisition of MRSA in hospitals is greater among patients who are transferred to or admitted to wards within the hospital where there are known patients who are MRSA positive (i.e. patients with MRSA colonisation or infection) compared to patients who are in wards with no known MRSA colonisation or infection; (ii) Is there any evidence that the probability of acquisition of MRSA in hospitals is affected by the duration of stay only in wards where MRSA is known to be present.

Chapter 7

Effect of Patient Movement between Wards on the Acquisition of MRSA.

7.1 Introduction.

In this chapter, we focus on the methodology of assessing the effect of patient movement in general hospitals on MRSA acquisition. The movement of patients with MRSA to other wards in the hospital and the movement of patients into wards where MRSA patients are already resident may bring hospital patients who do not have MRSA into closer proximity to patients who carry MRSA. In other words, we aim to investigate the effect of patient movement on the risk of a patient acquiring MRSA in hospital by using the movement data to generate other variables which measure the close proximity of patients without MRSA to patients with positive MRSA.

Ideally, if the ward of a patient on each day in the study is known, we would be able to map the dynamic patient movement. Therefore, a data matrix could be generated, which presents the number of patients and their relative MRSA status in each ward on each day in the study. We consider as an example five patients (denoted as patient A, B, C, D, E) each with their associated dates of admission, dates of discharge, dates of transfer to another ward, the wards that they have stayed in and the MRSA status on each day.

Note that the information for those five patients are established artificially as examples to demonstrate. For simplification, we assume that those five patients only moved through two wards while in hospital. Table 7.1 clearly illustrates the available data for those five patients.

Table 7.1: The example of the ideally available data.

Patient		Admission	Transfer to another ward	Discharge date
A	Date	1-2-2013	3-2-2013	4-2-2013
	Ward code	W1	W2	W2
	MRSA status	Negative	Negative	Negative
B	Date	3-2-2013		4-2-2013
	Ward code	W2		W2
	MRSA status	Positive		Positive
C	Date	1-2-2013	2-2-2013	3-2-2013
	Ward code	W1	W2	W2
	MRSA status	Negative	Negative	Negative
D	Date	2-2-2013	3-2-2013	4-2-2013
	Ward code	W1	W2	W2
	MRSA status	Negative	Negative	Positive
E	Date	2-2-2013	4-2-2013	4-2-2013
	Ward code	W2	W1	W1
	MRSA status	Negative	Negative	Negative

Patient A was admitted into ‘W1’ ward in the hospital on first of February, 2013 and then moved to ‘W2’ ward on third of February, 2013. This patient was discharged from ‘W2’ ward on fourth of February, 2013. Moreover, Table 7.1 also shows that patient A remained as MRSA negative while in hospital. Patient B was admitted into ‘W2’ ward and was negative for MRSA on third of February, 2013 but discharged, positive for MRSA, on fourth of February, 2013. Patient B stayed in the same ward all the time he or she was in hospital. The data for patients C, D, E can be interpreted similarly.

Based on those quantities, the data matrix presenting the number of patients in each ward on each day can be derived, where the columns correspond to the dates and the rows correspond to the ward codes. For example, on first of February, 2013, both patient A and patient C were staying in the ‘W1’ ward. Hence, if there are only these five patients in the population of patients,we can conclude that there are two patients in the 'W1' ward on first of February, 2013. Similarly, on second of February, 2013, patient C and patient E were staying in the same ward ('W2'). i.e. there are two patients in 'W2' ward on second of February, 2013. The full data matrix for these five patients is shown in Table 7.2.

Table 7.2: The example of data matrix presenting the number of patients in each ward on each day.

	Date			
Ward code	1-2-2013	2-2-2013	3-2-2013	4-2-2013
W1	2	2	0	1
W2	0	2	5	3

Since the MRSA status on each day for each patient was known, the data matrix corresponding to the number of positive MRSA patients in each ward on each day can also be generated (shown in Table 7.3). Clearly, patient B was MRSA positive on the third of February, 2013 and this patient was in 'W2' at that time whereas all the other patients who were staying in the same ward on the same day were MRSA negative. Therefore, there is only one positive MRSA patient in 'W2' ward on the third of February, 2013. Similarly, according to Table 7.1, there are three patients staying in 'W2' ward on the fourth of February, 2013, but only patient A was MRSA negative. In other words, there are two positive MRSA patients in 'W2' ward on the fourth of February, 2013

Table 7.3: The example of data matrix presenting the number of positive MRSA patients in each ward on each day.

	Date			
Ward code	1-2-2013	2-2-2013	3-2-2013	4-2-2013
W1	0	0	0	0
W2	0	0	1	2

Thus whether a patient was exposed to MRSA in a ward (referred to in this chapter as 'exposed to MRSA') can be calculated based on the data matrix. Furthermore, if a patient was potentially exposed to MRSA in a ward, then the corresponding number of days as well as the number of patient days which this patient had spent staying with other positive MRSA patient(s) in the same ward simultaneously can also be derived. The number of days that a patient is exposed to MRSA (referred to in this chapter as 'days exposed to MRSA') is the sum of the days that this patient was staying with other positive MRSA patients at the same ward. The number of patient days that a patient exposed to MRSA (referred to in this chapter as 'patient days exposed to MRSA') can be calculated by summing the number of other positive MRSA patients that this patient was staying with at the same ward for each day in hospital. Note that the number of patient days is a measure, derived from patient movement date, which is interpreted as a patient staying with two or more positive MRSA patients in the same ward at the same time has a higher risk of MRSA acquisition than staying with just one positive MRSA patient in the same ward. Taking the data illustrated in Table 7.1 as the examples, we are able to generate three variables of exposure to MRSA for each patient based on the data matrix derived above, which are: (1) exposed to MRSA, (2) days exposed to MRSA, and (3) patient days exposed to MRSA. The results are shown in Table 7.4.

Table 7.4: The example of the variables of exposure to MRSA.

Patient	MRSA status on admission	MRSA status on discharge	Exposed to MRSA	Days exposed to MRSA	Patient days exposed to MRSA
A	Negative	Negative	Y	2	3
B	Positive	Positive	Y	1	1
C	Negative	Negative	Y	1	1
D	Negative	Positive	Y	2	2
E	Negative	Negative	Y	1	1

Note that 'Y' means that the patient was exposed to MRSA while in hospital and on the other hand, 'N' means that the patient was not exposed to MRSA while in hospital. Note that patient B who is positive on admission is excluded in the analysis. Since there was no MRSA positive patient in 'W1' ward (shown in Table 7.3), patient A was not exposed to MRSA for the first two days after the admission. Then patient A moved to 'W2' ward on the third of February, 2013 (shown in Table 7.1) and was MRSA negative on that day. According to the data matrix in Table 7.3, there was one positive MRSA patient in ‘W2’ ward on the third of February, 2013. Hence we can conclude that patient A was exposed to MRSA while in hospital, i.e. the variable of exposed to MRSA for patient A is ‘Y’. Table 7.1 shows that patient A was discharged from ‘W2’ ward on the fourth of February, 2013 and remained MRSA negative. Table 7.3 shows that there are two positive MRSA patients in ‘W2’ ward on the fourth of February, 2013. Hence patient A was staying with two other positive MRSA patients on the discharge date. In summary, patient A was exposed to MRSA for two days while in hospital (i.e. third and fourth of February, 2013), i.e. the variable of days exposed to MRSA is two. Moreover, patient A was staying with one other positive MRSA patient on the third of February, 2013 and was staying with two positive MRSA patients on the fourth of February, 2013. Therefore, the patient days that patient A was exposed to MRSA (i.e. the variable of patient days exposed to MRSA) is three.

Based on the quantities derived from the data matrix, the effect of patient movements on MRSA acquisition can be assessed using logistic regression methods in three main aspects: (i) whether a patient exposed to MRSA while in hospital has higher risk of acquiring MRSA compared to an ‘unexposed’ patient, (ii) whether a patient exposed to MRSA for a longer time has a higher risk in acquiring MRSA compared to a patient exposed for a shorter duration, and (iii) whether a patient more heavily exposed to MRSA patients in the same ward at the same time has a higher risk of acquiring MRSA than a patient less heavily exposed to MRSA.

7.1.1 Available data.

In this chapter, the dataset that will be used for the analysis is from Aberdeen Royal Infirmary only. Compared to the other hospital, Crosshouse, which was also recruited in the MRSA Screening Project, the information collected from the Aberdeen Royal Infirmary showed more accuracy and integrity. In the preliminary analysis, we found out that there were only 1.4% of records suspected to have missing information or transcription errors in the number of wards or the ward codes in the Aberdeen Royal Infirmary dataset. On the other hand about 23.5% of the patient records were suspected to have mistakes in the ward data in Crosshouse hospital.

Before evaluating if an MRSA patient is in a specific ward, we check and correct the records of ward codes for each patient who was included in the study. One of the reasons causing the errors in the information on wards in the dataset involves the total number of wards that patients had moved through during their stay in hospital and the specific ward codes they had been to. For example, we found instances where a patient had been in three wards but only two codes were recorded and a patient with three ward codes but a total of two wards. This was due to data transcription and reading errors. Therefore, by comparing the number of non-empty ward code records for each patient in the dataset with the corresponding record of total number of wards that patient had been to, we adjusted the total number of wards on the basis of the corresponding amount of amended ward codes record in the study which were collated with the true ward codes listed in Aberdeen Royal Infirmary. For example, in the Aberdeen Royal Infirmary dataset, the ward code record 'I1' can be considered as a mistake in computer data entry, which should be '11' instead. After reasonably correcting the ward codes, we adjust the total number of wards by adding up the number of ward codes recorded for an individual patient. Two or more separate stays in a ward are considered separately, For example, if a patient was in Wards 10, 11 and then back to 10 this is recorded as three wards.

In the Aberdeen Royal Infirmary dataset, neither the individual length of stay that a patient had stayed for in each ward nor the duration of a patient carrying MRSA while in hospital were recorded. There are three cohorts involved in the dataset, which are the admission-discharge cohort, the admission only cohort and the discharge only cohort. These are serious drawbacks but ones which could not be addressed by going back to the original data. The admission-discharge cohort includes the information on date of admission, date of discharge, MRSA status on admission and on discharge, number of wards and length of stay. On the other hand, the admission only cohort includes the data on date of admission, MRSA status on admission and number of wards but there are missing records on length of stay, date of discharge and MRSA status on discharge. Similarly, the discharge only cohort includes the data on date of discharge, MRSA status on discharge and number of wards but there are missing records on length of stay, date of admission and MRSA status on admission. Table 7.5 shows clearly the availability of the data.

Table 7.5: The available data in the Aberdeen Royal Infirmary dataset.

Cohort	Date of admission	MRSA status on admission	Date of discharge	MRSA status on discharge	Number of wards and ward codes	Length of stay
Admission-discharge cohort (1,580)	Y	Y	Y	Y	Y	Y
Admission only cohort (4,748)	Y	Y	N	N	Y	N
Discharge only cohort (1,483)	N	N	Y	Y	Y	N

Note that in this table 'Y' means that the data are available in the dataset, 'N' means that the data are not available in the dataset. We assume that the duration of stay in each ward is positive and hence the records, where the number of wards is larger than the corresponding length of stay, would not be included in the study. There are 1,580, 4,748 and 1,483 patients in the admission-discharge cohort, admission only cohort and discharge only cohort of the Aberdeen Royal Infirmary respectively, giving the data shown in Table 7.5.

7.1.2 Method.

In order to investigate the dynamic situation of the patient movement, we construct the two-dimensional matrix of ward by date giving the total number of patients in a ward on a given study day (e.g. see in Table 7.2). The rows of this matrix correspond to ward codes and the columns correspond to the study day. For each patient we will impute the duration of stay for each corresponding ward he or she had been to (i.e simulate the days of moving wards) from his or her admission date to the discharge date according to the quantities in the dataset: length of stay, ward codes and the total number of wards. For example, for a patient who had stayed in hospital for three nights from the admission date 2010-06-01 and had been to two wards whose ward codes were, in order, '49' and '50', we aim to assign the three nights into two wards and hence obtain the duration of stay in ward '49' and '50' respectively. This patient could have had one day in '49' and two days in '50' or two days in '49' and one day in '50' as we assume that the duration for each ward is larger than zero. Based on this two-dimensional matrix, the timelines of MRSA infection and carriage pressure in each ward can be mapped for all patients under a further assumption of the duration of MRSA colonisation while in hospital.

For the admission-discharge cohort, once the individual length of stay in each ward is imputed, we are able to assess the dynamic patient movements by constructing the two-dimensional matrix and evaluate the effect of patient movements associated with MRSA acquisition by applying logistic regression. However, the exposure variables related to patient movement which are generated only for the admission-discharge cohort would be biased since the situations of all the other patients in the hospital are not included. Hence the data for the admission only and discharge only cohorts are also included for the calculation of the exposure variables in the analysis of dynamic patient movements. For the admission only cohort, we impute the missing length of stay based upon the number of wards patients had been in and then impute the individual length of stay for each ward. Similarly, for the discharge only cohort, we impute the missing length of stay that the patients had stayed for in hospital and then derive the corresponding date of admission and individual length of stay for each ward based on the imputation. Simultaneously, the MRSA status on admission is imputed for all the patients in the discharge only cohort and then the MRSA status for each patient in a given ward can be imputed. As a result, the dynamic patient movement data and the corresponding MRSA colonisation pressure can be mapped each day.

The problem of imputing the individual length of stay for each ward (i.e. moving days) can be addressed by starting to analyse the distribution of the length of stay and hence establish a reasonable distribution for this by testing hypotheses, which will be introduced in the next section. In Section 7.3, the imputation of the length of stay, MRSA status on admission and movement dates will be demonstrated. Moreover, the map of patients with positive MRSA in hospital can be derived and the calculation of the exposure variables of patient movements associated with MRSA acquisition will also be introduced in Section 7.3. Finally, the logistic regression for the exposure variables of patient movements will be demonstrated in Section 7.4, which will be bootstrapped to take into account the imputation.

7.2 The analysis of the distribution of length of stay.

First of all, we will analyse the distribution of observed length of stay which is the total time in hospital in the admission-discharge cohort only and investigate four different assumptions for the individual length of stay for each ward (i.e. moving days).

Obviously, the length of stay for a patient who had stayed in a ward is positive. For the cases that the length of stay equals zero, it implies that the patient moved into a ward and then moved out from there on the same day. For example, if the admission date for a patient was the same as his or her discharge date then the length of stay for that patient in hospital is treated as zero. On the other hand for the case that the discharge date for a patient was the day after the admission date, the length of stay becomes one. As we mentioned in the previous chapter, in this study, the patients with positive length of stay (i.e. the patients had stayed over one night) are included for the analysis. We first assess the distribution of length of stay. There is only one record with zero length of stay in the Aberdeen Royal Infirmary dataset, which is excluded from the analysis. According to the analysis in Section 6.4 of the previous chapter, we detected that the length of stay has a positive, approximately linear, relationship with the number of wards. Hence, we investigate the distribution of length of stay stratified by the number of wards.

The distribution of the length of stay for patients who had only been in one ward is shown in Figure 7.1.

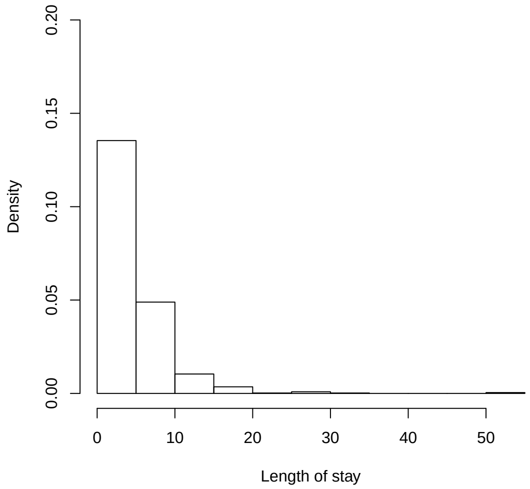


Figure 7.1: The distribution of length of stay (nights) for patients who had been in exactly one ward ($N = 904$).

The distribution in Figure 7.1 is strongly skewed to the right, indicating that the proportion of patients who had stayed five nights or under and only been to one ward during their stay in hospital is the highest, followed by the proportion of patients staying between six to ten nights. Only a few patients who had been to only one ward during hospitalisation had stayed over 15 nights.

The length of stay distribution for patients who had moved through two wards during their stay in hospital, is shown in Figure 7.2. This histogram shows a similar shape to

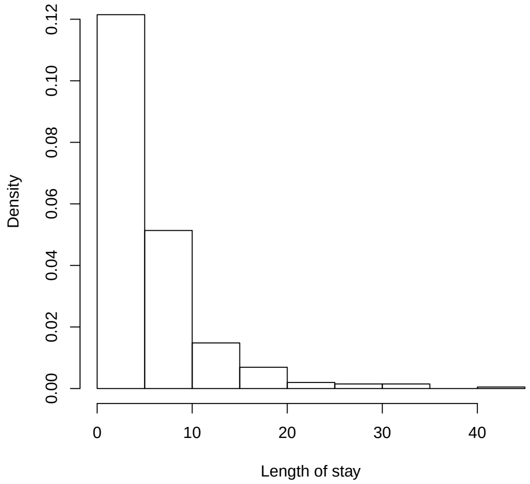


Figure 7.2: The distribution of length of stay (nights) among patients who had been to exactly two wards ($N = 405$).

the length of stay distribution for one ward cases, and is also highly skewed to the left.

Similarly, according to the distribution illustrated in Figure 7.3 patients who had moved through exactly three wards had relatively higher frequencies for shorter lengths of stay (i.e. the proportion of patients staying ten days or under was higher compared to those staying over 15 nights). Specifically, the relative frequency for the length of stay between one and five inclusive was slightly lower than the frequency for the length of stay between six and ten inclusive, which is reasonable since the total number of wards that patients had been to is three. From Figure 7.3, we can see that the majority of patients (approximately $(0.058 + 0.075) \times 5 = 67\%$) who had moved through three wards stayed in hospital for about ten days or less.

In comparison of Figure 7.1, 7.2 and 7.3, the density for 0 to 5 (i.e. (0,5]) is greater for exactly one ward and lower for exactly two wards and still lower for exactly three wards. The density for 5 to 10 (i.e. (5,10]) increases with increasing number of wards. This reflects that increasing length of stay is associated with increasing number of wards.

Analysis of the histogram of the length of stay for the cases with more than three wards reveals similar shapes to the distribution of the length of stay for patients moving

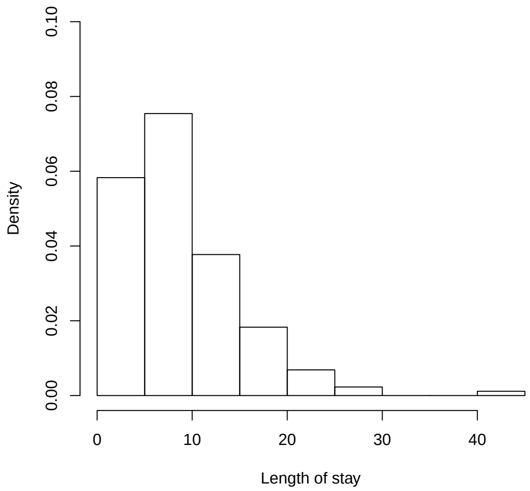


Figure 7.3: The distribution of length of stay (nights) for the cases that number of wards was three ($N = 175$).

through three wards. There are fewer patients and the histograms are not presented.

We need to impute the individual length of stay for each ward that a patient had moved through when patients had stayed in two or more wards. For each patient, we divide the total length of stay of that patient into several intervals so that the length of stay for each ward can be allocated. In the following subsections, four different assumptions of the distribution of the individual length of stay are proposed and the validity of assumptions are also investigated.

7.2.1 Assumption 1: the distribution of length of stay in multiple wards is the sum of lengths of stay in one ward.

Among patients where the total number of wards is larger than one, it is assumed that the distribution of the individual length of stay in each ward is the same as the distribution of length of stay for one ward only (displayed previously in Figure 7.1). We are able to compare the duration of stay calculated under this assumption to the true records. This was done by imputing the length of stay for each ward by simulating randomly from the observed distribution of length of stay in only one ward and then combining the lengths of stay for the two wards together so that the comparison of the imputation with the observed data (in Figure 7.1) can be made.

Another potential factor, which is also likely to have an effect on the length of stay is age. The corresponding two-way table (Table 7.6) shows that the relative frequencies of the length of stay are different in the four age groups. Especially, comparing the first column in Table 7.6 to the fourth column, the proportion of each length of stay for the age group 49 years or under is remarkably different from the corresponding proportions for the age group 80 years or over. For example, the proportions of patients staying for one night in the age group 49 years old or under (0.119) is relatively larger compared to the age groups 80 years or over (0.037).

Table 7.6: The two-way table of length of stay against age (one ward cases).

Length of stay	Age			
	≤49 years	50-64 years	65-79 years	≥80 years
1 night	0.119	0.064	0.069	0.037
2-3 nights	0.433	0.380	0.336	0.293
4-7 nights	0.320	0.410	0.390	0.367
≥ 8 nights	0.128	0.146	0.205	0.303
Number of patients	194	295	333	82

Investigating the two-way tables of length of stay against the other potentially related factors, gender and admission specialty, we find out that there is no strong evidence that the length of stay differs between gender or admission specialty. The two-way table regarding to gender, which is shown below in Table 7.7 as an example demonstrates that the length of stay does not differ substantially.

Table 7.7: The two-way table of length of stay against gender (one ward cases).

Length of stay	Gender	
	Male	Female
1 night	0.067	0.082
2-3 nights	0.385	0.351
4-7 nights	0.350	0.406
≥ 8 nights	0.198	0.161
Number of patients	431	473

Note that the ward codes listed in the one ward only case do not contain all the possible ward codes that patients had moved through for the two wards or more cases. This occurs as there are wards which no ‘one ward only’ patient stayed in but ‘two wards only’ patients did stay in; for example the ‘24’ ward. Hence for the purpose of investigating the current hypothesis that the distribution of individual lengths of stay for each ward is the same as that for the one ward case, we exclude the patients involving any different ward code compared to the ward codes contained in the one ward only case. For the exactly two wards case, we add up the imputed lengths of stay for the first and second wards to generate the corresponding distribution for total length of stay. Each imputed length of stay is sampled from the empirical distribution in the one ward only case with the ward code and age group corresponding to the ones recorded in the two wards case. For a patient aged 49 or under who had moved through two wards (where the ward codes were ‘HDU’ and ‘20’), we are able to simulate the individual length of stay for ward ‘HDU’ and ‘20’ respectively. Considering the simulated length of stay for ward ‘HDU’, it is sampled from the empirical distribution of length of stay for the patients aged 49 years old or under and also staying in ward ‘HDU’ in the one ward case. Note that ‘HDU’ is available as a one ward code. Similarly, we randomly sampled from the empirical distribution of length of stay for the patients aged 49 years old or under and also staying in ward ‘20’ in the one ward case to obtain the imputed length of stay for ward ‘20’.

The results (in Figure 7.4) show that the imputed distribution of total length of stay has a similar shape compared to the true distribution, which appears to support the assumption that the distribution of length of stay for each ward in the two wards case is the same as the corresponding distribution in the one ward only case. However, the relative frequency for the short length of stay (five nights or under) is lower than the true frequency. On the other hand, the relative frequencies for the longer lengths of stay (ten nights or over) are slightly higher than the true ones. One of the limits for this imputation approach is that there are only a limited number of observations for each combination

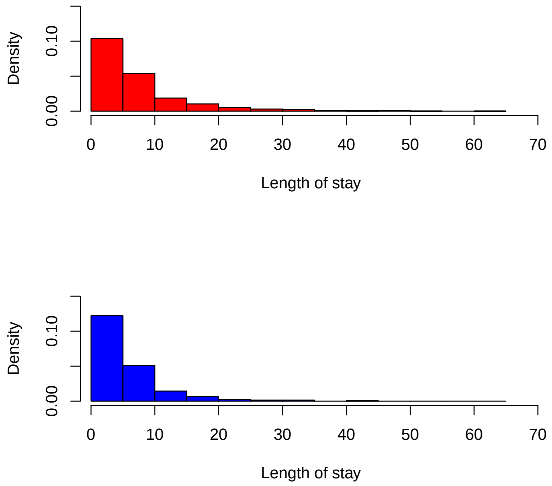


Figure 7.4: The comparison of length of stay (nights) between the imputed one (red histogram) and the observed one (blue histogram) for the case that number of wards was two. Both age and ward are used to select the appropriate empirical distribution. ($N = 1,600$)

of ward codes and age to generate the empirical distribution. For example, there is only one observation for a patient aged 49 years old or under and staying in ward ‘2’ in the one ward only case. This leads to problems in the imputation process in that there is not sufficient variability.

We attempted to apply this method to the exactly three wards case. In this case we are unable to simulate the length of stay due to the limited number of records in the one ward only case. The main problem is that we are unable to find some matched records in the one ward only case with the particular ward code and particular age group which are the same as those in the three wards case. Consequently, the corresponding empirical distribution for the total length of stay cannot be established.

Therefore, we consider constructing the empirical distribution of length of stay in each ward taking into account only one of the potential risk factors, namely ward code. We retain the same assumption that the distribution of length of stay for each ward is the same as the distribution of length of stay in the one ward only cases. The individual length of stay for a particular ward code is imputed from the corresponding empirical distribution of the length of stay associated with that ward code in the one ward only case.

The plot illustrating the distribution of the imputed total length of stay for the exactly two wards case together with the observed distribution is shown in Figure 7.5. The results show that the imputed distribution is clearly distinguished from the observed distribution, indicating that there is no evidence that the distribution of length of stay for each ward in two wards case is the same as the distribution of length of stay in the one ward only case. In Figure 7.5, the relative frequencies for the longer lengths of stay (over ten nights) in the red histogram (i.e. the imputed distribution for the exactly two wards case) appear much higher than the corresponding observed frequencies which are marked in colourless bars. Moreover, the density for the short length of stay (five nights or under) is clearly lower than the observed one.

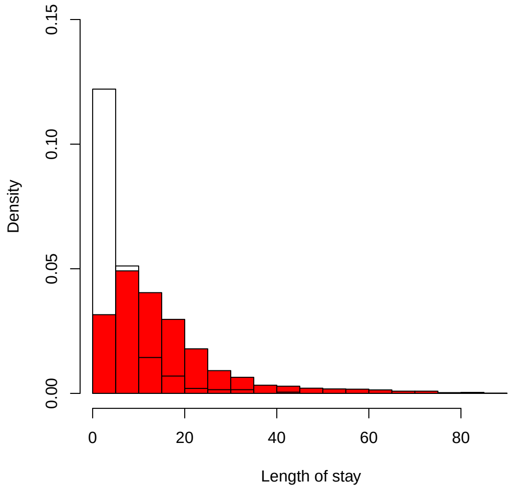


Figure 7.5: The comparison of length of stay (nights) between the imputed distribution (red histogram) and the observed distribution (colourless bars) for using patients who were in exactly two wards. Only ward code is used to select the empirical distribution. $(N = 2,014)$

We also compare the empirical distribution of the length of stay with the imputed one for patients who had been in exactly three wards. A similar conclusion can be drawn from Figure 7.6, which shows that the imputed distribution (i.e. the red histogram) does not coincide with the observed one (i.e. the colourless histogram). Therefore the assumption that the distribution of the length of stay in each ward within two and three ward stays is the same as that for one ward only is not valid. This is especially true when only ward code is used as a matching variable.

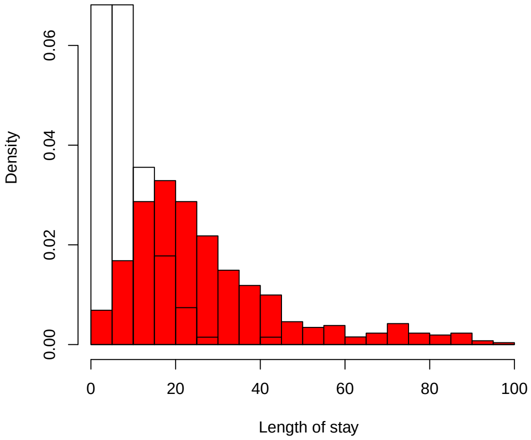


Figure 7.6: The comparison of length of stay (nights) between the imputed one (red histogram) and the observed one (colourless bars) for the case that where patients were in exactly three wards during their stay. Ward code used to select the empirical distribution ($N = 523$).

We also investigate the assumption that for the patients who had moved through two or more wards, the distribution of length of stay, stratified by the age groups, is the same as the distribution of length of stay in the one ward only case. However, the results also suggest that this assumption is invalid as the distribution of the imputed length of stay is quite different from the observed length of stay in the two wards case. The histogram of the imputed total lengths of stay is centred between 20-30 nights and the corresponding distribution tends to be much more symmetric (see the red histogram in Figure 7.7) than the observed distribution (colourless bars).

In this subsection, we investigated the imputation approach for lengths of stay per ward for patients who had been in multiple wards based on the empirical distribution of the observed length of stay for patients who had been in one ward only. However, when

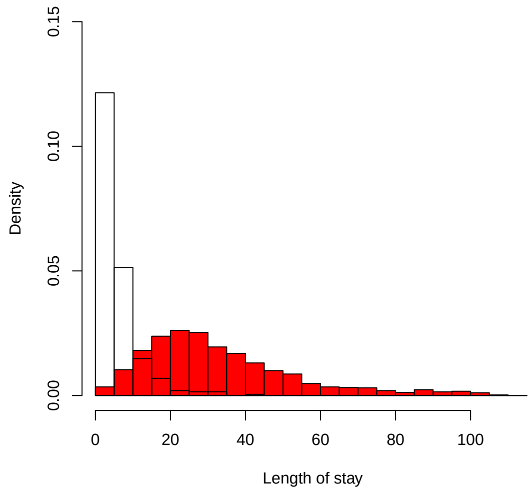


Figure 7.7: The comparison of length of stay (nights) between the imputed distribution (red histogram), where the empirical distribution is selected by the stratification on age, and the observed distribution (colourless bars) for patients who had been in exactly two wards ($N = 1,621$).

using these empirical distributions, the imputed total length of stay was not constructed to be the same as the observed one. This can be overcome by calculating the length of stay for the last ward by subtraction. For example, for two wards case, the length of stay for the first ward can be imputed based upon the corresponding empirical distribution whereas the length of stay for the second ward can be derived by subtracting the imputed length of stay for the first ward from the corresponding observed length of stay.

7.2.2 Assumption 2: Uniform distribution.

We do not have access to prior data about the distribution of length of stay in each ward when a patient is in two or more wards. As a first approximation we assume that the day a patient moves from one ward to another during the period from admission to discharge is equally likely for all days a patient is in hospital. Hence the assumption is made that the date of movement between wards for a patient follows a Uniform distribution. In other words, we assumed that the movement dates for each patient can be randomly sampled from the corresponding discrete Uniform distribution *U[1,los-1]* according to the quantities of the total number of wards (here *los* is the length of stay). For example, for a patient who had moved through two wards and stayed four days, we simulate the length of stay for the first ward which follows the discrete Uniform distribution with the domain $\Omega = \{1, 2, 3\}$ since the length of stay for each ward is a positive integer. If the result for the imputed length of stay for the patient staying in the first ward is two, it means that the date of movement for the patient moving from the first ward into the second one is two days after his or her admission. Hence the length of stay for the second ward in this example is now calculated to be two (i.e. *the observed length of stay - the imputed length of stay for the first ward*). Note that this investigation includes only the records which satisfy the criterion that the length of stay is larger than or equal to the corresponding number of wards since the length of stay for each ward is assumed to be positive. Thus there are 16 records excluded here as the recorded length of stay was smaller than the recorded number of wards that the patient was in. Those 16 patients had stayed for only one night but moved through two wards. Based on the above approach, the corresponding distribution of the length of stay for the first wards as well as for the second wards can be obtained individually for patients who had stayed in exactly two wards. The results are shown in Figure 7.8.

Comparing those two histograms with the one for the one ward only case in Figure 7.1, both histograms in Figure 7.8 show a similar shape to the distribution of length of stay for patients staying in only one ward. The proportion of patients staying in a ward for a smaller number of days is always relatively larger and the frequency declines dramatically as the length of stay increases.

For patients who had moved through exactly three wards, we impute two movements between the wards where the length of stay in each ward follows is imputed from the Uniform distribution *U[1,los-1]*, where *los* is the observed length of stay recorded in the dataset. Two integers randomly sampled from *U[1,los-1]* without replacement are sorted into an increasing order (i.e. the random integers are denoted as d_1, d_2 and $d_1 < d_2$). Also d_1 and d_2 are treated as the cumulative total length of stay that the patient had stayed

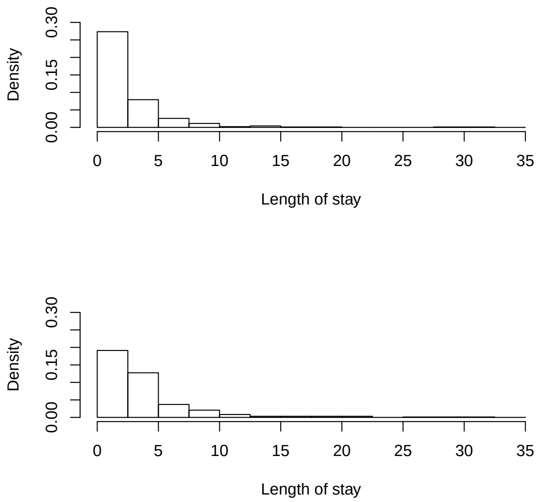


Figure 7.8: The distributions of length of stay (nights) for the first wards (the upper plot) and second wards (the bottom plot) under the hypothesis that the date of movement follows a Uniform distribution for the patients who were in exactly two wards.

in the different wards from the admission date. Namely d_1 is the length of stay for the patient in the first ward and d_2 is the duration of stay since the patient was admitted into the hospital until he or she was about to leave the second ward. The smaller value of the simulated results d_1 is regarded as the length of stay for the first ward and the length of stay for the second ward can be calculated by the subtraction of the length of stay for the first ward from the larger simulated integer (i.e d_2-d_1). Hence the length of stay for the third ward can also be generated by subtracting d_2 from the observed length of stay (i.e los - d_2).

For example, the dates of movements between wards for a patient who had moved through three wards and stayed for five nights are simulated without replacement from the Uniform distribution with the domain $\{1, 2, 3, 4\}$ in view of the necessity that the length of stay for a ward is guaranteed to be a non-negative integer. If the results of the simulation are three and one, we sort them in increasing order and then one is treated as the length of stay for the first ward that patient had been to and three is treated as the duration of stay in the first two wards that the patient had moved through since the date of admission (i.e the total length of stay for both the first and second wards). Thus the length of stay for the second ward is obviously two nights and the patient had moved into the third ward staying for another two nights before the discharge date.

From the individual histograms of length of stay for the first, second and third wards in the three wards cases (shown in Figure 7.9), we can see that the shape of all the simulated distributions are similar to the observed distribution for the one ward case in Figure 7.1. It is clear that those three histograms in Figure 7.9 are all strongly skewed to the right and decrease rapidly as the length of stay increases.

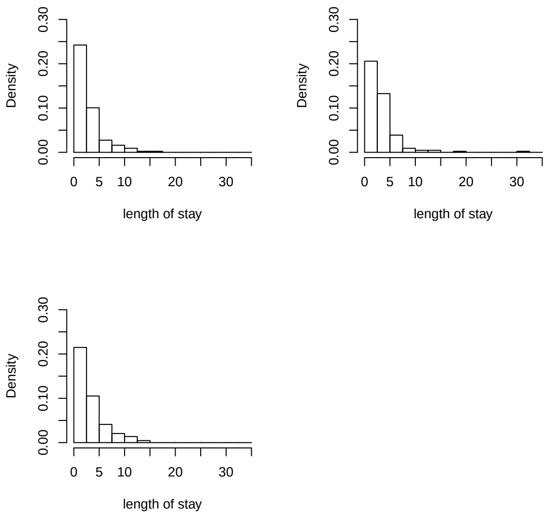


Figure 7.9: The distributions of length of stay (nights) for the first (plotted in the upper left), second (plotted in the upper right) and third wards (plotted in the bottom left) under the hypothesis that the dates of movement between wards follow a Uniform distribution for the case that the total number of wards was three.

Using the approach introduced above to impute the length of stay for each ward for the three wards case, the distribution of the imputed length of stay in the first ward is likely to be shorter because the length of stay for the first ward is imputed by a minimum of two random values sampled from the Uniform distribution without replacement. This can be seen in Figure 7.9 where the histogram for the length of stay in the first ward is more skewed to the right compared to the other two histogram. This imputation approach is appropriate since the anecdotal evidence is that patients who stayed in three wards in hospital were in the first ward for a relatively shorter time [116]. An alternative approach to impute the individual length of stay for three wards case is that the length of stay in the first ward (denoted as d_1) is sampled from $U[1, los - 2]$, where los is the observed length of stay, and then the length of stay in the second ward (denoted as d_2) is sampled from $U[1, los - d_1 - 1]$. Hence the length of stay in the third ward can be calculated by $los - d_1 - d_2$.

7.2.3 Assumption 3: Triangular distribution.

An alternative assumption which would appear quite reasonable is that patients are assigned into an initial ward for a short period when recently admitted into hospital. After making several examinations in a receiving ward, a patient might be moved into a specialist ward according to the diagnosis. Hence, there are good reasons to believe that the length of stay for the patients staying in the initial ward in hospital is highly likely to be short. This can be approximated by assuming that the length of stay for each ward follows a Triangular distribution, which means that for the ward into which the patient is initially admitted there is a high probability of a short length of stay. The Triangular distribution for the date of movement of a patient can be constructed based on the corresponding observed length of stay in the dataset which is denoted as los here. Note that the length of stay for each ward is assumed to be positive, hence only the records where the length of stay is larger than or equal to the corresponding number of wards are recruited for the analysis. As we mentioned in the previous subsection, there are 16 observations excluded in the analysis. The density function for the triangular distribution can be expressed as

$$f(x) = \frac{2 \times (los - x)}{(los - 1)^2} \quad \text{for } 1 \leq x \leq los.$$

Thus the probability of a patient staying for t nights in hospital can be calculated by

$$P(X = t) = \int_t^{t+1} \frac{2 \times (los - x)}{(los - 1)^2} dx,$$

$t = \{1, 2, \cdots, los - 1\}$. Here $P(X = los) = 0$ which means that the domain in the simulation remains to be $\Omega = \{1, 2, \cdots, los - 1\}$ and $P(X = 1) + P(X = 2) + \cdots + P(X = los - 1) = 1$ is guaranteed.

As long as we simulate the length of stay for the first ward, the length of stay for the second ward can be calculated based on the rest of nights that patients had stayed for. Using this procedure, we can simulate the length of stay, ward by ward, according to the relevant Triangular distributions which are constructed on the basis of the imputation for the previous wards. It is straightforward to assume that the length of stay for each ward always follows the Triangular distribution since the distribution of the length of stay for the latter wards that a patient had moved through can be assumed to be the same as the distribution for the initial ward that the patient had been admitted into (see Assumption 1). For example, regarding the two wards case we take as an example a patient who had moved through two wards and stayed for four nights, we use the Triangular distribution to simulate the date of moving where the density function can be expressed as $\frac{2\times(4-x)}{3^2}$. In order to simulate the length of stay for the first ward, we sample a random value from the domain $\{1, 2, 3\}$ with the probabilities for every point between 1 and 3 equalling 0.5556, 0.3333 and 0.1111 respectively (which can be calculated by $P(X = t) = \int_t^{t+1} \frac{2\times(4-x)}{3^2} dx$ where $t = 1, 2$ and 3). If the simulation result for the length of stay in the first ward is one, this indicates that the patient had moved into the second ward after one night stay in the first ward, and then left the hospital after another three nights stay in the second ward.

We are able to plot the histograms of the individual imputed lengths of stay for the first and second wards for the two wards only case and also compare those two imputed distributions with the distribution of the observed length of stay for the one ward cases. In Figure 7.10, the shape for both distributions appear similar to the one in Figure 7.1. Both distributions for imputed length of stay in the first and second wards show strong skewness to the right (i.e. the small length of stay in each ward has relatively high frequency). Moveover, the results show that the imputed length of stay in the first ward is shorter than the imputed length of stay in the second ward, which satisfies the anecdotal evidence.

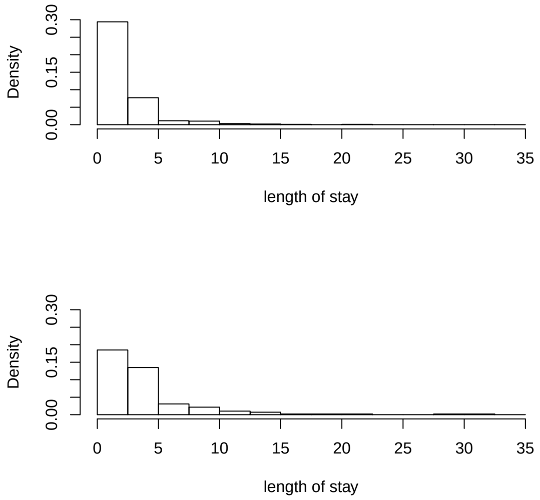


Figure 7.10: The distributions of length of stay (nights) for the first ward (in the upper plot) and second wards (in the bottom plot) under the hypothesis that the date of movement between wards follows a Triangular distribution for the case that the number of wards was two.

If a patient had stayed for five nights in total in three wards, we have to impute two movement dates from the Triangular distribution. The procedure of the simulation is similar to the two wards case that we introduced above. Obviously, the length of stay for the first ward is bounded above by three here since we need to guarantee the positive length of stay for the other two wards. Firstly, the length of stay for the first ward follows the Triangular distribution with the discrete density function expressed as $\frac{2\times(4-x)}{3^2}$ $(x = 1, 2, 3)$. Thus the probability that length of stay for the first ward is four or five equals zero (i.e. we ensure that the maximum length of stay for the first ward is three). Suppose that the randomly simulated result for the length of stay in the first ward is two, which yields the conclusion that the patient had stayed in the other two wards for three nights. Hence we construct a Triangular distribution for the date of movement when the patient left the second ward based on the first simulated date of movement we have obtained before. The corresponding density function for the length of stay in the second ward can be written as $\frac{2\times(3-x)}{2^2}$ $(x = 1, 2)$. This results in a random number from $\{1, 2\}$ where the corresponding probabilities are 0.75 and 0.25 when the simulation gave a random value of one. This corresponds to the length of stay in the second ward being one night and the patient being discharged from the hospital after staying in the third ward for another two nights. Using this method, we were able to yield the histograms of the simulated length of stay for the first, second and third wards respectively, which are shown in Figure 7.11.

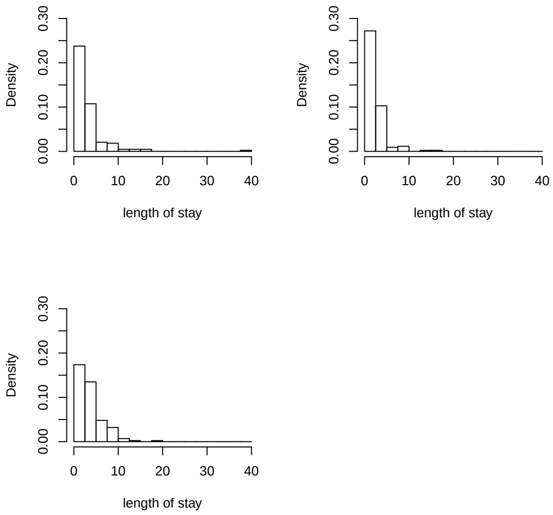


Figure 7.11: The distribution of length of stay (nights) for the first (plotted in upper left), second (plotted in upper right) and third wards (plotted in bottom left) under the assumption that the movement date follows a Triangular distribution for the case where the total number of wards was three.

All those three histograms are skewed to the right and decrease dramatically as the length of stay increases, showing a similar shape to the distribution of the observed length of stay for the one ward cases. The imputed length of stay in the third ward is relatively longer than the imputed lengths of stay in the first and second ward. This is because that we use the Triangular distribution to impute the length of stay for the first two wards respectively, which has a high probability of sampling a small random number. Thus the length of stay in the third ward, which is calculated by subtracting the imputed length of stay in the first two wards from the observed length of stay, is likely to be large.

Nonetheless, this satisfies the anecdotal evidence introduced above.

7.2.4 Assumption 4: Mixed distribution of Uniform and Triangular distributions.

Considering the case of patients who had moved through three or more wards, another possible assumption for the distribution for the movement date which is reasonable to consider is based on the idea that the length of stay for the first ward is usually short whereas afterwards the length of stay for the rest of wards that patients had been to follow the Uniform distribution. Specifically, we assume that the length of stay for the first ward is based on a Triangular distribution while the dates of movement for the rest of the wards follows a Uniform distribution. Taking the previous three ward case with length of stay equal to five as an example, the length of stay for the first ward can be still simulated by the Triangular distribution with density $\frac{2\times(4-x)}{3^2}$. Let x denote the length of stay in days. We obtain the simulated length of stay by using the calculated probabilities which are $P(X = 1) = 0.5556$, $P(X = 2) = 0.3333$ and $P(X = 3) = 0.1111$ based on the Triangular distribution, $P(X = t) = \int_t^{t+1} \frac{2\times(4-x)}{3^2} dx$ $(t = 1, 2, 3)$. We performed the simulation and found that the length of stay for the first ward was two nights. Based on this quantity, we are able to construct the Uniform distribution for simulating the date of movement between the remaining two wards with the corresponding domain $\{1, 2\}$. Randomly choosing an integer gives the result two, indicating that the length of stay for the second ward is two while as a result the length of stay for the third ward is one day.

Comparing the relative histograms of the simulated length of stay under this assumption for the first, second and third wards with the distribution of the observed length of stay for the one ward cases, we can conclude that the shape of all those three histograms in Figure 7.12 are similar to the one in Figure 7.1. In comparison of Figure 7.11 and Figure 7.12, the distribution for the first ward (i.e. the upper left histogram) is pretty identical but the last ward stay (shown in the bottom left histogram) is shorter in

Figure 7.12.

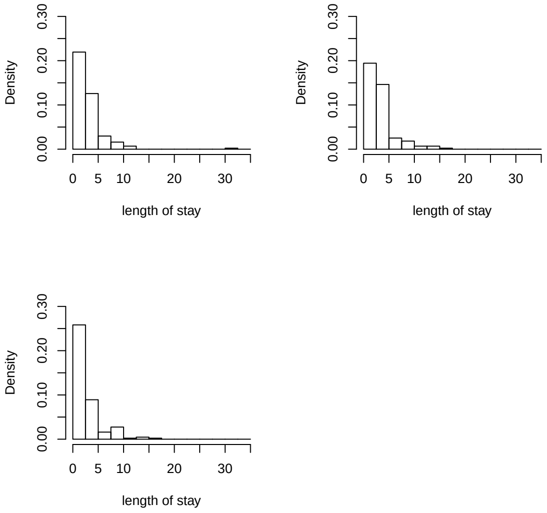


Figure 7.12: The distribution of length of stay (nights) for the first (plotted in upper left), second (plotted in upper right) and third wards (plotted in bottom left) under the hypothesis that the length of stay in the first ward follows a Triangular distribution and the length of stay in the second ward follows a Uniform distribution for the cases that number of wards was three.

7.2.5 Summary.

The first assumption (i) which assumes that the distribution of length of stay for each ward is the same as the distribution of observed length of stay for one ward only case is not valid because the imputed distribution of total length of stay is quite different from the observed one. The latter three assumptions provide results which are conditional on the observed total length of stay but there is no evidence from the data to select one assumption over the others. These are (ii) the assumption that the distribution of length of stay for each ward is based on the Uniform distribution; (iii) the assumption that the distribution of length of stay for each ward (apart from the last one) follows the Triangular distribution; (iv) the assumption that the distribution of length of stay for the first ward follows the Triangular distribution but the distribution of individual length of stay for the other wards is based on the Uniform distribution. It is impossible to validate those assumptions since there is no real data. In this thesis, we apply the Assumption 2 that the individual length of stay for each ward follows a Uniform distribution as an example for the analysis of patient movements as we have no reason to pick one method over another.

Based on the above analysis, we now aim to construct the two-dimensional matrix giving the number of patients staying in a ward on a particular study day. This will allow us to analyse the effect of patient movement based on the quantities derived from the length of stay and number of wards a patient has been in. Furthermore, using the MRSA status for each patient on admission and on discharge, we are also able to construct the data matrix giving the number of MRSA positive patients in a specific ward on each date in the study under an assumption about the duration of MRSA colonisation among hospital patients. As noted in Section 7.1.1, before the construction of the data matrix, the missing length of stay and MRSA status on admission and on discharge have to be imputed for the admission only and discharge only cohorts. Then the exposure variables based on patient movement can be calculated and the timeline of the patient movement as well as the MRSA colonisation pressure can be mapped. In the next section, we will demonstrate the methodology of this imputation.

7.3 Imputation of unknowns in admission only and discharge only cohorts.

According to the data which were provided by HPS (Health Protection Scotland) and collected from the Aberdeen Royal Infirmary, 1,598 out of 7,881 patients were recorded from the admission-discharge cohort where the complete information was collected including the length of stay, dates of admission and discharge, the measurement results of positive MRSA status both on admission and on discharge and other information. On the other hand, 4,800 out of 7,881 patients were from the admission only cohort, where the corresponding dates of discharge and the measurement results of positive MRSA status for patients on discharge were not collected. In the admission only cohort the majority of patients (3,956 out of 4,800) did not have any record on the length of stay. Similarly, there were 1,483 out of 7,881 patients recorded in the discharge only cohort without the dates of admission and the results of MRSA status for patients on admission. Moreover, the length of stay for 1,477 out of 1,483 patients in the discharge cohort was also missing. This lack of data was due to consent issues and if a patient did not give consent then admission or discharge data could not be collected. This incompleteness of the records in the dataset causes severe difficulty in trying to assess the effects of patient movement. Therefore, we aim to impute the missing length of stay as well as the dates of admission and the MRSA status results for patients on admission in the discharge only cohort and the dates of discharge for the patients in the admission only cohort.

7.3.1 The simulation of the length of stay.

In this subsection, we will introduce the methodology of imputing the length of stay for the admission only and discharge only cohorts. Two main simulation methods will be demonstrated here, which are (i) imputation based on the empirical density of the length of stay and (ii) imputation based on Negative Binomial regression.

7.3.1.1 Kernel smoothing empirical density of length of stay.

An investigation of the two-way table of the number of wards against the length of stay in the admission-discharge cohort, suggests that for the patients who had moved through a small number of wards, the corresponding length of stay that patient had stayed for tends to be short while for the patients who had moved through a large number of wards, the length of stay tends to be relatively long (shown in Section 6.4 of Chapter 6). Thus, there is evidence that the number of wards that patients had moved through is related to the length of stay. Hence for the purpose of imputing length of stay, we construct the empirical density function of the length of stay stratified by the different number of wards. Specifically, the patients in the admission-discharge cohort are divided into six groups according to the corresponding number of wards that patients had moved through, which are the 'one ward' only group, the 'two wards' group, the 'three wards' group, the 'four wards' group, the 'five wards' group and the 'greater than or equal to six wards' group. Then we estimate the the empirical density using kernel smoothing to generate the continuous density function for the stratified length of stay which was constructed using the observations of length of stay in each group. If we imputed the length of stay only by sampling from the observations, then lengths of stay which were not observed could not be imputed and this limits the range of possible lengths of stay in the imputation.

The six smoothed continuous empirical density functions stratified by the six categorised groups of number of wards are displayed in Figure 7.13. The results show

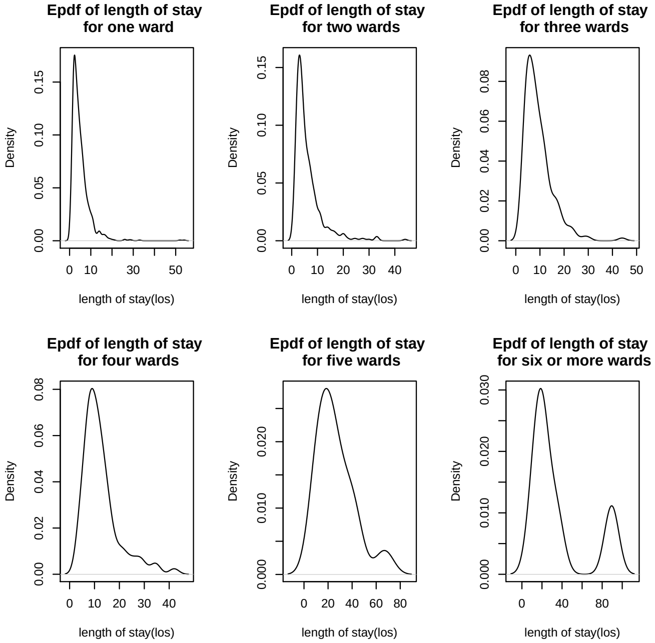


Figure 7.13: The distribution of length of stay for the different number of wards. ($N = 905$ for one ward; $N = 405$ for two wards; $N = 175$ for three wards; $N = 87$ for four wards; $N = 15$ for five wards; $N = 10$ for six or more wards.)

that roughly speaking, the peak of the distribution of length of stay moves to the right hand side slowly as the number of wards increases, which means that when the number of wards increases, the average length of stay generally increases. However for the patients who had moved through five and six or more wards, the empirical density function shows two peaks perhaps because this group contains a fairly small amount of observations on length of stay. For example, there are only ten patients in the group staying in six wards or more and for two of these in the patients stayed in ten wards. Those observations have relatively large length of stay. Also, there are only ten observations in this group. Hence a second peak is shown in the empirical distribution of length of stay for six or more wards. A slight peak showing around 65 days in the empirical distribution of length of stay for five wards can also be explained by the similar reasons that there is a small amount of observations ($N = 15$) and the length of stay for the patients who had stayed in five wards is likely to be long.

For all n patients with missing length of stay, we impute L_1, L_2, $\cdots$, L_n. From this we estimate the empirical density $f(x_j)$, indicating the proportion of patients with a length of stay of x_j. Repeat this whole process 100 times and we can get 100 empirical distributions. which are denoted as $f^1(x_j)$, $f^2(x_j)$, $\cdots$, $f^{100}(x_j)$. For each x_j, the corresponding densities of $f^1(x_j)$, $f^2(x_j)$, $\cdots$, $f^{100}(x_j)$ are assigned in order (i.e. $f^{(1)}(x_j)$, $f^{(2)}(x_j)$, $\cdots$, $f^{(100)}(x_j)$). This then gives the 2.5% and 97.5% percentiles, which is the simulation envelope for $\hat{f}(x_j)$. The plot shown in Figure 7.14 is the histogram of the simulated length of stay. Also, the observations of length of stay recorded in the admission and discharge cohort are plotted as a histogram in Figure 7.15 combined with the simulation envelope.

The histogram of one simulation of length of stay lies within the simulation envelope as we expected. The histogram of the simulated length of stay in Figure 7.14 indicates that the peak is around three days. The number of patients who had stayed over four nights decreases significantly as the length of stay increases. In addition, the simulation envelope represents the potential histograms of the length of stay that might be observed if the proposed simulation method is valid. Compared to the observations of length of stay recorded in the dataset (shown in Figure 7.15), we can also conclude that the relatively small observations of the length of stay (i.e. under two nights) lie above the upper bounds of the simulation envelope whereas the relatively large observations of the length of stay

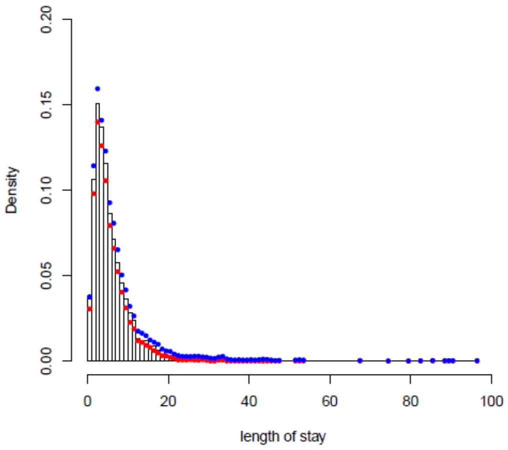


Figure 7.14: The distribution of simulated length of stay for the different number of wards in the admission only and discharge only cohorts, where the blue points are the upper bounds of the simulation envelope and the red points are the lower bounds for the simulated length of stay.

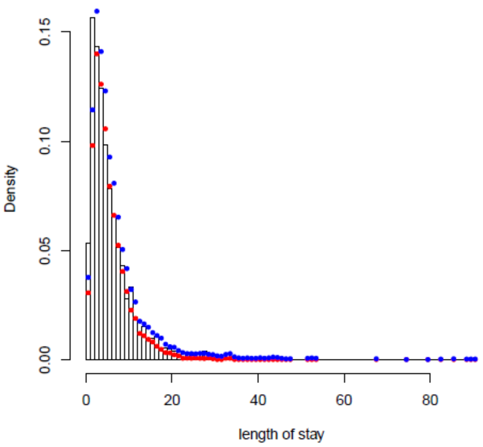


Figure 7.15: The histogram of the observed length of stay in the admission-discharge cohort with the simulation envelope. The blue points are the upper bounds of the simulation envelope and the red points are the lower bounds for the simulated length of stay. The simulation is based upon a stratification by number of wards.

(i.e. four to eight nights) lie under the lower bounds of the simulation envelope. This indicates that the simulation approach for the length of stay based on the kernel smoothed empirical density of observed length of stay stratified by number of wards does not work perfectly well. On the other hand, the simulated length of stay (in Figure 7.14) shows approximately the same distribution in shape and location as the observed length of stay in the admission and discharge cohort (in Figure 7.15), where the highest frequency is two nights.

7.3.1.2 Negative Binomial regression.

Since the method of simulation we introduced in 7.3.1.1 yields an imperfect prediction, specifically for the cases with length of stay under eight nights, it is worth applying and analysing a different simulation approach. By taking the potential risk factors such as age, ward code and gender into account, we attempt to build a model to fit the observed length of stay. The observed histograms of the length of stay for different 'wards' groups (in Figure 7.13) show heavy tails, which means that the variance of the length of stay in each wards group is relatively larger compared to the corresponding mean. Specifically, for the observed length of stay in the 'one ward' group with the mean 5.06 and variance 19.78 whose histogram shows an obvious heavy tail, it suggests an overdispersion in the observed length of stay. Hence the Negative Binomial regression model is proposed to be constructed individually for each different number of wards groups which are 'one ward', 'two wards', 'three wards', 'four wards' and 'five or more wards'. Note that we combine the categories of 'five wards' and 'six or more wards' together as a new group of 'five or more wards' here to avoid the limited sample size.

Considering the 'one ward' group, the trend test is used to detect the association between the observed length of stay and age. The low p-value (0.0116) implies a significant linear trend in the relative probability of having length of stay under two nights against having length of stay larger than three nights associated with age group.

We investigate the use of the fractional polynomial regression as a smoothing technique to investigate the nonlinear function between the length of stay and age. We build the one degree functions of length of stay associated with age with different powers (-2,-1,-0.5,0,0.5,1,2) as well as the corresponding two degree functions. Comparing the linear function (i.e the power of age is one) with the one degree function which involves a cubic age term and gives the smallest deviance among the other one degree functions with different powers, the Chi-square test generates a large p-value (0.66) and implies that there is no significant difference between the linear trend and the cubic trend in length of stay. Similarly, the two degree function with powers of age (-2,1) giving the model expressed as $E(los) = \beta_0 + \beta_1 \times age^{-2} + \beta_2 \times age$ is also not statistically significantly different from the linear function. Hence, this indicates that more complex trends over the linear trend of age do not adequately describe the variation in length of stay better. Moveover, this conclusion can be confirmed by treating age group as an ordered variable in the model. Hence generally speaking, it is unnecessary to include any nonlinear trend in age to describe the change in length of stay.

In order to simulate the length of stay for the 'one ward' group, first of all we test the full Negative Binomial model consisting of the three potential risk factors: age, ward code and gender. Note that the dependent variable in the regression model is treated as $los - 1$ to avoid the influence of the truncated Negative Binomial distribution for the observed length of stay. The results give the p-value for gender (0.876) which is nonsignificant. Therefore, there is no evidence for an the effect of gender on the length of stay. Next, the nested model which involves two potential risk factors: age and ward code is fitted since according to the previous analysis in Section 7.2.1, age and ward code are related to the length of stay. The results show significance of both age and ward code. The Chi-square test with a large p-value (0.876) suggests that there is no obvious difference between the full model and the nested one. Hence we can conclude that the nested model is more appropriate than the full model for goodness of fit since the nested model is more simple and also adequate enough for the modelling. Using this nested model, we are able to predict the missing length of stay for the 'one ward' group based on the corresponding records of ward code and age. Note that the missing length of stay in some cases where the ward codes are used as predictors in the simulations such as 'AMAU' and 'HDU' need to be simulated depending only on age since these ward codes are not involved in the set of only one ward code observations (i.e. the ward codes for the first wards) which were used to fit the model. This can be explained by the reason that a patient admitted into a high dependency unit (HDU) or an acute medical assessment unit (AMAU) is highly likely to move from the HDU or AMAU to another ward before discharge. Those cases lead to the problem of predicting the corresponding length of stay using the nested model due to the appearance of the new levels for the first ward code variable. In this circumstance, we use the model only involving age to simulate the length of stay. The new model here can be expressed as

$$log(E(los-1)) = \beta_0 + \beta_1 age.$$

Note that for the cases with missing age record on admission, we use the age recorded at discharge instead.

Under the assumption that the length of stay follows a Negative Binomial distribution, the mean of the response (i.e. length of stay -1) which is denoted as μ can be estimated for each case in the 'one ward' group according to the Negative Binomial model which is

$$E(los-1) = \mu, log(\mu) = \beta_0 + \beta_1 age + \boldsymbol{\beta}_2 \textbf{ward code 1}.$$

Note that **ward code 1** represents the ward code for the first ward that a patient had been in. Based on the estimated mean μ_i for each case, we can randomly simulate the $\xi_i = los_i - 1$ from the Negative Binomial distribution with the corresponding mean μ_i and variance $\mu_i + \mu_i^2/\theta$ where θ is the overdispersion parameter (here θ is estimated to be 2.633 in the fitted model). Clearly, the simulated length of stay for each case can be generated by $\xi_i + 1$. The estimates and the corresponding standard errors for this Negative Binomial regression model are shown in Table C.1 in Appendix C

The histogram of the imputed length of stay for the 'one ward' group and the histogram of the observed length of stay are compared in Figure 7.16. We can see that the simulated length of stay has a large proportion on one night. Specifically, the simulation approach which we applied gives lots of ones as the simulated length of stay and the corresponding frequency is overestimated. The simulated results show an approximately double proportion of patients staying for one night compared to the observed length of stay. On the other hand, the simulated length of stay for two nights or more is slightly underestimated.

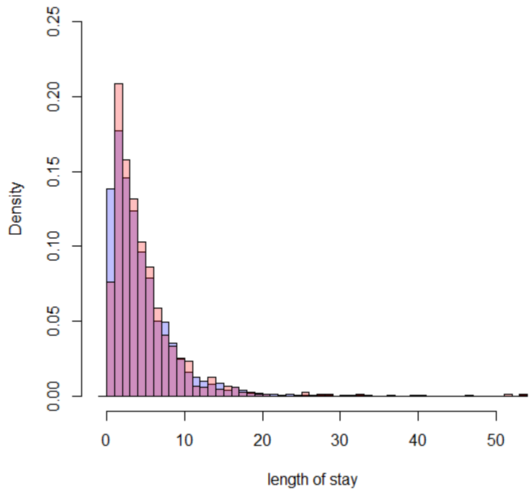


Figure 7.16: The blue histogram represents simulated length of stay and the pink one represents observed length of stay in the admission-discharge cohort. The overlap for those two histograms is marked as purple.

Using the same procedure to construct the Negative Binomial model for the two wards cases, three potential variables: ‘age’, ‘ward code 1’, ‘ward code 2’ show significant effect on length of stay with low p-values (which are 0.0004, 0.06 and 2.4×10^{-7} respectively) and hence they are considered to be involved into the final model. On the other hand, gender does not shows significant effect on length of stay with a high p-value (0.386). From a mathematical point of view, the Negative Binomial model in the ‘two wards’ group can be expressed as

$$E(los-1)=\mu, log(\mu)=\beta_0+\beta_1 age+\boldsymbol{\beta_2}\textbf{ward code 1}+\boldsymbol{\beta_3}\textbf{ward code 2}.$$

Note that the parameter estimates and the corresponding standard errors for this Negative Binomial model are shown in Table C.2 in Appendix C. For those patients with different ward codes from the set of observed ward codes, it is required to use a simple model which depends only on the age to simulate the corresponding missing length of stay. Moveover, for the prediction for patients with missing age on admission, we use age recorded at discharge instead. The estimated mean of the length of stay can be obtained from the fitted model using the records of age, 'ward code 1' and 'ward code 2'. Using the estimated overdispersion parameter in the fitted model (2.775), the imputed length of stay can be randomly generated from the fitted Negative Binomial distribution. By extending this approach for the other ward groups (three, four or five and more wards), we can obtain the imputed length of stay for the whole dataset where the original records did not have length of stay recorded.

In the groups of patients with a large number of wards (i.e. three or more wards), our investigation suggests that only the first three wards that patients had moved through are significant variables in the Negative Binomial model as well as age which also has significant effect on length of stay. The relatively small number of observations on the patients with a large number of wards ('four wards' and 'five or more wards' groups) causes difficulty in fitting a good model. One of the reasons is that the ward code variables as the predictors in the model have multiple levels, leading to relatively few observations in each combination of levels between the ward codes variables. Therefore, we re-categorised the ward codes into three levels according to the relative ward specialties, which are: 'surgical ward', 'medical ward' and 'mixed ward' and then the new categorised ward codes group variables are applied into the Negative Binomial regression model. Model selection is based on Chi-square tests and we only use the significant variables in the Negative Binomial regression models. Taking the 'four wards' group as an example, the Chi-square tests show that age and 'ward code group 1' are the significant factors with respect to length of stay due to the low p-values (< 0.05). The model can be expressed as

$$E(los - 1) = \mu, log(\mu) = \beta_0 + \beta_1 age + \boldsymbol{\beta_2}\textbf{ward code group 1}$$

and the estimated overdispersion parameter in this fitted model is 4.80. As to the 'larger than or equal to five wards' group, 'age', 'ward code group 1', 'ward code group 2' and 'ward code group 3' display the significant effects on length of stay since the corresponding p-values are less than 0.05. Hence those three potential risk factors are included into the fitted model. The model for patients with five or more wards is

$$\begin{aligned}E(los-1) =&\mu,\\ log(\mu) =&\beta_0+\beta_1 age+\boldsymbol{\beta_2}\textbf{ward code group 1}+\\ &\boldsymbol{\beta_3}\textbf{ward code group 2}+\boldsymbol{\beta_4}\textbf{ward code group 3}\end{aligned}$$

and the estimated overdispersion parameter is 6.25 in the fitted model. Using the fitted models, we can obtain the estimated mean of length of stay according to the records of age and recategorised ward codes. Hence the length of stay can be simply imputed from the Negative Binomial distribution with the estimated means and overdispersion parameters.

The histogram of the observed length of stay is compared to the histogram for the simulated length of stay in Figure 7.17 which shows that the simulated length of stay has a similar distribution to the observed one. Clearly, an overestimate occurs in the simulated length of stay of one night. The histogram in Figure 7.17 is truncated at 30 days for the purpose of demonstrating the comparison between imputed length of stay and observed length of stay clearly. This truncation is reasonable since the preliminary analysis shows that a small length of stay has relatively high frequency.

We use the bootstrap method to generate the simulation envelope which is displayed in Figure 7.18 combined with the observed length of stay. This figure shows clearly that the simulation approach using the Negative Binomial regression model overestimates the length of stay of one night but underestimates slightly the length of stay of two to four nights. Comparing the simulation envelope plot with the previous simulation approach (shown in Figure 7.15), there is no strong evidence that the simulation method using the Negative Binomial regression improves the results and the simulation results using the empirical density and the Negative Binomial regression show bias in different directions.

In this subsection, we used two different methods to impute the length of stay in admission only and discharge only cohorts using the observed data in the admission-discharge cohort. The results of imputation obtained by the empirical density

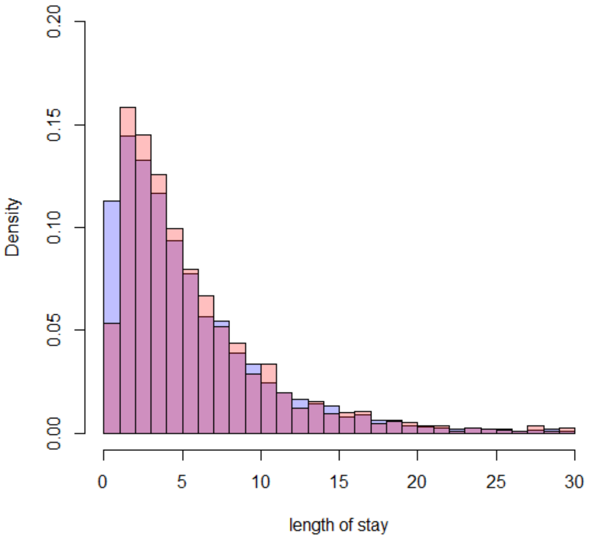


Figure 7.17: The blue histogram represents simulated length of stay and the pink one represents the observed length of stay. The overlap is marked as purple. This histogram is truncated at 30 days in order to demonstrate the comparison clearly.

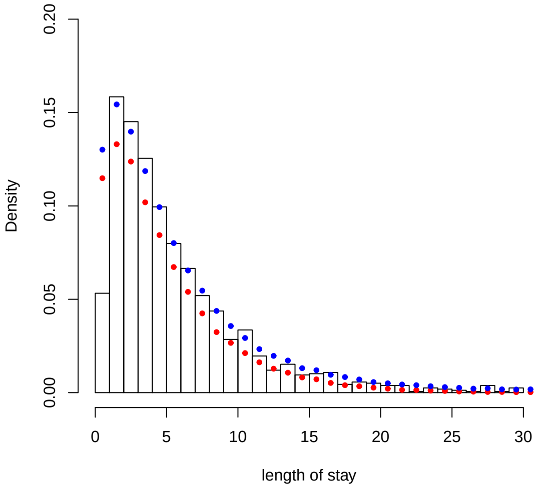


Figure 7.18: The histogram of observed length of stay in the admission-discharge cohort with the simulation envelope.

of observed length of stay show the underestimation in the length of stay of one to two days and the overestimation in the length of stay of three to five days. On the other hand, the imputed results obtained by the Negative Binomial regression show a bias in a different direction. However, there is no strong evidence that one imputation method is better than the other between the empirical density and the Negative Binomial regression and we have not been able to find a better fit.

Using the length of stay and the available data on the wards that a patient had been in, we now move on to impute the movement dates in order to construct the data matrix introduced in Section 7.1. The methodology of imputing the movement dates in the admission only and discharge only cohorts will be investigated in the next subsection.

7.3.2 The imputation of the date of movement between wards in the admission only and discharge only cohorts.

In the previous sections, we have investigated imputation of the dates of moving between wards in the admission-discharge cohort and length of stay in the admission only and discharge only cohorts. In this section, we use the imputed lengths of stay in the admission only and discharge only cohorts to impute dates of transferring between wards in those two cohorts. In order to include all the cases in the study, we pick a date which is long before the earliest admission date from the study in the admission-discharge cohort, or the admission only cohort. This date (here we choose 1st Dec, 2009) is treated as the very first day of the study (i.e. the baseline for evaluating the date of movement for each patient in a specific ward). This is just for convenience to avoid using dates. The admission date for each patient is considered as the ith day in the study where i is the time span between the baseline and the admission date of the patient. Take the patient in the admission-discharge cohort who was admitted into the hospital on 9th of February, 2010 as an example, we can generate the time lag between 09-02-2010 and the baseline date 01-12-2009. Therefore, the admission date for the patient is regarded as the 71st day in the study.

Since the discharge only cohort where the admission dates for the patients were unknown is also taken into account in the study, the corresponding admission date for each patient is calculated based on the recorded discharge date and imputed length of stay. Afterwards, a map of the admission dates can be derived by calculating the exact day in study based on the baseline (1st Dec, 2009) for the patients included in all the admission-discharge, admission only and discharge only cohorts.

After imputation each patient in all three cohorts now has an admission date and a length of stay (and consequently a discharge date). The movement dates (i.e. the period that the patient stayed in each ward he or she had moved through) can now be imputed under some reasonable assumptions. Recall that in this thesis, we apply the Uniform distribution as the assumption (introduced in Section 7.2) on estimating the length of stay in each ward; hence we can derive the movement dates from one ward to another.

As we mentioned before, the admission date for the patient can be converted into the ith day in the timeline of the study. Hence, the corresponding movement dates for that patient moving through n wards during his or her stay in hospital (denoted as d_1, d_2, $\cdots$, d_{n-1}), which are imputed based on the method we introduced above, can also be generated as the new movement dates d_1+i-1, d_2+i-1, $\cdots$, $d_{n-1}+i-1$ for the study. Then we are able to build the pattern of the movement, admission and discharge times where for each patient the allocation of the relative admission date and movement dates in the study can be decided with regard to the specific wards he or she had moved through. Therefore, the matrix showing the number of patients in a specific ward on a specific study day in hospital can be constructed. For example, we generated the data matrix of the number of patients in a specific ward on a specific study day for the admission-discharge cohort and plot the corresponding total number of patients on each study day. Generally speaking, the number of patients increases as the time goes by since the beginning of the study in the admission-discharge cohort (shown in Figure 7.19). This might be because some patients who were admitted into the hospital early in this study had stayed for a relatively long period and at the same time new patients were included constantly. Figure 7.19 shows that in the end of this one-year study, the number of patients drops dramatically. One of the reasons might be that a large number of patients who were admitted and included before the end of this study were not discharged yet and hence they were not included in the admission-discharge cohort.

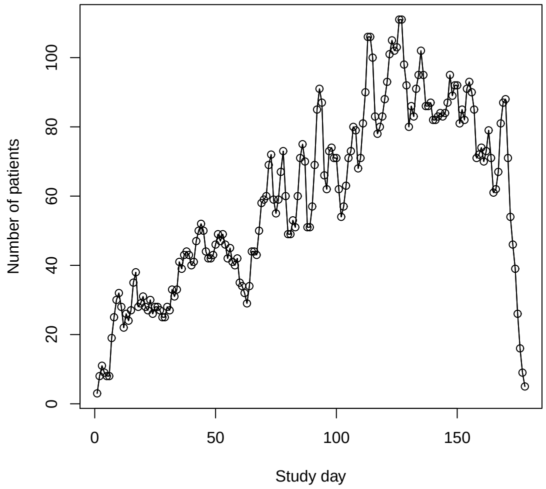


Figure 7.19: The plot of number of patients on each study day in the admission-discharge cohorts.

7.3.3 Imputation of the missing MRSA positive test results for the admission only and discharge only cohorts.

In order to investigate the effect of patient movement on MRSA acquisition in this subsection we focus on the number of MRSA positive patients in a ward on a given day. Knowledge of the number of MRSA positive patients in a ward enables us to estimate the probability that a patient who moves into that ward acquires MRSA.

By constructing the two-way table which reflects the MRSA measurement results on admission and on discharge in the admission-discharge cohort, we can see from Table 7.8 that 1,542 (98.78%) patients with negative MRSA on admission remained in the same negative condition on discharge. The proportion of patients with positive MRSA on admission but negative MRSA on discharge is 44.44%. Among the patients with positive MRSA on discharge almost half (48.72%) of the patients were measured as negative on admission.

In Table 7.8, 0 means MRSA negative and 1 means MRSA positive. Considering the

Table 7.8: Two-way table of MRSA on admission and on discharge for the admission-discharge cohort.

		MRSA on discharge	
		0	1
MRSA on admission	0	1542	19
	1	16	20

admission cohort and the lack of MRSA measurement results on discharge in this cohort, we randomly sample the value from {0,1} corresponding to the probabilities obtained by the two-way table of MRSA measurements which is shown in Table 7.8 as the result of MRSA on discharge. For patients who were MRSA negative on admission in the admission only cohort, we randomly sample the MRSA result on discharge from {0,1} with the corresponding probabilities being 98.78% and 1.22% respectively.

Similarly, considering the patients in the discharge only cohort, we sample the MRSA results on admission from {0,1} with the corresponding probabilities equalling 98.97% and 1.03% respectively for the patients who were MRSA negative on discharge whereas for patients who were MRSA positive on discharge we sample the MRSA results on admission from {0,1} with the corresponding probabilities equalling 48.72% and 51.28% respectively. The results from one simulation show that roughly 2.92% of patients were MRSA positive at discharge in the admission only cohort, which is similar to the observed percentage of patients who were MRSA positive on discharge in the admission-discharge cohort dataset (2.31%). With respect to the simulated MRSA results on admission in the discharge only cohort, there are about 2.10% of patients who were MRSA positive which is close to the observed proportion of MRSA positive patients at discharge in the admission-discharge cohort (2.25%).

Alternatively, a logistic regression model can be used to predict the MRSA status on admission and on discharge in the discharge only and admission only cohorts. However, we find no significant factors related to MRSA status on admission and on discharge here.

7.3.4 The total number of MRSA positive patients in each ward on each study day.

At this stage we know the admission and discharge dates and ward movement dates for all patients as well as MRSA status on admission and discharge, we do not know duration of MRSA carriage and this is the last piece of information needed to be able to construct the matrix representing the MRSA status for each patient on each individual study day. The duration of MRSA carriage varies and it could be persistent from days to years [4], [66], [128]. In this study, we assume that the duration of MRSA carriage depends on the observed length of stay in the admission-discharge cohort. In the admission-discharge cohort, a patient who is MRSA positive on admission but MRSA negative on discharge is most likely to stay in ten days whereas a patient who is MRSA negative on admission but MRSA positive on discharge is most likely to stay in hospital for seven days. Thus we assume that for patients who were MRSA positive on admission but MRSA negative on discharge, the MRSA bacteria persist for a maximum of ten days and for patients who were MRSA negative on admission but MRSA positive on discharge, the maximum time for being MRSA free is seven days. In other words, for patients who were confirmed as MRSA positive on admission but tested as negative on discharge, the carriage period for MRSA can be expressed as

$$t = \begin{cases} 10 & \text{if } los > 10, \\ los - 1 & \text{if } los \leq 10, \end{cases}$$

where *los* represents the length of stay. Similarly, for patients who were MRSA negative on admission but MRSA positive on discharge, the time before being colonising MRSA (i.e. the time for being MRSA free) is

$$t = \begin{cases} 7 & \text{if } los > 7, \\ los - 1 & \text{if } los \leq 7. \end{cases}$$

This means that most patients with a short length of stay who are colonised are regarded as colonised the day before discharge. Afterward, it is straightforward to build the pattern of the total number of MRSA positive patients in each ward on a specific study day. However, note that there is little justification for this assumption of duration of MRSA carriage due to the lack of preliminary information.

7.3.4.1 The analysis of the number of MRSA positive patients per day based on the length of stay obtained from Simulation Method 1: empirical density.

We display the plot of the simulated number of MRSA positive patients against the timeline of the study (shown in Figure 7.20) to highlight the fluctuation of the number of MRSA positive patients during the study, where the simulated lengths of stay were generated from the empirical density.

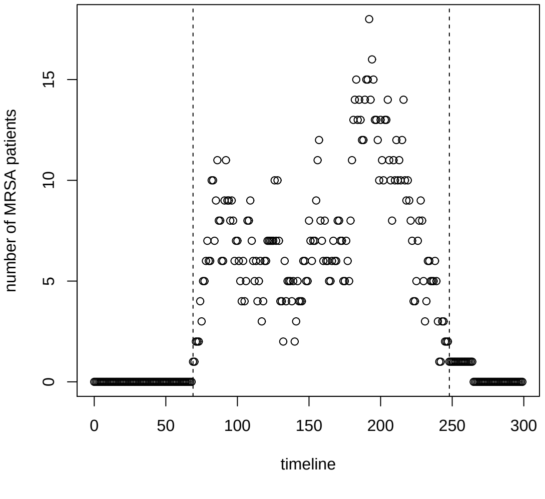


Figure 7.20: The plot of simulated number of MRSA positive patients as the time increases. The vertical lines are the start and the end of the study.

We focus on the period of the study of the Screening Pathfinder Programme, which began from the 69th study day (i.e the admission date of the first patient who had been recruited for the study is the 8th of February, 2010 and was in the admission only cohort) and ended on the 248th study day (i.e the last discharge date for the recruited patient is the 6th of August, 2010 and was in the discharge only cohort). Note that the last date of admission in the admission only cohort recorded in the dataset is 26th of July, 2010. Generally speaking, there are around 6-11 MRSA positive patients per day in the early period and 11-15 MRSA positive patients per day in the peak period. The number of MRSA positive patients increases with time until roughly the 200th day of the study (i.e. the peak of the number of MRSA positive patients), from when on the number of MRSA positive patients decreases significantly. The decline of the number of MRSA positive patients starting from the 200th day continues until the end of the study. One of the reasons is that the patients discharged after the 248th day did not give the information of MRSA status on admission. The recruitment to the study is also tailing off in the end. However, there are oscillations appearing during the increasing period. i.e. two valleys are displayed in the plot on around the 130th study day and the 175th study day.

We use the bootstrap method to construct the simulation envelope for the imputed number of positive MRSA patients in the admission-discharge, admission only and discharge only cohorts. Figure 7.21 shows that the width of simulation envelope becomes relatively large in the middle of the study (roughly between 100th day and 220th day). This indicates the large variances in the simulations for the length of stay, MRSA status and movement dates.

In order to validate the tenability of the simulation approach for the length of stay using the empirical density functions and the MRSA measurement results, we generate the simulation envelope of the percentage of MRSA positive patients in all three cohorts. Then we compare the simulation envelope to the simulated percentage of MRSA positive patients in the admission-discharge cohort which contains the complete information of length of stay and MRSA status. We assume that there is no selection bias in this study [106]. In other words, the percentage of MRSA positive patients in three cohorts would not be statistically different. The plot is shown in Figure 7.22 focusing on the period that patients were recruited in the admission-discharge cohort.

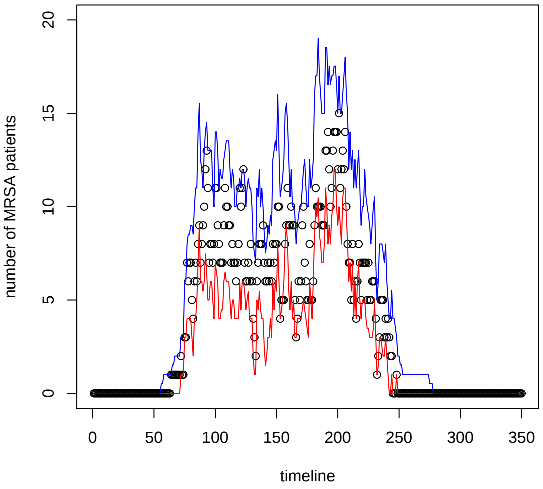


Figure 7.21: The plot of simulated number of MRSA positive patients with the simulation envelope. Here the red line is the lower bound of the simulation envelope while the blue one is the upper bound of the simulation envelope. The simulated number of MRSA positive patients are marked as black points.

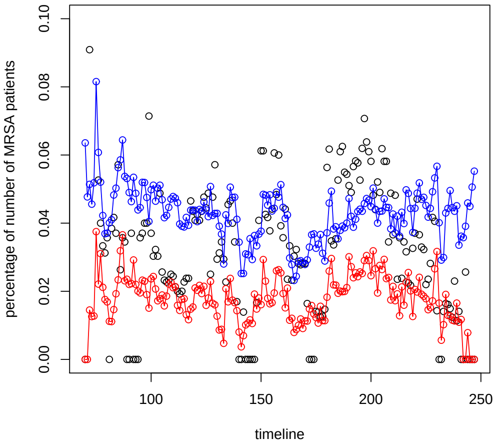


Figure 7.22: The plot of the percentage of simulated number of MRSA positive patients in the admission-discharge cohort with simulation envelope. Here the red points are the lower bounds of the simulation envelope while the blue ones are the upper bounds of the simulation envelope. The simulated percentage of number of MRSA positive patients are marked as black points.

Generally speaking, a large number of the imputed percentage of MRSA positive patients on a particular study day in the admission-discharge cohort lie outside the simulation envelope. Specifically, the imputed percentages of positive MRSA patients in the neighbourhood of the 200th study day, which is the peak of the distribution, exceed the simulated upper bounds slightly. This indicates that the theoretical distributions which we assumed previously and are used for the imputation of length of stay and MRSA status in the admission only and discharge only cohort might be biased.

Since the sample size of the discharge only cohort (1,452) is similar to the size of the admission-discharge cohort (1,580), we generate the simulation envelope for the number of MRSA positive patients in the discharge only cohort based upon the imputed length of stay and MRSA status on admission, and compare it with the simulated number of MRSA positive patients in the admission-discharge cohort. The results are displayed in Figure 7.23, where a large amount of the simulated number of MRSA positive patients in the admission-discharge cohort lie within the simulation envelope. However, the peak of the number of MRSA positive patients in the discharge only cohort occurs slightly earlier in the study compared to the peak in the admission-discharge cohort. On around the 200th study day, the simulated number of MRSA positive patients in the admission-discharge cohort is much larger than the simulated number of MRSA positive patients in the discharge only cohort. Comparing Figure 7.21 with Figure 7.23, we can see that the fluctuation of the number of MRSA positive patients is slightly more moderate in the discharge only cohort and the peak of the number of MRSA positive patients in the discharge only cohort is on about the 190th study day, which arises earlier than the peak in the admission-discharge cohort. As there is no real data of the dates of transfer between wards, it is difficult to validate the imputation approach for the movement dates.

Furthermore, according to the matrix of the number of MRSA positive patients in each ward on a specific study day, we are able to plot the pattern of MRSA positivity by ward displayed in Figure 7.24. This shows that the points concentrate in the middle of the timeline, implying that number of MRSA positive patients generally increases over the study whereas towards the end of the study, the number of MRSA positive patients decreased. Considering each single study day, a relatively large number of MRSA positive patients stayed in ward ‘49’. This can be explained by the fact that a reasonable number

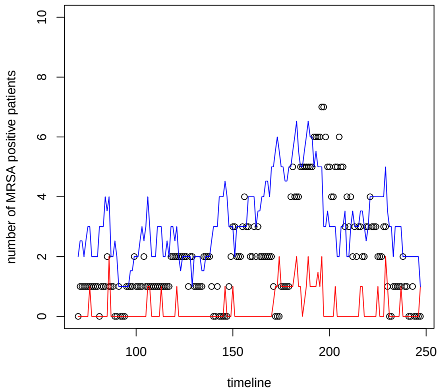


Figure 7.23: The plot of simulated number of MRSA positive patients in the admission-discharge cohort with simulation envelope based on the discharge only cohort. The blue points are the upper bound of the simulation envelope and the red points are the bottom bound.

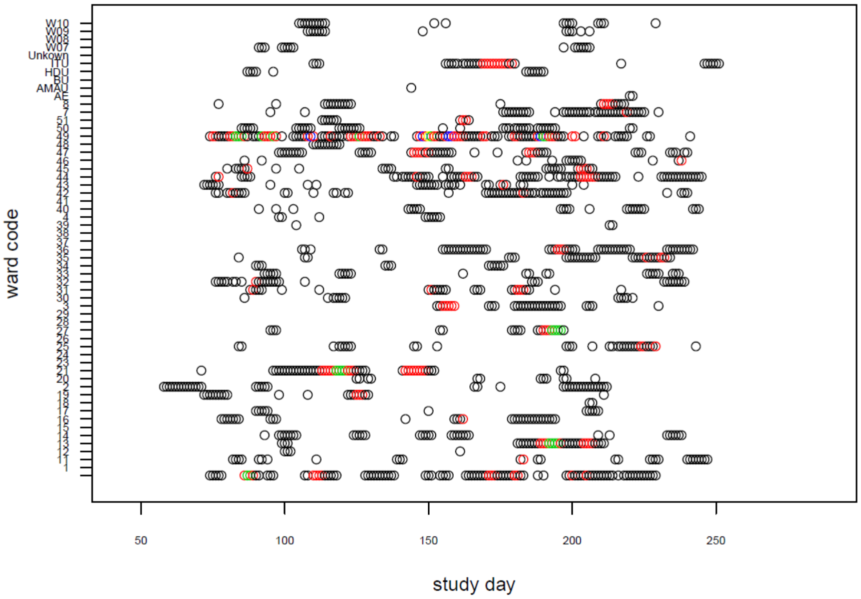


Figure 7.24: The plot of imputed number of MRSA positive patients (ward codes against timeline). Here the black points represent that there is one MRSA patient that day and the red, green, blue, yellow points represent two, three, four and five patients who are MRSA positive.

of patients were admitted or transferred into ward '49' which is a mixed specialty ward involving various types of treatment (i.e. types of specialty) such as 'Cardiac Surgery', 'Care of Elderly', 'Infectious Diseases', 'General Medicine', 'General Surgery' and other types of treatment. It is also evident that there are MRSA positive patients in virtually all wards though there are wards with larger exposure to MRSA than others.

7.3.4.2 The analysis of the number of MRSA positive patients based on the simulated length of stay obtained from the Simulation Method 2: Negative Binomial regression.

In order to investigate the sensitivity of using a different simulation approach for the length of stay, we build the matrix presenting the number of MRSA positive patients for each ward code on a given study day based on the simulated length of stay given by the fitted Negative Binomial regression. Also we apply the bootstrap method to generate the corresponding simulation envelope of the frequency of the population of MRSA positive patients.

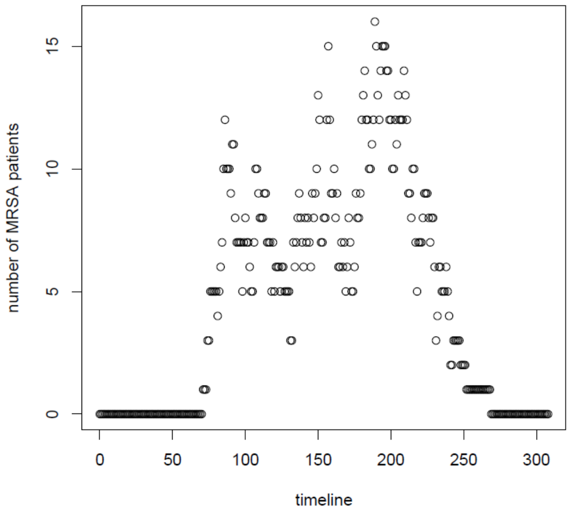


Figure 7.25: The plot of simulated number of MRSA positive patients based on the Negative Binomial simulation method.

The plot (Figure 7.25) of the total number of MRSA positive patients against the study day shows a similar pattern as the plot in Figure 7.20 where the imputed lengths of stay were obtained from the empirical density. This indicates that those two different imputation approaches for the length of stay do not markedly affect the imputation of the number of positive MRSA patients on a given study day.

The percentage of MRSA positive patients on a given study day in the

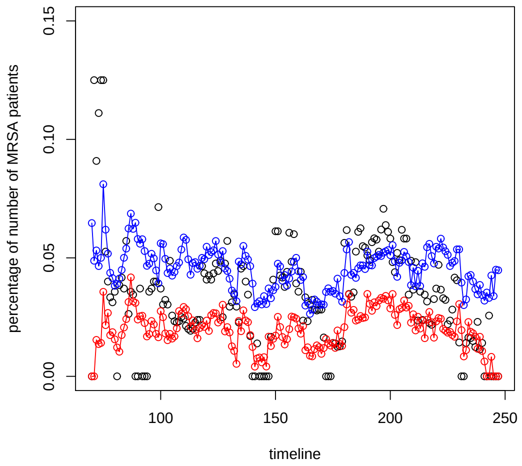


Figure 7.26: The plot of percentage of MRSA positive patients with the simulation envelope based on the Negative Binomial simulation method. The blue points are the upper bounds of the simulation envelope and the red points are the bottom bounds of the simulation envelope which is generated in the three cohorts. The black points are the imputed percentage of MRSA positive patients in the admission-discharge cohort.

admission-discharge cohort with the simulation envelope is shown in Figure 7.26. We can conclude that the observed percentage of positive MRSA patients exceeds the upper bound of the simulation envelope on a few occasions at the beginning of the study, which reflects an underestimation of the percentage of MRSA positive patients. The same situation occurs around the 200th study day where the percentage of positive MRSA patients in the admission-discharge cohort is also slightly higher than the upper bound of the simulation. Note that we set the beginning day of the study as 1st of October 2009 in order to guarantee to involve all the imputed admission dates for every patient in the discharge only cohort. Compared to the plot in Figure 7.23 where the empirical density was used when simulating the length of stay, there is no obvious evidence that the Negative Binomial regression for the simulation of the length of stay gives vastly different results for the total number of MRSA positive patients in a given ward on a given study day.

One of the aims of this study is to assess the effect of patient movement on the risk of MRSA acquisition. The matrix shown above in Figure 7.24 presents the location of the MRSA patients in the wards in hospital. Three indicators associated with patient movement which might be associated with the risk of acquiring MRSA are: (i) whether or not the patient is staying in a ward with MRSA present, (ii) the number of days that patient had been exposed to at least one MRSA positive patient and (iii) the total patient days of MRSA exposure. These indicators are considered as potential risk factors for MRSA acquisition and all can be generated from combining the matrix with the journey of each individual patient through the hospital. In the following section, we aim to analyse the effects of these three factors on the risk of MRSA acquisition based on imputation and logistic regression.

7.4 The logistic regression for exposure variables based on patient movements.

In Section 7.2 and Section 7.3, we investigated a methodology of using imputation to fill in the gaps in data availability brought about by consent and recording issues. This methodology was also investigated as an attempt to see if data which are relatively routinely recorded could be used to provide information on the effect of patient movement on the risk of acquiring MRSA. Firstly, we imputed the date of transfer from one ward to another in the admission-discharge cohort, and then imputed the length of stay (using empirical distribution approach in this study), and the date of transfer between wards in the admission only cohort, and also imputed the length of stay, the MRSA status on admission and then the date of transfer between wards in the discharge only cohort. After that, we derived the data matrix of the number of MRSA positive patients per ward per day. In this way three explanatory variables could be calculated for each person, namely (i) exposure to MRSA, (ii) days exposed to MRSA and (iii) patient days exposed to MRSA. As only the admission-discharge cohort has complete data on MRSA acquisition, the estimation of the effects of those three explanatory variables on MRSA acquisition can only be done in the admission-discharge cohort. A bootstrap method is applied to estimate the variability of the parameter estimates in the logistic regression model by repeating the whole process 100 times. Note that the repeated process includes the imputation of the length of stay, the dates of transfer between wards, the MRSA status and then the construction of data matrix and logistic regression analysis.

Logistic regression will be used in this section to estimate the effect of patient movement on MRSA acquisition. Particularly, three hypotheses are tested. These are (i) a patient exposed to MRSA has a higher risk of MRSA acquisition than a patient who was not exposed to MRSA (i.e. the variable exposure to MRSA). (ii) a patient exposed to MRSA for a longer length of stay has a higher risk of MRSA acquisition compared to a patient unexposed or exposed for a shorter duration (i.e. days exposed to MRSA). (iii) a patient exposed to MRSA for a larger number of patient days has a higher risk of MRSA acquisition than a patient with less exposure (i.e. the variable patient days exposed to MRSA). The analysis will only be carried out for the subset of the admission-discharge cohort who were MRSA negative on admission. The admission-discharge cohort contains the information on MRSA acquisition for each patient which is treated as the response variable in the logistic regression. Data from both the admission only and discharge only cohorts were only used to impute the data on MRSA carriage which contains the uncertainty. We take the variability induced by the imputation into account by using a bootstrap method within the logistic regression to provide the appropriate standard errors for the parameters associated with the imputed exposure variables. As we mention in Section 6.2 of the previous chapter, we use a 10% significance level for the univariate analysis to include the significant variables for the multivariable analysis.

Only the data for Aberdeen Royal Infirmary is used for analysis in this chapter. Hence it is a subset of the analysis in the previous chapter. The estimates from some of the models in Chapter 6 are repeated in this chapter so that they can be directly compared with the estimates from extended models which also included the exposure to MRSA variables.

7.4.1 The influence of being with MRSA positive patients in the same ward on MRSA acquisition.

Cohabitation, where a patient is in a ward at the same time as there are MRSA patients in the same ward, is treated as one of the important statistics inferred from the patient movement within the hospital. In the admission-discharge cohort, the point estimates show that there were 273 patients (17.28%) who had stayed in a ward with one or more than one MRSA positive patients whereas 1,307 patients had not been with any MRSA positive patient in hospital.

Univariate analysis

In order to investigate the statistical association between the risk of MRSA acquisition for a patient and the presence of MRSA positive patients in the same ward that patient had stayed for the admission-discharge cohort, we construct a logistic regression model. Firstly, univariate analysis is used to identify the important variables with respect to the risk of MRSA acquisition. Apart from the main risk factor in focus: whether or not patients are in wards with MRSA present, there are eight other potential risk factors involved in the univariate risk factor analysis for the MRSA acquisition. The results are shown in Table 7.9.

Table 7.9 illustrates that age has a significant effect on the risk of MRSA acquisition in this subset analysis. As the age increases the odds ratio increases significantly. Although the *p*-value (0.16) from the Wald test for age as a categorised potential risk factor does not demonstrate a significant difference between different age groups with respect to MRSA acquisition, the risk of acquiring MRSA for the patients 80 years old or over is approximately eight times as high as for the younger patients 49 years old or under. In addition, the results also show that the admission speciality as well as the length of stay are also significant potential risk factors. For patients admitted into renal or orthopedic speciality wards, the risk of acquiring MRSA is obviously different from that for the patient admitted into a ward in the 'Medicine group' since the corresponding odds ratios

Table 7.9: Univariate risk factor analysis for MRSA acquisition (N=1,580).

Variables	Categories	OR	p-value	95% CI	Combined p-value (Wald test)
Gender	Male	1			
	Female	1.46	0.418	(0.58,3.65)	na
Age (years)	≤ 49	1			
	$50-64$	2.83	0.35	(0.32,25.49)	
	$65-79$	5.48	0.11	(0.69,43.41)	
	≥ 80	8.21	0.056	(0.95,70.74)	0.16
Age (numeric)		1.05	0.0085	(1.01,1.09)	na
Admission specialty	Medicine	1			
	Renal	5.41	0.043	(1.06,27.76)	
	Cardiology	0.56	0.58	(0.07,4.59)	
	Oncology	1.23	0.85	(0.15,10.23)	
	Orthopedics	0.26	0.095	(0.05,1.26)	
	Surgery	0.45	0.18	(0.14,1.44)	0.044
Length of stay		1.04	0.0049	(1.01,1.07)	na
Patient admission type	Emergence	1			
	Elective	0.46	0.12	(0.18,1.23)	na
Number of wards		1.20	0.32	(0.84,1.70)	na
Number of wards (categorised)	1 ward	1			
	2 wards	1.46	0.51	(0.47,4.49)	
	≥ 3 wards	2.39	0.11	(0.82,6.95)	0.28
Co-morbidity: wounds/ulcers	No	1			
	Yes	3.06	0.05	(1.00,9.35)	na
Co-morbidity: renal failure	No	1			
	Yes	4.31	0.06	(0.96,19.29)	na
MRSA existence	No	1			
	Yes	1.72	0.3	(0.62,4.82)	na

for renal special wards and orthopaedic special wards are 5.41 and 0.26 with the p-values 0.043 and 0.095 respectively. The appearance of open wounds and renal failure also have potentially significant effects on the risk of MRSA acquisition. The admission speciality is a significant variable associated with MRSA acquisition in the analysis of a subset of Aberdeen Royal Infirmary dataset whereas it was insignificant in the univariate analysis of the previous chapter, which includes the data on both Aberdeen Royal Infirmary and Crosshouse. This indicates that the data on admission specialty in Crosshouse hospital weakens the effect on MRSA acquisition.

For the major potential risk factor under investigation in this chapter of whether or not the patient stayed in a ward with an MRSA patient, the risk of MRSA acquisition for a patient being in a ward with MRSA present is 1.72 times the risk for the patient who had never been exposed to MRSA while in hospital, with the corresponding 95% confidence interval (0.62, 4.82). However, there is no evidence that being in a ward with MRSA positive patients has a significant effect on the risk of MRSA acquisition due to the high p-value (0.30).

Multivariable analysis

A multivariable model is constructed using MRSA existence (as this is the main factor under investigation in this section) and the other significant univariate variables. The results are shown in Table 7.10.

Table 7.10: Multivariable analysis for MRSA existence (N=1,580).

Variables	Categories	Adjusted OR	p-value	95% CI	Combined p-value (Wald test)
Age (numeric)		1.05	0.015	(1.01,1.09)	na
MRSA existence	No	1			
	Yes	0.82	0.72	(0.26,2.54)	na
Admission specialty	Medicine	1			
	Renal	7.00	0.053	(0.98,50.31)	
	Cardiology	0.52	0.55	(0.06,4.38)	
	Oncology	1.40	0.76	(0.16,12.03)	
	Orthopedics	0.35	0.21	(0.07,1.80)	
	Surgery	0.57	0.35	(0.17,1.88)	0.12
Length of stay		1.04	0.041	(1.00,1.07)	na
Co-morbidity: wounds/ulcers	No	1			
	Yes	2.15	0.23	(0.61,7.60)	na
Co-morbidity: renal failure	No	1			
	Yes	1.12	0.90	(0.16,7.99)	na

The adjusted odds ratio for the MRSA existence displayed in Table 7.10 means that by taking the effect of age, admission specialty, length of stay, open wounds and renal failure into account, the risk of MRSA acquisition for patients staying in wards with MRSA present is 0.82 times the risk for patients who had never been exposed to MRSA before while in hospital. The corresponding p-value (0.72) implies that the effect of being in a ward with MRSA is not statistically significant on the risk of MRSA acquisition. Additionally, the adjusted effect of open wounds or ulcers and renal failure are non-significant on the risk of MRSA acquisition although the corresponding adjusted odds ratios still show that the patients with open wounds or renal failure are more likely to acquire MRSA. Compared to the results in the previous chapter, the higher p-values for these variables are possibly associated with the smaller sample size.

The two way interactions between the potential risk factors were investigated using the Chi-square test to compare the nested models where one of them consists of an interaction item. No interaction effects were significant.

7.4.2 The effect of the number of days exposed to MRSA on MRSA acquisition.

From one simulation for the admission-discharge cohort, the result shows that the majority of the patients (82.72%) had never been exposed to MRSA directly. Additionally, 103 out of 1,580 patients had been exposed to MRSA for exactly one day during their stay in hospital and 49 patients had exactly two days being exposed to MRSA. The histogram of the total number of days that patients had been exposed to MRSA is shown in Figure 7.27, indicating that the relative frequency of the number of days exposed to MRSA drops dramatically when the number of days exposed to MRSA becomes larger.

We can see from the histogram that the frequency of patients being exposed to MRSA for more than five days is relatively low. Hence it is reasonable to combine all the patients who had been exposed to MRSA for more than five days together and categorise the total number of days exposed to MRSA positive patients into six groups for the analysis: zero days, one day, two days, three days, four days and five or more days.

Univariate analysis

The results in Table 7.11 show that compared to patients who had not been exposed

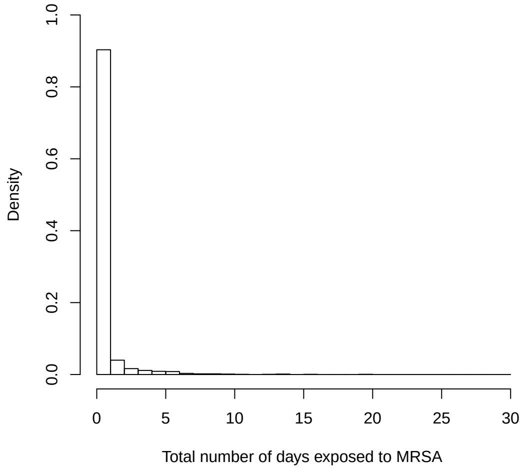


Figure 7.27: The histogram of the total number of days exposed to MRSA for one imputation in the admission-discharge cohort.

to MRSA, the risk of acquiring MRSA increases as the number of days exposed to MRSA increases but patients exposed to MRSA for more than five days have a relatively low odds ratio. A patient exposed to MRSA for four days in hospital is 4.20 times as likely to acquire MRSA as a patient never exposed to MRSA in hospital whilst a patient exposed to MRSA for one day in hospital is 0.91 times as likely to acquire MRSA as a patient never exposed to MRSA. However, the high overall p-value provides no evidence of a significant difference in the risk of MRSA acquisition between patients exposed to MRSA for a long time and a patient never exposed to MRSA while in hospital. The total number of days exposed to MRSA does not have significant effect on MRSA acquisition due to the high p-value (0.72) from the Wald test. The trend test also gives a high p-value (0.913), which shows that there is no linear trend in the odds ratio for the days exposed to MRSA.

Multivariable analysis

The results of the multivariable analysis are shown in Table 7.12. The adjusted odds ratio generally increases as the number of days exposed to MRSA increases but when the number of days exposed to MRSA is larger than three, the adjusted odds ratio does not increase further. For a patient who was exposed to MRSA for three days, the risk of

Table 7.11: The univariate analysis for total number of days exposed to MRSA positive patients ($N = 1,580$).

Variables	Categories	OR	p-value	95% CI	Combined p-value (Wald test)
The total number of days exposed to MRSA	0 days	1			
	1 day	0.91	0.92	(0.11,6.95)	
	2 days	1.92	0.53	(0.25,14.93)	
	3 days	2.50	0.38	(0.32,19.49)	
	4 days	4.20	0.18	(0.52,33.34)	
	$\geq$ 5 days	1.57	0.67	(0.20,12.11)	0.72

Table 7.12: Multivariable analysis for total number of days exposed to MRSA positive patients ($N = 1,580$).

Variables	Categories	Adjusted OR	p-value	95% CI	Combined p-value (Wald test)
Age (numeric)		1.05	0.012	(1.01,1.09)	na
The total number of days exposed to MRSA	0 days	1			
	1 day	0.49	0.52	(0.06,4.20)	
	2 days	1.81	0.58	(0.22,15.00)	
	3 days	1.58	0.67	(0.19,13.35)	
	4 days	1.09	0.94	(0.09,13.36)	
	$\geq$ 5 days	0.46	0.48	(0.05,3.94)	0.91
Admission specialty	Medicine	1			
	Renal	8.04	0.043	(1.07,60.64)	
	Cardiology	0.54	0.57	(0.06,4.58)	
	Oncology	1.55	0.69	(0.18,13.61)	
	Orthopedics	0.36	0.23	(0.07,1.88)	
	Surgery	0.58	0.38	(0.18,1.94)	0.11
Length of stay		1.04	0.037	(1.00,1.08)	na
Co-morbidity: wounds/ulcers	No	1			
	Yes	2.19	0.24	(0.60,8.00)	na
Co-morbidity: renal failure	No	1			
	Yes	1.13	0.91	(0.14,9.45)	na

MRSA acquisition is 1.58 times as high as the risk of MRSA acquisition for a patient who was never exposed to MRSA while in hospital by taking the other risk factors into account. On the other hand, for a patient who was exposed to MRSA for five days or more, the risk of MRSA acquisition is 0.46 times as high as a patient who was never exposed to MRSA while in hospital, after controlling for the effects of other risk factors. The corresponding high p-values for the adjusted odds ratios indicate that by adjusting for the effects of other risk factors of age, admission speciality, length of stay, wounds or ulcers and renal failure, the risk of MRSA acquisition between a patient exposed to MRSA for a larger number of days and a patient never exposed to MRSA is not statistically different. The combined p-value for the Wald test also provides the evidence that the adjusted effect of number of days exposed to MRSA has no significant effect on MRSA acquisition. Compared to the univariate analysis, there is a big change in the adjusted odds ratio for each category of the total number of days exposed to MRSA. Especially for the patients being exposed to MRSA for exactly four days, the adjusted odds ratio becomes 1.09 in the multiple logistic model while the crude odds ratio is 4.20 in Table 7.11. Besides, the adjusted odds ratios for the admission speciality, open wounds or ulcers and renal failure are different from the relative crude odds ratios in Table 7.9 respectively. Hence, it is possible that a confounding effect exists between those potential risk factors. The interaction tests revealed that there is no significant interaction term according to likelihood ratio tests for the nested models.

7.4.3 The effect of the total number of patient days exposed to MRSA.

Exposure can also be measured by the number of patient days that patient had been exposed to MRSA. Each patient day represents a unit of time during which the patient was in a ward at the same time as there was an MRSA patient in the same ward, counted according to the number of MRSA patients. Thus the patient staying in a ward with two MRSA positive patients for one day would represent two patient days. In other words, the number of patient days exposed to MRSA for each patient can be calculated by totalling the number of MRSA positive patients that the patient had stayed with on individual days while in hospital.

In one imputation, there are 1,307 out of 1,580 patients who had zero patient days being exposed to MRSA which comprises 82.72% of patients, followed by 5.63% of patients having exactly one patient day exposed to MRSA and 3.16% of patients having exactly two patient days exposed to MRSA. The histogram is displayed in Figure 7.28.

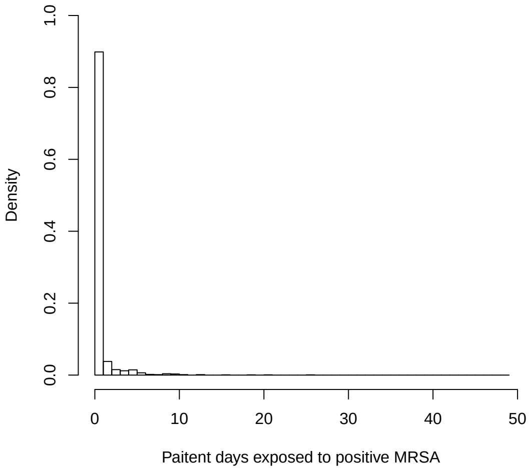


Figure 7.28: The histogram of the total patient days exposed to MRSA in one imputation.

Clearly, the histogram of the number of patient days exposed to MRSA shows a dramatic decrease, which has a similar distribution as the histogram for the number of days exposed to MRSA in Figure 7.27. The frequencies of patients exposed to MRSA for more than five patient days are relatively small compared to the frequency of the patients with zero patient days exposed to MRSA. Hence we build the new categorised variable involving four categories: zero days, one to two days, three to four days and five or more days.

Univariate analysis

The univariate analysis results are listed in Table 7.13. Compared to the zero patient days exposed to MRSA, the odds ratio for three to four patient days is the highest (3.42), followed by the odds ratio for five or more patient days (2.43). For the patients with one to two patient days, the risk of acquiring MRSA is only 0.67 times the risk of MRSA acquisition for the patients with zero patient days exposed to MRSA. The high p-value for the Wald test (>0.10) suggests that the effect of patient days exposed to MRSA is

Table 7.13: The univariate analysis for patient days exposed to MRSA ($N = 1,580$).

Variables	Categories	OR	p-value	95% CI	Combined p-value (Wald test)
The patient days exposed to MRSA	0 days	1			
	1-2 days	0.67	0.70	(0.08,5.13)	
	3-4 days	3.42	0.11	(0.76,15.43)	
	$\geq$ 5 days	2.43	0.25	(0.54,10.88)	0.27

not significantly associated with the risk of MRSA acquisition. In other words, there is no evidence that a patient staying with two or more positive MRSA patients in a ward at the same time has a higher risk of MRSA acquisition compared to a patient who stays with only one positive MRSA patient, or none.

Multivariable analysis

Table 7.14: Multivariable analysis for total number of days exposed to MRSA ($N = 1,580$).

Variables	Categories	Adjusted OR	p-value	95% CI	Combined p-value (Wald test)
Age (numeric)		1.05	0.010	(1.01,1.10)	na
The patient days exposed to MRSA	0 day	1			
	1-2 days	0.38	0.38	(0.05,3.21)	
	3-4 days	2.88	0.19	(0.59,14.02)	
	$\geq$ 5 days	0.59	0.55	(0.11,3.24)	0.38
Admission specialty	Medicine	1			
	Renal	7.72	0.042	(1.07,55.54)	
	Cardiology	0.55	0.58	(0.06,4.67)	
	Oncology	1.57	0.68	(0.18,13.71)	
	Orthopedics	0.35	0.21	(0.07,1.80)	
	Surgery	0.56	0.35	(0.17,1.86)	0.12
Length of stay		1.04	0.038	(1.00,1.07)	na
Co-morbidity: wounds/ulcers	No	1			
	Yes	2.20	0.23	(0.61,7.97)	na
Co-morbidity: renal failure	No	1			
	Yes	1.37	0.75	(0.19,9.79)	na

Table 7.14 shows that the risk of MRSA acquisition for a patient exposed to MRSA for three to four patient days is 2.88 times as high as the risk of MRSA acquisition for a

patient never exposed to MRSA after adjusting for the other risk factors of age, length of stay, wounds or ulcers and renal failure. On the other hand, by taking other risk factors into account, a patient exposed to MRSA for five or more patient days is 0.59 times as likely to acquire MRSA as a patient never exposed to MRSA and a patient exposed to MRSA for one to two patient days is 0.38 times as likely to acquire MRSA as a patient never exposed to MRSA. The p-value for the Wald test suggests that from a statistical point of view there is no significant adjusted effect of the number of patient days exposed to MRSA associated with the risk of MRSA acquisition in Table 7.13. Generally speaking, the adjusted odds ratio for each category of the patient days decreases compared to the corresponding crude odds ratio. Especially, for greater than or equal to five patient days, the adjusted odds ratio (0.59) is remarkably different from the corresponding crude odds ratio (2.43). Similarly, the adjusted odds ratios for the admission specificity, open wounds or ulcers and renal failure are obviously different from the corresponding crude ratios in Table 7.9, which reveals that there are possible confounding effects in the multivariable model. The results of the Chi-square tests show that there are no interaction terms which significantly affect MRSA acquisition.

7.4.4 The bootstrap method for the patient movement variables.

In the previous subsections, only one imputation was done to construct three explanatory variables of exposure to MRSA. Then the logistic regression models were fitted, yielding the point estimates of the odds ratios for those three explanatory variables. Obviously, with another set of imputations, different values for the exposure variables would be obtained and thus a logistic regression model would provide different estimates. In order to increase the variability of the estimation, we use multiple imputations to build up the distribution of the estimates. We apply the logistic regression within a bootstrap imputation to generate the 95% confidence intervals of the odds ratios for risk of MRSA acquisition associated with the patient being exposed to MRSA, for the number of days that the patient had been exposed to MRSA and for the total number of patient days exposed to MRSA respectively. This can be achieved by repeating the imputation and then the logistic regression model for 100 times. By iteratively imputing the dates of transfer between the wards, a set of data matrices could be constructed. After that, for each data matrix, three explanatory variables of exposure to MRSA can be calculated and then a set of those three variables is built up. We apply the logistic regression model on each set of three variables and then average the estimated odds ratios from the fitted models combined with the mean of the corresponding estimated standard variance. From a mathematical point of view, the 95% confidence interval for the odds ratio can be expressed as follows:

$$(exp(\overline{coef} - 1.96 \times (std(coef) + \overline{Std})), exp(\overline{coef} + 1.96 \times (std(coef) + \overline{Std})),$$

where $\overline{coef}$ is the mean of the estimated coefficients from the models filled to the bootstrap simulations, $std(coef)$ is the standard deviation of the estimated coefficients (i.e. $\sqrt{var(coef)}$ where $var(coef)$ is the variance of the estimated coefficients) from the fitted models and $\overline{Std}$ is the mean of the estimated standard deviation for the risk factor from the fitted models [62]. This technique is used for the admission-discharge cohort, imputed by the individual length of stay for each ward.

The results for the admission-discharge cohort in Table 7.15 show that a patient being concurrently in a ward with an MRSA patient present while in hospital does not affect the acquisition of MRSA significantly since one lies within the 95% confidence interval of the odds ratio for the risk factor that the patient was exposed to MRSA. Similarly, an increase in either the number of exposure days or patient days exposed does not raise the risk of acquiring MRSA. Comparing the 95% bootstrap confidence intervals with the corresponding confidence intervals for just one imputation, we can see that the width of the bootstrap confidence intervals are wider. Note that the confidence intervals for patient being exposed to MRSA, exposure days and patient days exposed for just one imputation are (0.272,2.572), (0.788, 1.195) and (0.872, 1.134) respectively.

Table 7.15: The 95% CI of the odds ratios for the three variables within the univariate analysis (admission-discharge cohort).

Variables	Estimate	95% CI
Patient being exposed to MRSA	0.692	(0.125,3.823)
Exposure days	1.003	(0.684,1.471)
Patient days exposed	0.993	(0.710,1.388)

For the admission-discharge cohort, using the two-way table for the number of patients who had acquired MRSA while in hospital against the number of patients being exposed to MRSA from the fitted logistic regression model within each bootstrap simulation, we are able to calculate the average proportion of patients acquiring MRSA but never being exposed to MRSA (i.e. patients who had acquired MRSA and had not stayed in a ward with any MRSA positive patient while in hospital) which is 0.765 with standard deviation 0.037. This implies that the majority of patients who acquired MRSA during their stay in hospital had not been exposed to MRSA. In reality, a patient can acquire MRSA by cross-transmission such as through the healthcare staff or from the environment [100], [106]. Similarly, the results of the average fraction of patients who had never acquired MRSA but stayed in a ward with MRSA present (i.e. patients who had not acquired MRSA and stayed in a ward with at least another MRSA positive patient while in hospital) is approximately 0.168 with standard deviation 0.0093, revealing that a majority of patients who had been exposed to MRSA while in hospital remain MRSA negative. Obviously, the average proportion of patients who had acquired MRSA and stayed in a ward with at least one MRSA positive patient which is 0.235 is slightly higher than the average proportion of patients who had not acquired MRSA but stayed in a ward with at least one MRSA positive patient (0.168).

We can conclude that this analysis suggests that the patient movement leading to MRSA exposure does not affect the risk of MRSA acquisition significantly. So the main risk factor associated with MRSA acquisition are still patient related ones, namely age, length of stay, open wounds or ulcers and renal failure.

7.5 Conclusion.

In this chapter, we mapped the timeline of MRSA infection and carriage pressure in each ward in Aberdeen Royal Infirmary for all patients in three cohorts. We mainly aimed to investigate the effect of a patient directly exposed to MRSA in a ward while in hospital (i.e. a patient stayed with other positive MRSA patients in the same ward at the same time) on MRSA acquisition. In addition, a bootstrapped logistic regression method was used with multiple imputations to investigate the performance of the analysis in presence of missing data.

We analysed the distribution of the observed length of stay in the dataset and investigated the distribution of length of stay in each ward. We also applied various methods of imputation for the missing length of stay, patient ward movement dates and MRSA measurement results.

The analysis of the distribution of the length of stay implies that there is no prominent difference among the assumption that the length of stay for each ward is based upon a Uniform distribution, the assumption that the length of stay for each ward (apart from the last one) follows a Triangular distribution and the assumption that the length of stay for each ward follows a mixed distribution (i.e. the length of stay for the first ward that a patient had been admitted into follows a Triangular distribution while the length of stay for the remaining wards is based on a Uniform distribution) with regards to the effectiveness of simulating the patient movement dates. In this study, the assumption that the length of stay for each ward is based on a Uniform distribution was applied. Additionally, further analysis, applying the other assumptions for the length of stay in the individual ward yields similar conclusions. This implies that there is no significant influence of using different distributions to simulate the length of stay for the individual ward on estimating the variables associated with patient movements and their further association with MRSA acquisition.

In order to simulate the missing length of stay in the dataset, two different methods, which are using the empirical density of the observed length of stay and using the Negative Binomial regression, were implemented. By comparing the simulated length of stay with the observed ones and assessing the histogram of the observed length of stay with the simulation envelope, it can be concluded that both simulation methods failed to estimate the length of stay perfectly (i.e. there was a the problem of underestimation or overestimation, especially for the patients who had stayed in hospital for less than five days). With regard to the simulation of length of stay using Negative Binomial regression, the QQ plot for the fitted residuals shows heavy tails in the ends although the mean of the observed length of stay is obviously less than the variance. Hence Gaussian regression or quasi-Poisson regression can also be applied for possible improvements. However, the results for Gaussian regression also give a heavy tailed QQ plot, which means that there is no improvement. Furthermore, one of the advantages of using the Negative Binomial method for the simulation of the length of stay compared to the empirical density method is that it does not limit the range of the simulated length of stay. On the other hand, by sampling the simulated length of stay randomly from the corresponding empirical density, the results are guaranteed to lie within the domain of the empirical density.

An assumption was made to impute the MRSA status for a patient on each day in hospital. However, there is no reliable a priori knowledge on the duration of MRSA carriage. This assumption for the imputation of the MRSA status per day might be biased since a large number of patients had stayed in hospital for a short time and those patients were assumed to be colonised with MRSA the day before discharge. Therefore, it might lead to an overestimation on the days that the patients were colonised with MRSA while in hospital and the days that a patient exposed to MRSA.

In addition, we built the pattern of the patient movements along with the timeline of the study based on imputation so that the number of positive MRSA patients per day can be elucidated. The plot revealing the number of positive MRSA patients per day was shown in Figure 7.20. There was a drop from 200th day to the end of the study. One potential effect of this drop is that patients admitted later have a bias on exposure to MRSA, which is underreported in the study. In this study, there are 539 patients who were admitted into the hospital after 200th day in the admission-discharge cohort. The effect of patient movement on MRSA acquisition through exposure to MRSA in a ward can be detected by the logistic regression. The results show that the exposure to MRSA does not have significant influence. In other words, the risk of MRSA acquisition in this study, which is considered as a representative of the general Scottish in-patient population, is not statistically significantly associated with the patient movement. There were 34 patients who acquired MRSA whilst in hospital. This analysis here suggests that other sources of transmission such as sporadic MRSA which includes the transmission from outside the ward or via hospital staff or environment are possibly more important in the general hospital population.

The application of the Bayesian network analysis may lead to an understanding of the pattern dynamics of epidemic MRSA transmission [134]. Our study also reveals that the different simulation methods for the length of stay do not significantly affect the conclusion of the estimated effect of the patient movements on the risk of MRSA acquisition. Moveover, the results of the bootstrap analysis also indicate that the majority of patients who had been exposed to MRSA did not acquire MRSA while in hospital.

There are many limits in this study. Our study does not address practices on comparing the simulated results with the real observed data. Further study is needed as an implementation study by collecting the data of patient movement dates between wards as a comparison with the simulation results obtained here. If the data on the dates of transfer between wards is available, no imputation would have been needed. However, the collection of the data, which might be available from hospital records, would be difficult and expensive. In this study, we used cohabitation as a surrogate for exposure to MRSA, which essentially implies that a patient is more likely to be exposed to MRSA if there is an MRSA patient in the ward. In fact, it is still possible to be exposed to MRSA if there is no patient in the ward. For example, the MRSA bacterium could be left by a positive MRSA patient the day before a new patient is admitted into the ward.

This study could be underpowered since the effect of MRSA exposure is too small to be detected. One of the reasons is that we may not have data on all MRSA patients. Most patients consented and were included in one of the three cohorts. We assumed that those three cohorts are representative but there could be bias in practice.

Some further work regarding to the analysis in this study can be considered. For example, we can investigate that whether the method of bootstrapped logistic regression is appropriate. This can be implemented by a replicated process. First we take one imputation for the exposure variable and then use this imputed variable to construct the logistic regression model associated with MRSA acquisition in the admission-discharge cohort. According to the fitted model, we can simulate the response of MRSA acquisition in the admission-discharge cohort, but with a big effect size for the exposure variable (e.g. take variable of days exposed to MRSA as 10). Then we get a new imputation for the exposure variable and applied the same process to estimate a new response of MRSA acquisition. By comparing those two estimated responses of MRSA acquisition, we can assess that if the process of bootstrapped logistic regression can be replicated. If those two estimated responses are close, then we can conclude that this process can be replicated. In addition, our work also raises the questions such as the re-admittance of the patients who were colonised with MRSA on discharged and the risk of transmission for the patient colonising MRSA at discharge to other household members, which could be addressed by further work.

Chapter 8

Conclusion and Discussion

In this thesis, we have used imputation approaches and modelling methods to investigate two problems associated with HIV and MRSA. Particularly, two main questions have been discussed in two separate parts of this thesis: (i) the quantification of replication in HIV anonymous test reports and (ii) the effect of patient movement between wards on the acquisition of MRSA. The first part of this thesis (i.e. Chapters 2, 3 and 4) focused on the replication problem in HIV reports and the second part (i.e. Chapters 5, 6 and 7) was about the analysis of the association between the patient movement between wards and the risk of MRSA acquisition while in hospital. In Chapter 1, we generally described the main objectives of this thesis and briefly introduced the methods that would be used for the analyses.

8.1 The discussion for the HIV replication study

The first objective was to estimate the true number of individuals who were recorded in the PHLS dataset and the amount of replication in this dataset. The HIV dataset consisted of the birth dates of the individuals but the names of patients were not recorded due to confidentiality. Hence there were multiple records with the same birth dates and it was not known if they are the same person or not. In the first part of this thesis, we used the maximum likelihood method to estimate the amount of replication. The second objective of this thesis was to estimate the impact of patient movement within a hospital on MRSA acquisition. The data were collected from two hospitals in Scotland. However, missing data caused severe problems when attempting to estimate the effect of patient movement. Hence, in the second part of the thesis, we used imputation to make inferences in the presence of missing data such as the movement dates of each patient. Then based upon logistic regression modelling, we assessed the impact of patient movement between the wards in hospital on MRSA acquisition.

8.1.1 The summary of the HIV replication study

Firstly, we investigated the replication problem in HIV reports in the first part of this thesis. In Chapter 2, we did the literature review of the replication in HIV reports. The background of HIV was introduced. HIV infection is considered as a pandemic disease in public health and it can cause impact on societal and economic well-being. Thus it is important to know the reliable information on the number of HIV infected individuals in order to improve public health. In the 1991 and 1994 datasets of HIV anonymous test reports given by the PHLS, only birth dates of individuals were held and there were repeated records of the same individual which were addressed in the previous researches by Greenhalgh et al. [27], [51], [52]. The PHLS was interested in estimating the duplication in both 1991 and 1994 datasets. Greenhalgh et al. used various statistical approaches to assess the replication in the PHLS dataset, but only for a few birth years in the 1991 dataset. In this thesis, we used the maximum likelihood method to estimate the percentage of replication present in the 1991 and 1994 datasets.

In Chapter 3, we introduced the methodology of estimating the replication in PHLS HIV reports and developed the bootstrap method for calculating the 95% confidence interval for the estimate of the proportion of replicated records of birth dates. The maximum likelihood method was the main technique used for estimating the replication.

Based upon the observed sample in a given birth year, the potential true replication vectors were derived iteratively. We derived a pattern for generating the potential replication vectors, which mapped the movements of the tuples. A theorem was given to calculate the upper bound for the number of potential true replication vectors so that the iterative computation for the potential replication vectors was able to be done efficiently using statistical software R and programming language C. We constructed the likelihood function for each potential replication vector based upon a theorem of probability distribution of a replication vector (given by Greenhalgh et al. [51]), which involved the potential replication vector, the potential true number of distinct individuals and the unknown probability distribution. Note that the unknown probability distribution gave the probabilities that the individuals having HIV tests once, twice, three times and so on. Using the 'alabama' library in the software R, the maximum likelihood estimate of the probability of individuals having had HIV tests repeatedly were calculated. Thus, the true number of distinct individuals was estimated for a given birth year and the corresponding percentage of the replication was calculated according to (3.5.1) in Section 3.5 of Chapter 3. In this chapter, we also proposed another approach based upon the estimated probability distribution of the individuals having had repeated HIV tests to calculate the percentage replication. Both methods gave virtually the same answer and thus in this thesis, we chose only the former approach to calculate the percentage replication. Since a point estimate of the replication is of limited use, we used the bootstrap method to calculate 95% confidence intervals. For a given birth year, according to the estimated probability distribution for the replication, we generated a set of samples of replication vectors using the statistical software R. For each sample, we used the same technique to estimate the corresponding true amount of replication. Then we deduced the 95% bootstrap confidence interval for that birth year.

In Chapter 4, we presented the results of the estimated percentage of replication and the corresponding 95% bootstrap confidence interval in the 1991 and 1994 datasets using the program written in R and C according to the method introduced in the previous chapter. The 'alabama' package in R was used to obtain the maximum likelihood estimates since the likelihood function of the potential replication vectors given the true sample size was generally nonlinear. The program in 'alabama' used the augmented Lagrangian algorithm. For the 1994 dataset which contained records of birth years with large sample sizes, we used the program written in C based upon the optimisation package nag_opt_nlp since the running time was cut down. The program written in R was introduced briefly in this chapter, which was the same as the program written in C. Using the program written in R (and in C), we estimated the true number of distinct individuals, the probabilities of an individual having a certain number of HIV tests and the corresponding maximum value of the likelihood function for each birth year in the 1991 and 1994 datasets. We also constructed the 95% and 99% confidence intervals for each birth year in both datasets. The results showed that the replication present in the 1991 dataset was 3.37% and in the 1994 dataset the replication decreased to 0.58%. The replication was smaller in the 1994 database than the 1991 database as expected because in the more recent years the establishment of the surname Soundex code used in recording the data provides better identification of duplicate reporting of the same individual. We found that the years where replication was estimated to be present by the method used here were the same as the ones identified by the matching pairs method [51]. The same conclusion was achieved by two different methods, ensuring confidence in the results.

In the previous analysis addressed by Greenhalgh et al. [27], [51], [52], only a few birth years in the 1991 dataset were analysed. Those birth years had very small sample sizes. In this thesis, we extended this work based upon the maximum likelihood technique. The entire 1991 and 1994 datasets were analysed where the sample sizes increased substantially. The confidence intervals for the percentage replication were also calculated using the parametric bootstrap method.

8.1.2 The limits and further works of the HIV replication study

The major limit of the HIV replication study is that for the large datasets, the computation efficiency is poor. i.e. the running time for calculating the amount of replication estimate is long for the large datasets. Moreover, the accuracy of the estimation depends on the available information. The improvements in providing better identification of duplicate reporting of the same individual would reduce the chance of replication in the datasets. In this study, the leap year is not considered when we construct the likelihood functions. We took $n = 365$ in the likelihood function (shown in (3.2.7)) for all the birth years in the datasets. In order to obtain the more accurate estimation, the leap year should be considered and n should be 366 for those leap birth years in the datasets.

Although more advanced recording techniques for the HIV reports have been used such as Soundex code using the surname of the patient to improve the accuracy, some replication will still exist. There may be different reasons for this such as after women are married they are likely to change their surname, and foreign patients, who were registered in the disease reporting system of the UK, are likely to have the same surname. Hence this method can also be used for estimating the duplication in the dataset. In this study, a uniform distribution for the birth dates of the patients was used to construct the likelihood function. Different assumptions for the distribution of the birth dates can also be considered as further work.

The method for estimating the replication can be applied in the other areas. For example, the NHS 24 dataset which contains the data of the age, gender, postal code and calling time of individuals to the NHS 24 helpline may have the duplicate records of the same individual. For example, an individual called the NHS 24 twice successively and the information of that individual was recorded repeatedly. By applying the same principles as we did in this study, the amount of duplication can be estimated, which indicates the reliability of the dataset. However, the likelihood function would be different. In this study, we assumed that the distribution of the birth dates for the patients is a uniform distribution whereas the distribution for the calling time in NHS 24 dataset may not be a uniform distribution.

8.2 The discussion for the MRSA acquisition study

Then we moved to the second part of this thesis, which focused on a particular hospital-acquired infection caused by MRSA. This worldwide spread of disease is considered as a global problem of public health. The main aim in this part of thesis was to estimate the effect of patient movement between wards on MRSA acquisition while in hospital from a one-year MRSA screening pilot study. We used the imputation to make inferences in the presence of the unknown data as the dates of movement from one ward to another. The second part of this thesis consisted of three chapters (i.e. Chapters 5, 6 and 7), which were literature review of MRSA, the risk factor analysis for MRSA acquisition in the hospital and the analysis of the effect of patient movements between wards on acquisition of MRSA.

8.2.1 The summary of the MRSA acquisition study

In Chapter 5, we briefly introduced four main aspects, which were (i) the medical, biological and economic background of MRSA, (ii) the published studies on MRSA (iii) the introduction of the MRSA Screening Programme launched in Scotland and (iv) the method of collecting data within the MRSA Screening Programme that was used for further analysis in the following chapters. MRSA is reported as the most frequently isolated organism in skin and soft tissue Healthcare Associated Infections (HAIs) [28] and it also causes bone, joint and surgical HAIs [72]. Although the incidence of MRSA has declined recently in several European countries, MRSA infections, which are resistant to the antibiotic methicillin, remain a major cause of morbidity and mortality in patients admitted to hospital, particularly those in intensive care unit (ICUs) [100].

In order to reduce the risk of cross-infections of MRSA in hospital, an MRSA Screening Programme has been implemented since 2007 and continues to be improved with the aim of establishing an efficient and economic prevention strategy. Recently, much research has been published on MRSA. However, the majority focused on the investigation of MRSA acquisition in presumed 'high risk' wards such as ICUs. These researches showed that isolation of MRSA positive patients may not be directly associated with interruption of the spread of MRSA. Several studies modelled the dynamic transmission process in ICUs using different methods. These studies took parameter uncertainty into account. However, those proposed models were limited to only a small population. In our study, we aimed to investigate the potential risk factors for MRSA acquisition, which may indicate suitable parameters to use in a dynamic transmission model.

A few studies were also published on MRSA acquisition in general hospitals, but to date, there is limited research focusing on the association between patient movement such as number of wards where the patient resided per hospital stay, and the risk of MRSA acquisition. The data collected from the one-year MRSA Screening pilot were also introduced briefly in this review chapter. Patient admission information was collected from the hospital Patient Administration System and the Clinical Risk Assessment and consent form were scanned into a holding database. All consenting patients were swabbed on admission and the results of MRSA colonisation were taken from the laboratory and recorded in the database. Generally speaking, data on demographics and risk factors for MRSA acquisition were collected for the analysis. The data used in our study consists of (i) an admission only database (7,181 patients) (ii) a discharge only database (2,432 patients) and (iii) a combined admission-discharge cohort (2,792 patients).

In Chapter 6, we used the data in the admission-discharge cohort to rework the investigation of the association between the risk of MRSA acquisition and the potential risk factors as the work done by Velzen et al., with the addition of an in depth evaluation of the role of the number of wards a patient was resident in. In this study, we found that 34 patients were MRSA negative on admission and MRSA positive on discharge.

These patients were considered as acquiring MRSA in hospital. We used the categorical logistic regression in the univariate risk factor analysis to estimate the effect of potential risk factors on MRSA acquisition using a 10% significance level. The results showed that three out of twelve potential risk factors for acquiring MRSA were identified, which were age, open wounds or ulcers and renal failure. The categorical variable of the number of wards did not show a significant effect on MRSA acquisition. Considering the association between the potential risk factors, we used the χ^2 test. The number of wards was strongly associated with age, length of stay, open wounds or ulcers and renal failure. In order to investigate the multiplicative interaction between the risk factors and the potential confounding effects, a stratified analysis was applied. The results indicated that there was no multiplicative interaction between the potential risk factors. Moreover, there was no significant trend in risk of MRSA acquisition as the number of wards increases. Based upon the CMH method, we identified the confounders in the multivariable analysis. A categorical multivariable logistic model was constructed.

In order to increase the power of the analysis, we also investigated the multivariable model using the numerical covariant values rather than categories. Generalised logistic regression was applied to test the nonlinearity in the numeric multivariable model. The results showed that a linear logistic model using numeric variables was adequate to estimate the effects of risk factors on MRSA acquisition. Comparing the categorical logistic model with the numeric one, there was no evidence to suggest that one of those two models is more reliable or predictive based upon the corresponding ROC curves. The findings in this chapter provided evidence that the cross-transmission of MRSA still takes place in Scottish hospitals and hence implementing contact precaution and infection control in the hospital is also important to prevent the cross-transmission. This conclusion was the same as the work done by Velzen et al.

In Chapter 7, we developed the methodology of assessing the effect of patient movement in general hospital on MRSA acquisition. We mapped the timeline of MRSA infection and carriage pressure in each ward in Aberdeen Royal Infirmary for all patients in three cohorts based upon imputation of transfer dates from one ward to another. In this chapter, various imputation methods were applied for making inference in the presence of the missing data on length of stay, patient movement dates and MRSA measurement results. The performance of different imputations were evaluated. In this study, we used a Uniform distribution to estimate the patient movement dates in the three cohorts. An assumption was also made to impute the duration of MRSA carriage but it might be biased and lead to an overestimate on the days that a patient was colonised with MRSA while in hospital. Our study also revealed that the different simulation methods do not significantly affect the conclusion of the estimated effect of the patient movements on the risk of MRSA acquisition.

Three exposure variables (i) whether a patient was exposed to MRSA in a ward, (ii) the number of days that a patient was exposed to MRSA and (iii) the number of patient days which this patient had spent staying with other positive MRSA patient or patients in the same ward simultaneously were calculated. Patient movement was measured as a volume indicator in terms of those three variables. In the more sophisticated analysis, those three variables were considered as the risk factors for MRSA acquisition and we investigated their effects on MRSA acquisition based upon the logistic regression method. The results showed that the effect of patient movement between wards was not significant on MRSA acquisition. This indicated that there were other transmission sources affecting MRSA acquisition in the general hospitals. For example, MRSA bacteria might be transmitted via hospital staff. Moreover, we used the bootstrapped logistic regression method with multiple imputations to investigate the performance of the analysis in presence of missing data. The results showed that the majority of patients who had possibly been exposed to MRSA did not acquire MRSA while in hospital.

8.2.2 The limits and further work of the MRSA acquisition study

However, there were some major limits in this study. This one-year MRSA screening pilot did not have all patients in the hospitals. In other words, selection bias existed. For example, relatively healthier patients may be more likely to sign a consent form to be recruited. In this study, only the data in the admission-discharge cohort were used for the logistic regression analysis since we needed both the MRSA measurements on admission and on discharge. This may lead to an underestimation of the proportion of patients acquiring MRSA. In addition, since acquisition of MRSA is not likely, the proportion of patients acquiring MRSA in hospital was small compared to the study population. Thus the power of the study is low. The results obtained in this study are equivocal since there were a lot of missing data. In addition, the method we used in this study is also feasible due to the missing data. The better data is required as the imputation is not totally successful and it does not yield robust conclusions. If the data on the dates of transfer between wards were available, no imputation would have been needed. However, the collection of the data would be difficult and expensive. For example, if many patients do not consent, data will not be available. The limited data is one reason for the low power to evaluate the effect of MRSA exposure. In this study, we assumed that a patient was exposed to MRSA when he or she was staying with other MRSA positive patients in the same ward simultaneously. However, in fact, there is a possibility that the MRSA bacterium could be left by a MRSA positive patient the day before a new patient is admitted into the ward.

Some further work can be considered regarding the study in the MRSA Screening Pathfinder project. If there are better data available such as the admission data in hospital including all transfers information, an investigation on comparing the simulated results with the real observed data can be done. Our work also raises the questions such as the re-admittance of the patients who were colonised with MRSA on discharge and the risk of transmission from the patient colonised with MRSA at discharge to other household members. In addition, the estimation of the effect of isolation on MRSA acquisition in the general hospitals can be investigated within the MRSA Screening Pathfinder Project when the other contact precautions have been taken into account.

Bibliography

[1] E.D. Acheson. *Medical Record Linkage.* Oxford University Press: London, 1976.

[2] A.E. Ades, J. Walker, B. Botting, S. Parker, D. Cubitt, and R. Jones. Effect of the world wide epidemic on HIV prevalence in the United Kingdom: record linkage in anonymous neonatal seroprevalence surveys. *AIDS*, 13:2437–2443, 1999.

[3] Health Protection Agency. HIV in the UK: 2011 annual report. Technical report, 2011.

[4] Health Protection Agency, Royal College of Nursing, and Infection Prevention Society. MRSA information for patient. Technical report, 2010.

[5] R. Aldrich and G. Wotherspoon. *Who's Who in Gay and Lesbian History.* London: Routledge, 2001.

[6] R.M. Anderson and R.M. May. *Infectious Diseases of Humans: Dynamics and Control.* Oxford University Press: Oxford, 1991.

[7] M. Arellano and G. Weber. Issues in indetification and linkage of patient records across an integrated delivery system. *J Health Care Infor Manag*, 12(3):43–52, 1993.

[8] L. Barabesi and M. Marcheselli. Parameter estimation in the clasical occupancy model. *Stat Methods Appl*, 20:304–327, 2011.

[9] F. Barre-Sinoussi, J. Chermann, F. Rey, M. Nugeyre, S. Chamaret, J. Gruest, C. Dauguet, C. Axler-Blin, F. Vezinet-Brun, C. Rouzioux, W. Rozenbaum, and

L. Montagnier. Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). *Science*, 220(4599):868–871, 1983.

[10] R. Batra, B.S. Cooper, C. Whiteley, A.K. Patel, D. Wyncoll, and J.D. Edgeworth. Efficacy and limitation of a Chlorhexidine-based decolonization strategy in preventing transmission of Methicillin-Resistant Staphylococcus Aureus in an Intensive Care Unit. *Clin Infect Dis*, 50:210–217, 2010.

[11] S. Benenson. *Control of Communicable Diseases in Man. 16th ed.* American Public Health Association: Washington D.C., 1990.

[12] V. Berridge and P. Strong. AIDS Policies in the United Kingdom: A Preliminary Analysis. In Fee E and Fox DM, editors, *AIDS: The Marking of a Chronic Disease*, pages 300–321. University of California Press, 1991.

[13] T. Blakley and C. Salmond. Probabilistic record linkage and a method to calculate the positive predictive value. *Int J Epidemiol*, 31:1246–1252, 2002.

[14] L.G. Bode, J.A. Kluytmans, H.F. Wertheim, D. Bogaers, C.M. Vandenbroucke-Grauls, R. Roosendaal, A. Troelstra, A.T. Box, A. Voss, I. Van der Tweel, A. Van Belkum, H.A. Vervrugh, and M.C. Vos. Preventing surgical-site infections in nasal carriers of Staphylococcus Aureus. *N Engl J Med*, 362:9–17, 2010.

[15] M.J.M. Bonten, D.J. Austin, and M. Lipsitch. Understanding the spread of antibiotic resistant pathogens in hospitals: mathematical models as tools for control. *Clin Infect Dis*, 33:1739–1746, 2001.

[16] J.M. Boyce, B. Cookson, K. Christiansen, S. Hori, J. Vuopio-Varkila, S. Kocagoz, A.Y. Oztop, C.M. Vandenbroucke-Grauls, S. Harbarth, and D. Pittet. Meticillin-resistant Staphlococcus Aureus. *Lancet Infect Dis*, 5(10):653–663, 2005.

[17] S.L. Bronzwaer, U. Buchholz, J.L. Kool, J. Monen, and P. Schrijnemakers. EARSS activities and results: update. *Euro Surveill*, 6:2–5, 2001.

[18] J.A. Cepeda, T. Whitehouse, B. Cooper, J. Hails, K. Jones, F. Kwaku, L. Taylor, S. Hayman, B. Cookson, S. Shaw, C. Kibbler, M. Singer, G. Bellingan, and A.P. Wilson. Isolation of patients in single rooms or cohorts to reduce spread of MRSA in intensive-care units: prospective two-centre study. *Lancet*, 365(9456):295–304, 2005.

[19] C. Chaix, I. Durand-Zaleski, C. Alberti, and C. Brun-Buisson. Control of endemic methicillin-resistant Staphylococcus Aureus: a cost-benefit analysis in an intensive care unit. *JAMA*, 282:1745–1751, 1999.

[20] A. Christensen, O. Scheel, K. Urwitz, and K. Bergh. Outbreak of methicillin-resistant Staphylococcus Aureus in a Norwegian hospital. *Scand J Infect Dis*, 33:663–666, 2001.

[21] D.E. Clark and D.R. Hahn. Comparison of probabilistic and deterministic record linkage in the development of a statewide trauma registry. In R.M. Gardner, editor, *Proc Annu Symp Comput Appl Med Care*, pages 397–401, New Orleans, Louisiana, Oct, 1995. Hanley & Belfus, Inc.

[22] B. Cooper and M. Lipsitch. The analysis of hospital infection data using hidden Markov models. *Biostatistics*, 5(2):223–237, 2004.

[23] B.A. Cunha. Methicillin-resistant Staphylococcus Aureus: clinical manifestations and antimicrobial therapy. *Clin Microbiol Infect*, 11(Suppl.4):33–42, 2005.

[24] A. Cunningham, H. Donaghy, A. Harman, M. Kim, and S. Turville. Manipulation of dendritic cell function by viruses. *Curr Opin Microbiol*, 13(4):524–529, 1993.

[25] K. Devlin. Every MRSA case costs NHS an extra £9,000. *The Telegraph*, 25th, June, 2008.

[26] M.T. Doyle. *Modelling Some Aspects of the AIDS Pandemic.* Ph.D. thesis, University of Strathclyde: Glasgow, UK, 1994.

[27] M.T. Doyle, D. Greenhalgh, and J. Mortimer. Three statistical tests for detecting overcounting of individuals in serological test data. *Appl Stochastic Models Data Anal*, 13:307–314, 1998.

[28] M. Dryden. Complicated skin and soft tissue infections caused by methicillin-resistant Staphylococcus Aureus: epidemiology, risk factors, and presentation. *Surg Infect*, 9(Suppl 1):3–10, 2008.

[29] A.K. Elmagarid, P.G. Ipeiritos, and V.S. Verykios. Duplicate record detection: a survey. *IEEE Trans Knowl Data Eng*, 19:1–16, 2007.

[30] Health Protection England. The HIV and AIDS Reporting System (HARS). Technical report, 2013.

[31] B.G. Evans, D.J. Howitt, and J. Mortimer. Surveillance of HIV infection and AIDS in the UK: an overview from the PHLS AIDS Centre. *PHLS Microbiol Digest*, 10:141–143, 1993.

[32] F.R. Falkiner. The consequences of antibiotic use in horticulture. *J Antimicrob Chemother*, 41:429–431, 1998.

[33] M. Farrington, C. Redpath, C. Trundle, S. Coomber, and N.M. Brown. Winning the battle but losing the war: methicillin-resistant Staphylococcus Aureus (MRSA) infection at a teaching hospital. *QJM*, 91:539–548, 1998.

[34] M. Fearon. The laboratory diagnosis of HIV infections. *Can J Infect Dis Med Microbiol*, 16(1):26–30, 2005.

[35] W. Feller. *An Introduction to Probability Theory and Its Applications, 3rd edn.* Wiley, New York, USA, 1967.

[36] J.T. Fishbain, J.C. Lee, H.D. Nguyen, J.A. Mikita, C.P. Mikita, C.F. Uyehara, and D.R. Hospenthal. Nosocomial transimission of methicillin-resistant Staphylococcus Aureus: a blinded study to establish baseline acquisition rates. *Infect Control Hosp Epidemiol*, 24(6):415–421, 2003.

[37] National Institute for Health and Clinical Excellence. Increasing the uptake of HIV testing to reduce undiagnosed infection and prevent transmission among black African communities living in England. Technical report, 2011.

[38] National Institute for Health and Clinical Excellence. Quality improvement guide: Preventation and control of healthcare-associated infections. Costing report, implementing NICE guidance. Technical report, National Institute for Health and Clinical Excellence and National Health Service, 2011.

[39] M.L. Forrester, A.N. Pettitt, and G.J. Gibson. Bayesian inference of hospital-acquired infectious diseases and control measures given imperfect surveillance data. *Biostatistics*, 8:383–401, 2007.

[40] A.L. Frank, J.F. Marcinak, P.D. Mangat, and P.C. Schreckenberger. Increase in community-acquired methicillin-resistant Staphylococcus Aureus in children. *Clin Infect Dis*, 29(4):935–936, 1999.

[41] K.F. Froslie, J. Roislien, P. Laake, T. Henriksen, E. Qvigstad, and M.B. Veierod. Categorisation of continuous exposure variable revisited. A response to the Hyperglycaemia and Adverse Pregnancy Outcome (HAPO) study. *BMC Med Res Methodol*, 10(103), 2010.

[42] R.C. Gallo, P.S. Sarin, E.P. Gelmann, M. Robert-Guroff, E. Richardson, V.S. Kalyanaraman, D. Mann, G.D. Sidhu, R.E. Stahl, S. Zolla-Pazner, J. Leibowitch, and M. Popovic. Isolation of human T-cell leukemia virus in Acquired Immune Deficiency Syndrome (AIDS). *Science*, 220(4599):865–867, 1983.

[43] A. Gigli and A. Verdecchia. Uncertainty of AIDS incubation time and its effects on back-calculation estimates. *Stat Med*, 19(2):175–189, 2000.

[44] P.B. Gilbert, I.W. McKeague, G. Eisen, C. Mullins, A. Gueye-Ndiaye, S. Mboup, and P.J. Kanki. Comparison of HIV-1 and HIV-2 infectivity from a prospective cohort study in Senegal. *Stat Med*, 22(4):573–893, 2003.

[45] P.E. Gill, S.J. Hammarling, W. Murray, M.A. Saunders, and M.H. Wright. User's guide for LSSOL (Verion 1.0): a Fortran package for constrained linear least-square and convex quadratic programming. Report SOL 86-1. Technical report, Department of Operations Research, Stanford University, 1986.

[46] G.S. Gottlieb, V.G. Fowler, L.K. Kang, R.S. McClelland, A.K. Gopal, K.A. Marr, J. Li, D.J. Sexton, D. Glower, and G.R. Corey. Staphylococcus bacteraemia in the surgical patient: a prospective analysis of 73 postoperative patients who developed Staphylococcus Aureus bacteraemia at tertiary care facility. *J Am Coll Surg*, 190:50–57, 2000.

[47] A. Goubar, A.E. Ades, D.D. Angelis, C.A. McGarrigle, C.H. Mercer, P.A. Tookey, K. Fenton, and O.N. Gill. Estimates of human immunodeficiency virus prevalence and proportion diagnosed based on Bayesian multiparameter synthesis of surveillance data. *J R Stat Soc*, 171:541–580, 2008.

[48] S.J. Grannis, J.M. Overhage, Siu Hui, and C.J. McDonald. Analysis of a probabilistic record linkage technique without human review. *AMIA Annu Symp Proc*, 2003:259–263, 2003.

[49] W.R. Gransden, S.J. Eykyn, and I. Phillips. Staphylococcus Aureus bacteraemia: 400 episodes in St Thomas' Hospital. *Br Med J (Clin Res Ed)*, 288:300–303, 1984.

[50] R. Greener. State of The Art: AIDS and Economics. In *The Impact of AIDS*, pages 49–55. United Nations Publication, 2002.

[51] D. Greenhalgh and M.T. Doyle. A test to detect replication in HIV serological data labelled by birth date based on the number of matching pairs in a sample. *Stat Med*, 18:1641–1656, 1999.

[52] D. Greenhalgh, M.T. Doyle, and J. Mortimer. A partial ranking method for identifying repeated inclusion of individuals in anonymized HIV infection reports. *Biometrics*, 55:165–173, 1999.

[53] J. Hails, F. Kwaku, A.P. Wilson, G. Bellingan, and M. Singer. Large variation in MRSA policies, procedures and prevalence in English intensive care units: a questionnaire analysis. *Intensive Care Med*, 29:481–483, 2003.

[54] S. Harbarth, C. Fankhauser, J. Schrenzel, J. Christenson, P. Gervaz, C. Bandiera-Clerc, G. Renzi, N. Vernaz, H. Sax, and D. Pittet. Universal screening for methicillin-resistant Staphylococcus Aureus at hospital admission and nonsocomial infection in surgical patients. *JAMA*, 299(10):1149–1157, 2008.

[55] K. Hardy, C. Price, A. Szczepura, S. Gossain, R. Davies, N. Stallard, S. Shabir, C. McMurray, A. Bradbury, and P.M. Hawkey. Reduction in the rate of methicillin-resistant Staphylococcus Aureus acquisition in surgical wards by rapid screening for colonization: a prospective, cross-over study. *Clin Microbiol Infect*, 16:333–339, 2010.

[56] C. Horner, P. Parnell, D. Hall, A. Kearns, J. Heritage, and M. Wilcox. Meticillin-resistant Staphylococcus Aureus in elderly residents of care homes: colonization rates and molecular epidemiology. *J Hosp Infect*, 83(3):212–218, 2013.

[57] G.R. Howe and J. Lindsay. A generalised record linkage computer system for use in medical follow-up studies. *Comput Biomed Res*, 14:327–340, 1981.

[58] A.B. Hutchinson, P.G. Farnham, H.D. Dean, D.U. Ekwueme, C. del Rio, L. Kamimoto, and S.E. Kellerman. The economic burden of HIV in the United

States in the era of Highly Active Antiretroviral Therapy: evidence of continuing racial and ethnic differences. *J Acquir Immune Defic Syndr*, 43(4):451–457, 2006.

[59] R.J. Hyndman and Y. Fan. Sample quantiles in statistical packages. *Am Stat*, 50(4):361–365, 1996.

[60] M.A. Jaro. Probabilistic linkage of large public health data files. *Stat Med*, 14:491–498, 1995.

[61] M.P. Jevons. Celbin-resistant staphylococci. *BMJ*, 1:124–125, 1961.

[62] N.P. Jewell. *Statistics for Epidemiology.* Chapman & Hall/CRC, 2004.

[63] D. Jeyaratnam, C.J. Whitty, K. Phillips, D. Liu, C. Orezzi, U. Ajoku, and G.L. French. Impact of rapid screening test on acquisition of meticillin resistant Staphylococcus Aureus: cluster randomised crossover trial. *BMJ*, 336:927–930, 2008.

[64] E. De Jonge, M.J. Schultz, L. Spanjaard, P.M. Bossuyt, M.B. Vroom, J. Dankert, and J. Kesecioqlu. Effects of selective decontamination of digestive tract on mortality and acquisition of resistant bacteria in intensive care: a randomised controlled trial. *Lancet*, 362(9389):1011–1016, 2003.

[65] R.M. Klevens, M.A. Morrison, J. Nadle, S. Petit, K. Gershman, S. Ray, L.H. Harrison, R. Lynfield, G. Dumyati, J.M. Townes, A.S. Craig, E.R. Zell, G.E. Fosheim, K.D. McDougal, R.B. Carey, and S.K. Fridkin. Invasive methicillin-resistant Staphylococcus Aureus infections in the United States. *JMAM*, 298(15):1763–1771, 2007.

[66] J. Kluytmans, A.V. Belkam, and H. Verbrugh. Nasal carriage of Staphylococcus Aureus: epidemiology, underlying mechanisms, and associated risks. *Clin Micobiol Rev*, 10:505–520, 1997.

[67] T. Kypraios, P.D. O'Neill, S.S. Huang, S.L. Rifas-Shiman, and B.S. Cooper. Assessing the role of undetected colonization and isolation precautions in reducing Methicillin-Resistant Staphylococcus Aureus transmission in intensive care units. *BMC Infect Dis*, 10(29), 2010.

[68] S.O. Larsen. Estimation of the number of people in a register from the number of birthdates. *Stat Med*, 13:177–183, 1994.

[69] W.C.C. Leung. The effect of Methicillin-resistant Staphylococcus Aureus (MRSA) prevention bundle in reducing hospital-acquired MRSA rate in an acute public hospital in Hong Kong. M.Phil Thesis, The Hong Kong Polytechinic University, 2012.

[70] J.A. Levy. HIV pathogenesis and long-term survival. *AIDS*, 7(11):1401–1410, 1993.

[71] H. Libman and R. Arbeit. Complications associated with Staphlococcus Aureus bacteremia. *Arch Intern Med*, 144:541–545, 1984.

[72] B.A. Lipsky, J.A. Weigelt, V. Gupta, A. Killian, and M.M. Peng. Skin, soft tissue, bone, and joint infections in hospitalized patients: epidemiology and microbiological, clinical, and economic outcomes. *Infect Cont Hosp Ep*, 28(11):1290–1298, 2007.

[73] J.C. Lucet, S. Chevert, I. Durand-Zaleski, C. Chastang, and B. Régnier. Prevalence and risk factors for carriage of Methicillin-resistant Staphylococcus Aureus at admission to the Intensive Care Unit. *Arch Intern Med*, 163:181–188, 2003.

[74] S. Mandalia, R. Mandalia, G. Lo, T. Chadborn, P. Sharott, M. Youle, J. Anderson, G. Baily, R. Brettle, M. Fisher, M. Gompels, G. Kinghorn, M. Johnson, B. McCarron, A. Pozniak, A. Tang, J. Walsh, D. White, I. Williams, B. Gazzard, and E.J. Beck. Rising population cost for treating people living with HIV in the UK, 1997-2013. *PlosONE*, 5(12):15,677–15,684, 2010.

[75] R.R. Marples and S. Reith. Methicillin-resistant Staphylococcus Aureus in England and Wales. *Commun Dis Rep CDR Rev*, 2:25–29, 1992.

[76] C. Marshall, G. Harrington, R. Wolfe, C.K. Fairley, S. Wesselingh, and D. Spelmen. Acuqisition of Methicillin-resistant Staphylococcus Aureus in a large intensive care unit. *Infect Cont and Hosp Epidemiol*, 24:322–326, 2003.

[77] A. McHenry, N. Macdonald, K. Sinka, J. Mortimer, and B. Evans. National assessment of prevalent diagnosed HIV infections. *Commun Dis Public Health*, 3:277–281, 2000.

[78] M. Melzer, S.J. Eykyn, W.R. Gransden, and S. Chinn. Is methicillin-resistant Staphylococcus Aureus more virulent than methicillin-susceptible S. Aureus? A comparative cohort study of British patients with nosocomial infection and bacteremia. *Clin Infect Dis*, 37:1453–1460, 2003.

[79] A. Mocroft, S. Vella, T.L. Benfield, A. Chiesi, V. Miller, P. Gargalianos, A. d'Arminio Monforte, I. Yust, J.N. Bruun, A.N. Phillips, and J.D. Lundgren. Changing patterns of mortality across Europe in patients infected with HIV1. *Lancet*, 352:1725–1730, 1998.

[80] R.D. Moore and R.E. Chaisson. Natural history of HIV infection in the era of combination antiretroviral therapy. *AIDS*, 13:1933–1942, 1999.

[81] M. Morgan. Methicillin-resistant Staphylococcus Aureus and animals:zoonosis or humanosis? *J Antimicrob Chemother*, 62(6):181–187, Dec 2008.

[82] J. Mortimer and J.A. Salathiel. 'Soundex' codes of surnames provide confidentiality and accuracy in a national HIV database. *Commun Dis Rep*, 5:183–186, 1995.

[83] A.G. Muse, J. Mikl, and P.F. Smith. Evaluating the quality of anonymous data linkage using deterministic procedures with the New York State AIDS Registry and a hospital discharge file. *Stat Med*, 14:499–509, 1995.

[84] H.B. Newcombe. *Handbook of Record Linkage Methods for Health and Statistical Studies, Administration and Business.* Oxford University Press: London, 1988.

[85] K. Nigam, A.K. McCallum, S. Thrun, and T. Mitchell. Text classification from labeled and unlabelled documents using EM. *Mach Learn*, 39:103–134, 2003.

[86] NINSS. NINSS report on surgical site infection and hospital-acquired bacteremia. *Commun Dis Rep CDR Wkly*, 10:213–216, 2000.

[87] G.A. Noskin, R.J. Rubin, J.J. Schentag, J. Kluytmans, E.C. Hedblom, C. Jacobson, M. Smulders, E. Gemmen, and M. Bharmal. Nation trends in Staphylococcus Aureus infection rrates: Impact on economic burden and mortality over a 6-year period (1998-2003). *Clin Infect Dis*, 45:1132–1140, 2007.

[88] University of California. What kind of HIV screening tests are available in the United States? Technical report, 2011.

[89] National Audit Office. The management and control of hospital-acquired infection in acute NHS trusts in England: report by the Comptroller and Auditor General. Technical report, London: Stationary Office, 2000.

[90] F.J. Palella, K.M. Delaney, A.C. Moorman, M.O. Loveless, J. Fuhrer, G.A. Satten, D.J. Aschman, and S.D. Holmberg. Declining morbidity and mortality among patients with advanced Human Immunodeficiency Virus infection. HIV Outpatient Study Investigators. *N Engl J Med*, 338(13):835–860, 1998.

[91] I. Pelupessy, M.J.M. Bonten, and O. Diekmann. How to assess the relative importance of different colonization routes of pathogens within hospital settings. *Proc Natl Acad Sci USA*, 99:5601–5605, 2002.

[92] L. Perrin and A. Telenti. HIV treatment failure: testing for HIV resistance in clinical practice. *Science*, 280:1871–1873, 1998.

[93] R. Plowman. The Socio-economic burden of Hospital Acquired Infection. *Euro Surveill*, 5(4):49–50, 2000.

[94] A.M. Presanis, D.de. Angelis, A. Goubar, O.N. Gill, and A.E. Ades. Bayesian evidence synthesis for a tranmission dynamic model for HIV among men who have sex with men. *Biostatistics*, 12(4):666–681, 2011.

[95] G. Gopal Rao, P. Michalczyk, N. Nayeem, G. Walder, and L. Wigmore. Prevalence and risk factors for meticillin-resistant Staphylococcus Aureus in adult emergency admission- a case for screening all patients? *J Hosp Infect*, 66(1):15–21, 2007.

[96] B.D. Rice, D.C. Delpech, T.R. Chadborn, and J. Elford. Loss to follow-up among adults attending Human Immunodeficiency Virus services in England, Wales and Northern Ireland. *STD*, 38:685–690, 2011.

[97] S. Rieg, G. Peyerl-Hoffmann, K. De With, C. Theilacker, D. Wagner, J. Hubner, M. Dettenkofer, A. Kaasch, H. Seifert, C. Schneider, and W.V. Kern. Mortality of S.Aureus bacteremia and infectious disease specialist consultation A study of 521 patients in Germany. *J Infect*, 59(4):232–239.

[98] C. Rioux, Armand-Lefever, W. Guerinot, A. Andermornt, and J.C. Lucet. Acquisition of methicillin-resistant Staphylococcus Aureus in the acute care setting: incidence and risk factor. *Infect Control Hosp Epidemiol*, 28(6):733–736, 2007.

[99] K. Ritchie, I. Bradbury, J. Craig, J. Eastagte, L. Foster, H. Kohli, K. Iqbal, K. MacPherson, T. McCarthy, H. Mclntosh, E. Nic Lochlainn, M. Reid, and J. Taylor. The clinical and cost effectiveness of screening for meticillin-resistant Staphylococcus Aureus (MRSA). Technical report, Edinburgh: NHS Quality Improvement Scotland, 2007.

[100] J.V. Robotham, N. Graves, B.D. Cookson, A.G. Barnett, J.A. Wilson, J.D. Edgeworth, R. Batra, B.H. Cuthbertson, and B.S. Cooper. Screening, isolation, and

decolonisation strategies in the control of meticillin resistant Staphylococcus Aureus in intensive care units: cost effectiveness evaluation. *BMJ*, 343(7827):5694–5706, 2011.

[101] J.K. Rochstroh, S. Bhagani, Y. Benhamou, R. Bruno, S. Mauss, L. Peters, M. Puoti, V. Soriano, C. Tural, and the EACS Executive Committee. European AIDS Clinical Society (EACS) guidelines for the clinical management and treatment of chronic hepatitis B and C coinfection in HIV-infected adults. *HIV Med*, 9:82–88, 2008.

[102] L.L. Roose and A. Wajda. Record linkage strategies. Part I: Estimating information and evaluating approaches. *Method Inform Med*, 30(2):117–123, 1991.

[103] P. Royston, D.G. Altman, and W. Sauerbrei. Dichotomizing continuous predictors in multiple regression: a bad idea. *Stat Med*, 25:127–141, 2006.

[104] J.B. Sarma, B. Marshall, V. Cleeve, D. Tate, and T. Oswald. Impact of universal screening on MRSA bacteremias in a single acute NHS organisation (2006-12): interrupted time-series analysis. *Antimicrob Res Infect Cont*, 2:2–12, 2012.

[105] Health Protection Scotland. NHS Scotland MRSA Screening Pathfinder Programme: Interim report. Technical report, Glasgow: Health Protection Scotland, 2009.

[106] Health Protection Scotland. Discharge testing for MRSA in scottish hospitals: MRSA acquisition, description of acquired strains and risk factors for acquisition of MRSA in the hospital. Technical report, Glasgow: Health Protection Scotland, 2010.

[107] Health Protection Scotland. The value of nasal swabbing versus full body screening or clinical risk assessment to detect MRSA colonisation at admission to hospital. Technical report, Glasgow: Health Protection Scotland, 2010.

[108] Health Protection Scotland. The annual report of healthcare associated infection report January-Decemeber 2009. Technical report, Glasgow: Health Protection Scotland, 2011.

[109] Health Protection Scotland. Final report volume 1: an investigation of the clinical effectiveness of universal MRSA screening. Technical report, Glasgow: Health Protection Scotland, 2011.

[110] Health Protection Scotland. NHS Scotland MRSA Screening Pathfinder Programme: Economic analyses. Technical report, Glasgow: Health Protection Scotland, 2011.

[111] Health Protection Scotland. NHS Scotland MRSA Screening Pathfinder Programme: Final Report Executive Summary. Technical report, Glasgow: Health Protection Scotland, 2011.

[112] Health Protection Scotland. NHS Scotland MRSA Screening Pathfinder Programme: SBAR Report to Scottish Government Health Directorates: policy implications of further research studies for national rollout of MRSA screening. Technical report, Glasgow: Health Protection Scotland, 2011.

[113] Health Protection Scotland. Quarterly report on the surveillance of Staphylococcus Aureus bacteraemias in Scotland, July-September 2012. Technical report, Glasgow: Health Protection Scotland, 2012.

[114] J.M. Scriven, P. Silva, R.A. Swann, M.M. Thompson, A.R. Naylor, P.R. Bell, and N.J. London. The acquisition of methicillin-resistant Staphylococcus Aureus (MRSA) in vascular patients. *Eur J Vasc Endovasc Surg*, 26(2):147–151, 2003.

[115] National Health Service. HIV and AIDS. `http://www.nhs.uk/Conditions/HIV/Pages/Introduction.aspx`, Accessed Nov,2013.

[116] National Health Service. NHS hospital services. `http://www.nhs.uk/NHSEngland/AboutNHSservices/NHShospitals/Pages/going-into-hospital.aspx`, Accessed Nov,2013.

[117] UK Health Services. Annual UK HIV treatment and care costs could reach £750 million by 2013. `http://www.aidsmap.com/Annual-UK-HIV-treatment-and-care-costs-could-reach-750-million-by-2013/page/1618137/`, Accessed Dec,2013.

[118] I.P. Shevchenko, J.T. Lynch, A.S. Mattie, and L.I. Reed-Fourquet. Verification of information in a large medical database using linkages with an external database. *Stat Med*, 14:511–530, 1995.

[119] A. Shorr and T.P. Lodise. ISMR Update: Burden of Methicillin-resistant Staphylococcus Aureus on healthcare cost and resource utilization. Technical report, University of Kentucky Colleges of Medicine and Pharmacy and the International Society of Microbial Resistance, 2006.

[120] R. Song, T. Green, M. McKenna, and M.K. Glynn. Using occupancy models to estimate the number of duplicate cases in a data system without unique identifiers. *J Data Sci*, 5:53–66, 2007.

[121] D.C. Speller, A.P. Johnson, D. James, R.R. Marples, A. Charlett, and R.C. George. Resistance to methicillin and other antibiotics in isolates of Staphylococcus Aureus from blood and cerebrospinal fluid, England and Wales, 1989-95. *Lancet*, 350:323–325, 1997.

[122] NNIS System. National Nosocomial Infections Surveillance System Report, data summary from January 1992 through June 2004, issued October 2004. *Am J Infect Control*, 32:470–485, 2004.

[123] D. Thompson. Methicillin-resistant Staphylococcus Aureus in a general intensive care unit. *J R Soc Med*, 97:521–526, 2004.

[124] National AIDS Trust. Latest UK Statistics. http://www.nat.org.uk/HIV-Facts/Statistics/Latest-UK-Statistics.aspx, Accessed Nov,2013.

[125] NHS Foundation Trust. MRSA Screening, March Accessed Dec,2013.

[126] K. Trzcinski, B.S. Cooper, W. Hryniewicz, and C.G. Dowson. Expression of resistance to tetracyclines in strains of methicillin-resistant Staphylococcus Aureus. *J Antimicrob Chemother*, 45:763–770, 2000.

[127] UNAIDS. UNAIDS World AIDS Day Report 2012. Technical report, Joint United Nations Programme on HIV/AIDS, 2012.

[128] M.F.Q. Vandenbergh, E.D. Yzerman, A.V. Belkum, H.A.M. Boelens, M. Sijmons, and H.A. Verbrugh. Follow-up of Staphylococcus Aureus nasal carriage after 8 years: redefining the persistent carrier state. *J Clin Microbiol*, 37(10):3133–3140, 1999.

[129] R. Varadhan and G. Grothendieck. Package 'alabama'. http://cran.r-project.org/web/packages/alabama/alabama.pdf, Accessed Sep,2012.

[130] E.V.H.V. Velzen, J.S. Reilly, K. Kavanagh, A. Leanord, G.F.S. Edwards, E.K. Girvan, I.M. Gould, F.M. MacKenzie, and R. Masterton. A retrospective cohort study into acquisition of MRSA and associated risk factors after implementation of universal screening in scottish hospitals. *Infect Control Hosp Epidemiol*, 32(9):889–896, 2011.

[131] T.W. Victor and R.M. Mera. Record linkage of healthcare insurance claims. *J Am Med Inf Assoc*, 8(3):281–288, 2001.

[132] A. Voss, D. Milatovic, C. Wallrauch Schwarz, V.T. Rosdahl, and I. Braveny. Methicillin-resistant Staphylococcus Aureus in Europe. *Eur J Clin Invest*, 13:50–55, 1994.

[133] J. Wang, M. Wang, Y. Huang, M. Zhu, Y. Wang, J. Zhou, and X. Lu. Colonization pressure adjusted by degree of environmental contamination: A better indicator

for predicting methicillin-resistant Staphylococcus Aureus acquisition. *Am J Infect Control*, 39(9):763–769, 2011.

[134] M. Waterhouse, A. Morton, K. Mengersen, D. Cook, and G. Playford. Role of overcrowding in meticillin-resistant Staphylococcus Aureus transmission:Bayesian network analysis for a single public hosptial. *J Hosp Infect*, 78(2):92–96, 2011.

[135] W.E. Winkler. Using the EM algorithm for weight computation in the Fellegi-Sunter Model for record linkage. In *Proceedings of the Section on Survey Research Methods*, pages 667–671, New Orleans, Louisiana, Aug, 1988. American Statistical Association, The Association.

[136] S.N. Wood. *Generalized additive models*. Chapman & Hall/CRC, 2006.

Appendix A

The datasets of HIV diagnoses held by the PHLS

Table A.1: 1991 dataset sent to us by PHLS. Birth years are tabulated in ascending order and leap years are indicated with an asterisk.

Year of birth	Number of individuals	S_1	S_2	S_3	S_4	S_5	S_6
1929	28	26	1	-	-	-	-
1930	25	23	1	-	-	-	-
1931	26	19	2	1	-	-	-
1932∗	27	23	2	-	-	-	-
1933	44	38	3	-	-	-	-
1934	50	22	14	-	-	-	-
1935	54	40	5	-	1	-	-
1936∗	52	48	2	-	-	-	-
1937	68	57	4	1	-	-	-

Continued on next page

Table A.1 – continued from previous page

1938	78	66	6	-	-	-	-
1939	99	67	13	2	-	-	-
1940∗	87	71	8	-	-	-	-
1941	83	63	10	-	-	-	-
1942	124	86	13	4	-	-	-
1943	113	69	17	2	1	-	-
1944∗	176	104	24	6	-	-	1

Table A.2: 1994 dataset sent to us by PHLS. Birth years are tabulated in ascending order and leap years are indicated with an asterisk.

Year of birth	Number of individuals	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}	S_{11}
1901	0	-	-	-	-	-	-	-	-	-	-	-
1902	0	-	-	-	-	-	-	-	-	-	-	-
1903	1	1	-	-	-	-	-	-	-	-	-	-
1904∗	0	-	-	-	-	-	-	-	-	-	-	-
1905	2	2	-	-	-	-	-	-	-	-	-	-
1906	0	-	-	-	-	-	-	-	-	-	-	-
1907	0	-	-	-	-	-	-	-	-	-	-	-
1908∗	1	1	-	-	-	-	-	-	-	-	-	-
1909	0	-	-	-	-	-	-	-	-	-	-	-
1910	0	-	-	-	-	-	-	-	-	-	-	-
1911	2	2	-	-	-	-	-	-	-	-	-	-
1912∗	4	4	-	-	-	-	-	-	-	-	-	-

Continued on next page

Table A.2 – 1994 dataset sent to us by PHLS. Continued from previous page

1913	5	5	-	-	-	-	-	-	-	-	-	-
1914	10	10	-	-	-	-	-	-	-	-	-	-
1915	5	5	-	-	-	-	-	-	-	-	-	-
1916∗	4	2	1	-	-	-	-	-	-	-	-	-
1917	7	5	1	-	-	-	-	-	-	-	-	-
1918	6	6	-	-	-	-	-	-	-	-	-	-
1919	10	8	1	-	-	-	-	-	-	-	-	-
1920∗	6	6	-	-	-	-	-	-	-	-	-	-
1921	3	3	-	-	-	-	-	-	-	-	-	-
1922	11	7	2	-	-	-	-	-	-	-	-	-
1923	13	13	-	-	-	-	-	-	-	-	-	-
1924∗	19	17	1	-	-	-	-	-	-	-	-	-
1925	33	28	1	1	-	-	-	-	-	-	-	-
1926	20	17	-	1	-	-	-	-	-	-	-	-
1927	24	22	1	-	-	-	-	-	-	-	-	-
1928∗	30	26	2	-	-	-	-	-	-	-	-	-
1929	41	39	1	-	-	-	-	-	-	-	-	-
1930	43	35	4	-	-	-	-	-	-	-	-	-
1931	52	37	6	1	-	-	-	-	-	-	-	-
1932∗	59	51	4	-	-	-	-	-	-	-	-	-
1933	74	68	3	-	-	-	-	-	-	-	-	-
1934	82	60	8	2	-	-	-	-	-	-	-	-
1935	78	59	8	1	-	-	-	-	-	-	-	-
1936∗	95	69	10	2	-	-	-	-	-	-	-	-
1937	118	86	16	-	-	-	-	-	-	-	-	-

Continued on next page

Table A.2 – 1994 dataset sent to us by PHLS. Continued from previous page

1938	129	96	12	3	-	-	-	-	-	-	-	-
1939	156	94	25	4	-	-	-	-	-	-	-	-
1940*	143	106	17	1	-	-	-	-	-	-	-	-
1941	149	105	19	2	-	-	-	-	-	-	-	-
1942	212	101	43	7	1	-	-	-	-	-	-	-
1943	202	115	28	9	1	-	-	-	-	-	-	-
1944*	280	127	49	14	2	1	-	-	-	-	-	-
1945	279	118	55	11	2	2	-	-	-	-	-	-
1946	320	130	56	18	6	-	-	-	-	-	-	-
1947	411	133	69	35	6	1	1	-	-	-	-	-
1948*	392	128	66	27	9	3	-	-	-	-	-	-
1949	418	150	66	29	11	1	-	-	-	-	-	-
1950	430	131	68	36	10	3	-	-	-	-	-	-
1951	444	137	78	33	6	3	1	1	-	-	-	-
1952*	515	131	91	41	13	3	2	-	-	-	-	-
1953	485	133	75	35	11	6	1	-	1	1	-	-
1954	591	110	78	53	28	7	2	1	-	-	-	-
1955	624	88	104	57	24	11	1	-	-	-	-	-
1956*	648	130	92	61	8	3	3	1	-	-	-	-
1957	724	103	99	57	30	15	2	4	1	1	-	-
1958	770	84	107	62	40	15	5	3	-	-	-	-
1959	798	86	91	75	37	18	4	5	1	-	-	-
1960*	890	87	92	71	37	25	8	7	2	1	-	1
1961	858	79	96	72	48	19	9	2	2	-	-	-
1962	929	68	107	55	47	29	13	4	3	1	1	-

Continued on next page

Table A.2 – 1994 dataset sent to us by PHLS. Continued from previous page

1963	880	82	80	75	47	25	12	4	-	-	-	-
1964∗	856	80	87	76	42	18	12	5	-	1	-	-
1965	703	107	92	71	28	9	7	-	-	-	-	-
1966	639	109	96	52	22	15	2	1	-	-	-	-
1967	508	114	88	38	16	8	-	-	-	-	-	-
1968∗	380	123	68	24	11	1	-	-	-	-	-	-
1969	294	112	63	12	2	1	-	1	-	-	-	-
1970	221	107	41	9	-	1	-	-	-	-	-	-
1971	125	92	12	3	-	-	-	-	-	-	-	-
1972∗	76	62	7	-	-	-	-	-	-	-	-	-
1973	35	31	2	-	-	-	-	-	-	-	-	-

Appendix B

The analysis for the possible confounding effects, multiplicative interactions and the trend in risk of MRSA acquisition

The association between the length of stay and MRSA acquisition, assuming age is a potential confounder

Since age is also highly associated with length of stay (noted in Table 6.4 in Section 6.3), it is plausible to assume that there is a potential confounding effect on length of stay by age. Similarly, we apply the procedure described in Section 6.5.

First of all, stratify by age group and thus the association between length of stay and MRSA acquisition in different strata of age can be analysed. Here we use the eight or more nights as the baseline for the estimation of the stratum-specific odds ratios for the length of stay since there are limited records of patients who acquired MRSA and had stayed for a short time. This causes zero cells in the stratified two-way table of length of stay associated with MRSA acquisition, which can be explained by one of the reasons that patients (especially the younger patients aged 49 or under) who had stayed for a relatively long time in hospital are likely to acquire MRSA. The results are illustrated below (see Table B.1):

Table B.1: Stratified risk analysis of length of stay by age groups.

Age	Length of stay	MRSA acquisition		OR	p-value
		No	Yes		
≥80 years old	1 night	29 (93.55%)	2 (6.45%)	1.47	0.647
	2-3 nights	65 (100.00%)	0 (0%)	0	0.989
	4-7 nights	116 (97.48%)	3 (2.52%)	0.55	0.408
	≥8 nights	128 (95.52%)	6 (4.48%)	1	
65-79 years old	1 night	101 (100.00%)	0 (0%)	0	0.989
	2-3 nights	243 (98.38%)	4 (1.62%)	2.26	0.348
	4-7 nights	335 (97.38%)	9 (2.62%)	3.69	0.096
	≥8 nights	275 (99.28%)	2 (0.72%)	1	
50-64 years old	1 night	103 (99.04%)	1 (0.96%)	0.55	0.610
	2-3 nights	252 (100.00%)	0 (0%)	0	0.992
	4-7 nights	258 (99.23%)	2 (0.77%)	0.44	0.374
	≥8 nights	171 (98.28%)	3 (1.72%)	1	
≤49 years old	1 night	127 (100.00%)	0 (0%)	0	0.998
	2-3 nights	229 (100.00%)	0 (0%)	0	0.995
	4-7 nights	165 (100.00%)	0 (0%)	0	0.996
	≥8 nights	93 (97.89%)	2 (2.11%)	1	

In general, Table B.1 shows that except for the categories with zero cells which lead to the zero odds ratios and correspond to a meaningless statistical interpretation, the risk of MRSA acquisition increases as the length of stay increases for the young patients (64 years or under) whereas for the elderly patients (65 years or over), the risk of MRSA acquisition is higher when those patients had a short stay in hospital. For example, for the patients aged 50-64, the risk of MRSA acquisition for those patients who also had stayed only for one night is 0.44 compared to those who had stayed for eight nights or larger. On the other hand, for the patients 80 or over, the risk of MRSA acquisition is 1.47 times as high for those patients who had stayed for one night compared to those who had stayed for eight nights or longer. The corresponding p-values are considerably high, which indicates that the risk of MRSA acquisition does not vary with different length of stay in each stratum of age.

Using the adjusted χ^2-test to investigate the trend of length of stay in the risk of MRSA acquisition stratified by age, we use the logistic regression model with age as a categorical variable and length of stay taking values 0, 1, 2. The p-value (0.111) reflects that there is little convincing evidence that there is a linear trend in MRSA acquisition associated with length of stay.

In order to assess the multiplicative interaction between age and length of stay, we use Woolf's method to detect the homogeneity of the odds ratios for length of stay associated with MRSA acquisition across the strata of age. It yields a high p-value (0.728), indicating that the stratum-specific odds ratios do not vary from each other statistically and there is no effect modification for MRSA acquisition associated with length of stay.

By using the MH method, we estimate the adjusted odds ratios for each category of length of stay with greater than or equal to eight nights as a baseline. The results are shown in Table B.2.

Table B.2: Comparison of every two different length of stay groups, stratified by different groups.

	p-value (CMH method)	$\mathrm{OR}_{\mathrm{MH}}$	95% CI
1 night vs. $\geq$8 nights	0.33	0.56	(0.16,1.94)
2-3 nights vs. $\geq$8 nights	0.044	0.35	(0.11,1.05)
4-7 nights vs. $\geq$8 nights	0.81	0.91	(0.43,1.96)

Clearly, the risk of MRSA acquisition for the patients staying for two to three nights is 0.35 times the risk of MRSA acquisition for the patients staying for eight nights or longer, controlling for the possible confounding effect of age. The patients staying for eight nights or longer have a slightly higher risk of acquiring MRSA compared to the patients staying for four to seven nights. Similarly, the risk of MRSA acquisition for the patients staying eight nights or longer is 1.78 (i.e. 1/0.56) times as high as the risk of MRSA acquisition for the patients staying for one night. Moreover, the corresponding high p-values based on the CMH method reflect that there is no statistical difference between the stratum-specific odds ratios of four to seven nights against eight nights or longer across the strata of age and also the stratum-specific odds ratios of one night against eight nights or longer equal to one. In other words, the risk of MRSA acquisition for the patients staying for one or four to seven nights is statistically the same as the risk of MRSA acquisition for the patients staying for eight or more nights, by taking the effect of age into account. On the other hand, the risk of MRSA acquisition for the patients staying for eight nights or longer is significantly different with the risk of MRSA acquisition for the patients staying for two to three nights.

Table B.3: The crude odds ratio of the length of stay.

Length of stay	Crude OR
1 night vs. ≥ 8 nights	0.42
2-3 nights vs. ≥ 8 nights	0.26
4-7 nights vs. ≥ 8 nights	0.82

Comparing the adjusted odds ratio (0.56) for the length of stay of 1 night against larger than or equal to eight nights associated with MRSA acquisition (in Table B.2) to the corresponding crude odds ratio which is 0.42 (in Table B.3), it is obvious that the crude odds ratio is much less ($> 10\%$) than the adjusted one. Similarly, the adjusted odds ratios for the length of stay of two to three nights and four to seven nights (which are 0.35 and 0.91 respectively) are more than 10% greater than the corresponding crude ones (i.e. 0.26, 0.82 respectively). Therefore, there is strong evidence that age has a confounding effect on the length of stay associated with MRSA acquisition.

The association between open wounds or ulcers and MRSA acquisition, assuming that age is a potential confounder

Now we shall use the same procedure to analyse the potential confounder of age associated with the co-morbidity: open wounds or ulcers. The results for the stratified analysis are shown in Table B.4.

The results demonstrate that for the patients 80 years old or older, the risk of MRSA acquisition for those patients with open wounds or ulcers is almost 10 times as high as the patients without open wounds or ulcers when admitted into hospital. Furthermore, the corresponding trivial p-value also implies a strong association between open wounds

Table B.4: Stratified risk analysis of open wounds or ulcers by age groups.

Age	Wounds/ulcers	MRSA acquisition		OR	p-value
		No	Yes		
≥80 years old	No	308 (97.78%)	7 (2.22%)	1	
	Yes	18 (81.82%)	4 (18.18%)	9.78	0.0007
65-79 years old	No	861 (98.40%)	14 (1.60%)	1	
	Yes	65 (98.48%)	1 (1.52%)	0.95	0.958
50-64 years old	No	706 (99.44%)	4 (0.56%)	1	
	Yes	51 (96.23%)	2 (3.77%)	6.92	0.028
≤49 years old	No	523 (99.62%)	2 (0.38%)	1	
	Yes	70 (100.00%)	0 (0%)	0	0.996

or ulcers and MRSA acquisition for the patients 80 years and older. A similar conclusion can be drawn for the stratum of age of 50-64 years, where statistically open wounds or ulcers cause 6.92 times as likely risk of MRSA acquisition in comparison with no open wounds or ulcers. The corresponding p-value (0.028) also reveals that open wounds or ulcer is strongly associated with MRSA acquisition for the patients 50-64 years old. On the other hand, for the patients whose age ranged from 65 to 79 years, the substantial p-value (> 0.9) shows no significance in the effect of open wounds or ulcers with respect to MRSA acquisition. Note that there is no record of patients who acquired MRSA and had open wounds or ulcers, and who were aged 49 years and under, leading to the statistically meaningless interpretation for the corresponding stratum-specific odds ratio.

The homogeneity test by Woolf's method yields a large p-value (0.23). Therefore we can conclude that there is no multiplicative interaction between age and open wounds or ulcers. i.e. the odds ratios for MRSA acquisition associated with open wounds or ulcers are constant across the strata of age.

Using the MH method, the adjusted odds ratio for open wounds or ulcers is 3.48 with the corresponding 95% confidence interval (1.49,8.12). The CMH test statistic yields a p-value of 0.0021, showing striking evidence that open wounds or ulcers are strongly related to the risk of the acquisition of MRSA. Compared to the corresponding crude odds ratio 3.04 obtained in Table 6.3, the adjusted odds ratio is slightly higher ($> 10\%$) giving the conclusion that age is a confounder associated with open wounds or ulcers.

• **Number of wards as a potential confounder**

Since the trivial p-values (<0.001) in Table 6.4 of Section 6.3 imply high correlations between number of wards and age, length of stay, open wounds or ulcers and renal failure respectively and the number of wards which is the main risk factor that we are interested in analysing is included into the multivariable analysis, it is possible that there is a confounding effect caused by the number of wards with respect to the other risk factors.

The association between age and MRSA acquisition, assuming that number of wards is a potential confounder

In order to analyse the plausible confounder effect of number of wards, the stratified analysis can be applied. Firstly, we assess the effect of age associated with MRSA acquisition, controlling for the possible confounding effect of number of wards. The results of the stratum-specific odds ratios for age associated with MRSA are demonstrated in Table B.5. Here the category of age of 80 years or older is treated as a baseline for each stratum of number of wards since there is a zero record for the patients aged 49 or under who had acquired MRSA while staying in one ward in hospital.

Table B.5: Stratified risk analysis in different number of wards.

		MRSA acquisition		
No. of wards	Age	No	Yes	OR
1 ward	≤49 years old	398 (100.00%)	0 (0%)	0
	50-64 years old	473 (99.58%)	2 (0.42%)	0.20
	65-79 years old	504 (98.24%)	9 (1.76%)	0.85
	≥80 years old	143 (97.95%)	3 (2.05%)	1
2 wards	≤49 years old	161 (99.38%)	1 (0.62%)	0.28
	50-64 years old	204 (98.55%)	3 (1.45%)	0.67
	65-79 years old	310 (98.41%)	5 (1.59%)	0.74
	≥80 years old	137 (97.86%)	3 (2.14%)	1
≥ 3 wards	≤49 years old	55 (98.21%)	1 (1.79%)	0.21
	50-64 years old	107 (99.07%)	1 (0.93%)	0.11
	65-79 years old	138 (99.28%)	1 (0.72%)	0.08
	≥80 years old	58 (92.06%)	5 (7.94%)	1

In general, the stratum-specific odds ratio increases as the age becomes older in each stratum. Particularly for the stratum where the patients had stayed in one ward, the risk of MRSA acquisition for the patients 65-79 years old is around 0.85 times that of the patients aged 80 or over while the stratum-specific odds ratio for the patients 50-64 years old decreases to 0.20 in comparison with the patients aged 80 or over. With regard to the stratum of two wards, the stratum-specific odds ratio increases steadily as the age increases. On the other hand, for the stratum of three or more wards, the stratum-specific odds ratio decreases when age increases from the group of 49 or under to the group of 65-79. For the patients who had moved through three wards or more, the risk of MRSA acquisition appears to be around 0.11 times as low for the patients aged 50-79 years as the risk for the patients 80 or over and the risk of MRSA acquisition for the patients 49 or under becomes 0.21 times as low as the risk of MRSA acquisition for the patients 80 or over.

An adjusted χ^2 test is used to investigate the trend of age in the risk of MRSA acquisition adjusted for number of wards based upon a logistic regression model with number of wards as a categorical variable and age taking values 0, 1, 2, 3. It gives a low p-value (0.00026), which reveals that the risk of MRSA acquisition increases as the age increases under the control of the possible confounding effect of the number of wards.

The homogeneity test for the consistency of the effect measures (i.e. odds ratio) across strata gives a high p-value (0.26), which indicates that there is little evidence that the odds ratios for age associated with MRSA acquisition are modified by number of wards. In other words, there is an absence of the multiplicative interaction between number of wards and age, which has also been confirmed in the analysis of confounding effect of age with respect to the number of wards previously in Section 6.5.

Considering the association between age and MRSA acquisition, stratification by the number of wards for the estimation of the odds ratio for age associated with MRSA acquisition is applied to remove the possible confounding effect of number of wards. Based on the MH method, we are able to assess the effect of age on MRSA acquisition across the strata of number of wards. The results are shown in the following table (see Table B.6):

Table B.6: Comparison of each pair of different age groups, stratified by number of wards.

	p-value (CMH method)	$\mathrm{OR}_{\mathrm{MH}}$	95% CI
50-64 years vs. $\leq$ 49 years	0.33	2.12	(0.44,10.28)
65-79 years vs. $\leq$ 49 years	0.021	4.35	(1.04,18.12)
$\geq$80 years vs. $\leq$ 49 years	0.0029	6.49	(1.50,28.08)

From the table, it is clear that controlling for the possible confounding effect of number of wards, the patients 65 years old or over have relatively higher risk of MRSA acquisition than the younger patients under 49 since the adjusted odds ratios for the patients 65-79 and the patients 80 or over are 4.35 and 6.49 respectively. The corresponding low p-values also indicate that the patients 65 years or over have a significant difference in the risk of MRSA acquisition from the patients aged 49 years old or under. On the other hand, the risk of MRSA acquisition for the patients aged 50-64 is not statistically higher than the risk of MRSA acquisition for the patients aged 49 or under due to the p-value of 0.33 although the corresponding adjusted odds ratio is 2.12.

Comparing the crude odds ratios for the age without adjusting by the potential confounding effect of the number of wards (shown in Table 6.3) and the adjusted odds ratios (in Table B.6) which is calculated by the MH method, we can conclude that there is a slightly significant confounding effect on age by the number of wards since the adjusted odds ratios decrease by about 9.79%, 9.93% and 35.04% respectively.

The association between length of stay and MRSA acquisition, assuming that number of wards is a potential confounder

Now we investigate the potential confounding effect of number of wards associated with length of stay using the same procedure we applied above. Table B.7 illustrates the number of patients who had acquired MRSA and spent different numbers of nights while in hospital which is stratified by different number of wards.

Due to the limited records for the patients who had stayed in two wards or more and spent three or less nights, there are a few zero cells in the table producing zero odds ratios. One of the reasons is that the patients who had moved through a large number of wards

Table B.7: Stratified risk analysis of length of stay in different number of wards.

No. of wards	Length of stay	MRSA acquisition		OR	p-value
		No	Yes		
1 ward	1 night	314 (99.05%)	3 (0.95%)	2.04	0.540
	2-3 nights	536 (99.26%)	4 (0.74%)	1.59	0.679
	4-7 nights	455 (98.70%)	6 (1.30%)	2.81	0.340
	≥8 nights	213 (99.53%)	1 (0.47%)	1	
2 wards	1 night	44 (100.00%)	0 (0%)	0	0.995
	2-3 nights	237 (100.00%)	0 (0%)	0	0.988
	4-7 nights	309 (98.10%)	6 (1.90%)	0.72	0.571
	≥8 nights	222 (97.37%)	6 (2.63%)	1	
≥ 3 wards	1 night	1 (100.00%)	0 (0%)	0	0.998
	2-3 nights	16 (100.00%)	0 (0%)	0	0.993
	4-7 nights	109 (98.20%)	2 (1.80%)	0.71	0.677
	≥8 nights	232 (97.48%)	6 (2.52%)	1	

usually had spent a relatively long period in hospital. Hence the length of stay over eight nights is chosen as a reference group. The stratum-specific odds ratio of four to seven nights in the stratum of two wards is less than one, indicating that the risk of MRSA acquisition for the patients who had stayed for eight nights or over and stayed in exactly two wards is slightly higher than the patients who had stayed for four to seven nights and stayed in exactly two wards. A similar situation occurs in the third stratum where the patients had stayed in three wards or more. i.e. for the patients who had stayed in three wards or more while in hospital, the risk of MRSA acquisition for those patients who had also stayed for four to seven nights is 0.71 times the risk of MRSA acquisition for the patients staying for eight nights or over. However, for the patients staying in one ward, the stratum-specific odds ratios for the length of stay are all above one. Especially, in the stratum of one ward, the stratum-specific odds ratio for four to seven nights is the largest, followed by the stratum-specific odds ratio for one night while the stratum-specific odds ratio for two to three nights is smallest. The high p-values in each stratum of the number of wards indicate that the stratum-specific odds ratios are consistent across the different lengths of stay.

Applying the trend test based upon a logistic regression model with number of wards as a categorical variable and length of stay taking values 0, 1, 2, 3, for the investigation of the linear relationship between length of stay and MRSA acquisition adjusted by the effect of number of wards, it generates a p-value (0.091), leading to the conclusion that there is no obvious evidence that the length of stay is associated with MRSA acquisition across the strata of number of wards.

We combine the first three length of stay categories together due to the extremely small number of records with MRSA acquisition for the patients staying for seven nights or less. Thus a 2×2 table in each stratum of number of wards can be constructed (shown in Table B.8) so that the CMH test as well as the MH method can be carried out to assess the effect of length of stay on the risk of MRSA acquisition by controlling for the potential confounding effect of number of wards. The stratified analysis based on the new categorised length of stay is illustrated in Table B.8.

Table B.8: Results of combined length of stay groups, stratified by number of wards.

		MRSA acquisition		
No. of wards	Length of stay	No	Yes	OR
1 ward	≤7 nights	1305	13	1
	≥8 nights	213	1	0.47
2 wards	≤7 nights	590	6	1
	≥8 nights	222	6	2.66
≥ 3 wards	≤7 nights	126	2	1
	≥8 nights	232	6	1.63

In general, the stratum-specific odds ratio for the patients who had stayed in exactly two wards is the largest, followed by the stratum-specific odds ratio for the patients who had stayed in three or more wards. The patients who had moved through three or more wards and stayed for eight nights or over are 1.63 times as likely to acquire MRSA as the patients who had also moved through three or more wards but stayed for seven nights or less. On the other hand, for the patients who had stayed in only one ward, the risk of MRSA acquisition for those patients who had also stayed for eight nights or longer is 0.47 times the risk of MRSA acquisition for the patients who had stayed in only one ward but stayed for seven nights or less.

Thus using Woolf's method to detect the homogeneity, the p-value (0.40) implies that there is no multiplicative interaction between length of stay and number of wards.

Based on the MH method, we can obtain the adjusted odds ratio associated with the 95% confidence interval. The results including the p-value adjusted for the possible confounding effect of number of wards are demonstrated as follows:

Table B.9: Results of combined length of stay, adjusted by number of wards.

Length of stay	p-value (CMH method)	$\mathrm{OR}_{\mathrm{MH}}$	95% CI
$\geq$8 nights vs. $\leq$7 nights	0.296	1.54	(0.69,3.40)

The high p-value of the adjusted χ^2 test based on the CMH method reflects that the stratum-specific odd ratios differ from one in an inconsistent manner, i.e. the length of stay is independent of MRSA acquisition controlling for the confounding effect of number of wards. The adjusted odds ratio along with the corresponding 95% confidence interval, which contains one, also indicates that there is no strong difference in the risk of MRSA acquisition between the various lengths of stay that the patients had spent in hospital after controlling for the possible confounding effect of the number of wards. Comparing the adjusted odds ratio for length of stay which is 1.54 (in Table B.9) with the corresponding crude odds ratio which is obtained in Table B.10 (1.88), we can conclude that there exists a confounding effect on length of stay by the number of wards since the adjusted odds ratio decreases by 18.09% in comparison with the corresponding crude odds ratio.

Table B.10: The two-way table for the new categorised length of stay.

	MRSA acquisition		
Length of stay	No	Yes	OR
$\leq$7 nights	2,023	21	1
$\geq$8 nights	667	13	1.88

The association between open wounds or ulcers and MRSA acquisition, assuming that number of wards is a potential confounder

Similarly, we investigate the potential confounding effect of the number of wards on open wounds or ulcers by a stratified analysis due to the strong association between those two risk factors. First of all, we assess the effect of open wounds or ulcers on MRSA acquisition by controlling for the possible confounding effect of the number of wards, which is done by applying the stratified analysis.

Table B.11: Stratified risk analysis of open wounds or ulcers in different number of wards.

		MRSA acquisition			
No. of wards	wounds/ulcers	No	Yes	OR	p-value
1 ward	No	1342 (99.11%)	12 (0.89%)	1	
	Yes	129 (98.47%)	2 (1.53%)	1.73	0.715
2 wards	No	738 (98.80%)	9 (1.20%)	1	
	Yes	44 (93.62%)	3 (6.38%)	5.59	0.012
≥ 3 wards	No	316 (98.14%)	6 (1.86%)	1	
	Yes	31 (93.94%)	2 (6.06%)	3.40	0.144

The results in Table B.11 show that for the patients who had stayed in two wards, the risk of MRSA acquisition for those patients who also had open wounds or ulcers is 5.59 times as high as the risk of MRSA acquisition for the patients without open wounds or ulcers. The corresponding lower p-value 0.012 indicates that there is a strong significance in the effect of open wounds or ulcers with respect to MRSA acquisition for the patients who had stayed in two wards. However, the odds ratio in the stratum of one ward does not significantly vary from one. A similar situation occurs in the stratum of three or more wards, where although the stratum-specific odds ratio is 3.40, the corresponding p-value (0.144) is quite large, indicating that the stratum-specific odds ratio for the patients staying in three or more wards is not significantly different from one.

Using Woolf's method to investigate the homogeneity, the p-value (0.523) indicates that there is little convincing evidence that the odds ratio of MRSA acquisition associated with open wounds or ulcers is modified by number of wards.

Using the MH method, the adjusted odds ratio 3.09 associated with the 95% confidence interval (1.33, 7.18) in Table B.12 reflects that the patients with open wounds or ulcers are 3.09 times as likely to acquire MRSA as the patients without open wounds or ulcers, controlling for the possible confounding effect of number of wards. Moreover, the p-value using the CMH method also implies that open wounds or ulcers are strongly related with MRSA acquisition when adjusted by number of wards. Comparing the adjusted odds ratio for open wounds or ulcers (3.09) shown in Table B.12 with the crude odds ratio 3.04 shown in Table 6.3, we can conclude that there is no confounding effect of the number of wards associated with open wounds or ulcers since the adjusted odds ratio is around 1.64% higher than the corresponding crude odds ratio.

Table B.12: Adjusted results of open wounds or ulcers, adjusted by number of wards.

Wounds/ulcers	p-value (CMH method)	$\mathrm{OR}_{\mathrm{MH}}$	95% CI
Yes vs. No	0.0057	3.09	(1.33,7.18)

The association between renal failure and MRSA acquisition, assuming that number of wards is a potential confounder

Considering another co-morbidity risk factor: renal failure, we apply the stratified analysis in order to assess the effect of renal failure on MRSA acquisition by removing the possible confounding effect of number of wards and the results are shown in Table B.13.

Table B.13: Stratified risk analysis of renal failure in different number of wards.

		MRSA acquisition			
No. of wards	Renal failure	No	Yes	OR	p-value
1 ward	No	1431 (99.10%)	13 (0.90%)	1	
	Yes	35 (97.22%)	1 (2.78%)	3.14	0.276
2 wards	No	749 (98.55%)	11 (1.45%)	1	
	Yes	21 (95.45%)	1 (4.55%)	3.24	0.271
$\geq$ 3 wards	No	324 (98.18%)	6 (1.82%)	1	
	Yes	17 (89.47%)	2 (10.53%)	6.35	0.030

It is obvious that the stratum-specific odds ratio increases as the number of wards that the patients had stayed in increases. i.e. the risk of MRSA acquisition for the patients with renal failure increases when the number of wards that the patients had stayed in increases. Specifically, the stratum-specific odds ratio becomes 6.35 for the patients staying in three or more wards, which almost doubles the stratum-specific odds ratio in the stratum of one ward. Moreover, in each stratum of the number of wards, the risk of MRSA acquisition for those patients with renal failure is always higher than the risk of MRSA acquisition for the patients without renal failure although the high p-values for the first two strata of the number of wards imply nonsignificant effect of renal failure.

In order to examine the homogeneity of consistency of stratum-specific odds ratios, Woolf's method can successfully generate the corresponding p-value (0.901), which was significantly high. Hence we can conclude that there is no multiplicative interaction between renal failure and number of wards associated with MRSA acquisition.

Now we use the adjusted odds ratio based on the MH method to assess the relationship between renal failure and MRSA acquisition, controlling for the potential confounding effect of number of wards.

Table B.14: Adjusted results of renal failure, adjusted by number of wards.

Rental failure	p-value (CMH method)	OR_{MH}	95% CI
Yes vs. No	0.005	4.22	(1.43,13.49)

The low p-value in Table B.14 reflects that renal failure and MRSA acquisition are associated, adjusting for the possible confounding effect of the number of wards. Moreover, the adjusted odds ratio indicates that patients with renal failure are nearly 4.22 times as likely to acquire MRSA compared to the patients without renal failure across the strata of number of wards. With regard to the adjusted odds ratio of MRSA acquisition associated with renal failure (which is 4.22 shown in Table B.14), it decreases slightly by 7.86% from the corresponding crude odds ratio (4.58 shown in Table 6.3), which implies that there is no significant confounding effect of number of wards on renal failure.

- **Open wounds or ulcers as a potential confounder**

The association between age and MRSA acquisition, assuming that open wounds or ulcers is a potential confounder

According to the two-way tables (in Table 6.4 of Section 6.4), there is a strong association between open wounds or ulcers and age, number of wards and renal failure respectively. Moreover, the univariate analysis also provides a significant result of the effect of open wounds or ulcers associated with MRSA acquisition. Therefore, there is convincing evidence that open wounds or ulcers can be considered as a potential confounder. Using the same procedure we introduced above, Table B.15 demonstrates the results of the association between age and MRSA acquisition, controlling for the possible confounding effect of open wounds or ulcers by means of stratification.

Table B.15: Stratified analysis of age in different wounds/ulcers groups.

		MRSA acquisition			
Wounds/ulcers	Age	No	Yes	OR	p-value
Yes	$\leq$ 49 years	70 (100.00%)	0 (0%)	0	0.993
	50-64 years	51 (96.23%)	2 (3.77%)	0.18	0.056
	65-79 years	65 (98.48%)	1 (1.52%)	0.07	0.020
	$\geq$ 80 years	18 (81.82%)	4 (18.18%)	1	
No	$\leq$ 49 years	523 (99.62%)	2 (0.38%)	0.17	0.027
	50-64 years	706 (99.44%)	5 (0.56%)	0.25	0.028
	65-79 years	861 (98.40%)	14 (1.60%)	0.72	0.474
	$\geq$ 80 years	308 (97.78%)	7 (2.22%)	1	

By treating the age group of 80 years or over as a reference level, in general, the stratum-specific odds ratio decrease as the age is decreasing. In other words, the elderly patients have a higher risk of acquiring MRSA in each stratum. Using the trend test to detect the trend in risk of MRSA, as age increases, is consistent from stratum to stratum, a relatively low p-value (< 0.001) indicates that the risk of MRSA increases as age increases after controlling for the possible confounding effect of open wounds or ulcers.

Considering the homogeneity test, we obtain the p-value (0.926) by Woolf's method, indicating that there is no modified effect of open wounds or ulcers on age (i.e. there is no evidence that multiplicative interaction between open wounds or ulcers and age is associated with the risk of MRSA acquisition).

We estimate the pairwise adjusted odds ratios for the categorised age across the strata of open wounds or ulcers by applying the MH method. In addition, the CMH method is also applied for the assessment of the dependence of age and MRSA acquisition, controlling for the possible confounding effect of open wounds or ulcers.

Generally speaking, Table B.16 shows that by adjusting for the possible confounding

Table B.16: Comparison results of age groups across wounds strata.

Age	p-value (CMH method)	$\mathrm{OR}_{\mathrm{MH}}$	95% CI
50-64 years vs. $\leq$ 49 years	0.23	2.48	(0.52,11.82)
65-79 years vs. $\leq$ 49 years	0.023	4.67	(1.07,20.35)
$\geq$ 80 years vs. $\leq$ 49 years	< 0.001	10.09	(2.38,42.88)

effect of open wounds or ulcers, the risk of MRSA acquisition is relatively larger for the elderly patients (say over 65 years) compared to the young patients 49 or under. Specifically, the risk of MRSA acquisition is more than ten times as high for the patients 80 years and older compared with the patients 49 years and younger. The relatively small p-values corresponding to the adjusted odds ratios of 80 years or over and 65-79 years against 49 or under also indicate that both the stratum-specific odds ratios for the patients 80 years or over and the patients aged between 65 and 79 years vary significantly from one consistently across the strata of open wounds or ulcers respectively. On the other hand, there is no significant difference in the risk of MRSA acquisition between the patients 49 years or under and the patients aged between 50-64 years across the strata of open wounds or ulcers.

Comparing the adjusted odds ratios (which are 2.48, 4.67, 10.09 respectively shown in Table B.16) with the crude odds ratios (which are 2.35, 4.83, 9.99 shown in Table 6.3), we can conclude that there is no confounding effect of open wounds or ulcers on age of patient associated with MRSA since the adjusted odds ratios are roughly 5.53%, 3.31% and 1.00% larger than the corresponding crude odds ratios.

The association between number of wards and MRSA acquisition, assuming that open wounds or ulcers is a potential confounder

Now we investigate the possible confounding effect of open wounds or ulcers associated with number of wards following the same procedure that we applied above. Firstly, the stratified analysis is implemented in order to remove the possible confounding effect for assessing the effect of the number of wards on MRSA acquisition.

In general, in each stratum the risk of MRSA acquisition for the patients who

Table B.17: Stratified analysis associated with wounds/ulcers.

Wounds/ulcers	Number of wards	MRSA acquisition		OR	p-value
		No	Yes		
Yes	1 ward	129 (98.47%)	2 (1.53%)	1	
	2 wards	44 (93.62%)	3 (6.38%)	4.40	0.111
	$\geq$ 3 wards	31 (93.94%)	2 (6.06%)	4.16	0.162
No	1 ward	1342 (99.11%)	12 (0.89%)	1	
	2 wards	738 (98.80%)	9 (1.20%)	1.36	0.484
	$\geq$ 3 wards	316 (98.14%)	6 (1.86%)	2.12	0.135

had stayed in two or more wards is higher than the risk of MRSA acquisition for the patients who had stayed in exactly one ward. From Table B.17, we can see that both stratum-specific odds ratios of two wards and three or more wards against one ward are greater than four in the stratum of patients with open wounds or ulcers though the stratum-specific odds ratio of three or more wards is slightly lower than the stratum-specific odds ratio of two wards. As to the stratum of patients without open wounds or ulcers, the stratum-specific odds ratio increases gently as the number of wards increases. The trend test yields a low p-value (0.036) which reveals that there is a slightly increasing trend in the odds ratio of number of wards associated with MRSA acquisition across the strata of open wounds or ulcers.

Using Woolf's method to investigate the homogeneity, the high p-value (0.799) indicates that there is no multiplicative interaction between open wounds or ulcers and number of wards associated with MRSA acquisition.

The adjusted odds ratios for the pairwise categories of the number of wards is generated for detecting the significance of the effect of number of wards associated with MRSA acquisition by controlling for the possible confounding effect of open wounds or ulcer. The results are shown in the Table B.18.

Table B.18: Comparison results of number of wards stratified by wounds.

Number of wards	p-value (CMH method)	$\mathrm{OR}_{\mathrm{MH}}$	95% CI
2 wards vs. 1 ward	0.17	1.68	(0.78,3.64)
$\geq$ 3 wards vs. 1 ward	0.043	2.42	(1.00,5.81)

Clearly, by treating one ward as the baseline the odds ratio of three or more wards is significantly different from one across the strata of open wounds or ulcers whereas the risk of MRSA acquisition for the patients staying in one ward does not significantly vary from the risk of MRSA acquisition for the patients who had stayed in two wards, adjusting for the possible confounding effect of open wounds or ulcers.

Since the adjusted odds ratios (i.e. 1.68 and 2.42) are fairly close to the corresponding crude odds ratios (i.e. 1.60, 2.42 shown in Table 6.2), we can conclude that there is no confounding effect of open wounds or ulcers associated with number of wards.

The association between renal failure and MRSA acquisition, assuming that open wounds or ulcers is a potential confounder

Considering the potential confounding effect of open wounds or ulcers on renal failure, we construct the two-way table presenting the number of patients with renal failure who had acquired MRSA while in hospital stratified by open wounds or ulcers. The results of the evaluation of the stratum-specific odds ratios for MRSA acquisition associated with renal failure are shown in Table B.19.

Table B.19: Stratified analysis for the renal failure, assuming wounds/ulcers is a potential confounder.

		MRSA acquisition			
Wounds/ulcers	Renal failure	No	Yes	OR	*p*-value
Yes	No	188 (97.41%)	5 (2.59%)	1	
	Yes	15 (88.24%)	2 (11.76%)	5.01	0.07
No	No	2309 (98.93%)	25 (1.07%)	1	
	Yes	58 (96.67%)	2 (3.33%)	3.18	0.121

The stratum-specific odds ratio for the patients with renal failure in comparison with the patients without renal failure is larger than one in each stratum, which means that by controlling for the possible confounding effect of open wounds or ulcers, the risk of MRSA acquisition is higher for the patients with renal failure. For the patients with open wounds or ulcers, the risk of MRSA acquisition for the patients who also had renal failure is 5.01 times as high as the risk of MRSA acquisition for the patients who did not have renal failure, but the difference in the risk of MRSA acquisition between the patients with both open wounds or ulcers and renal failure and the patients with open wounds or ulcers but without renal failure is not significant due to the slightly high p-value (0.07). On the other hand, the odds ratio of renal failure for the patients without open wounds or ulcers is 3.18, indicating that the risk of MRSA acquisition for the patient with renal failure but without open wounds or ulcers is 3.18 times as high as the risk of MRSA acquisition for the patients without renal failure or open wounds or ulcers.

The test of the homogeneity by Woolf's method provides the conclusion that there is no modified effect of open wounds or ulcers on renal failure associated with MRSA acquisition since the corresponding adjusted p-value equals 0.735. i.e. there is no multiplicative interaction between open wounds or ulcers and renal failure associated with MRSA acquisition.

Using the MH method to control the possible confounding effect of open wounds or ulcers, we generate the adjusted odds ratio as well as the corresponding 95% confidence interval. Table B.20 includes all the adjusted estimates of the effect of renal failure associated with MRSA acquisition.

Table B.20: Adjusted results.

Renal failure	p-value (CMH method)	$\mathrm{OR}_{\mathrm{MH}}$	95% CI
Yes vs. No	0.010	3.86	(1.28,11.61)

The small p-value obtained by the CMH method presents that the stratum-specific odds ratios of renal failure varies from one consistently across the strata, which means that the renal failure is associated with MRSA acquisition after adjustment for the possible confounding effect of open wounds or ulcers. The adjusted odds ratio shown in Table B.20 is 3.86, which means that the risk of MRSA acquisition for the patients with renal failure is about 3.86 times as high as the risk of MRSA acquisition for the patients without renal failure after controlling for the possible confounding effect of open wounds or ulcers. Compared to the crude odds ratio for renal failure (4.58 shown in Table 6.3), the corresponding adjusted odds ratio decreases by 15.72%. Hence, we can conclude that there is a confounding effect caused by open wounds or ulcers with respect to renal failure.

- **Renal failure as a potential confounder**

The two-way tables in Table 6.4 of Section 6.4 also reflect that renal failure is strongly associated with number of wards and open wounds or ulcers. As we mentioned above, it is reasonable to consider renal failure as a potential confounder since renal failure is also significantly related to MRSA acquisition according to the results in the univariate analysis (see Section 6.3).

The association between number of wards and MRSA acquisition, assuming that renal failure is a potential confounder

In order to remove the potential confounding effect of renal failure, the estimation of the odds ratio for number of ward associated with MRSA acquisitions is carried out by stratification of renal failure. The results are shown in Table B.21.

Table B.21: Stratified analysis associated with renal failure.

Renal failure	Number of wards	MRSA acquisition		OR	p-value
		No	Yes		
Yes	1 ward	35 (97.22%)	1 (2.78%)	1	
	2 wards	21 (95.45%)	1 (4.55%)	1.67	0.723
	$\geq$ 3 wards	17 (89.47%)	2 (10.53%)	4.12	0.261
No	1 ward	1431 (99.10%)	13 (0.90%)	1	
	2 wards	749 (98.55%)	11 (1.45%)	1.62	0.244
	$\geq$ 3 wards	324 (98.18%)	6 (1.82%)	2.04	0.152

Generally speaking, the odds ratio in each stratum shows an increasing trend when the number of wards that the patient had moved through increases. For example, compared to the patients who had stayed in one ward and had renal failure, the patients who had stayed in two wards and had renal failure are 1.67 times as likely to acquire MRSA while the risk of MRSA for the patients who stayed in three or more wards and had renal failure increases by a factor of 4.12. However the corresponding p-values for the stratum-specific odds ratios show non-significance. The trend test produces a p-value (0.054), indicating that there is no significant increasing linear trend in the stratum-specific odds ratio as the number of wards increases across the strata of renal failure.

Similarly, using Woolf's method to detect the homogeneity, the p-value (0.838) implies that there is no multiplicative interaction between renal failure and number of wards associated with MRSA acquisition.

The adjusted effect of number of wards on MRSA acquisition can be estimated using the MH method by controlling for the potential confounding effect of renal failure. The results are illustrated in Table B.22.

Table B.22: Comparison results of number of wards stratified by wounds.

Number of wards	p-value (CMH method)	OR_{MH}	95% CI
2 wards vs. 1 ward	0.219	1.62	(0.75,3.52)
$\geq$ 3 wards vs. 1 ward	0.067	2.28	(0.93,5.57)

This shows that the adjusted odds ratio for three or more wards is 2.28 in comparison with one ward with the corresponding p-value of 0.067 using the CHM method which indicates that the stratum-specific odds ratios of three or more wards against one ward are statistically equal to one across the strata of renal failure. On the other hand, the patients who had stayed in two wards are 1.62 times as likely to acquire MRSA compared to the patients who had stayed in one ward, controlling for the potential confounding effect of renal failure.

Comparing the adjusted odds ratios (i.e. 1.62 and 2.28) with the corresponding crude odds ratios (i.e. 1.60 and 2.42 shown in Table 6.2), there is no confounding effect of renal failure on the number of wards associated with the MRSA acquisition since the differences between the adjusted odds ratios and the crude odds ratios are less than 10%.

The association between open wounds or ulcers and MRSA acquisition, assuming that renal failure is a potential confounder

Finally, we investigate the possible confounding effect of renal failure on open wounds or ulcers. We similarly apply the stratified analysis to estimate the odds ratios for the MRSA acquisition associated with open wounds of ulcers, controlling for the effect of renal failure.

Table B.23: Stratified analysis of wounds/ulcers associated with renal failure.

		MRSA acquisition			
Renal failure	Wounds/ulcers	No	Yes	OR	p-value
Yes	No	58 (96.67%)	2 (3.33%)	1	
	Yes	15 (88.24%)	2 (11.76%)	3.87	0.194
No	No	2309 (98.93%)	28 (1.07%)	1	
	Yes	188 (97.41%)	5 (2.59%)	2.46	0.070

In Table B.23, the stratum-specific odds ratio indicates that the risk of MRSA acquisition is higher for the patients with open wounds or ulcers. Particularly, the patients with both open wounds or ulcers and renal failure are 3.87 times as likely to acquire MRSA as the patients with renal failure but without open wounds or ulcers.

Using Woolf's method, it yields a p-value (0.735) which gives the underlying conclusion that there is no multiplicative interaction between renal failure and open wounds or ulcers associated with MRSA acquisition.

Next, the adjusted odds ratios for MRSA acquisition associated with open wounds or ulcers can be estimated across the strata of renal failure using the MH method. The results are shown in Table B.24.

Table B.24: Adjusted results.

Wound/ulcers	p-value (CMH method)	OR_{MH}	95% CI
Yes vs. No	0.021	2.70	(1.13,6.44)

Table B.24 illustrates that the patients with open wounds or ulcers are 2.70 times as likely to acquire MRSA compared to the patients without open wounds or ulcers, adjusting for the possible confounding effect of renal failure. The small p-value of 0.021 based on the CMH method also implies that open wounds or ulcers is strongly associated with MRSA acquisition, controlling for the possible confounding effect of renal failure. Since the crude odds ratio for MRSA acquisition associated with open wounds or ulcers (3.04 in Table 6.3) is 11.18% larger than the corresponding adjusted odds ratio (2.70 in Table B.24), we can conclude that there is a slightly confounding effect of renal failure associated with open wounds or ulcers which weakens the effect of open wounds or ulcers with respect to MRSA acquisition.

- **Length of stay as a potential confounder**

Based on the strong association between length of stay and MRSA acquisition (according to the results of the univariate risk factor analyisis shown in Table 6.3 of Section 6.3) and high correlation between length of stay and age and number of wards respectively (obtained from Table 6.4 in Section 6.4), the risk factor of length of stay is reasonable to be considered as a potential confounder. However there were rarely records of the patients acquiring MRSA while in hospital when they had stayed for less than three nights. One of the modifications is that for the stratification analysis with respect to the length of stay as a potential confounder, the stratum of the length of stay which is less than three nights can be combined as a new stratum so that the records in this new stratum can be expanded. Regarding to the possible confounding effect of length of stay associated with age, the results of the stratification analysis for the new categorised length of stay is shown in Table B.25.

Table B.25: Stratified risk analysis in different categories of length of stay.

		MRSA acquisition		
Length of stay	Age	No	Yes	OR
1-3 nights	≤49 years old	356 (100.00%)	0 (0%)	0
	50-64 years old	355 (99.72%)	1 (0.28%)	0.55
	65-79 years old	344 (98.85%)	4 (1.15%)	0.13
	≥80 years old	94 (97.92%)	2 (2.08%)	1
4-7 nights	≤49 years old	165 (100.00%)	0 (0%)	0
	50-64 years old	258 (99.23%)	2 (0.77%)	1.04
	65-79 years old	335 (97.38%)	9 (2.62%)	0.30
	≥80 years old	116 (97.48%)	3 (2.52%)	1
≥ 8 nights	≤49 years old	93 (97.89%)	2 (2.11%)	0.46
	50-64 years old	171 (98.28%)	3 (1.72%)	0.37
	65-79 years old	275 (99.28%)	2 (0.72%)	0.16
	≥80 years old	128 (95.52%)	6 (4.48%)	1

In Table B.25, there is no record of younger patients aged 49 or under who had acquired MRSA and stayed for seven nights or less which leads to zero cells in the table. Hence we use eight or more nights as the reference group for the estimation of the stratum-specific odds ratios. Generally speaking, the stratum-specific odds ratio decreases as age increases but then it increases dramatically for the patient aged 80 years old or older in each stratum of length of stay. A small p-value (< 0.01) obtained from the trend test indicates that there is a linear trend of age associated with the risk of MRSA acquisition adjusted for the strata of length of stay.

In order to estimate the effect of age on MRSA acquisition across the strata of length of stay based on the MH method, we combined the first two categories of age (≤ 49 years and 50-64 years) together as a new category since there are few records of patients aged 49 years old or under in the stratum. The Woolf's method for testing for homogeneity yields a high p-value of 0.093, which implies that there is no multiplicative interaction between length of stay and age associated with MRSA acquisition. The results for the adjusted odds ratio for the new categorised age and the corresponding p-value, controlling for the possible confounding effect of length of stay are shown in Table B.26.

Table B.26: Results of combined length of stay.

Age	p-value (CMH method)	OR_{MH}	95% CI
65-79 years vs. ≤ 64 years	0.037	2.32	(1.00,5.37)
≥ 80 years vs. ≤ 64 years	0.0017	3.77	(1.51,9.42)

The results shows that the risk of MRSA acquisition for the patients aged 65-79 years old is 2.32 times as high as the risk of MRSA acquisition for patients aged 64 years or under, adjusting for the possible confounding effect of length of stay. The small p-value obtained from the CMH method also indicates that the odds ratios of the age of 65-79 years in each stratum of length of stay differ from one consistently. The patients aged 80 years old or over are about 3.77 times as likely to acquire MRSA in hospital as the patients aged 64 years old or under.

Comparing the crude odds ratios for age (i.e. 2.75 and 5.68) shown in Table B.27 and the corresponding adjusted odds ratio of 2.32 and 3.77 (shown in Table B.26), we can conclude that there is a confounding effect of length of stay on age associated with

MRSA acquisition since the adjusted odds ratios for age decrease by 19.06% and 33.63% from the crude one.

Table B.27: The crude odds ratio for the new categorised age.

Age	Crude odds ratio
$\leq$ 64 years	1
65-79 years	2.75
$\geq$ 80 years	5.68

Similarly, we use the same procedure to investigate the potential confounding effect of length of stay on number of wards associated with MRSA acquisition. Firstly, the results for the stratification analysis are illustrated in Table B.28. Note that since the records for the patients who had acquired MRSA in hospital and stayed at two or more wards for three nights or less are zeros, we combined the first three strata of length of stay here.

Table B.28: Stratified risk analysis in different categories of length of stay.

		MRSA acquisition		
Length of stay	Number of wards	No	Yes	OR
1-7 nights	1 ward	1350 (99.01%)	13 (0.99%)	1
	2 wards	590 (98.99%)	6 (1.01%)	1.02
	$\geq$3 wards	126 (98.44%)	2 (1.56%)	1.59
$\geq$ 8 nights	1 ward	213 (99.53%)	1 (0.47%)	1
	2 wards	222 (97.37%)	6 (2.63%)	5.75
	$\geq$3 wards	232 (97.48%)	6 (2.52%)	5.51

Generally speaking, the stratum-specific odds ratio increases as the number of wards increases. Especially, for the patients who had stayed at two wards or more for eight nights or over, the risk of MRSA is more than five times higher in comparison with the patients who had stayed at one ward for eight nights or over. The trend test provide a high p-value of 0.15, indicating that there is no significant linear trend of number of wards in the risk of MRSA acquisition adjusting for the possible confounding effect of length of stay.

The homogeneity test using Woolf's method gives the conclusion that there is no multiplicative interaction between length of stay and number of wards due to the high p-value (0.86).

Table B.29: Results of new categorised number of wards.

	p-value (CMH method)	OR_{MH}	95% CI
2 wards vs. 1 ward	0.303	1.54	(0.69,3.49)
$\geq$ 3 wards vs. 1 ward	0.081	2.81	(0.91,8.67)

Table B.29 show that the risk of MRSA acquisition for patients staying at three wards or more is 2.81 times as high as the risk of MRSA acquisition for patients staying in one ward, which is adjusted for the possible confounding effect of length of stay. In addition, the corresponding high p-value (0.081) based on CMH methods demonstrates that the stratum-specific odds ratio of three or more wards against one ward is statistically equal to one consistently. Similarly, the patients who had stayed in two wards is 1.54 times as likely to acquire MRSA as the patients who had stayed in one ward.

Compared to the crude odds ratios for the number of wards which is 1.60 and 2.42 respectively (shown in Table 6.2), the corresponding adjusted odds ratios decrease by 16.12% and 3.75%, indicating that there is a confounding effect of length of stay on number of wards associated with MRSA acquisition.

Appendix C

The results of the Negative Binomial regression models

Table C.1: The estimates of the Negative Binomial regression models for one ward cases.

Variable		Estimate	Standard Error	p-value
Intercept (β_0)		1.258943	0.767416	0.100902
Ward code 1	1	0.096337	0.775522	0.901140
	11	-0.848932	0.790683	0.282971
	12	-0.506235	1.102517	0.646117
	13	-0.394752	0.822303	0.631187
	14	0.350917	0.785323	0.654987
	16	0.317375	0.841164	0.705947
	17	-0.505317	0.773624	0.513639
	19	-0.074593	0.814100	0.926995
	2	-0.316642	0.790967	0.688918
	20	1.220860	0.830169	0.141395
	21	-0.251031	0.857627	0.769748

Continued on next page

Table C.1 – continued from previous page

Variable		Estimate	Standard Error	p-value
	25	0.149274	0.796376	0.851315
	27	0.834269	0.834582	0.317492
	29	0.316170	0.817463	0.698927
	3	-0.934858	1.009832	0.354573
	30	-0.236847	0.795013	0.765767
	31	-0.738036	0.775344	0.341157
	32	-1.037492	0.792772	0.190640
	33	-0.420780	0.804742	0.601061
	34	-0.781715	0.806905	0.332654
	36	0.019104	0.782154	0.980514
	4	-1.703802	1.400088	0.223633
	40	-0.332728	0.773270	0.666987
	42	-0.467312	0.768649	0.543211
	43	-0.210884	0.778244	0.786411
	44	-0.344128	0.768077	0.654125
	45	-0.107353	0.788919	0.891762
	46	-0.071220	0.769883	0.926295
	47	0.181354	0.768727	0.813499
	48	-0.453402	0.798681	0.570247
	49	-1.601715	0.899741	0.075044
	50	0.011845	0.779161	0.987870
	7	-1.892438	0.849157	0.025840
	8	-1.511116	0.844667	0.073614
	W07	-1.181617	0.768846	0.124325
	W08	-1.697072	1.127058	0.132130
Ward code 1				Continued on next page

Table C.1 – continued from previous page

Variable		Estimate	Standard Error	p-value
	W09	-0.384960	0.766651	0.615574
	W10	0.039794	0.765648	0.958549
Age		0.006740	0.001838	0.000246

Table C.2: The estimates of the Negative Binomial regression models for two wards cases.

Variable		Estimate	Standard Error	p-value
Intercept (β_0)		2.838814	0.528682	< 0.001
Ward code 1	13	-0.853984	0.856839	0.318926
	14	-1.568288	0.689879	0.023009
	16	-2.832724	1.540072	0.065864
	17	-2.539149	1.077559	0.018454
	18	-2.116110	0.691506	0.002212
	19	-1.555685	1.488852	0.296074
	20	-1.837788	0.910735	0.043600
	21	-1.734243	0.524366	0.000942
	25	-1.362057	1.106619	0.218388
	3	-2.616592	1.081613	0.015557
	31	-1.575428	0.563406	0.005170
	32	-1.201436	0.569365	0.034847
	33	-1.162515	0.671743	0.083524
	34	-1.187229	0.647080	0.066543
	35	0.304371	0.918348	0.740318
	38	-2.821055	1.083626	0.009232

Continued on next page

Table C.2 – continued from previous page

Variable		Estimate	Standard Error	p-value
	42	-0.678162	0.717573	0.344619
	43	-2.889476	1.367636	0.034622
	44	-2.355419	0.956835	0.013829
	45	-2.556684	0.785824	0.001140
	46	-1.335451	0.726639	0.066085
	47	-0.885478	0.541721	0.102140
	49	-1.324207	0.478934	0.005694
	50	-1.620076	0.890290	0.068802
	8	-1.768056	0.831578	0.033491
	AE	-1.212188	1.046103	0.246552
	W07	-2.087251	0.936794	0.025875
	W08	-2.269145	0.979024	0.020462
	W09	-3.419151	0.972793	0.000440
	W10	-3.577045	1.543513	0.020478
Ward code 2	11	-0.475708	0.353357	0.178220
	12	-0.565596	0.360801	0.116971
	13	-0.534036	0.400890	0.182819
	14	-0.008064	0.272717	0.976411
	17	-0.460650	0.562514	0.412837
	19	-0.105661	0.274399	0.700191
	2	-0.022755	0.341413	0.946860
	20	0.083285	0.289250	0.773397
	21	-0.590040	1.152712	0.608741
	25	0.276755	0.317279	0.383058
	27	-0.003633	0.233262	0.987572

Continued on next page

Table C.2 – continued from previous page

Variable		Estimate	Standard Error	p-value
	29	-1.033095	0.854259	0.226530
	3	-0.274697	0.297570	0.355938
	30	-0.532065	0.367809	0.148015
	31	-1.327961	0.426548	0.001850
	32	-1.267066	0.486264	0.009168
	33	-0.978135	0.375235	0.009141
	34	-0.948038	0.487052	0.051597
	39	0.082720	0.706042	0.906733
	4	-0.065896	0.284852	0.817057
	40	-0.321023	0.318439	0.313401
	42	-0.109014	0.333163	0.743510
	43	-0.970669	0.213701	< 0.001
	44	0.261049	0.353126	0.459754
	45	-1.312059	0.550828	0.017220
	46	-1.131518	0.453438	0.012581
	47	-0.801787	0.583696	0.169554
	50	-0.685253	0.267992	0.010558
	7	-1.222455	0.652774	0.061109
	8	-0.570296	0.904834	0.528513
	W08	-1.298137	1.404011	0.355178
	W09	0.138762	0.869227	0.873166
Age		0.008816	0.002491	0.000402