Statistical Inference for Population Based Measures of Risk Reduction

A thesis presented in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Statistics

at Massey University, Palmerston North, New Zealand.

Sarah Pirikahu

2019

Abstract

Epidemiologists and public health practitioners often wish to determine the population impact of an intervention to remove or reduce a risk factor. Population attributable type measures, such as the population attributable risk (PAR) and population attributable fraction (PAF), provide a means of assessing this impact in a way that is accessible for a non-statistical audience. To apply these concepts to real-world data, the calculation of estimates and confidence intervals for these measures should take into account the study design and any sources of uncertainty.

We provide a Bayesian approach for estimating the PAR and its credible interval, from cross-sectional data resulting in a 2×2 table, and assess its Frequentist properties. With the Bayesian approach proving superior this model is then extended by incorporating uncertainty due to the use of an imperfect diagnostic test for exposure. The resulting model is under-identified which causes convergence problems for common MCMC samplers, such as Gibbs and Metropolis-Hastings. An alternative importance sampling method performs much better for these under-identified models and can be used to explore the limiting posterior distribution. However, this comes at the cost of needing to identify an appropriate transparent parameterisation, which can be difficult. We provide an adaptation of the Metropolis-Hastings random walk sampler which, in comparison to other MCMC samplers, more efficiently explores the posterior ridge of an under-identified model for large sample sizes.

Often data used to estimate these population attributable measures may include multiple risk factors in addition to the one being considered for removal. Uncertainty regarding the distribution of these risk factors in the population affects the inference for PAR and PAF. To allow for this uncertainty we propose a methodology where the uncertainty in the joint distribution of the response and the covariate is accommodated.

Acknowledgments

There are many people over the years who have provided me with support, without whom completion of this thesis would not have been possible. First and foremost, I would like to thank my primary supervisor Professor Geoff Jones for his valuable contributions and support throughout this research project. Geoff's endless patience (even when explaining a concept to me for the nth time, where n is large), vast knowledge and methodical approach to problem solving provided me with the most rewarding PhD journey. Geoff was always willing to go the extra mile and his incredible attention to detail has improved my skills as a researcher in many ways.

I am also very grateful to my co-supervisors Professor Martin Hazelton and Professor Cord Heuer. Martin's statistical expertise, attention to detail and creative way of describing complex ideas has helped enhance many aspects of this research. I am also thankful to Martin for first encouraging me into the field of statistics. Without your continued support over the years I would not be where I am today. I also wish to express my thanks to Cord, who provided the motivation for this research project at the outset and introduced me to the field of epidemiology. As a student who tends to enjoy the more theoretical aspects of statistics, I was greatly assisted by Cord in learning the many challenges that come with real-world applications (plus many new facts about cows). Cord provided valuable ideas and data which without this thesis would not have been possible. Geoff, Martin, Cord, I don't think I could have asked for a more amazing supervisory team.

I wish to acknowledge that this research would not have been possible with out the financial assistance of a Massey University Doctoral Scholarship and employment as a Massey University graduate teaching assistant. My time as a graduate teaching assistant would not have been the same without the guidance of Debbie Leader and Anne Lawrence. I would like to extend my thanks to both Anne and Debbie for helping improve my confidence as a public speaker and igniting my passion for teaching. I also wish to thank all the staff of the Massey University

Statistics department who have provided me with support over this PhD journey. I also would like to acknowledge my manager Rob Lake and mentor Beverley Horn at ESR who not only provided me with my first position as a statistician, but were extremely flexible in allowing me time to complete my PhD over this final year.

I also want to thank the statistics and mathematics postgraduate students who have gone through parts of this PhD journey with me. I have made many friendships which I hope will last a life time. My special thanks goes to Sih-Jing Liao, Ellie Johnson and Simon Brady who were always there to lend an ear, celebrate in my successes and provide positive reinforcement in the tough times. You guys listened to my talks millions of times, proof read things when you need not, went along with my sometimes crazy ideas and I love you all! Finally I want to thank my family for their support, love and patience. You will only have to listen to me say over the phone for a short while longer that I am almost finished with my studies. My family is of course not complete without the addition of my furry friends Leo, Sebastian and Sybil, who provided many cuddles, demanding meows for attention and fun times over my studies.



Contents

List of Figures

List of Tables

Chapter 1

Introduction

The implementation of public health initiatives affects the health and wellbeing of communities both now and in the future. In today's society there exist a large number of known risk factors for diseases that result from potentially modifiable behaviors. Being able to actively reduce or remove these risk factors from the population will be beneficial to overall public health. Implementing these interventions though requires both time and money, making it important for decision makers to know in advance the impact the proposed intervention will have on the population. This concept of assessing the change in risk as a result of reducing or removing a risk factor from the population was introduced in the early 1950s, with the novel paper by Levin (1953) being perhaps the most well known. Since Levin's publication a number of authors (MacMahon et al., 1960; Cole and MacMahon, 1971; Miettinen, 1974) re-discovered and expanded on the concept of a population attributable measure. However, the similarity in nomenclature of these measures, and sometimes lack of clear mathematical definition or assumptions being made, has resulted in confusion in the literature as addressed by Greenland and Robins (1988), Rockhill et al. (1998) and Uter and Pfahlberg (2001). In fact, the structured literature search performed by Uter and Pfahlberg (2001) on 334 papers between 1966 and 1996 showed that 64.5% of authors provided no exact definition for their attributable measure used and only 19.3% provided confidence intervals.

Throughout this thesis we focus on two attributable measures: the population attributable risk (PAR) defined by MacMahon and Pugh (1970) and the measure proposed by Levin (1953) which we define as the population attributable fraction (PAF). Analogues of these measures have also been proposed for rates by MacMahon and Trichopoulos (1996) which we will adopt later in this thesis (see Chapter 4). The PAR allows one to describe the change in the risk (or rate) of disease in the population due to the removal of a risk factor, whereas the PAF provides the proportion of the diseased individuals which can be attributed to that risk factor. This chapter aims to provide an overview of the literature surrounding these population attributable measures. Section 1.1 presents their definition and requirements for estimation. Given the importance of the underlying study design on estimation, the statistical models for commonly used epidemiological studies are outlined in Section 1.2. A discussion of previous work in regards to deriving the uncertainty for the PAR and PAF is given in Section 1.3. We then conclude this chapter with the objectives and structure of this thesis in Section 1.4.

1.1 Population attributable measures

Greenland and Robins (1988) claim that "the number of terms for the attributable risk is perhaps the largest of any concept in epidemiology." This opinion is shared by Gefeller (1990) who cites sixteen different synonymous terms for the original definition of the attributable measure provided by Levin (1953) including: population attributable risk (MacMahon and Pugh, 1970), attributable fraction (Ouellet et al., 1979), population attributable risk percent (Cole and MacMahon, 1971), excess fraction (Miettinen, 1974), etiological fraction (Kleinbaum et al., 1982) and others. In Levin's original paper he defines his attributable measure for case-control data as:

$$\frac{b(r-1)}{b(r-1)+1}, \tag{1.1}$$

where r represents the relative risk (RR) and b the proportion of the general population which is exposed to the risk factor. In terms of naming this measure, Levin (1953) simply refers to

it in context as the "the maximum proportion of lung cancer attributed to smoking". In this thesis we refer to Levin's measure as the population attributable fraction (PAF). After Levin's inception of the PAF it was over a decade later before it was used in published literature again, by MacMahon and Pugh (1970) and Miettinen (1974). However, an equivalent measure was independently derived by Markush and Seigel (1968) and Seigel and Markush (1968) for use in mortality studies, which was later proven to be equivalent to Levin's PAF (Markush, 1977). For an excellent history of the population attributable fraction see Poole (2015).

Several common parameterisations of the formula (1.1) can be seen in the literature. The parameterisation

$$PAF = \frac{P(D^+) - P(D^+|E^-)}{P(D^+)}, \tag{1.2}$$

where $D^+(D^-)$ represent those that have the disease (or not), and $E^+(E^-)$ those who are exposed (or not) to the risk factor being considered for removal (Kleinbaum et al., 1982), makes clear that the PAF can take on values from $-\infty$ to 1. A negative value for PAF implies a protective exposure, whereas a positive PAF a harmful exposure. As an example, suppose we estimated the PAF as 0.5; this value can be interpreted as indicating that 50% of cases in the population would no longer be cases if the exposure could be removed. In contrast, if the PAF were estimated as -9 this would suggest that complete removal of the exposure would result in 9 times as many cases as are currently in the population.

The equivalence between the formulae (1.1) and (1.2) can be demonstrated using simple algebraic manipulation as shown by Leviton (1973). Another commonly seen parameterisation, proposed by Miettinen (1974), is

$$PAF = pd \times \frac{r-1}{r} \tag{1.3}$$

where pd is the proportion of cases exposed to the risk factor being considered for removal. Equivalent formulae have also been proposed by Taylor (1977) and Levin and Bertell (1978) for case-control studies, and Fleiss (1979) for cross-sectional studies. Hanley (2001) points out that often only estimation formulae for the PAF (i.e. $\widehat{PAF}$) are provided in the literature, as

formal definitions expressed as an integral can be difficult for practitioners to use in practice (see Chapter 4, Section 4.3-4.4 for formal definitions of the PAR and PAF). Hanley (2001) takes a heuristic approach to describing the formulae for the PAF in order to aid understanding. Although there are many equivalent parameterisations, expressing the PAF in terms of the relative risk as done in (1.1) and (1.3) provides computational advantages when it comes to deriving its variance formula (Walter, 1976). This is because the moments of the RR, which can be approximated by the odds ratio (OR) when the proportion of disease in the exposed and not exposed groups is small, are well known (Gart, 1962; Gart and Thomas, 1972) and may be used to derive approximate moments for the maximum likelihood estimator of the PAF.

An extension of the PAF which takes into account multiple levels of exposure (e.g $2 \times k$ tables) has been proposed for case-control studies by Walter (1976). The PAF in this case can be defined as

$$PAF = \frac{\sum_{i=1}^{k} \theta_i (r_i - 1)}{1 + \sum_{i=1}^{k} \theta_i (r_i - 1)} = 1 - \frac{1}{\sum_{i=1}^{k} \theta_i r_i}, \tag{1.4}$$

where θ_i is the proportion of the population exposed at level i for $i = 1, \ldots, k$ and r_i the relative risk for exposure at level i. An important limitation of all the parameterisations of the PAF mentioned thus far is that they assume no confounding effects are present. Rockhill et al. (1998), who provides a review on the use and misuses of the PAF, warns that these formulae can be found widely in epidemiological texts often with no warning about their invalidness when confounding factors exist. Additionally, when estimating population attributable measures in general the relationship between the risk factor being considered for removal and the disease must be assumed to be causal, rather than just statistically associated (Walter, 1976).

If confounding variables are present then estimating the PAF without accounting for them can provide substantially different estimates, as shown by Whittemore (1983). Miettinen (1974) extends the PAF (1.3) to adjust for confounding by incorporating the adjusted relative risks. Whittemore (1982, 1983) implements similar extensions, without reference to Miettinen (1974),

for the PAF (1.2) for case-control studies as follows to adjust for confounding:

$$PAF = \frac{P(D^+) - \sum_k P(C_k)P(D^+|E^-, C_k)}{P(D^+)}, \tag{1.5}$$

where C_k is the fraction of all subjects in the kth stratum of the confounding factor. If the confounder C has only one level then (1.5) collapses to (1.2). Whittemore (1982, 1983) derives the maximum likelihood estimate for the PAF (and its corresponding standard error) based on 2×2 tables representing each stratum of the confounder. Deriving the maximum likelihood estimate in this way limits one to datasets with only a few strata. No discussion of regression models for estimation, which would also allow for the addition of more than one confounding variable, was included in her papers.

Later Bruzzi et al. (1985) provided a more generalised multivariate setting for estimation of the PAF from case-control data, which takes into consideration multiple levels of exposure and multiple confounding variables. They define this extension of the PAF as follows:

$$PAF = \sum_{i=1}^{k} pd_i \times \frac{r_i - 1}{r_i}, \tag{1.6}$$

where pd_i is the proportion of cases in the ith exposure level. Bruzzi et al. (1985) make use of generalised linear regression models (such as the logistic regression model) to allow for easier estimation of the relative risk within each stratum, as well as the opportunity to incorporate important interactions between factors. However, as Bruzzi et al. (1985) use the model to provide estimates of the relative risk only and not the covariate distribution, the estimate provided for the PAF is not the maximum likelihood estimate based on the model (Greenland and Drescher, 1993). Greenland and Drescher (1993) derive the maximum likelihood estimate for the adjusted PAF (1.6) based on the logistic regression model for both cohort and case-control studies along with their corresponding variance estimators.

So far when estimating the PAF we have only discussed the complete removal of a risk

factor from the population. However, rather than complete absence of exposure to the risk factor one may have a specified target or reference value which they wish to achieve. For example, in a population where a certain proportion of the individuals smoke it maybe very difficult to implement an intervention where all subjects stop smoking. In practice decreasing the proportion of smokers to say 5% of the population maybe more realistic. Morgenstern and Bursic (1982) extend the adjusted PAF (1.5) to determine the proportion of cases that can be attributed to the risk factor of interest under intervention programs of varying success.

The PAF has been praised as "important to the public health administrator" (Lilienfeld, 1973), due to its ability to assess the impact of a risk factor on the disease. It is also a measure which is easily understood by a non-specialist audience. In comparison, the population attributable risk proposed by MacMahon and Pugh (1970) is seldom used. This measure was first defined in early work by MacMahon et al. (1960), where it went by the name "attributable community risk" and was defined as the following difference in population parameters:

$$PAR = P(D^+) - P(D^+|E^-). \tag{1.7}$$

MacMahon and Pugh (1970) later went on to rename this measure as the population attributable risk. It should be noted that the PAR is simply the numerator of the PAF given by (1.2), and that the PAR only takes on values between -1 and 1. An alternative way of expressing the PAR is

$$PAR = e(p - q), \tag{1.8}$$

where p, q and e are population parameters given by $p = P(D^+|E^+)$, $q = P(D^+|E^-)$ and $e = P(E^+)$. Since $P(D^+) = P(D^+|E^+)P(E^+) + P(D^+|E^-)P(E^-) = pe + q(1 - e)$ by the Law of Total Probability, (1.8) follows immediately from (1.7). The PAR can be interpreted as the reduction in the risk of disease that could occur if exposure to the risk factor were removed.

MacMahon and Trichopoulos (1996) also extend the PAR (and PAF) to account for rates

as follows:

$$PARate = I_T - I_O, \tag{1.9}$$

where I_O and I_T are the disease incidence rates (i.e. the expected number of new cases in a reference time period) among the unexposed and total population respectively. To avoid confusion, throughout this thesis we choose to call the PAR proposed for rates the population attributable rate (PARate) and the PAF proposed for rates the population attributable rate fraction (PARF). No discussion of extending the PAR (or PARate) to account for confounding was provided in MacMahon et al. (1960), MacMahon and Pugh (1970) or MacMahon and Trichopoulos (1996). Newson (2013), the only author to provide confidence intervals for the PAR, estimates an adjusted version of the PAR. However, his paper lacks a clear explanation of how this estimate is calculated.

In order to provide estimates for both the PAR and the PAF it is important to note that we require either *a-priori* knowledge about the probability of exposure and disease in the population, or the ability to estimate this information from the data. Whether or not these probabilities for the population can be estimated from the data, depends largely on the study design. If the study is cross-sectional, and the sample assumed representative of the population, then estimation of the population probability of exposure and disease from the data may be adequate. When a case-control study has been carried out, Levin (1953) suggests estimating the population exposure from the control group, under the assumption that it is representative of the general population not having the disease. However, direct estimation of the population disease prevalence from the data is not possible due to the probability of disease being fixed by design. Cohort studies suffer a similar problem as the probability of exposure is fixed by design. In these situations an estimate of the population parameters must be provided, possibly from a previous study. Alternatively, a Bayesian approach can be adopted where an informative prior is assigned to the population parameters based on expert knowledge. The informative prior could be based on a previous study, allowing the uncertainty in the estimate from the previous study

	Diseased		
Exposed	D^+	D^-	Total
E^+	x_{11}	x_{12}	$x_{11} + x_{12}$
E^-	x_{21}	x_{22}	$x_{21} + x_{22}$
Total	$x_{11} + x_{21}$	$x_{12} + x_{22}$	n

Table 1.1: An example of a general 2×2 contingency table, where x_{11}, x_{12}, x_{21} and x_{22} represent the observed counts and n the total sample size

to be incorporated into the analysis. A Bayesian analysis was explored by Graham (2000) to estimate the adjusted PAF for a cohort study.

1.2 Epidemiological study designs

Epidemiological studies are often observational by nature, which can give rise to various limitations regarding causation, sources of variation and identifiability. Some of the most common epidemiological studies concerned with discrete data are the cross-sectional, case-control and cohort designs. In their simplest form where only a single dichotomous risk factor is being considered, data are often presented in a 2×2 table similar to Table 1.1. To provide estimates for the PAR and PAF a model must be defined based on the way in which the data is collected. The focus of this section is on defining the underlying statistical models associated with each of the study designs mentioned, and specifically when the data is in the form of a 2×2 table. Extensions of this data structure can be made to account for multiple risk factors, stratification and clustering, but we defer detailed discussions of these extensions until Chapter 4.

1.2.1 Cross-sectional studies

A cross-sectional study is conducted by taking a random and representative sample, of size n, from the population at a single point in time. As Rothman (2002) puts it, studies of this type are essentially a "snapshot" of the exposure and disease status of that population at a particular

time. The main advantage of the cross-sectional study is that it allows for estimation of the population prevalence of disease and exposure, both of which are required for estimation of the PAR and PAF.

Each of the four different exposure-disease combinations in Table 1.1 have a certain probability of occurring in the population. Given that we have not sampled the entire population, these probabilities are unknown. We denote these unknown probabilities by π_{11}, π_{12}, π_{21} and π_{22}. These probabilities can then be expressed in the following way:

$$\begin{aligned} \pi_{11} &= pe & \pi_{12} &= (1-p)e \\ \pi_{21} &= q(1-e) & \pi_{22} &= (1-q)(1-e), \end{aligned} \tag{1.10}$$

where we recall that p, q and e are population parameters given by $p = P(D^+|E^+)$, $q = P(D^+|E^-)$ and $e = P(E^+)$. Since n is fixed for cross-sectional studies, the appropriate statistical model is

$$(X_{11}, X_{12}, X_{21}, X_{22}) \sim \text{Multinomial}(n, \pi), \tag{1.11}$$

where π is the vector of probabilities $(\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22})$ and X_{ij} the random variable representing the possible number of observations in the ith row and jth column of the table.

1.2.2 Case-control studies

Case-control studies have received particular attention in the literature surrounding attributable measures. The literature search performed by Uter and Pfahlberg (2001) found 49.1% of the papers relating to attributable measures were from case-control studies. Case-control studies prove particularly useful when the disease of interest is rare. This is because rather than taking a random sample from the population, which could result in very few subjects with the disease, it is instead decided in advance how many diseased (and non-diseased) subjects should be sampled. That is, if we let $x_{11} + x_{21}$ and $x_{12} + x_{22}$ from Table 1.1 equal n_1 and n_2 respectively, then n_1

and n_2 are considered fixed under the case-control design. If we again let the random variable X_{ij} represent the possible number of observations in the ith row and jth column of the table, then the appropriate statistical model is the product of the independent binomial distributions

$$X_{11} \sim \text{Binomial}(n_1, \phi_1) \quad \text{and} \quad X_{12} \sim \text{Binomial}(n_2, \phi_2), \tag{1.12}$$

where $\phi_1 = P(E^+|D^+)$ and $\phi_2 = P(E^+|D^-)$. An obvious disadvantage of this approach is that an estimate of the population disease prevalence and exposure based on the data will be inflated, due to there being a larger proportion of those diseased in the sample than would be seen in the population. However, the population proportions of those exposed and not exposed are often estimated from the control group, under the assumption that it is representative of the population which is not diseased.

1.2.3 Cohort studies

Cohort studies measure the occurrence of disease within a group of subjects which have been followed over a period of time. Typically each of these groups are made up of subjects sharing a similar characteristic, such as being exposed or not exposed to a certain risk factor. The advantage of following cohorts over a period of time is that it allows for the calculation of the incidence, that is the number or rate of new cases of disease that occur within a period of time. Incidence calculations can not be carried out for case-control and cross-sectional studies.

When implementing a cohort study the number of subjects in each cohort may be decided in advance. That is, if we let $x_{11}+x_{12}$ and $x_{21}+x_{22}$ from Table 1.1 equal m_1 and m_2 respectively, then m_1 and m_2 are fixed under the cohort design. It follows that the appropriate statistical model is the product of the following binomial distributions

$$X_{11} \sim \text{Binomial}(m_1, p) \quad \text{and} \quad X_{21} \sim \text{Binomial}(m_2, q), \tag{1.13}$$

recalling that $p = P(D+|E+)$ and $q = P(D+|E-)$. Having fixed the row totals of the 2×2 table, the probability of exposure (and non-exposure) in the population cannot be estimated from the data. Therefore, it is not possible to estimate the PAR and PAF in this situation without additional information.

1.3 Uncertainty for population attributable measures

When applying the concept of a population attributable measure to epidemiological data it is desirable to provide not only a point estimate, but also a corresponding measure of uncertainty such as a confidence or credible interval. Much effort has gone into quantifying the uncertainty for the PAF and its generalisations. Walter (1975, 1976) derived the asymptotic variance for the maximum likelihood estimate of the PAF (1.1) under cross-sectional, case-control and prospective cohort studies, making use of the $\ln(1 - \text{PAF})$ transformation. Simplifications to Walters derivations were made by Fleiss (1979) for the cross-sectional study by re-parameterizing the PAF. Leung and Kupper (1981) later provided asymmetric confidence intervals for the PAF (1.1), for cross-sectional, case-control and prospective cohort studies, whilst making use of the logit transformation. Their simulations showed their intervals were narrower and closer to nominal coverage than those proposed by Walter (1976). An extensive comparison of all three authors' confidence intervals was carried out by Lui (2001), which allowed for more detailed recommendations, and revealed that Leung and Kupper (1981) intervals can be inaccurate when the relative risk equals 1. Denman and Schlesselman (1983) later provided a variance estimate for the PAF from case-control data in the form of a $2 \times k$ table.

This early theoretical work was then extended to provide approximations of the variance for the adjusted PAF. Whittemore (1982, 1983) provides an asymptotically unbiased estimate of the large-sample standard error for the adjusted PAF (1.5) calculated from case-control data. Comparing her log and logit transformed confidence intervals to that of the simple symmetric interval through simulation, Whittemore (1983) suggests using the logit transform for values

of PAF (1.5) in the range 0.21 - 0.79 and the simple symmetric interval outside this range. Greenland (1987) uses the Mantel-Haenszel estimator of the adjusted relative risk to derive a maximum likelihood estimate for the adjusted PAF (1.3) that is suitable for sparse data. He then derives a formula for the variance of this estimate using the delta method (discussed in Chapter 2, Section 2.2.1).

Benichou and Gail (1990) also make use of the delta method to provide a variance formula for the adjusted PAF (1.6). Their formulae are particularly complex, so the non-parametric bootstrap approach was adopted by Kooperberg and Petitti (1991) to allow for simpler estimation of the variance. An extensive review of interval estimation for the adjusted PAF for case-control studies is provided by Benichou (1991). For cross-sectional studies Lehnert-Batar et al. (2006) provide a comparison of the bootstrap, Bayesian bootstrap and jackknife methods for variance estimation of the adjusted PAF (1.6). This paper is one of only two to apply a Bayesian approach, although the Bayesian bootstrap is not what one would consider a fully Bayesian method: a posterior distribution is generated but no prior information is required (Rubin, 1981). A jackknife re-sampling approach was also utilized by Wagner et al. (2001) to approximate the variance of the adjusted PAF (1.6) for complex survey designs.

Each of the authors mentioned focus solely on the PAF or its generalisation. Only one author, Newson (2013), attempts to quantify the uncertainty for the PAR. His methodology is based on fitting a logistic regression model to the data and using the covariance matrix of the parameter estimates. However, he does not take into consideration all possible sources of uncertainty, and does not clearly address the underlying study design. This leaves a large gap in the current body of knowledge and a need for statistical methods to provide the uncertainty surrounding the PAR. Such methods need to be developed not only for the common epidemiological studies mentioned here, but also for more complex designs. It is also found that most studies encountered in the literature adopt the standard Frequentist approach to estimation. Given that case-control and cohort studies require population estimates for the probability of disease or exposure, which we sometimes do not have, it would make sense to take the Bayesian

approach and provide this information in the form of a prior distribution. The Bayesian framework also allows for a more natural extension to complex designs than the standard Frequentist approach.

1.4 Thesis aims and structure

The overarching aim of this thesis is to develop methods to quantify the uncertainty for population attributable measures, taking into consideration all sources of uncertainty and the underlying study design. Our focus lies predominately with Bayesian methodologies due to the need to incorporate external information about population parameters, and in some cases measurements concerning the accuracy of diagnostic tests. As population attributable measures are of most relevance to the fields of epidemiology and public health, we concentrate our efforts on the most commonly used study and experimental designs in those fields. We also select real-world data from the epidemiological literature in order to illustrate the performance of our proposed methodologies. To achieve our aim this work is broken down into three main objectives:

1. To develop and compare Frequentist and Bayesian methodologies for estimating the PAR from a cross-sectional study with a single dichotomous risk factor.

2. To develop Bayesian methodologies to estimate the PAR and PAF for situations where the statistical model is under-identified (e.g. case-control and cohort studies).

3. To extend the Bayesian methodologies developed to account for more complex designs and to estimate attributable rates.

Chapter 2 of this thesis addresses the first of these objectives, beginning with a discussion of a study of leptospirosis in New Zealand (Dreyfus et al., 2014) which identifies the lack of a standard method for estimating the confidence interval for the PAR. This knowledge gap led us to applying standard Frequentist approaches for variance estimation and confidence interval

construction such as the delta, bootstrap and jackknife methods, which we discuss in detail in Sections 2.2.1-2.2.3. Due to a Bayesian approach being necessary for situations where the probability of disease or exposure may not be accurately estimated from the data (e.g. case-control and cohort studies), a Bayesian approach is also adopted for estimating the PAR and its credible interval from a cross-sectional study (Chapter 2, Section 2.3). Each of these methods is then compared through simulation based on their Frequentist properties (Section 2.4). With the Bayesian method proving superior, additional exploration into the effect of the prior on the performance of the Bayesian method is also carried out (see Section 2.5).

In many epidemiological studies imperfect diagnostic tests are used to determine the disease or exposure status of an individual. The leptospirosis study is an example of this as the imperfect Microscopic Agglutination Test (MAT) was used to determine the exposure status of each individual. Extending the analysis performed in Chapter 2 for the cross-sectional study, to incorporate the uncertainty associated with an imperfect diagnostic test, results in a model which is under-identified, which is the focus of Chapter 3. Section 3.1 begins with a discussion of the meaning of identification and the consequences associated with using a model which is under-identified.

For the leptospirosis example we compare the performance of several well-known MCMC samplers (described in Section 3.2.3) with the novel importance sampling approach for under-identified models (described in Section 3.2.5) proposed by Gustafson (2015). Although simulation showed Gustafson's approach to be vastly superior in terms of efficiency and effective sample size (see Section 3.2.6), implementation of this approach for high dimensional problems can prove challenging due to the need for an appropriate "transparent parameterisation". We therefore, develop several novel adaptations of the Metropolis-Hastings random walk approach, described in Section 3.2.3.5, which encourage the sampler to take steps along the posterior ridge formed by the under-identified model. These approaches have the advantage that they do not require transparent re-parameterisation. We then show that these methods can outperform standard MCMC samplers used on under-identified models when the sample size is large (Sec-

tion 3.2.7). Case-control and cohort studies can also be considered as under-identified models due to the need to specify $P(D^+)$ or $P(E^+)$ in advance. We conclude Chapter 3 by outlining a Bayesian approach for estimating the PAR, PAF and their credible intervals for these studies (see Section 3.3 and 3.4).

Chapter 4 addresses the final objective of this thesis by extending the approaches developed for estimating the PAR, PAF and their credible intervals over Chapters 2 and 3 in several ways. We begin with a simple extension from the 2×2 table studied in previous chapters by exploring a dataset concerned with low birth weights in newborn babies (discussed in Section 4.1), which can be expressed as a $2 \times 2 \times 3$ table. This dataset was explored by Newson (2013), the only author found to provide a methodology for confidence interval estimation for the PAR, to demonstrate his approach. Adopting a similar logistic regression approach to Newson (2013), although under the Bayesian framework (see Section 4.6.1 for discussion), we compare through simulation the Frequentist coverage achieved when the joint distribution of exposure and the covariate race (i.e. (E, x), where x represents the covariates not considered for removal) is considered as either fixed or variable (Section 4.7). We show that Newson's approach does not take into consideration the uncertainty in (E, x), which results in intervals with less than nominal coverage for large PAR (Section 4.7.3). It is also shown that our Bayesian approach where (E, x) is represented by a Dirichlet posterior distribution results in superior coverage, especially when PAR is large.

In Chapter 4 we also formally define the population attributable rate and population attributable rate fraction, both unadjusted and adjusted for confounding (see Section 4.4). We then use these measures to describe the change in the rate of mastitis per year in cows due to the use of monensin treatment (see Section 4.2 for a description of the experiment which results in the mastitis dataset). The mastitis dataset provides additional extensions to our previous work due it having been collected as part of an experiment, rather than an observational study. Furthermore, the model and the way in which the joint distribution of (E, x) is specified must be extended to take into consideration the effect of cows clustering into herds. We explore how the choice of model, and the way of specifying the joint distribution of (E, x), affects the Frequentist

coverage of the credible intervals for PARate and PARF through three simulations.

The first of these simulations assumes no clustering in the rate of disease and covariates. A Poisson regression model (described in Section 4.6.2) is used to estimate the rate of disease with fixed weights (i.e. $1/n$) for the joint distribution of (E, x) compared to using variable weights assigned via Bayesian bootstrap re-sampling (described in Section 4.5). We show that describing (E, x) using Bayesian bootstrap re-sampling results in improved coverage of the PARate intervals (see Section 4.8.1.2). The second simulation considers clustering in the covariates only and the Bayesian bootstrap is extended to take into consideration the effect of clustering (see Section 4.5.1 for cluster Bayesian bootstrap description). Use of the cluster Bayesian bootstrap to describe the joint distribution (E, x) provides greatly superior coverage for the PARate in comparison to using the standard Bayesian bootstrap or fixed values, especially as the number of clusters increases (see Section 4.8.2.1). The final simulation considers clustering in both the rate of disease and the covariates. A Poisson mixed effects model, with herd treated as a random effect, is adopted and the Frequentist coverage of the credible intervals for the PARate and PARF compared for different numbers of clusters (see Section 4.8.3).

We conclude this thesis with a summary of novel results achieved in Chapter 5 and outline the proposed avenues for future research. Furthermore, all R code associated with the methods developed in this thesis can be found at the authors Github page (`https://github.com/spirikahu`).

Chapter 2

A comparison of Frequentist and Bayesian intervals for cross-sectional studies[1]

Frequentist and Bayesian analyses are two types of statistical inference which have been the subject of much debate over the last century. The fundamental difference between these methods of inference is that in Bayesian statistics the unknown parameters of interest (θ), conditional on the observed data (x), are described by a probability distribution known as the posterior, $p(\theta|x)$. Conversely, in Frequentist analysis the unknown parameters are treated as fixed unknowns. The posterior incorporates information from the likelihood for the data (similarly to Frequentist analysis), but any prior information experts may have can also be incorporated in the form of a prior distribution, $p(\theta)$. Bayesian and Frequentist inference also differ in terms of their statistical interpretation. Gelman et al. (2004) suggest that the Bayesian way of thinking is actually a more common-sense approach to statistical interpretation. Perhaps the most

[1]This chapter was the basis of the published work: Pirikahu, S., Jones, G., Hazelton, M. L., and Heuer, C. (2016). Bayesian methods of confidence interval construction for the population attributable risk from cross-sectional studies. *Stat. Med.*, 35(18):3117-3130.

important difference in interpretation that concerns us is that of the Bayesian credible interval, which is the Bayesian equivalent of a confidence interval. If we provide a 95% Bayesian credible interval for a measure, say θ, we interpret this credible interval by saying the probability that θ lies within this interval is 0.95. Now say a Frequentist provided a 95% confidence interval for θ. When interpreting this interval all we can say is that the true value of θ lies in this interval in 95% of repeated samples. Although the interpretations are different, in order to compare Bayesian and Frequentist intervals researchers often simply explore the Frequentist properties (e.g. coverage) of the Bayesian interval through simulation. In this chapter we make use of both Frequentist and Bayesian approaches to quantify the uncertainty for the PAR. We then compare these methods via simulation and apply them to a real-world application.

2.1 A study of leptospirosis in New Zealand

As an illustrative example we consider the cross-sectional study performed by Dreyfus et al. (2014) concerning leptospirosis in New Zealand. In New Zealand human leptospirosis is a common occupational hazard for abattoir workers and livestock farmers (Thornley et al., 2002), with up to 81% of adult deer herds, 97% of adult beef cattle herds, and 97% of adult sheep flocks having seropositive animals (Dreyfus et al., 2014). Exposure to these infected animals or contact with their blood or urine (commonly through water contamination) can result in mild flu-like symptoms such as fever, headache, lethargy, nausea and photo-sensitivity. In rare cases more severe symptoms can present, resulting in hospitalisation. Vaccination of ruminant livestock against leptospirosis was introduced in the 1990s in an attempt to reduce the number of human cases. This intervention resulted in a reduction in the number of cases in the farming industry from 234 per 100,000 in 1990-1992 to 90 per 100,000 in 1996-1998 (Thornley et al., 2002). However, less than 10% of deer, sheep and beef farmers currently use the vaccination (Dreyfus et al., 2011). This presents serious concerns for New Zealand's public health.

The study by Dreyfus et al. (2014) investigates the association between exposure to two

	Diseased		
Exposed	D^+	D^-	Total
E^+	22	25	47
E^-	82	251	333
Total	104	276	380

Table 2.1: Data from the New Zealand leptospirosis study for sheep abattoirs performed by Dreyfus et al. (2014)

of the most commonly found serovars of *Leptospira* in New Zealand livestock: *Leptospira borgpetersenii* and *Leptospira interrogans*, and flu-like symptoms in abattoir workers. Their data for sheep abattoirs is presented in Table 2.1, where $E^+(E^-)$ represents exposure (or not) of an abattoir worker to at least one of the leptospira bacteria and $D^+(D^-)$ the presence of flu-like symptoms (or not). The authors wished to assess the population impact of removing or decreasing exposure to these *Leptospira* serovars by estimating the PAR, PAF and their confidence intervals. For this dataset the PAF is 10% (95% CI 2-16%) and the PAR 2.7%, but it is reported that "confidence intervals for the PAR could not be provided as a variance formula for PAR was not readily available in the literature" (Dreyfus et al., 2014). Therefore, what proceeds in the remainder of this chapter is the presentation of several standard Frequentist approaches for calculating a confidence interval of the PAR from a cross-sectional study. We also introduce a Bayesian approach for estimating the PAR and its corresponding credible interval, and then compare these approaches in terms of their Frequentist properties through simulations, based on the data provided in Table 2.1.

2.2 Frequentist methods for confidence interval calculation

There are a number of different Frequentist approaches that can be used to calculate an approximate variance for the estimate of a function f, of the parameter vector θ, $f(\theta)$, and provide confidence intervals. Three which we concern ourselves with include the delta, jackknife and

bootstrap methods, due to been commonly used approaches for estimating confidence intervals which were also adopted by authors estimating confidence intervals for the PAF (Greenland, 1987; Benichou and Gail, 1990; Kooperberg and Petitti, 1991; Lehnert-Batar et al., 2006). Given that we can provide an estimate of the true value of $f(\theta)$, i.e. $f(\hat{\theta})$, a confidence interval can be calculated with the following equation:

$$f(\hat{\theta}) \pm z_{\alpha/2} \times SE(f(\hat{\theta})), \tag{2.1}$$

where $z_{\alpha/2}$ is a critical value from the standard normal distribution for a given significance level α and $SE(.)$ is the standard error. An interval calculated in this way will be symmetric, but relies on the assumption that the sampling distribution of $f(\theta)$ is normal. For large samples the sampling distribution should be approximately normal (according to the Central Limit Theorem), but this assumption may not hold for small samples. If the likelihood surface of $f(\theta)$ is not symmetric, then an interval constructed in this way may provide less than nominal coverage. In this situation an asymmetric interval may be more appropriate and can be obtained by using either the percentile bootstrap approach or by applying transformations.

2.2.1 The delta method

The delta method is the oldest technique for calculating the standard error of complex statistical functions, predating the jackknife and bootstrap by over 150 years (Efron, 1990). The delta method approximates the moments of a function of random variables through the use of a Taylor series approximation. Taylor series are commonly used in the field of mathematics to approximate non-linear functions by polynomials. When used in statistical application though, lower order Taylor series approximations (e.g. first and second order) are of most interest. If $\hat{\theta}$ is the maximum likelihood estimate of the parameter vector θ and $g(\theta)$ a differentiable function,

then the first order Taylor series expansion of g about $\hat{\theta}$ is

$$g(\theta) \approx g(\hat{\theta}) + \frac{\partial g(\hat{\theta})}{\partial \theta}(\theta - \hat{\theta}). \tag{2.2}$$

Taking expectations on both sides of (2.2) results in an approximately unbiased estimate of the mean, assuming it exists, for $g(\theta)$

$$E[g(\hat{\theta})] \approx g(\theta). \tag{2.3}$$

Assuming the variance of the estimator $g(\hat{\theta})$ exists it can be approximated by

$$Var[g(\hat{\theta})] \approx \left(\frac{\partial}{\partial \theta} g(\hat{\theta})\right)^t Var(\hat{\theta}) \left(\frac{\partial}{\partial \theta} g(\hat{\theta})\right). \tag{2.4}$$

This approach can become challenging as the function being considered becomes increasingly complex, but it has the advantage that no re-sampling or simulation is required if an analytical solution for the variance can be reached.

To apply the delta method to cross-sectional data we first need to recall that the underlying statistical model (1.11) for this type of study is multinomial. Given the complex functional form of the probabilities in this model, maximisation in its current form is inherently difficult due to the dense covariance matrix for the maximum likelihood estimate. To simplify the algebra we make use of the multinomial-Poisson transform presented by Baker (1994). Let X be the vector of independent random variables $X_1, \ldots X_n$, where $X_i \sim \text{Poisson}(\lambda_i)$. The sum of these random variables will follow the Poisson distribution, $\sum_{i=1}^n X_i \sim \text{Poisson}(\sum_{i=1}^n \lambda_i)$, but the conditional distribution of X given $\sum_{i=1}^n X_i$ is:

$$P\left(X = x | \sum_{i=1}^n X_i = \sum_{i=1}^n x_i\right) = \frac{P\left(X = x \cap \sum_{i=1}^n X_i = \sum_{i=1}^n x_i\right)}{P(\sum_{i=1}^n X_i = \sum_{i=1}^n x_i)} \tag{2.5}$$

$$= \frac{(\sum_{i=1}^n x_i)!}{(\sum_{i=1}^n \lambda_i)^{\sum_{i=1}^n x_i} e^{-\sum_{i=1}^n \lambda_i}} \prod_{i=1}^n \frac{\lambda_i^{x_i} e^{-\lambda_i}}{x_i!} \tag{2.6}$$

$$= \frac{(\sum_{i=1}^n x_i)!}{\prod_{i=1}^n x_i!} \prod_{i=1}^n \left(\frac{\lambda_i}{\sum_{i=1}^n \lambda_i}\right)^{x_i} \tag{2.7}$$

which can be identified as the Multinomial$(n, \lambda_i / \sum_{i=1}^{n} \lambda_i)$ distribution. What has just been shown here is that the unconditional distribution of $(X_1, \ldots, X_n)$ can be factored into two independent distributions: a Poisson for $\sum_{i=1}^{n} X_i$ and a multinomial for X given $\sum_{i=1}^{n} X_i$. With this transformation we have the powerful result that any likelihood-based inferences about the parameters of a multinomial distribution are the same regardless of whether our random variable X is sampled from n independent Poisson distributions or from a single multinomial distribution. Therefore, we let x_{11}, x_{12}, x_{21} and x_{22} from Table 1.1 be independently Poisson-distributed with rates λ_1, λ_2, λ_3 and λ_4. The use of this alternative Poisson model not only simplifies the algebra required for the delta method, but also allows for both the cross-sectional study and studies where none of the totals from Table 1.1 are fixed by design to be considered together.

Under the Poisson model we can redefine the PAR as

$$\text{PAR} = \frac{\lambda_1 + \lambda_3}{\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4} - \frac{\lambda_3}{\lambda_3 + \lambda_4}, \tag{2.8}$$

which can be estimated by

$$\widehat{\text{PAR}} = \frac{x_{11} + x_{21}}{x_{11} + x_{12} + x_{21} + x_{22}} - \frac{x_{21}}{x_{21} + x_{22}}. \tag{2.9}$$

Let $\theta = (\lambda_1, \ldots, \lambda_4)^t$, and the MLE $\hat{\theta} = (x_{11}, x_{12}, x_{21}, x_{22})^t$. Since the components of $\hat{\theta}$ are independent Poisson-distributed variables, $\text{Var}[\hat{\theta}] = \text{diag}(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$ which can be approximated by $\text{diag}(x_{11}, x_{12}, x_{21}, x_{22})$. The partial derivatives, the components of $\partial g(\theta)/\partial\theta$, are given by

$$\begin{aligned}
\frac{\partial \text{PAR}}{\partial \lambda_1} &= \frac{\lambda_2 + \lambda_4}{n^2} \\
\frac{\partial \text{PAR}}{\partial \lambda_2} &= -\frac{\lambda_1 + \lambda_3}{n^2} \\
\frac{\partial \text{PAR}}{\partial \lambda_3} &= \frac{\lambda_3}{(\lambda_3 + \lambda_4)^2} - \frac{(\lambda_1 + \lambda_3)}{n^2} + \frac{1}{n} - \frac{1}{(\lambda_3 + \lambda_4)}
\end{aligned}$$

$$\frac{\partial \text{PAR}}{\partial \lambda_4} = \frac{\lambda_3}{(\lambda_3 + \lambda_4)^2} - \frac{(\lambda_1 + \lambda_3)}{n^2}.$$

Using (2.4) and applying a square-root gives an estimated standard error from which a symmetric confidence interval can be constructed using (2.1). Using the delta method to calculate a 95% confidence interval for the PAR (2.7%) from the leptospirosis data, given in Table 2.1, gives the interval $(0.75\%, 4.74\%)$, which has an interval width of 3.99%.

2.2.2 The jackknife method

Jackknife re-sampling is a non-parametric approach originally proposed by Quenouille (1949) to reduce the bias in serial correlated estimators. This is done by splitting the observed data into smaller sub-samples where some observations are excluded. At first Quenouille (1949) began by splitting the data into only two half-samples. He then later generalised this approach to allow for the splitting of the data into k groups of size $h = n/k$ (Quenouille, 1956). Much research has been focused on the special case of the jackknife where $k = n$ and $h = 1$, as it eliminates any arbitrariness in the formation of groups (Miller, 1974). This special case is often referred to as the "leave one out" approach, as each group (or jackknife sample) is formed by excluding a single observation from the original sample as follows:

$$x_{(i)} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n), \tag{2.10}$$

where $i = 1, 2, \ldots, n$. For each of these n jackknife samples an estimate of the statistic of interest, $\hat{\theta}_{(i)}$, can be calculated. Tukey (1958), responsible for the name "jackknife", was the first to propose the use of this technique for variance estimation after identifying that the jackknife pseudovalues defined by $n\hat{\theta}_{(.)} - (n-1)\hat{\theta}_{(i)}$, where $\hat{\theta}_{(.)} = \sum_{i=1}^{n} \hat{\theta}_{(i)}/n$ is the mean of the jackknife replications, can be treated as approximately independent and identically distributed (iid) in

many situations. The jackknife estimate of the variance is defined by

$$Var(\hat{\theta}) = \frac{n-1}{n}\sum_{i=1}^{n}(\hat{\theta}_{(i)} - \hat{\theta}_{(.)})^2. \tag{2.11}$$

The beauty of this method for approximating the variance is that it can be applied to any statistic regardless of its mathematical complexity (Efron and Tibshirani, 1993).

To estimate the jackknife variance for the PAR, for each of the n jackknife samples we calculate the ith jackknife replicate of the PAR, $\widehat{\text{PAR}}_{(i)}$. In the case where the data are in the form of a 2×2 table, due to the inherent structure of the table the jackknife variance becomes a weighted average of four possible values:

$$\begin{aligned}\widehat{Var}_{2\times 2jack}(\widehat{\text{PAR}}) = \Bigg[\frac{n-1}{n}\Big(&x_{11}(\widehat{\text{PAR}}_{(x_{11})} - \widehat{\text{PAR}}_{(.)})^2 + x_{12}(\widehat{\text{PAR}}_{(x_{12})} - \widehat{\text{PAR}}_{(.)})^2 \\ &+ x_{21}(\widehat{\text{PAR}}_{(x_{21})} - \widehat{\text{PAR}}_{(.)})^2 + x_{22}(\widehat{\text{PAR}}_{(x_{22})} - \widehat{\text{PAR}}_{(.)})^2\Big)\Bigg]\end{aligned} \tag{2.12}$$

where $\widehat{\text{PAR}}_{(x_{11})}$ represents the estimator of the jackknife sample where x_{11} is replaced by $x_{11}-1$ and $\widehat{\text{PAR}}_{(.)} = \sum_{i=1}^{n}\widehat{\text{PAR}}_{(i)}/n$, is the mean of all the jackknife replicates. Taking the square root gives an estimated standard error from which a symmetric confidence interval is constructed as $\widehat{\text{PAR}} \pm z_{1-\alpha/2}\text{SE}(\widehat{\text{PAR}})$. Using the jackknife method to calculate a 95% confidence interval for the PAR (2.7%) from the leptospirosis data, given in Table 2.1, gives the interval $(0.74\%, 4.75\%)$, which has an interval width of 4.01%.

2.2.3 The bootstrap method

The bootstrap method was inspired by its predecessor the jackknife, and was first introduced by Efron (1979) as a computer-based method for determining the standard error of a statistic through re-sampling. The bootstrap approximates the sampling distribution of a statistic using repeated samples either from an assumed distribution (parametric), or from a given data set (non-parametric). Provided we have either a model or data set which is representative of the

true population, the bootstrap approach can be implemented by sampling N times (for large N) from the known distribution, or with replacement from the dataset, to provide N bootstrap samples of size n. Sub-sampling using repeated samples of size m, where m is less than n, has been considered by Politis (1998) and Politis et al. (1999). The variance of the bootstrap samples can be calculated using the formula

$$Var(\hat{\theta}) = \frac{1}{N-1}\sum_{b=1}^{N}[\hat{\theta}^*_{(b)} - \hat{\theta}^*_{(.)}]^2 \tag{2.13}$$

where $\hat{\theta}^*_{(b)}$ is our statistic of interest for the bth bootstrap sample and $\hat{\theta}^*_{(.)}$ the mean of the bootstrap replications. Note that the only difference between this variance formula (2.13) and the variance for the jackknife (2.11) is the constant factor. In fact the jackknife can be regarded as a linear approximation to the bootstrap, which Efron (1980) shows performs less well in most cases.

For cross-sectional data represented by a 2×2 table the probability of selection for each classification when sampled according to the non-parametric bootstrap is equal to its estimated value in the parametric model. Therefore, the non-parametric bootstrap for a 2×2 contingency table is the same as the parametric bootstrap. To perform the bootstrap in this case we begin by taking N bootstrap samples of size n from the model (1.11), with p_{11} replaced by $\hat{p}_{11} = x_{11}/n$, p_{12} replaced by $\hat{p}_{12} = x_{12}/n$, p_{21} replaced by $\hat{p}_{21} = x_{21}/n$ and p_{22} replaced by $\hat{p}_{22} = x_{22}/n$. For the bth bootstrap sample, where $b = 1, \ldots, N$, we can calculate the bth bootstrap replicate of the PAR using (2.9). A confidence interval can then be obtained either by using (2.1) and the bootstrap variance estimator (2.13), or by taking the 100αth and $100(1-\alpha)$th percentiles of the distribution for the bootstrap replicates of the PAR. The latter approach is known as the percentile bootstrap method and will provide an asymmetric interval. Many other bootstrap confidence intervals have been discussed in the literature. Some of the more well known intervals are the bias-corrected percentile and bootstrap-t interval (Efron, 1981), the bias-corrected, accelerated (BC_a) intervals (Efron, 1987) and approximate bootstrap confidence

(ABC) intervals (Diciccio and Efron, 1991). See Efron and Tibshirani (1993) for an accessible introduction to these intervals and others. Calculating a 95% confidence interval for the PAR (2.7%) from the leptospirosis data, Table 2.1, using a symmetric bootstrap interval and asymetric interval via the percentile bootstrap gives the intervals $(0.76\%, 4.73\%)$ and $(1.01\%, 4.90\%)$, of width 3.98% and 3.90% respectively.

2.2.4 Transformations

Each of the methods described above, excluding the percentile bootstrap, will result in a symmetric confidence interval. If the underlying sampling distribution is not symmetric this may result in the coverage of the confidence interval being significantly less than nominal. An appropriate transformation, in which the transformed variable has a more symmetrical sampling distribution, can be used to produce an asymmetric confidence interval with better coverage. Calculation of a confidence interval for a transformation of the parameter vector, $f(\theta)$, is given by (2.1). A confidence interval for θ can then be calculated by taking the inverse of the lower and upper bounds of the interval produced, that is $(f^{-1}[f(\hat{\theta}) - z_{\alpha/2}SE(f(\hat{\theta}))], f^{-1}[f(\hat{\theta}) + z_{\alpha/2}SE(f(\hat{\theta}))])$.

As the PAR is in the range of $[-1, 1]$ the hyperbolic arctangent transform, also known as the Fisher z-transformation, would be an appropriate choice and was applied by Newson (2013). Given that the arctanh function takes values in the interval $[-1, 1]$ and maps them onto the interval $[-\infty, \infty]$, it is often associated with the correlation coefficient (Silverman and Young, 1987; Young, 1988). It can be seen in Figure 2.1 that this transformation has a larger effect on values closer to -1 and 1, whilst only having a small effect on those values close to 0. In practice large values of PAR are rare, so it is unclear whether this transformation will provide much benefit.

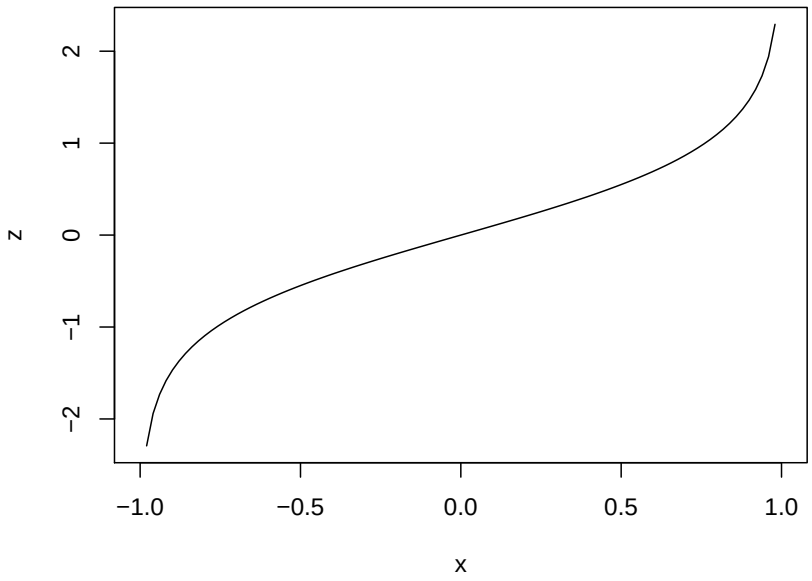

Figure 2.1: Fisher z-transformation curve

2.3 Bayesian inference for population attributable measures

Bayesian statistics and the use of Markov chain Monte Carlo (MCMC) methods have become increasingly popular with advancements in computational technology (Christensen et al., 2010). As mentioned earlier, Bayesian inference deviates from the standard Frequentist inference as parameters of interest are described in terms of probability distributions, rather than as fixed but unknown values. Under this framework it is also possible to incorporate any *a priori* knowledge available about the parameters through a prior distribution. Given a prior distribution on the parameter set, θ, and some observed data, x, Bayes theorem can be applied in order to update our prior beliefs about θ to posterior beliefs. In other words, the posterior distribution is simply the conditional distribution of our parameters given the data as expressed by

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{\int p(\theta)p(x|\theta)d\theta}. \tag{2.14}$$

In simple situations it is possible to determine the posterior distribution analytically. However, as the problem becomes higher dimensional, direct calculation of the posterior becomes too difficult. Specifically, the integration step required to provide the normalising constant in the denominator of (2.14) becomes increasingly challenging to calculate in higher dimensions. To avoid having to explicitly evaluate the normalising constant, Markov Chain Monte Carlo methods can be adopted and will be discussed in more detail in Chapter 3.

For cross-sectional data where $\theta = (\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22})$ and $x = (x_{11}, x_{12}, x_{21}, x_{22})$ the probability mass function for our model (1.11) is given by

$$p(x|\theta) = \frac{n!}{x_{11}!x_{12}!x_{21}!x_{22}!}\pi_{11}^{x_{11}}\pi_{12}^{x_{12}}\pi_{21}^{x_{21}}\pi_{22}^{x_{22}}, \tag{2.15}$$

where n is the sample size. For convenience, we make use of the fact that the conjugate prior for a multinomial distribution is the Dirichlet distribution (Gelman et al., 2004). We select our prior on θ, $p(\theta)$, to be Dirichlet$(1, 1, 1, 1)$ which is often considered a standard reference prior since $p(\theta)$ is constant. The posterior density is $p(\theta|x) \propto f(x|\theta)p(\theta)$, and due to the conjugacy relationship we find

$$\theta|x \sim \text{Dirichlet}(x_{11} + 1, x_{12} + 1, x_{21} + 1, x_{22} + 1). \tag{2.16}$$

As we are able to represent the posterior analytically, rather than resorting to MCMC simulation which can be computationally expensive, we are instead able to sample directly from this known posterior. Using the `rdirichlet` function, from the package `MCMCpack` (Martin et al., 2011) in the statistical software R (R Core Team, 2017), we can generate N independent samples (we use $N = 10,000$), from the known posterior at little computational expense. The PAR can then be calculated for each sample to obtain the posterior distribution of the PAR directly. The credible interval can then be obtained by taking the 100αth and $100(1-\alpha)$th percentiles which represent the lower and upper bounds of the interval respectively. It would also be easy to obtain the posterior distribution of PAF simultaneously, if required.

p	0.01	0.05	0.1	0.2	0.3	0.35	0.4	0.45	0.5
q	0.001	0.05	0.1	0.2	0.3	0.35	0.4	0.45	0.5
e	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9

Table 2.2: Parameter input values for Simulation 1

2.4 Simulation study

In order to explore the performance of the Frequentist and Bayesian methods described in the preceding sections, we first select fixed known parameters p, q, e and sample size n. For these chosen p, q, e and n parameters we then generate $10,000$ contingency tables from the model (1.11). For each of the contingency tables a 95% confidence interval is calculated for the Frequentist methods, the Fisher z-transform of the delta and jackknife methods, and the Bayesian approach. For simplicity we use the following notation to denote these methods: Bayes = Bayesian, Boot = Bootstrap, Delta = Delta, FDelta = Fisher z-transformed delta, Jack = jackknife, FJack = Fisher z-transformed jackknife, and PBoot = Percentile bootstrap. The performance of each of these methods is measured by the overall percent coverage achieved taking note of the interval length, because if two methods both achieve nominal coverage the shorter interval is preferable.

Simulation 1: This simulation is based on the New Zealand leptospirosis data in Table 2.1, using a fixed sample size $n = 380$. Table 2.2 provides a plausible range of p, q and e values, as suggested by an epidemiologist, for studies of this type. Each of the 729 possible combinations of the values given in Table 2.2 are used in this simulation to explore the parameter space of our problem. These parameter combinations result in PAR values in the range $[-0.45, 0.45]$.

For all methods some combinations of the input values in Table 2.2 result in the actual coverage being less than nominal. Overall Figure 2.2 shows that the Bayesian method performs better in terms of percent coverage, sometimes even over-covering, with significantly less variability in the coverage over the explored parameter space. It was seen that for the cases in

which the actual coverage was less than 90%, p and q were small (< 0.1) resulting in small expected counts (< 5) for x_{11} and x_{21}. These small counts occur mainly due to the restriction of the sample size, and may not be representative of a realistic study design. However, it is not impossible that a study may result in a table containing low counts. It was found that when coverage was less than nominal for all methods the average Bayesian interval length was largest, but when the nominal coverage was achieved there was little difference in interval length across all methods; see Figure 2.2.

Simulation 2: We hypothesise that having low counts at any position in the table could detrimentally affect the coverage of some methods. To investigate the effect of low expected values, x_{11} and x_{21} were set equal to 5 counts each, whilst n remained at 380. Then 100 sequential values of e were selected from the interval $[0.1, 0.9]$ and the remaining parameters p and q calculated using the following equations

$$p = \frac{x_{11}}{en} \tag{2.17}$$

$$q = \frac{x_{21}}{(1-e)n}. \tag{2.18}$$

Selecting parameter inputs in this way results in PAR values in the range $[-0.1, 0.1]$. PAR in this range are most likely in practice for typical situations where exposure has a small to moderate effect on the individual risk of getting a disease, or when clinical disease is moderately rare so cannot be reduced or increased substantially. Figure 2.3, which has been smoothed using a cubic spline, shows that the Bayesian method has greater than nominal coverage for all values, whilst all other methods suffer from less than nominal coverage. The over coverage seen for the Bayesian method is due to a shrinkage effect caused by the prior, which we discuss in more detail in Section 2.5. An obvious peak can also be seen at $e = 0.5$ for the methods where the confidence interval is symmetric by construction (i.e using $\pm z_{1-\alpha/2}$SE). We conjecture that this is related to the asymmetry in the sampling distribution. Figure 2.3 also shows there was

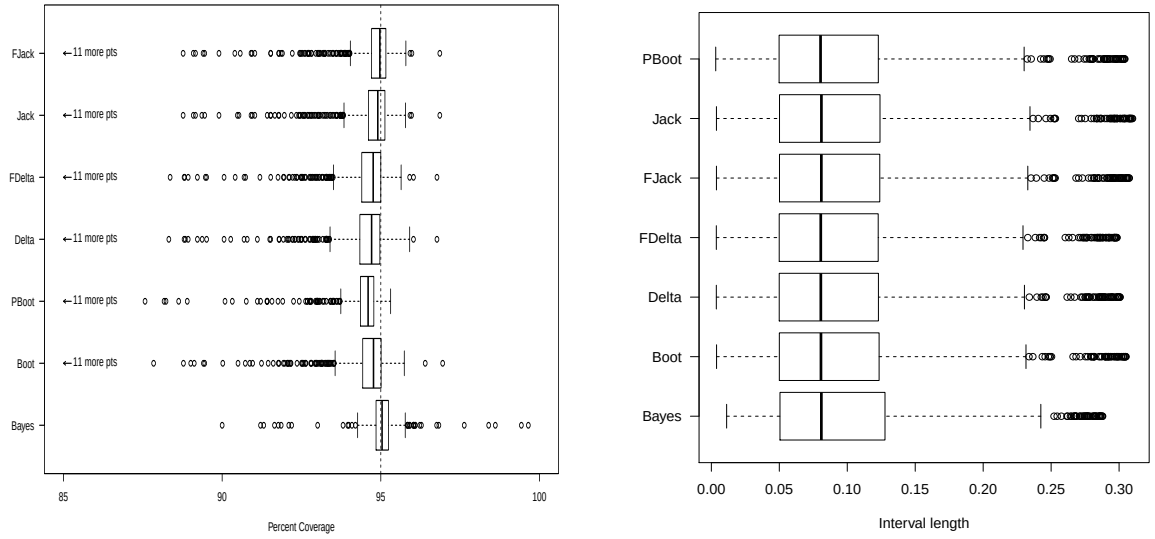

Figure 2.2: Percent coverage and interval length of Simulation 1 where Bayes = Bayesian method, Boot = bootstrap method, Delta = delta method, FDelta = Fisher transformed delta method, FJack = Fisher transformed jackknife method, Jack = jackknife method and PBoot = percentile bootstrap method. Note that 11 points for each Frequentist method have being omitted due to having a percent coverage less than 87%.

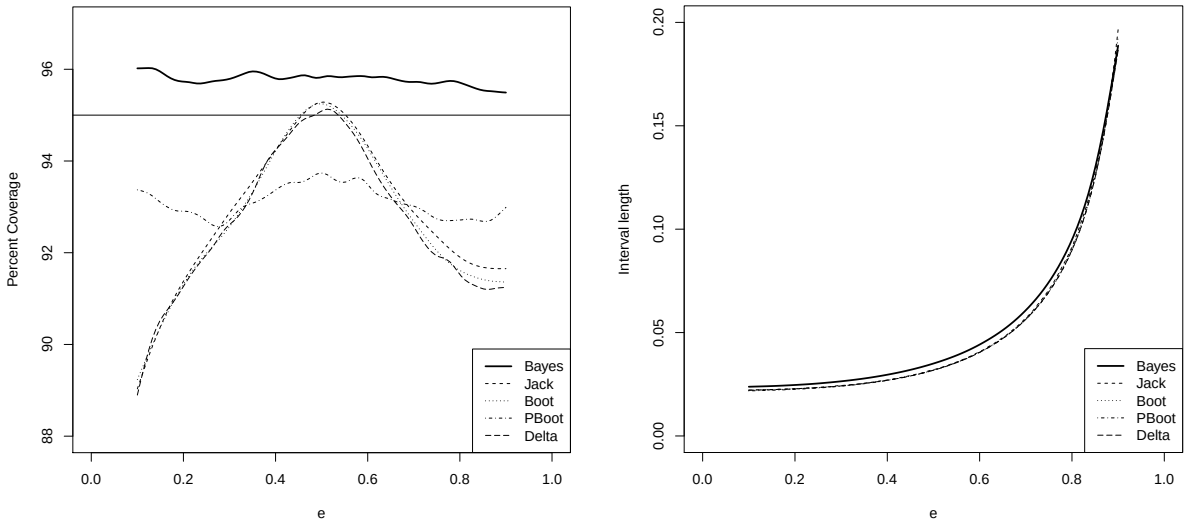

Figure 2.3: Percent coverage (left) and interval length (right) for e in the interval $[0.1, 0.9]$ for all untransformed methods with $n = 380$.

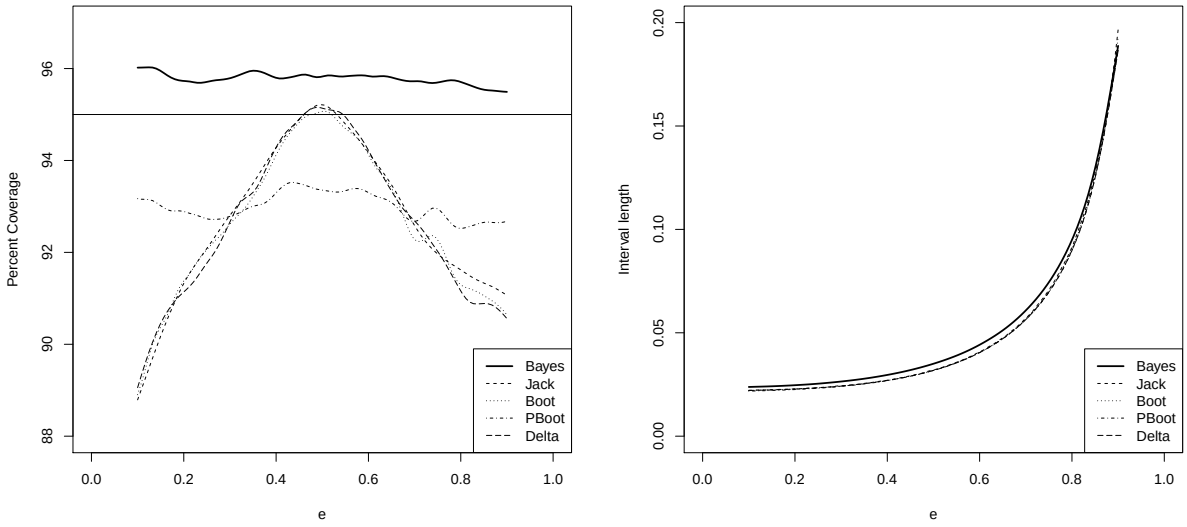

Figure 2.4: Percent coverage (left) and interval length (right) for e in the interval $[0.1, 0.9]$ for all untransformed methods with $n = 1000$.

very little difference in interval length across all methods. When the simulation was re-run with the sample size increased to $n = 1000$ (and all other variables remaining the same) there was very little change in the percent coverage and interval length (see Figure 2.4). The Fisher z-transformed methods have been omitted here due to their being almost identical to the standard delta and jackknife methods. This similarity is because the Fisher z-transformation only has a significant effect on the extreme values of PAR close to 1 or -1, and very little effect on those values close to zero.

Simulation 3: To determine how large the counts x_{11} and x_{21} must be for each method to achieve nominal coverage, a simulation was run with $n = 380$ and $e = 0.2$ (an area of lesser coverage as seen in Figure 2.3 and 2.4) and integer values of x_{11} and x_{21} in the interval $[1, 25]$. All other parameters were then calculated using equations (2.17-2.18).

The contour plots given in Figure 2.5, which have been smoothed using a two-dimensional spline, show that the delta method achieves only 90% coverage when both x_{11} and x_{21} reach

x_{11}	x_{21}	Bayes	Boot	PBoot	Delta	FDelta	Jack	FJack
5	5	0.0247	0.0228	0.0226	0.0228	0.0228	0.0230	0.0230
6	5	0.0265	0.0249	0.0246	0.0249	0.0249	0.0251	0.0251
7	5	0.0282	0.0268	0.0266	0.0268	0.0268	0.0269	0.0269
8	5	0.0298	0.0287	0.0284	0.0287	0.0287	0.0287	0.0287
9	5	0.0314	0.0303	0.0301	0.0303	0.0303	0.0305	0.0304
10	5	0.0329	0.0319	0.0316	0.0319	0.0319	0.0321	0.0320
5	6	0.0248	0.0229	0.0227	0.0229	0.0229	0.0230	0.0230
6	6	0.0267	0.0251	0.0248	0.0250	0.0250	0.0252	0.0252
7	6	0.0283	0.0269	0.0267	0.0268	0.0268	0.0270	0.0270
8	6	0.0300	0.0287	0.0284	0.0287	0.0287	0.0289	0.0289
9	6	0.0314	0.0304	0.0301	0.0304	0.0304	0.0305	0.0305
10	6	0.0329	0.0320	0.0317	0.0319	0.0319	0.0322	0.0322
5	7	0.0249	0.0229	0.0227	0.0230	0.0230	0.0231	0.0231
6	7	0.0267	0.0250	0.0247	0.0252	0.0252	0.0253	0.0253
7	7	0.0284	0.0271	0.0268	0.0270	0.0270	0.0271	0.0271
8	7	0.0301	0.0288	0.0285	0.0288	0.0288	0.0288	0.0288
9	7	0.0315	0.0304	0.0301	0.0304	0.0304	0.0305	0.0305
10	7	0.0329	0.0320	0.0318	0.0319	0.0319	0.0321	0.0321
5	8	0.0250	0.0230	0.0228	0.0232	0.0232	0.0232	0.0232
6	8	0.0268	0.0251	0.0249	0.0252	0.0252	0.0254	0.0254
7	8	0.0284	0.0271	0.0269	0.0271	0.0271	0.0271	0.0271
8	8	0.0302	0.0288	0.0285	0.0288	0.0288	0.0290	0.0290
9	8	0.0315	0.0304	0.0301	0.0304	0.0304	0.0305	0.0305
10	8	0.0330	0.0320	0.0318	0.0321	0.0321	0.0321	0.0321
5	9	0.0252	0.0233	0.0231	0.0232	0.0232	0.0233	0.0233
6	9	0.0269	0.0252	0.0250	0.0252	0.0252	0.0254	0.0254
7	9	0.0286	0.0271	0.0268	0.0272	0.0272	0.0273	0.0273
8	9	0.0302	0.0289	0.0286	0.0289	0.0289	0.0289	0.0289
9	9	0.0316	0.0304	0.0302	0.0305	0.0305	0.0307	0.0307
10	9	0.0330	0.0320	0.0318	0.0320	0.0320	0.0321	0.0321
5	10	0.0253	0.0234	0.0232	0.0234	0.0234	0.0235	0.0235
6	10	0.0270	0.0254	0.0252	0.0254	0.0253	0.0254	0.0254
7	10	0.0286	0.0272	0.0270	0.0272	0.0272	0.0273	0.0273
8	10	0.0302	0.0289	0.0287	0.0289	0.0289	0.0290	0.0290
9	10	0.0317	0.0305	0.0303	0.0305	0.0305	0.0307	0.0307
10	10	0.0332	0.0320	0.0318	0.0321	0.0321	0.0322	0.0322

Table 2.3: Average interval length of each method for small counts x_{11} and x_{21} in Simulation 3

10, but close to nominal coverage when x_{11} and x_{21} equal 20. This trend is similar for both the jackknife and bootstrap methods, so not all the contour plots are presented here. The contour plot for the percentile bootstrap method shows faster improvement in coverage than the delta method as the counts increase. The Bayesian method however, achieves nominal coverage for all values of x_{11} and x_{21} even those with counts less than 5. Inspection of the average interval lengths reveals slightly wider confidence intervals for the Bayesian method than those of the Frequentist methods, perhaps accounting for the superior coverage. As the counts increased though the interval lengths became more similar across all methods (see Table 2.3).

For the three simulations investigated the Bayesian approach outperforms all other methods in terms of percent coverage, except in a few cases. The Frequentist methods perform adequately if all the cell counts are large (≥ 20). This leads us to suspect that if an appropriate sample size determination is implemented in the experimental design phase, the Frequentist methods, in particular the delta method, could be a computationally cheaper alternative that enjoys similar coverage to that of the Bayesian approach. A variable sample size simulation (not presented here) was conducted using sample size determination, see Fleiss (1973) for a discussion about sample size determination. However, the Bayesian method was still superior in terms of coverage particularly when, as for very extreme values of e (> 0.7) and q (< 0.1), tables with at least one of the cell counts being small were still produced.

2.5 Alternative priors

The range of PAR values produced by the parameter values considered for the simulations performed previously, were those deemed by expert opinion to be most realistic in practice. There may however be particular situations where values closer to $+1$ or -1 could be obtained. Consider for example the case where a mass outbreak of water-borne disease has occurred, say due to contamination of a city's water supply. If the population is defined as all those individuals within the city, and a large proportion of those individuals consume the water from the city

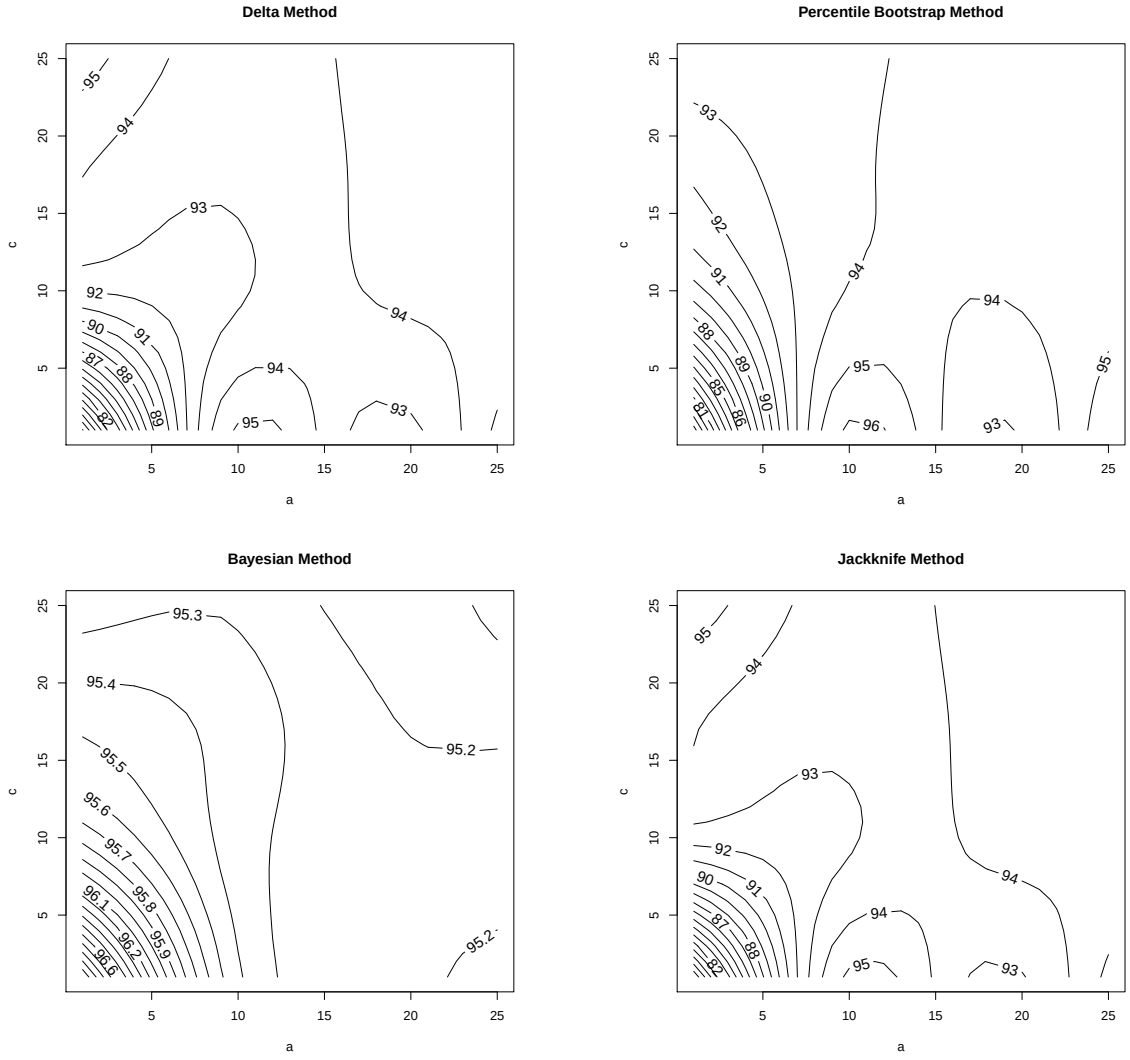

Figure 2.5: Contours of the percent coverage of all integer combinations of $a = x_{11}$ and $c = x_{21}$ in the interval $[1, 25]$ for fixed $e = 0.2$, for the delta, percentile bootstrap, jackknife and Bayesian methods.

	PAR	Bias (10^{-6})	Variance (10^{-4})	MSE (10^{-4})
MLE	0.902	9.65	1.0789	1.0788
Posterior Mean	0.880	-21266.81	0.8078	5.3305
Posterior Median	0.884	-17366.18	0.8710	3.8868

Table 2.4: Comparison of the MLE (delta method) with the mean and median estimates of the posterior for the Bayesian method with Dirichlet$(1, 1, 1, 1)$ prior. The true PAR for this case, with $n = 1000$, is 0.902, and the percent coverage is 94.6 and 83.5 for the delta and Bayesian methods respectively.

supply, the probability of being exposed to the contaminant, $e = P(E^+)$, will be large. In this situation those exposed to the contaminant will be at a much greater risk of becoming diseased, that is $p = P(D^+|E^+)$ will be large. Furthermore, those not exposed to the contaminant will be at a lower risk of becoming diseased, meaning that $q = P(D^+|E^-)$ will be small. In this situation the resulting PAR might be close to 1, and significantly larger than any of the values considered previously. Alternatively, we could consider the case where exposure is treated as protective, for example vaccination, which could result in a PAR value close to -1. To investigate the the performance of the Bayesian method in these extreme cases we simulate a toy example based on the outbreak case described above. We let $p = 0.95$, $e = 0.95$, $q = 0.001$ and $n = 1000$ to achieve a PAR value of 0.902. For this example the average percent coverage for the delta, Fisher z-transformed delta and Bayesian methods were 94.6%, 95.2% and 83.5% respectively with the corresponding interval lengths of 0.04, 0.04 and 0.08. This dramatic reduction in the coverage and increased interval width for the Bayesian method suggests that our chosen prior may be having a larger effect on our posterior than is desirable.

The Dirchlet$(1, 1, 1, 1)$ prior has been used in all the simulations performed so far, due to being a standard reference prior for the multinomial distribution. Figure 2.6 shows that on the PAR scale this prior, rather than producing a flat distribution, is bell-shaped with mean 0. The shape of this distribution leads to shrinkage of the point estimates towards 0, resulting in better coverage for those intervals when the true PAR is close to 0. This shrinkage effect can be seen in Table 2.4, which provides the bias, variance and mean square error (MSE) for our

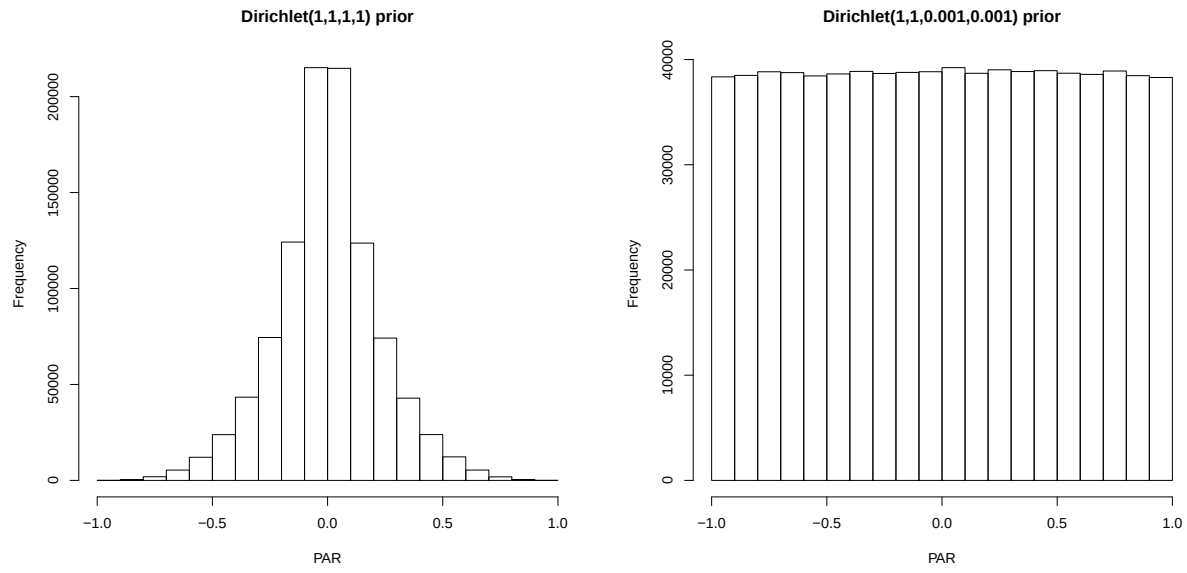

Figure 2.6: *Left:* The distribution of the Dirichlet(1,1,1,1) prior on the PAR scale. *Right:* The distribution of the Dirichlet$(1, 1, \epsilon, \epsilon)$ prior on the PAR scale, where $\epsilon = 0.001$.

toy example. The posterior mean and median were both considered as viable point estimates for the Bayesian method and were compared to the maximum likelihood estimate (MLE) given by the delta method. It can be clearly seen from Table 2.4 that the posterior mean (0.880) and median (0.884) of the PAR have been shrunk away from the true estimate (0.902). The greater the amount of shrinkage the more bias that is introduced, accompanied by a reduction in the variance following the usual bias-variance trade-off. An alternative non-informative prior derived according to $p(\theta) \propto \sqrt{|I(\theta)|}$, where $I(.)$ represents the Fisher information matrix, as proposed by Jeffreys (1945) (i.e. the Jeffreys prior) follows a Dirchlet$(0.5, 0.5, 0.5, 0.5)$ distribution for this particular example. Using the Jeffreys prior gave improved but still less than nominal coverage (see Figure 2.7).

For most practical purposes the introduction of bias may not be problematic as the PAR would be expected *a priori* to be moderate to small. However, for these extreme values of PAR close to $+1$ or -1 our current prior would not provide a satisfactory confidence interval. In situations where PAR might be expected to be extreme it would be useful to have a prior which

is uniform on the PAR scale. We have found that the prior distributions Dirichlet$(1, 1, \epsilon, \epsilon)$ for ϵ sufficiently small (we have used $\epsilon = 0.001$), Dirichlet$(1, \epsilon, \epsilon, 1)$ and Dirichlet$(\epsilon, 1, 1, \epsilon)$, give priors for PAR that are approximately equivalent to the Uniform[-1,1], Uniform[0,1] and Uniform[-1,0] respectively. Figure 2.6 shows the simulated distribution for the Dirichlet$(1, 1, \epsilon, \epsilon)$ with $\epsilon = 0.001$. Note that we let $\epsilon = 0.001$, rather than $\epsilon = 0$, as when $\epsilon = 0$ we no longer have a proper probability distribution (i.e. $\int f(x) = 1$, where $f(x)$ represents the probability density function) that can be sampled from. The limiting probability density function (pdf) for the PAR using a Dirichlet$(1, 1, \epsilon, \epsilon)$ as $\epsilon \to 0$ can be derived analytically, as follows.

Theorem 2.5.1 *If* $(X_{11}, X_{12}, X_{21}, X_{22}) \sim$ *Dirichlet*$(1, 1, \epsilon, \epsilon)$ *and* $Z = \frac{X_{11}+X_{12}}{X_{11}+X_{12}+X_{21}+X_{22}} - \frac{X_{21}}{X_{21}+X_{22}}$ *then as* $\epsilon \to 0$, $Z \xrightarrow{d}$ *Uniform*$(-1, 1)$

Proof For simplicity and due to dependence drop $X_{12} = 1 - X_{11} - X_{21} - X_{22}$ and denote the joint pdf of the prior distribution on the multinomial probabilities by

$$f_{X_{11},X_{21},X_{22}}(\pi_{11}, \pi_{21}, \pi_{22}) = \frac{\Gamma(2(1+\epsilon))}{\Gamma(\epsilon)\Gamma(\epsilon)} \pi_{21}^{\epsilon-1} \pi_{22}^{\epsilon-1}. \tag{2.19}$$

We wish to find the joint pdf of the transformed variables

$$Y_{11} = Z = X_{11} + X_{21} - (X_{21}/(X_{21} + (X_{22})), \quad Y_{21} = X_{21}, \quad Y_{22} = X_{22} \tag{2.20}$$

and hence the marginal distribution of Y_{11}. Let g be the function which maps $(\pi_{11}, \pi_{21}, \pi_{22})$ onto (z, π_{21}, π_{22}) by

$$z = \pi_{11} + \pi_{21} - (\pi_{21}/(\pi_{21} + \pi_{22})), \qquad \pi_{21} = \pi_{21}, \qquad \pi_{22} = \pi_{22}. \tag{2.21}$$

The inverse, g^{-1}, maps (z, π_{21}, π_{22}) onto $(\pi_{11}, \pi_{21}, \pi_{22})$ by

$$\pi_{11} = z - \pi_{21} + (\pi_{21}/(\pi_{21} + \pi_{22})), \qquad \pi_{21} = \pi_{21}, \qquad \pi_{22} = \pi_{22}. \tag{2.22}$$

The Jacobian, $J = \partial g^{-1}(\pi_{11}, \pi_{21}, \pi_{22})/\partial(z, \pi_{21}, \pi_{22})$, is then given by

$$|J| = \begin{vmatrix} 1 & \frac{1}{\pi_{21}+\pi_{22}} - \frac{\pi_{21}}{(\pi_{21}+\pi_{22})^2} - 1 & -\frac{\pi_{21}}{(\pi_{21}+\pi_{22})^2} \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{vmatrix} = 1. \tag{2.23}$$

It follows that the joint pdf of Y_{11}, Y_{21} and Y_{22} is given by (with support to be calculated)

$$\begin{aligned} f_{Y_{11},Y_{21},Y_{22}}(z, \pi_{21}, \pi_{22}) &= |J|\, f_{X_{11},X_{21},X_{22}}(g^{-1}_{\pi_{11}}(z, \pi_{21}, \pi_{22}), g^{-1}_{\pi_{21}}(z, \pi_{21}, \pi_{22}), g^{-1}_{\pi_{22}}(z, \pi_{21}, \pi_{22})) \\ &= \frac{\Gamma(2(1+\epsilon))}{\Gamma(\epsilon)\Gamma(\epsilon)} \pi_{21}^{\epsilon-1} \pi_{22}^{\epsilon-1}. \end{aligned} \tag{2.24}$$

In order to find the marginal distribution of z, we must integrate out π_{21} and π_{22}; that is

$$f_{Y_{11}}(z) = \int\int f_{Y_{11},Y_{21},Y_{22}}(\pi_{11}, \pi_{21}, \pi_{22})\, d\pi_{21}\, d\pi_{22}. \tag{2.25}$$

To determine the bounds on π_{21} and π_{22} it is convenient to consider $z \leq 0$; we can then use symmetry to infer the result for $z > 0$. Recall that $0 \leq \pi_{11} \leq 1$, where $\pi_{11} = z - \pi_{21} + (\pi_{21}/(\pi_{21} + \pi_{22}))$. It follows that

$$\begin{aligned} & 0 \leq z - \pi_{21} + \frac{\pi_{21}}{(\pi_{21}+\pi_{22})} \leq 1 \\ \implies & 0 \leq -\pi_{21}^2 + \pi_{21}(z - \pi_{22} + 1) + z\pi_{22} \leq \pi_{21} + \pi_{22}. \end{aligned} \tag{2.26}$$

Solving the right hand side of this inequality for π_{21} we have

$$\begin{aligned} & -\pi_{21}^2 + \pi_{21}(z - \pi_{22} + 1) + z\pi_{22} \leq \pi_{21} + \pi_{22} \\ \Rightarrow & \pi_{21}^2 + \pi_{21}(\pi_{22} - z) + \pi_{22}(1 - z) \geq 0. \end{aligned} \tag{2.27}$$

As $0 \leq \pi_{21} \leq 1$ and $0 \leq \pi_{22} \leq 1$, equation (2.27) is satisfied for all values of π_{21} and π_{22}. Solving

the left hand side of equation (2.26) for π_{21} we have,

$$\begin{aligned} 0 &\leq -\pi_{21}^2 + \pi_{21}(z - \pi_{22} + 1) + z\pi_{22} \\ \Rightarrow \pi_{21}^2 - \pi_{21}(z - \pi_{22} + 1) - z\pi_{22} &\leq 0 \qquad (2.28) \\ \Rightarrow (\pi_{21} - r_1)(\pi_{21} + r_2) &\leq 0 \qquad (2.29) \end{aligned}$$

where

$$r_{1,2} = \frac{(1 + z - \pi_{22}) \pm \sqrt{(\pi_{22} - 1 - z)^2 + 4z\pi_{22}}}{2} \qquad \text{by the quadratic formula.} \qquad (2.30)$$

If $(\pi_{22} - 1 - z)^2 + 4z\pi_{22} \geq 0$, that is $(\pi_{22})^2 - 2(1 - z)\pi_{22} + (z + 1)^2 \geq 0$, then r_1, $r_2 \in \mathbb{R}$. If r_1, $r_2 \in \mathbb{R}$, then equation (2.29) will be satisfied when $r_1 \leq \pi_{21} \leq r_2$. In order for r_1, $r_2 \in \mathbb{R}$ it is required that $\pi_{22} \leq r_1^*$ or $\pi_{22} \geq r_2^*$, where $r_1^* = 1 - z - 2\sqrt{-z}$ and $r_2^* = 1 - z + 2\sqrt{-z}$. Since $r_2^* > 1$, the following inequality must be satisfied $\pi_{22} \leq r_1^*$. Therefore, the bounds on π_{21} and π_{22} are $r_1 \leq \pi_{21} \leq r_2$ and $0 \leq \pi_{22} \leq r_1^*$ respectively.

Substituting our known bounds into (2.25) we now have the following equation that we wish to solve

$$f_{Y_{11}}(z) = \lim_{\epsilon \to 0} \frac{\Gamma(2(1+\epsilon))}{\Gamma(1+\epsilon)\Gamma(1+\epsilon)} \int_0^{r_1^*} \int_{r_1}^{r_2} \epsilon \pi_{21}^{\epsilon-1} \epsilon \pi_{22}^{\epsilon-1} \, d\pi_{21} \, d\pi_{22}. \qquad (2.31)$$

Integrating out π_{21} we have

$$\begin{aligned} f_{Y_{11}}(z) &= \lim_{\epsilon \to 0} \frac{\Gamma(2(1+\epsilon))}{\Gamma(1+\epsilon)\Gamma(1+\epsilon)} \int_0^{r_1^*} \epsilon \pi_{22}^{\epsilon-1} \left(r_2^\epsilon - r_1^\epsilon\right) \, d\pi_{22} \\ &= \lim_{\epsilon \to 0} \int_0^{r_1^*} \epsilon \pi_{22}^{\epsilon-1} r_2^\epsilon \, d\pi_{22} - \lim_{\epsilon \to 0} \int_0^{r_1^*} \epsilon \pi_{22}^{\epsilon-1} r_1^\epsilon \, d\pi_{22}. \qquad (2.32) \end{aligned}$$

In order to simplify the preceding integration we introduce the following lemma.

Lemma 2.5.2 *Let c and K be positive constants. If $g(x)$ is continuous in $[0, K]$ with $g(0) > 0$*

then it follows that

$$\lim_{\epsilon \to 0} \int_0^K \epsilon x^{c\epsilon - 1} g(x)^\epsilon d\, x = \frac{1}{c}. \tag{2.33}$$

Proof If $g(x)$ is continuous in $[0, K]$ and $g(0) > 0$, then there exists $k > 0$ and $m_1 > 0$ such that $g(x) \geq m_1$ in $[0, k]$. Let

$$I = \int_0^k \epsilon x^{c\epsilon - 1} g(x)^\epsilon dx + \int_k^K \epsilon x^{c\epsilon - 1} g(x)^\epsilon dx = I_1 + I_2. \tag{2.34}$$

As $g(x)$ is continuous in $[0, k]$ it is bounded above, $g(x) \leq m_2$. Therefore,

$$\begin{aligned} \int_0^k \epsilon x^{c\epsilon - 1} m_1^\epsilon dx \leq I_1 \leq & \int_0^k \epsilon x^{c\epsilon - 1} m_2^\epsilon \\ \Rightarrow m_1^\epsilon \frac{k^{c\epsilon}}{c} \leq I_1 \leq & \, m_2^\epsilon \frac{k^{c\epsilon}}{c}. \end{aligned} \tag{2.35}$$

It follows that $\lim_{\epsilon \to 0}\ I_1 = \frac{1}{c}$.

Since

$$\mid I_2 \mid = \epsilon \int_k^K \mid x^{c\epsilon - 1} g(x)^\epsilon \mid dx \leq \epsilon \int_k^K x^{c\epsilon - 1} |g(x)^\epsilon| dx \tag{2.36}$$

and $g(x)$ is continuous in $[k, K]$ it follows that $g(x)$ is bounded in absolute value, $|g(x)| \leq M$. So $|g(x)^\epsilon| = |g(x)|^\epsilon \leq M^\epsilon$ and integrating similarly to (2.35) it can be seen $\lim_{\epsilon \to 0}\ I_2 = 0$ and $\lim_{\epsilon \to 0}\ I = \frac{1}{c}$. ∎

Since r_2 is positive and differentiable with respect to π_{22}, by Lemma 2.5.2 it follows that

$$f_{Y_{11}}(z) = 1 - \lim_{\epsilon \to 0} \int_0^{r_1^*} \epsilon \pi_{22}^{\epsilon - 1} r_1^\epsilon \, d\pi_{22}. \tag{2.37}$$

In order to evaluate the remaining integral we re-write r_1 as a power series expansion about π_{22},

$$r_1 = \frac{(1 + z - \pi_{22}) - (1 + z)\sqrt{1 + \frac{2z-2}{(1+z)^2}\pi_{22} + \frac{1}{(1+z)^2}\pi_{22}^2}}{2} \tag{2.38}$$

where

$$\sqrt{1+\frac{2z-2}{(1+z)^2}\pi_{22}+\frac{1}{(1+z)^2}\pi_{22}^2}=1+\frac{1}{2}\left(\frac{2z-2}{(1+z)^2}\pi_{22}\right)+\pi_{22}^2h(\pi_{22}), \tag{2.39}$$

where $h(\pi_{22})$ is continuous. Substituting this expansion into (2.38) gives

$$\begin{aligned} r_1 &= \frac{(1+z-\pi_{22})-(1+z)\left[1+\frac{1}{2}\left(\frac{2z-2}{(1+z)^2}\pi_{22}\right)+\pi_{22}^2h(\pi_{22}\right]}{2} \\ &= \frac{(1+z-\pi_{22})-(1+z)+\frac{1-z}{z+1}\pi_{22}+\pi_{22}^2h(\pi_{22})}{2} \\ &= \pi_{22}\left(\frac{1-z}{2(1+z)}-\frac{1}{2}\right)+\pi_{22}^2h(\pi_{22}). \end{aligned} \tag{2.40}$$

The power series expansion for r_1 is then substituted back into (2.37) to give

$$f_{Y_{\pi_{11}}}(z)=1-\lim_{\epsilon\to 0}\left(\frac{1-z}{2(1+z)}-\frac{1}{2}\right)^{\epsilon}\int_0^{r_1^*}\epsilon\pi_{22}^{2\epsilon-1}(1+\pi_{22}h(\pi_{22}))^{\epsilon}d\,\pi_{22}. \tag{2.41}$$

Let $1+\pi_{22}h(\pi_{22})=g(\pi_{22})$, then by Lemma 2.5.2 we arrive at $f_{Y_{11}}=1/2$, which we identify as the pdf of the uniform$[-1,1]$ distribution. For $z>0$, we note that interchanging the columns in Table 1.1 (re-labeling disease status) changes the sign of PAR, and our prior is symmetric for the two columns, so the result still holds. ∎

Implementing our toy example with the Dirichlet$(1,1,\epsilon,\epsilon)$ prior, we see from Table 2.5 that there are no signs of shrinkage, with the point estimates being close to the true value of PAR. There is also no significant change in the MSE from that of the MLE. The average percent coverage and interval length for this method were 95.11% and 0.04 respectively. Further exploration of the alternative priors revealed that the Dirichlet$(1,1,\epsilon,\epsilon)$, and the Jeffreys prior do not perform quite as well as the Dirichlet$(1,1,1,1)$ in general (see Figure 2.7). The Dirichlet$(1,1,\epsilon,\epsilon)$ also tends to under-cover when x_{11} and x_{21} are small (see Figure 2.8) since adding 1 to x_{11} but ϵ to x_{21} tends to introduce bias. Therefore, this prior should only be considered when x_{11} and x_{21} are large. If there is a strong prior belief that the PAR is positive, and x_{11} and x_{21} are not small, then the Dirichlet$(1,\epsilon,\epsilon,1)$ could be used since this implies a Uniform[0,1] for PAR. However if

	PAR	Bias (10^{-6})	Variance (10^{-4})	MSE (10^{-4})
MLE	0.902	9.65	1.0789	1.0788
Posterior Mean	0.901	-913.18	1.0564	1.0647
Posterior Median	0.901	-419.02	1.0012	1.0028

Table 2.5: Comparison of the MLE (delta method) with the mean and median estimates of the posterior for the Bayesian method with Dirichlet$(1, 1, \epsilon, \epsilon)$ prior. The true PAR for this case, with $n = 1000$, is 0.902, and the percent coverage is 94.63 and 95.11 for the delta and Bayesian methods respectively.

there is prior information available it is perhaps better incorporated by setting the prior for π_{11}, π_{12}, π_{21} and π_{22} based on pseudo-data for the contingency table (a data augmentation prior). Alternatively, the practitioner may prefer to specify independent beta priors on the probabilities $P(E^+)$, $P(D^+|E^+)$ and $P(D^+|E^-)$, perhaps using a graphical approach (Jones and Johnson, 2014); in this case the posterior can no longer be derived analytically and Markov Chain Monte Carlo methods would be required. Regardless of the prior adopted for π it is recommended that the practitioner investigate the implication of this prior on the PAR graphically.

2.6 Concluding remarks

The population attributable risk is widely used in the epidemiological literature. However, we suggest that the ambiguity in the literature surrounding the algebraic definition of the PAR, and confusion with similar but distinct concepts, has resulted in a lack of standard methodology for calculation of its confidence interval. Here we have not only examined the performance of some standard Frequentist approaches to confidence interval calculation for the PAR, but also presented a Bayesian alternative which allows for incorporation of *a priori* knowledge.

Use of this Bayesian approach with the standard reference prior Dirichlet$(1, 1, 1, 1)$ results in nominal coverage for almost all practically realistic values of PAR, and is recommended especially when there are small or moderate observed counts in the table. Because this method is based on random samples from the posterior, the given confidence interval can appear slightly

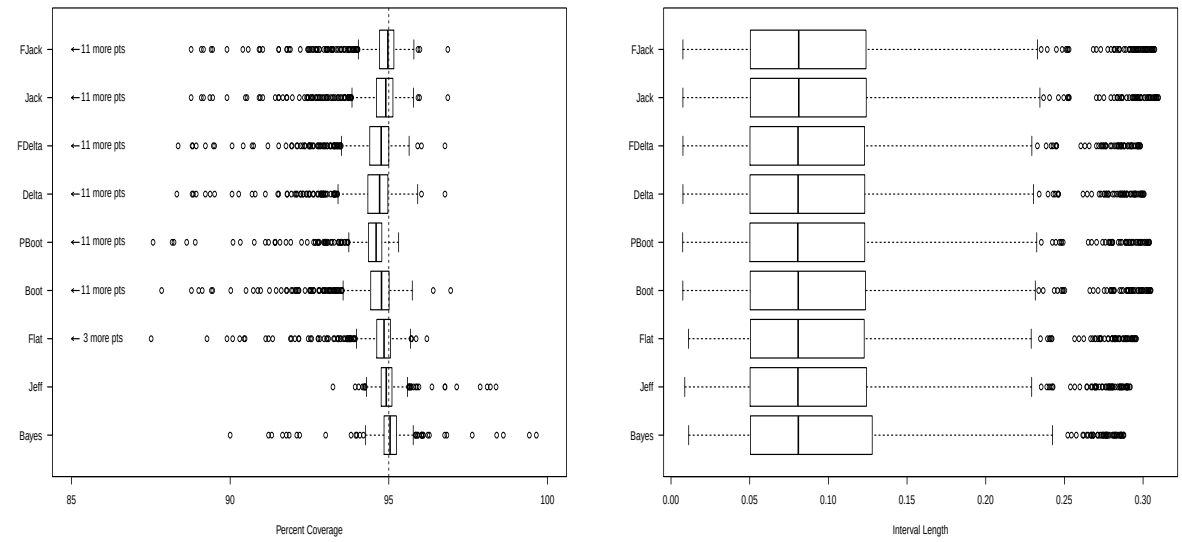

Figure 2.7: Percent coverage (left) and interval length (right) of each method, following Simulation 1 methodology, including a Bayesian approach with Dirichlet$(1, 1, \epsilon, \epsilon)$ prior and Jeffrey's prior. Note that 11 points for each Frequentist method and 3 for the newly added Bayesian approaches have being omitted due to having a percent coverage less than 87%.

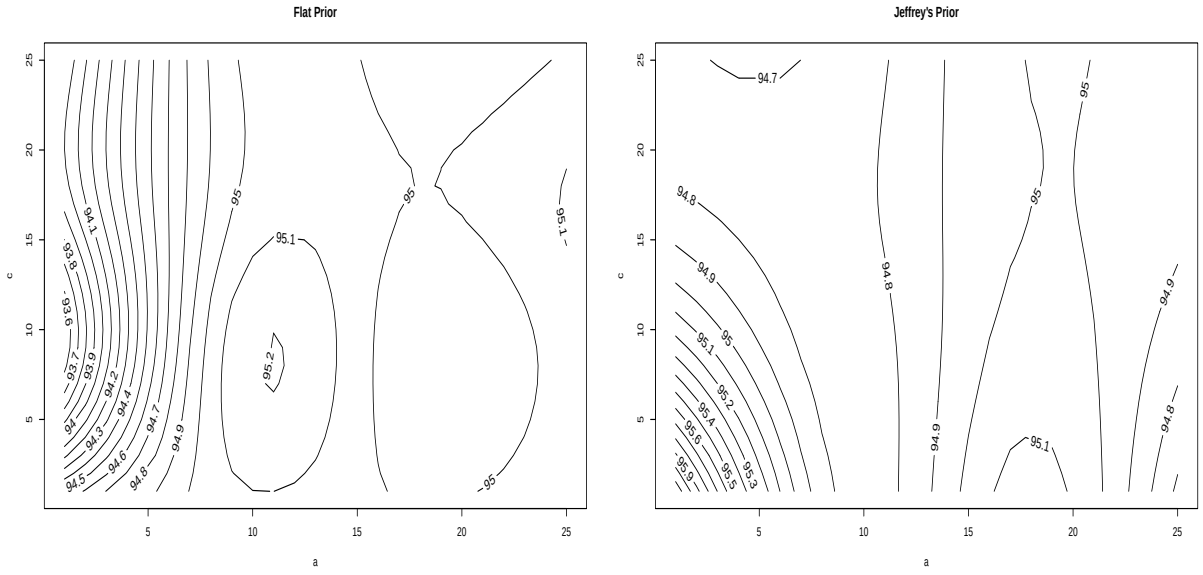

Figure 2.8: Contour plots of the percent coverage for the Bayesian approach with Jeffrey's and flat prior respectively, where all integer combinations of $a = x_{11}$ and $c = x_{21}$ in the interval $[1, 25]$ are explored for fixed $e = 0.2$ (Simulation 3).

different each time the method is applied. This can be overcome by using a very large number of simulations, at the expense perhaps of increased computational run time. The delta method is the only method where a re-sampling or simulation procedure is not required. Provided the observed counts in the table are large, this computationally cheap alternative gives close to nominal coverage, with the Fisher-z transformation improving coverage only when PAR is particularly large. In extreme situations such as a mass outbreak of disease, where the PAR is close to $+1$ (or alternatively close to -1, as in the case of protective exposures like vaccination), use of the Dirichlet$(1, 1, 1, 1)$ prior results in inadequate coverage. For these extreme cases, we have found that the Dirichlet$(1, 1, \epsilon, \epsilon)$ prior provides a flat distribution over the parameter space of the PAR and also provides close to nominal coverage.

One advantage of the Bayesian approach implemented here is that it is very straightforward to extend to other measures. For example, a confidence interval for the PAF can be produced simultaneously by applying equation (1.2) to each random draw from the posterior of $(\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22})$. This can be done similarly for a number of different functions, such as the odds ratio or relative risk. The Bayesian framework also allows us to add any existing prior information into our model, for example information on the population risk of disease or exposure, which would be particularly useful for other study designs such as a case-control or cohort study.

An R-shiny app that calculates the Bayesian interval using the Dirichlet$(1, 1, 1, 1)$ prior is available at the web page `https://spirikahu.shinyapps.io/PARCIs/`. The delta method approach for calculating confidence intervals has also being included in the R package `EpiR` (Stevenson et al., 2017).

Chapter 3

Bayesian Inference for Population Attributable Measures from Under-identified Models

Calculation of the PAR and PAF requires population estimates for the prevalence of disease and exposure. In the previous chapter, where we were concerned with cross-sectional studies, we were fortunate enough to be able to calculate these estimates directly from the data. However, estimation of these parameters from the data might not be possible for some study designs. When this is the case we require extra information about these parameters, and the underlying statistical model is considered under-identified (see Section 3.1 for a formal definition). Case-control and cohort studies, discussed in Section 1.2, are examples which lead to under-identified models requiring either additional assumptions or the population prevalence of disease and exposure to be known. Taking the Frequentist approach to estimation in these situations means nominating fixed values for the population prevalences, which many authors have done (Walter, 1975, 1976; Leung and Kupper, 1981; Newson, 2013). The disadvantage of this approach is that no information regarding the uncertainty of these parameters is provided. Failing to take into

account all the possible sources of uncertainty when calculating confidence intervals can result in intervals which have less than nominal coverage. Alternately, a Bayesian methodology can be adopted where information about the unknown parameters can be provided in the form of a prior distribution. Given that the prior is a proper probability distribution (i.e. $\int f(\theta) = 1$ and $f(\theta) \geq 0 \; \forall \; \theta$), then along with the likelihood arising from the statistical model and the data, Bayes theorem can be applied to generate the posterior distribution. This is the case regardless of whether the model is identified or not (Gustafson, 2015). However, although Bayesian inference is applicable in the case where the model is under-identified it is not without flaws. Through simulation we demonstrate (see Section 3.2.7) that the standard MCMC methods, discussed in Section 3.2.3, often have poor convergence or efficiency when the model lacks identifiability.

Another example of an under-identified model emerges when analysing data from a 2×2 table where either the disease or exposure status is based on an imperfect diagnostic test. The experiment resulting in the leptospirosis dataset (Table 2.1), which was studied in detail in the previous chapter, made use of the Microscopic Agglutination Test to determine whether an individual was exposed (or not) to at least one of the *leptospira* bacteria. To take into consideration the uncertainty in the results due to the use of this imperfect diagnostic test, we incorporate prior information on its sensitivity (Se), the probability of getting a true positive result, and its specificity (Sp), the probability of getting a true negative result. Given that a 2×2 contingency table provides us with three degrees of freedom, but we require estimates of more than three parameters (p, q, e, Se and Sp) to calculate the PAR and PAF in this situation, the model is under-identified.

The main aim of this chapter is to explore the performance of different Bayesian approaches for estimating parameters from an under-identified model, so that a recommendation can be made as to which sampler should be applied to provide estimates for the PAR, PAF and their credible intervals. Through this process we develop alternative MCMC samplers aimed at exploring the parameter space for an under-identified model more efficiently. Section 3.1 begins by providing a formal definition of identifiability as well as an overview of how identification (or

lack of) has been explored and defined in the literature. We then discuss each of the three examples suggested above in turn, beginning with the incorporation of an imperfection diagnostic test in Section 3.2, case-control study in Section 3.3 and cohort study in Section 3.4.

3.1 Identifiability

Statistical models allow us to make inference about unknown parameters for a population by assuming a certain probability distribution for our observed variables. Intuitively, if we obtain more information about the observed variables we expect to learn more about the unknown parameters for the population. This can be represented pictorially by the left hand side of Figure 3.1. Let the area enclosed by the solid line represent the parameter space and the dashed lines the contours of the likelihood surface. What we expect to see is that as the sample size increases these likelihood contours will shrink towards a point. This point represents the true value for our parameter, which we would get if we had an infinite amount of data. Models that possess this property are considered to be identifiable. More formally this property of identification relates to the mapping from the parameter space to the set of possible distributions for the observed values. If this mapping is invertible, that is multiple points in the parameter space do not map to the same distribution of the observed values, i.e. $p(x|\theta_1) = p(x|\theta_2)\ \forall\ x$ only when $\theta_1 = \theta_2$, then the model is considered to be identifiable (Gustafson, 2015, p. 15). When there exist multiple values for the parameter vector θ, which produce the same probability distribution for the observed values, the model is classified as under-identified. In many under-identified situations, the parameters θ_1 for which $p(x|\theta_1) \equiv p(x|\theta_2)$ where $\theta_1 \neq \theta_2$, form a continuum in the parameter space (see Figure 3.1 right). As a result of this one could collect an infinite amount of data, but not arrive at a point estimate for the parameters of interest. Consequentially, standard Frequentist methods for parameter estimation are not a viable option for a model which is under-identified.

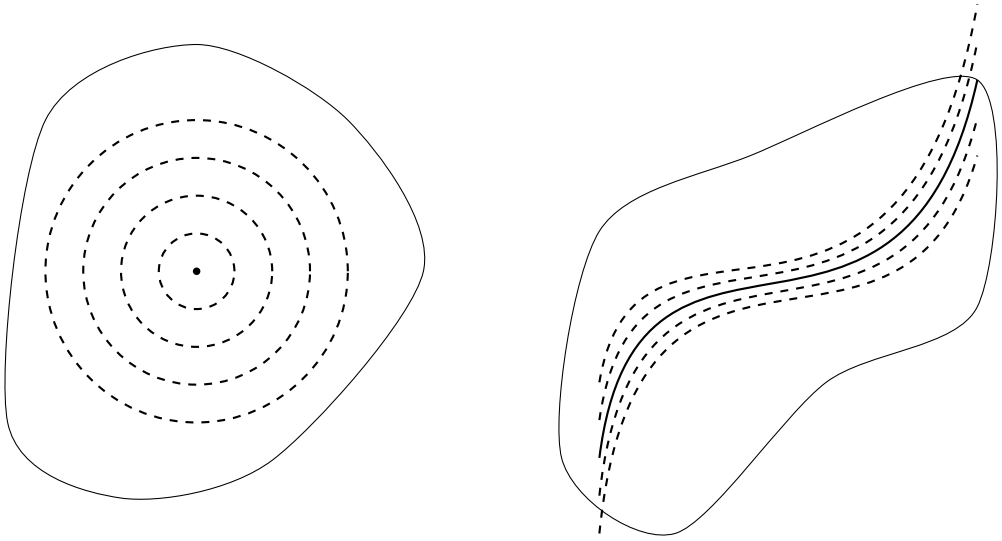

Figure 3.1: *Left:* Identifiable model where the central dot represents the maximum likelihood estimate and dashed lines the likelihood contours. *Right:* Non-identifiable model where the solid line represents the set of values which have the same maximum likelihood and the dashed lines the likelihood contours.

Goodman (1974) discusses the concept of identification whilst analysing m-way contingency tables, which arise when exploring the relationship between variables with three or more distinct categories, using a multinomial latent class model. After outlining an iterative Gibbs sampling procedure to estimate the parameter vector of interest θ, Goodman questions if the maximum likelihood estimate, $\hat{\theta}$, is unique. If a unique solution for $\hat{\theta}$ can be determined from the model, $f(\theta)$, then $\hat{\theta}$ is identifiable. However, if a unique solution for $\hat{\theta}$ can only be found for $f(\theta)$ within some neighborhood of θ, then $\hat{\theta}$ is considered locally identifiable. Taking partial derivatives of the model $f(\theta)$ with respect to each element of θ, results in the Jacobian matrix $J = \frac{\partial f(\theta)}{\partial \theta}$. If the rank of the Jacobian matrix J is equal to the number of parameters in θ, then the model will be locally identifiable (Goodman, 1974). If the number of parameters in the vector θ exceeds the number of degrees of freedom, then the model will be under-identified and have $|J| = 0$ (Goodman, 1974).

It is often assumed that if the degrees of freedom for a data set is equal to the number of parameters in the model, then the model is identifiable (Jones et al., 2010). Johnson et al.

(2001) and Johnson and Hanson (2005) provide a counter example to this claim by considering three independent binary tests for disease for three sampled populations. This results in three 2×2 tables and nine degrees of freedom, an extension of work by Hui and Walter (1980). Despite having the required degrees of freedom to estimate the prevalence of disease for each population, as well as the sensitivity and specificity for each test, they show the model has infinitely many solutions and thus lacks identification. Since global identification is difficult to prove, authors will often investigate local identification in a neighborhood around a point θ_i instead (Rothenberg, 1971; Goodman, 1974; Jones et al., 2010). Gustafson (2010) explores models which are partially identifiable. Gustafson (2010) describes that in Bayesian terminology this means, "as the sample size goes to infinity, the support of the marginal posterior distribution on the target converges to a set which is smaller than the corresponding prior support but larger than a single value." The shape of the marginal posterior distribution for θ in the limit (as $n \to \infty$) could provide additional inferential insight, as well as sample size guidance for future studies. A novel approach for generating the posterior distribution for a partially identifiable model was recently proposed by Gustafson (2015) which is based on the use of importance sampling. We later show via simulation (see Section 3.2.7) that this approach is superior to MCMC sampling methods in terms of its efficiency for the models we are considering.

In practice if a statistical model with Θ parameters seems appropriate for the situation, but the model is under-identified, a simpler sub-model with fewer than Θ parameters can sometimes be selected to achieve identifiability. This process however, can result in a model which is reliant on unrealistic assumptions (e.g. by assuming a fixed value for a parameter). Less intuitively the model could be expanded by incorporating multiple populations which share a common measure. For example Hui and Walter (1980) consider a model for two independent diagnostic tests of the same accuracy, on two sampled populations. Their data formed two 2×2 tables resulting in a total of six degrees of freedom. This allowed for the estimation of the desired prevalences of disease for each population as well as the sensitivity and specificity measures for each diagnostic test. Gustafson (2005) questions whether contraction or expansion of a model for the sake of

identification is a good idea. Through simulation he shows that the use of an under-identified model with good prior information, and an intelligent choice of parameterisation (discussed in Section 3.2), can lead to more reasonable estimates than a model that has been altered to achieve identifiability.

3.2 An example concerning measurement error

In order to obtain information regarding a subject's disease or expose status it is sometimes necessary for practitioners to perform diagnostic tests. Diagnostic tests, even those considered to be gold standard (i.e. assumed to be of perfect classification), are often imperfect (Rothman, 2002). That is, these tests can sometimes provide false positive or false negative results. The accuracy of a diagnostic test is described by its sensitivity and specificity. For the leptospirosis data (Table 2.1) an imperfect diagnostic test was used to determine the exposure status of an individual. To estimate the PAR and PAF, taking into consideration the uncertainty associated with the diagnostic test, we require estimates for p, q, e, Se and Sp, where $Se = P(T^+|E^+)$, $Sp = P(T^-|E^-)$ and T^+ represents those testing positive to the risk factor of interest and T^- those testing negative. Given the data can be represented in a 2×2 table the resulting model is under-identified, as we have five parameters (p, q, e, Se and Sp) we wish to estimate, but only three degrees of freedom available. A similar example is explored in the Bayesian framework by Joseph et al. (1995) who makes use of latent class analysis and Gibbs sampling (which we implement in Section 3.2.3.4) to estimate the population prevalence.

For the leptospirosis example, let η_{ij}, for $i, j \in \{1, 2\}$, represent the observed probabilities (calculated from the data) for each cross classification in Table 3.1 (left). For example, η_{11} represents the probability of testing positive for exposure to the leptospira bacteria (T^+) and exhibiting flu-like symptoms (D^+). Additionally, let π_{ij} for $i, j \in \{1, 2\}$ represent the true, but unknown, probability for each cross classification in Table 3.1 (center). For example, π_{11} represents the true probability that an individual is exposed to the leptospira bacteria and

exhibits flu-like symptoms. In practice the true probabilities $\pi = (\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22})$ can be estimated by the observed probabilities from the data if Se and Sp are known (see Equation 3.1). To express the observed probabilities in Table 3.1 (left) in terms of the true parameters in Table 3.1 (center), the probability that the diagnostic test accurately assigns an individual to an exposure category needs to be taken into consideration (see Table 3.1, right). For example, to determine the probability of observing an individual who is exposed and diseased (η_{11}) we sum the true probability of testing positive for exposure when actually exposed (i.e. $\pi_{11}Se$) with the probability of receiving a false positive test result (i.e. $(1 - Sp)\pi_{21}$). The formulation for each of the observed probabilities can be defined as follows:

$$\begin{aligned} \eta_{11} &= Se\pi_{11} + (1 - Sp)\pi_{21} & \eta_{12} &= Se\pi_{12} + (1 - Sp)\pi_{22} \\ \eta_{21} &= Sp\pi_{21} + (1 - Se)\pi_{11} & \eta_{22} &= (1 - Se)\pi_{12} + Sp\pi_{22}. \end{aligned} \tag{3.1}$$

Note that since $\eta_{22} = 1 - \eta_{11} - \eta_{12} - \eta_{21}$, it is not independent of η_{11}, η_{12} and η_{21}, so only three equations are actually needed. As the parameters p, q and e are required to provide estimates for the PAR and PAF, the above equations (3.1) can be re-expressed as:

$$\begin{aligned} \eta_{11} &= Se\, pe + (1 - Sp)q(1 - e) & \eta_{12} &= Se(1 - p)e + (1 - Sp)(1 - q)(1 - e) \\ \eta_{21} &= Sp(1 - e)q + (1 - Se)pe, \end{aligned} \tag{3.2}$$

recalling the relationship between π and p, q, e from (1.10). It just so happens, that in this example the formulation (3.2) has been nicely separated into an identifiable part $\eta = (\eta_{11}, \eta_{12}, \eta_{21}, \eta_{22})$ and a non-identifiable part consisting of the terms p, q, e, Se and Sp.

Gustafson (2015) discusses transforming models which are partially identifiable into an identified part, such that the distribution for the data D depends only on identified parameters, Φ_I, and a non-identified part Φ_N. Gustafson (2015) calls this reparameterization from the parameter vector of interest θ to $\Phi = (\Phi_I, \Phi_N)$ according to $h(\theta)$, where $h()$ is a smooth, invertible

	Diseased	
Test	D^+	D^-
T^+	η_{11}	η_{12}
T^-	η_{21}	η_{22}

	Diseased	
Exposed	D^+	D^-
E^+	π_{11}	π_{12}
E^-	π_{21}	π_{22}

	Exposed	
Test	E^+	E^-
T^+	Se	$(1-Sp)$
T^-	$(1-Se)$	Sp

Table 3.1: *Left:* Observed probabilities, *Center:* True probabilities and *Right:* Diagnostic test accuracy

function and $\dim(\theta) = \dim(\Phi_I) + \dim(\Phi_N)$, a "transparent reparameterization". Transparent parameterizations may not be unique and the choice of Φ can be rather arbitrary. Additionally, as the dimension of θ increases finding a transparent parameterisation becomes more difficult (Gustafson, 2015). The advantage of finding a transparent parameterisation is that it makes clear the structure of the limiting posterior distribution (discussed in more detail in Section 3.2.2), and allows for use of Gustafson's novel approach for deriving the posterior distribution in the presences of a finite amount of data.

Returning to the leptospirosis example, the Jacobian matrix, $J = \frac{\partial \eta}{\partial \theta}$, for (3.2) is given by,

$$\begin{bmatrix} eSe & (Sp-1)(e-1) & pSe + q(Sp-1) & ep & q(e-1) \\ -eSe & -(Sp-1)(e-1) & -Se(p-1) - (Sp-1)(q-1) & -e(p-1) & -(e-1)(q-1) \\ -e(Se-1) & -Sp(e-1) & -qSp - p(Se-1) & -ep & -q(e-1) \end{bmatrix} \tag{3.3}$$

where the rows correspond to the observed probabilities η_{11}, η_{12}, η_{21} and columns the parameters in the following order: p, q, e, Se and Sp. Since J is rectangular it is rank deficient and thus $|J| = |J^T J| = 0$, where J^T represents the transpose of J. That is the model (3.2) is under-identified as expected. Although, this model lacks identification, insights may still be gained by exploring the null space of J, denoted $\mathcal{N}(J)$, using algebraic singular value decomposition as done by Jones et al. (2010).

3.2.1 Exploring identification via the null space

Jones et al. (2010) provides a geometric approach of Goodman's technique (Goodman, 1974), in the context of diagnostic testing. Jones et al. (2010) examined the rank of J using algebraic singular value decomposition in the mathematical software Matlab. The authors note that if any of the singular values, that is the square roots of the eigenvalues of J, are identically zero then the model is under-identified. Whereas, if the singular values of J are all nonzero, then J is of full column rank and the model locally identifiable in any region where those singular values are nonzero. Moreover, Jones et al. (2010) provides further insights by considering the null singular vectors of J.

The null space of J is defined as the set of non-zero solutions to the equation, $Jx = 0$. When J is a square $n \times n$ matrix of full rank, meaning we have the same number of degrees of freedom as parameters we wish to estimate, then the $\mathcal{N}(J)$ can be calculated using standard eigenvalue decomposition. The case of most interest to us is when J is a rectangular $m \times n$ matrix, where $m < n$. In this situation the $\mathcal{N}(J)$ is calculated by taking the singular value decomposition of J. However, it should be noted that since $\mathcal{N}(J) = \mathcal{N}(J^T J)$, when J is an $m \times n$ matrix (Smith and Teo, 1989, p. 68), eigenvalue decomposition on $J^T J$ could be used to determine the $\mathcal{N}(J)$. Similarly, if J where an $n \times m$ matrix eigenvalue decomposition on JJ^T could be used to determine the $\mathcal{N}(J)$. Singular value decomposition of J takes on the form $J = ULV^T$, where U is an $m \times m$ matrix containing the right singular vectors of J, L the diagonal $m \times n$ matrix of singular values and V the $n \times n$ matrix with left singular vectors of J. The null space is comprised of the right null singular vectors, corresponding to the zero singular values. Geometrically the null space of J represents the tangent space to the contours of equal likelihood in the parameter space. For our model (3.2) this can be seen by expressing the vector η, as the first order Taylor series approximation $\eta \approx f(\theta_0) + J(\theta_0)(\theta - \theta_0)$ evaluated at the point θ_0. Now if $\theta - \theta_0 \subseteq \mathcal{N}(J(\theta_0))$ then $\eta(\theta) = \eta(\theta_0)$. That is, our observed probabilities will be the same as the probability evaluated at the parameter vector θ_0. Therefore, if any of

the parameters in the vector θ have components which are all identically zero in the null space, then that element of θ will be estimable (Rothenberg, 1971).

In the previous section we confirmed that our model (3.2) is under-identified, as J was not of full rank therefore $|J^T J| = 0$. We now wish to explore the null space of this model to determine for which situations, if any, the model is locally identifiable. Making use of the Matlab program outlined by Jones et al. (2010) we find that the null space of J for model 3.2 is spanned by the columns of:

$$\begin{bmatrix} 0 & (p-q)(e-1)/[e(Se+Sp-1)] \\ -e(p-q)/[(e-1)(Se+Sp-1)] & 0 \\ -e/(Se+Sp-1) & -(e-1)/(Se+Sp-1) \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$$

where the rows represent the parameters in the order: p, q, e, Se and Sp. As there is no parameter with a zero in both columns, none of the parameters are estimable.

In practice it is commonly assumed that $Se + Sp > 1$ (Jones et al., 2010), which asserts that classification via the diagnostic test is better than random allocation. If we consider the case where we have perfect specificity ($Sp = 1$, therefore no false positives) then the null space is reduced to:

$$\begin{bmatrix} 0 \\ -e(p-q)/[Se(e-1)] \\ -e/Se \\ 1 \end{bmatrix}$$

where the rows represent the following parameter order p, q, e and Se. It can be seen that the null space contains zero for the parameter p, meaning it is possible to estimate p when $Sp = 1$. This occurs because the bottom row of Table 3.1 (left) is now being observed without error, since there are no false positive results. An explicit solution for p can be determined by solving the system of equations given by (3.2), as functions of Se and the observed quantities η. The solution vector for this situation is given by:

$$\begin{bmatrix} \eta_{11}/(\eta_{11}+\eta_{12}) \\ [\eta_{21}-(1-Se)/Se\,\eta_{11}]/[1-(\eta_{11}+\eta_{12}/Se)] \\ (\eta_{11}+\eta_{12})/Se \\ Se \end{bmatrix}$$

We can see that p is the only parameter expressed in terms of identifiable quantities (η), whereas solutions for all other parameters contain the non-identifiable quantity Se.

A similar situation occurs, if we assume perfect sensitivity ($Se = 1$, therefore no false negatives). The null space is reduced to:

$$\begin{bmatrix} -[(p-q)(e-1)]/eSp \\ 0 \\ -(e-1)/Sp \\ 1 \end{bmatrix}$$

where the rows now represent the following parameter order p, q, e and Sp. The null space now contains a zero for the parameter q meaning that it is estimable when $Se = 1$. In terms of Table 3.1 (left) the top row is now being observed without error, since there are no false negative results. An explicit solution for q can be found similarly to before by solving the

system of equations given by (3.2), as functions of Sp and the known quantities η. The solution vector for this situation is given by:

$$\begin{bmatrix} 1 - \{\eta_{12} - (1 - Sp)(1 - [\eta_{21}/(\eta_{21} + \eta_{22})]) + (\eta_{21} + \eta_{22})/Sp\}/e \\ \eta_{21}/(\eta_{21} + \eta_{22}) \\ 1 - (\eta_{21} + \eta_{22})/Sp \\ Sp \end{bmatrix}$$

where q is expressed only in terms of identifiable quantities. Interestingly, as a consequence of q being estimable the PAR also becomes estimable, as it can be expressed by the difference between the $P(D^+)$ and q. The $P(D^+)$ can be estimated directly from the data as it is not effected by the uncertainty in e. Unfortunately the MAT test used in this example is likely to have a larger specificity than sensitivity. In some situations the sensitivity can be increased at the expense of the specificity, by using a combination of tests or symptoms (Rothman, 2002). Rothman (2002) demonstrates this with an example which considers two independent diagnostic tests each with a sensitivity of 80% and specificity of 90%. If a positive result is required for both tests in order to indicate the presence of disease then the sensitivity decreases ($Se = 0.8 \times 0.8 = 0.64$), whereas the specificity increases ($Sp = 0.9 + (0.1 \times 0.9) = 0.99$). However, if a positive result for only one of the two tests is required to indicate the presence of disease then the sensitivity increases ($Se = 0.8 + (0.2 \times 0.8) = 0.96$) whereas the specificity decreases ($Sp = 0.9 \times 0.9 = 0.81$). As another example, analytical test sensitivity can also be increased for continuous outcome tests such as quantitative polymerase chain reaction (qPCR) by adjusting the threshold cycle cutoff value (Caraguel et al., 2011). The above results suggest that increasing the sensitivity at the cost of the specificity maybe a useful strategy if the PAR is to be estimated.

The methods described here have allowed us to analyze and provide estimates for some parameters in the models based on their identifiable part only. Ideally we want to be able to

make inference about each of the parameters in the model and in particular the PAR. As mentioned earlier, this is not possible using Frequentist methods. The simple solution to parameter estimation for under-identified models is to adopt a Bayesian approach, because regardless of whether the model is identifiable, as long as good prior information is available for at least some of the parameters, an estimate for the parameter of interest can be obtained (Gustafson, 2005). It is then a matter of determining whether that parameter estimate is useful or not, based on the resulting credible interval.

3.2.2 Bayesian approach for estimating uncertainty in population attributable measures when incorporating measurement error

In Chapter 2 (Section 2.3) we introduced how Bayes theorem can be used in order to update prior beliefs about a parameter set θ to posterior beliefs. Previously we were fortunate enough to be able to explicitly state the posterior distribution for our example, due to the conjugacy relationship. For this example where we incorporate information about the uncertainty of our diagnostic test it is not possible to determine the posterior distribution analytically so we must resort to MCMC simulation. We begin by assigning the priors $p, q \sim \text{Beta}(1,1)$, $e \sim \text{Beta}(2,2)$, $Se \sim \text{Beta}(25,3)$ and $Sp \sim \text{Beta}(30,1.5)$. The priors on Se and Sp were assigned based on expert opinion, whereas no prior knowledge was available for p, q and e. The priors on p, q and e are the marginal distributions induced by specifying a flat Dirichlet$(1,1,1,1)$ prior on the vector $(\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22})$. Applying a change of variables from $g(\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22})$ to $g(p, q, e)$, where $g()$ is given by (1.10), it can be seen that $p \sim \text{Beta}(\alpha_1, \alpha_2)$, $q \sim \text{Beta}(\alpha_3, \alpha_4)$ and $e \sim \text{Beta}(\alpha_1 + \alpha_2, \alpha_3 + \alpha_4)$ as stated above where $\alpha_1, \ldots, \alpha_4 = 1$ in this case. The model for this example is still multinomial

$$(x_{11}, x_{12}, x_{21}, x_{22}) \sim \text{Multinomial}(n, \eta), \tag{3.4}$$

but now based on the transparent parameterisation (3.2). Given our priors are specified on the original parameterisation this induces a prior density $f(\Phi_I, \Phi_N)$ in the transparent parameterisation.

Markov-chain Monte-Carlo methods using Gibbs or Metropolis-Hastings updating are commonly used to derived posterior distributions. However, it has been found that when a model lacks identifiability this can often result in poor convergence of these updaters due to "ridges" in the likelihood function for θ (Gustafson, 2015). These "ridges" are the product of there being a set of values of equal likelihood for θ (in the limit of infinite data) rather than a single value. In the Bayesian framework this set of values is referred to as the limiting posterior distribution (LPD), which is the probability distribution induced on the ridge by the prior distribution on the original parameters. The LPD will be more concentrated than the specified prior distribution, but will be larger than a single value (Gustafson, 2010). In practice where we only have a finite amount of data, having some knowledge about the LPD can help provide insights about the sample size required to obtain a posterior distribution which is similar to that of the LPD (Gustafson, 2010).

Recently, Gustafson (2015) proposed a novel approach for determining the posterior distribution for under-identified models using the general strategy of importance sampling. However, this method relies on being able to find an appropriate transparent parameterisation for the model, which as previously stated may not be easy if the problem is high dimensional. With standard MCMC samplers though, we need not find such a parameterisation. Over the following sections we provide an overview of Gustafuson's importance sampling approach for under-identified models as well as several MCMC updaters including Metropolis-Hastings, Gibbs, Metropolis-adjusted Langavin algorithm (MALA) and Hamiltonian Monte Carlo algorithm (HMC). Based on these approaches we then develop a new updater which takes into consideration the shape of the posterior ridge to encourage improved convergence. Through simulation we then compare these MCMC methods to the importance sampling approach, in terms of their efficiency.

3.2.3 Markov Chain Monte Carlo Methods

Markov Chain Monte Carlo methods provide us with a means of evaluating the posterior distribution for complex models. As the name suggests these algorithms rely on the construction of a Markov Chain. Markov chains consist of a sequence of random variables, say $\theta^1, \theta^2, \ldots, \theta^t$, where the distribution of θ^t, for any t, depends only on the previous value, θ^{t-1}, in the sense that $p(\theta^t|\theta^{t-1}, \theta^{t-2}, \ldots) = p(\theta^t|\theta^{t-1})$. For large t the distribution of θ^t will settle down to a stationary distribution , i.e. the targeted posterior, $p(\theta|x)$, regardless of the chain's initial state. The way in which our sequence of random variables is derived is dependent on the updating procedure adopted. Brooks et al. (2011) provides an introduction to a large number of different procedures for updating a Markov chain.

3.2.3.1 Metropolis-Hastings Algorithm

Metropolis et al. (1953) first developed this sampling algorithm in order to model the phase transitions of a substance. The method was later generalised by Hastings (1970) so that it could be used in the field of statistics to construct posterior distributions and has since been called the Metropolis-Hastings algorithm. The Metropolis-Hastings algorithm is an iterative method which begins with selection of a proposal distribution, $Q()$. From this distribution we then propose a new candidate value which we either accept, with probability α and move to, or reject, with probability $(1 - \alpha)$, meaning we remain at our current position. A more detailed outline of the iterative processes is as follows:

1. Define an arbitrary initial value θ^0 and initialise an iteration counter at $t = 0$.
2. Generate a candidate value, $\theta^\dagger$, from the proposal distribution $Q(\theta^\dagger|\theta^t)$, where θ^t is the value of the Markov chain at iteration t.

3. Compute the probability of acceptance

$$\alpha = \min\left\{1, \frac{p(\theta^\dagger)L(\theta^\dagger)Q(\theta^t|\theta^\dagger)}{p(\theta^t)L(\theta^t)Q(\theta^\dagger|\theta^t)}\right\}, \tag{3.5}$$

   where $p(\theta)$ represents the prior for θ and $L(\theta)$ the likelihood.

4. Take a random draw, U, from a Uniform$(0, 1)$ distribution. If $U \leq \alpha$ set $\theta^{t+1} = \theta^\dagger$, otherwise set $\theta^{t+1} = \theta^t$.

5. Set the iteration counter from t to $t + 1$.

6. Repeat steps 2-5 until the desired number of iterations is reached.

Implementation of this algorithm can be done either by updating each component of the vector θ separately (component-wise updating), or by updating the entire vector θ at once (block updating). Component-wise updating means that each parameter will have its own specific proposal distribution, whereas for block updating a single multivariate proposal distribution is selected for the vector θ. The choice of this proposal distribution can be rather arbitrary. Ideally we wish to have a proposal distribution which is similar to that of our target, but in practice we often don't know what this target is. One possibility is to simply let the proposal distribution be the same as the prior distribution, as was originally done by Metropolis et al. (1953). This approach is known as independence sampling and results in a simplification of the acceptance probability to $\alpha = L(\theta^\dagger)/L(\theta^t)$. However, this is not always a particularly efficient sampling approach as many candidate values are often rejected due to being in areas of low likelihood.

A popular alternative to independence sampling is the random walk approach. Mathematically a random walk is described as $\theta^\dagger = \theta^t + e$, where e is a random variable with a distribution independent of the chain, centered on zero (Gamerman and Lopes, 2006). The most common choice of proposal in this case when sampling component-wise is $Q(\theta^\dagger|\theta^t) \sim \text{Normal}(\theta^t, \sigma^2)$, where σ^2 is a scalar multiple c of a fixed estimate for the posterior variance σ^{2*}. For choice of σ^{2*} Christensen et al. (2010) suggest selecting an arbitrary σ for the first k iterations of the

chain, after which the sample variance σ^{2*} for the first k iterations can be calculated and a scalar multiple of this determined to achieve optimal acceptance. When block updating a common choice of proposal is the multivariate normal distribution, $Q(\theta^{\dagger}|\theta^{t}) \sim \text{MVN}(\theta^{t}, \Sigma)$, where the covariance matrix Σ is a scalar multiple c of the current estimate of the posterior covariance, Σ^{*} (Muller, 1994). This scalar multiple c is called a tuning parameter, which can be altered in order to achieve the desired acceptance rate.

The larger the variance of $Q()$ the greater the distance between θ^{t} and the proposed value $\theta^{\dagger}$. This may result in a low acceptance rate due to many of the proposed values being rejected. If the variance of $Q()$ were smaller, then the closer θ^{t} would be to $\theta^{\dagger}$. This may increase the acceptance rate, but the time it takes for the chain to converge to the desired stationary distribution may also increase. Theoretically the optimal acceptance rate for high-dimensional problems with normal proposals is 24% (Roberts et al., 1997), which was demonstrated using simulation by Gelman et al. (1996). In applied contexts authors Bennett et al. (1996) and Besag et al. (1995) indicate acceptance rates should be between 20% to 50%. Choosing the appropriate tuning parameter to achieve this acceptance rate can sometimes simply be a process of trial and error. Tierney (1994) suggests the tuning parameters of $c = 1$ and $c = 0.5$ seem to work well in a number of examples.

3.2.3.2 Metropolis-Adjusted Langevin Algorithm (MALA)

A problem that can arise with the random walk approach described in the previous section, is that the sampler can spend large amounts of time revisiting regions of the parameter space which it has already explored, or that are of low likelihood (Robert and Casella, 2010). In order to more effectively explore the parameter space of a model, Besag (1994) proposed a modification to the random walk algorithm which aims to steer the proposed values towards areas of high posterior probability density. He achieves this by incorporating gradient information for the

target distribution into the proposal distribution as follows:

$$Q(\theta^{\dagger}|\theta^{t}) \sim N\left(\theta^{t} + c\nabla \log(p(\theta^{t}|x)), 2cI\right), \tag{3.6}$$

where $\log(p(\theta|x))$ represents the log-likelihood for the joint posterior distribution on θ, c is some small positive constant (often referred to as the scale factor) and I the identity matrix. The addition of the gradient term $(\nabla \log p(\theta^{t}|x))$ in the proposal distribution comes at an increased computational cost per iteration of the algorithm. However, given that the MALA is likely to take steps in the direction of increasing likelihood, the chain is likely to converge to the posterior distribution faster than a random walk. Moreover, the optimal acceptance probability for MALA will be larger than that of a random walk (Roberts and Rosenthal, 1998). Roberts and Rosenthal (1998) show that optimal scaling for the MALA can be achieved in high dimensions (i.e. as $n \to \infty$) when the acceptance probability is 57.4%.

3.2.3.3 Hamiltonian Monte Carlo Algorithm (HMC)

The Hamiltonian (or hybrid) Monte Carlo algorithm developed by Duane et al. (1987) (for application in physics) is another MCMC sampler which allows for more effective exploration of the parameter space, in comparison to a random walk, by incorporating gradient information about the target distribution. In fact the MALA discussed in the previous section arises as a special case of the HMC algorithm where only a single "leapfrog step", which we discuss soon, is carried out. The HMC algorithm was first adapted for statistical application by Neal (1996) in the context of neural network modeling. Neal (2011) also provides an extensive introduction to MCMC using Hamiltonian dynamics, which has been used as a basis for this section. To implement the HMC algorithm a basic understanding of the components of the Hamiltonian function is necessary.

A Hamiltonian is a function of a position vector q and momentum vector p. Note that in this section alone p and q are not used to represent the $P(D^{+}|E^{+})$ and $P(D^{+}|E^{-})$ respectively,

as they are in other sections throughout this thesis. In non-physical applications of HMC the position vector q corresponds to the parameters of interest (i.e. θ), whereas the elements of the momentum vector p are artificially introduced auxiliary variables, which are typically described by independent Gaussian distributions. Taking the partial derivatives of the Hamiltonian allows us to describe how q and p change over time, t, and results in the Hamiltonian equations of motion:

$$\frac{dq_i}{dt} = \frac{\partial H}{\partial p_i} \tag{3.7}$$

$$\frac{dp_i}{d_t} = -\frac{\partial H}{\partial q_i}, \tag{3.8}$$

for $i = 1, \ldots, d$, where d represents the number of elements in q and p. For the HMC algorithm we make use of Hamiltonian functions that can be written as

$$H(q, p) = U(q) + K(p), \tag{3.9}$$

where $U(q)$ represents the potential energy, which can be described by minus the log posterior density of the distribution for q that we wish to sample, and $K(p)$ the kinetic energy. In its simplest form the kinetic energy can be described by

$$K(p) = \sum_{i=1}^{d} \frac{p_i^2}{2m_i}, \tag{3.10}$$

where $m_i = 1$. With this form for $H(q, p)$ the Hamiltonian equations are defined as follows:

$$\frac{dq_i}{dt} = p_i \tag{3.11}$$

$$\frac{dp_i}{dt} = -\frac{\partial U}{\partial q_i}. \tag{3.12}$$

Ultimately we wish to solve the system of equations (3.11-3.12) for the position vector q, corresponding to our parameters of interest. However, it may not always be possible to find an

analytic solution for q, in which case a numerical method must be adopted. Many numerical methods exist for solving differential equations. Within our algorithm however, we adopt the leapfrog method (see Neal, 2011).

The HMC algorithm is made up of two steps. First a change in the momentum vector only, then a Metropolis update using Hamiltonian dynamics to propose candidate values for both the momentum and position vectors. The procedure can be outlined as follows:

1. Define an arbitrary initial condition for the position vector $q(t = 0)$, for $t = 0, \ldots, N$ where N is the total number of iterations in the chain, and set an iteration counter at $t = 0$.

2. Generate the momentum vector $p(t)$ by taking a draw from a standard Gaussian distribution, i.e. MVN$(0, I)$.

3. Begin the Hamiltonian Metropolis update, using the leapfrog method with L steps, by making a half step for momentum. That is,

$$p_i(t + \epsilon/2) = p_i(t) - \frac{\epsilon}{2}\nabla U(q(t)), \tag{3.13}$$

   where ϵ represents the step size for the algorithm and i, for $i = 1, \ldots, L$, the current leapfrog iteration.

4. Perform L leapfrog steps where a full step is taken for position according to

$$q_i(t + \epsilon) = q_i(t) + \epsilon p_i(t + \epsilon/2), \tag{3.14}$$

   for $i = 1, \ldots, L$. Then a full step is taken for momentum when $i \neq L$ according to

$$p_i(t + \epsilon) = p_i(t + \epsilon/2) - \epsilon\nabla U(q_i(t + \epsilon)). \tag{3.15}$$

5. After completing L leapfrog steps a final half step for momentum is taken according to

$$p_{i=L}(t+\epsilon) = p_i(t+\epsilon/2) - \frac{\epsilon}{2}\nabla U(q_{i=L}(t+\epsilon)), \tag{3.16}$$

   and the momentum vector negated (i.e. $p(\dagger) = -p_{i=L}(t+\epsilon)$) to provide a symmetric proposed state $\{q(\dagger), p(\dagger)\}$, where $q(\dagger) = q_{i=L}(t+\epsilon)$.

6. Evaluate the potential energy, $U(q)$, and kinetic energy, $K(p)$, for $\{q(t), p(t)\}$ and $\{q(\dagger), p(\dagger)\}$.

7. Compute the probability of acceptance

$$\alpha = \min\left\{1, \exp(-U(q(\dagger)) + U(q(t)) - K(p(\dagger)) + K(p(t))\right\}. \tag{3.17}$$

8. Take a random draw, U, from a Uniform$(0,1)$ distribution. If $U \leq \alpha$ set $q(t+1) = q(\dagger)$, otherwise set $q(t+1) = q(t)$.

9. Set the iteration counter from t to $t+1$

10. Repeat steps 2-9 until desired number of iterations is reached.

Similarly to other MCMC algorithms the HMC approach requires appropriate tuning. However, Neal (2011) states that tuning the HMC algorithm can be difficult, as it is a lot more sensitive to tuning than the simple Metropolis-Hastings random-walk. Specifically, tuning the HMC algorithm requires choosing a leapfrog step size, ϵ, and the number of leapfrog steps, L, to perform. The choice of L can be made independently of ϵ. The wrong choice of ϵ however, can have a detrimental effect on the acceptance rate and compute time of the algorithm. Too large a step size will result in a low acceptance rate, whereas too small a step size will mean the algorithm will take a very long time to explore the parameter space. As we may have little *a priori* knowledge about the posterior for the parameters of interest (i.e. position vector q) a common approach for selecting L and ϵ is to simply perform some preliminary runs using different values for L and ϵ, then selecting L and ϵ based on which run provides an acceptance

rate between 20-50% (Neal, 2005). Alternatively, an adaptive approach could be adopted so not to waste compute time on ϵ's and L's that are much too large. Adaptive tuning is discussed by Neal (2005) and can be implemented within the RStan software (Stan Development Team, 2017).

3.2.3.4 Gibbs sampling

The Gibbs sampler proposed by Geman and Geman (1984) is of particular use for problems where the full conditional posterior distribution for a component θ_i, given all the other components of θ, can be specified. For each iteration of the algorithm, for each component of θ_i, a sample is drawn from the conditional posterior distribution $p(\theta_i|\theta_1^t, \ldots, \theta_{i-1}^t, \theta_{i+1}^{t-1}, \ldots, \theta_n^{t-1}, y)$, such that each θ_i being updated is based on the latest values of the other components of θ. In this situation since we are sampling directly from the conditional posterior distribution every draw is retained. The Gibbs sampler can thus be viewed as a special case of the Metropolis-Hastings algorithm where $\alpha = 1$.

3.2.3.5 An adaptation of the Metropolis-Hastings random walk sampler

When a model is under-identified the posterior distribution will contain a set of values of equal maximum likelihood forming a continuum or "ridge". We wish to be able to adapt the Metropolis-Hastings algorithm in order to more efficiently sample from along the posterior ridge. One way we hypothesize this could be done is by specifying the covariance matrix, Σ^*, for the multivariate normal proposal distribution, using the Jacobian matrix $J = \partial\eta/\partial\theta$ as in (3.3). This allows for the incorporation of information about the direction for which the posterior likelihood remains constant. As shown in Section 3.2.1, the null singular values of the Jacobian matrix give the direction of the posterior ridge. Recall that the singular value decomposition of $J = ULV^T$ and that if J is an $m \times n$ matrix, where $m < n$, there will be at least one zero singular value. The null space of J, which corresponds to the right null singular vectors in V,

provides the direction for which the posterior likelihood remains constant. Since the covariance matrix must always be symmetric, to incorporate the direction of constant likelihood the dimensions of J need to be considered. If J is an $m \times n$ matrix, where $m = n$, identifiability is not necessarily guaranteed, but J will be symmetric. When $m < n$ the model is under-identified and J not symmetric. In this case we consider the matrix $J^T J$ when specifying the covariance for the proposal.

When J is an $m \times n$ matrix and our model under-identified, to ensure we take large steps in the directions specified by the null space we want to invert the matrix $J^T J$ to get Σ^*. However, $J^T J$ is singular and therefore can not be inverted. To circumvent this problem we adopt the approach used in ridge regression analysis, by Hoerl and Kennard (1970), of adding a small positive quantity to the diagonal of the matrix $J^T J$. The addition of this small quantity has very little effect on the singular vectors and provides a matrix which can be inverted. Given this information we propose the following covariance matrix for the multivariate normal proposal distribution:

$$\Sigma^* = [\tau I + J^T J]^{-1} \tag{3.18}$$

where I is the identity matrix and τ a small positive quantity added to achieve an invertible matrix. With this particular formulation the components of the covariance matrix may be small as sample size has not been taken into consideration. Consequently steps made in directions other than along or parallel to the ridge will be minuscule. Alternatively, the standard asymptotic approximation to the covariance, $I_E(\hat{\theta}) = E\left[-\frac{d^2 l(\theta)}{d\theta d\theta^T}\right]$, the expected Fisher's information could be used (Bishop et al., 1975). This results in the covariance matrix

$$\Sigma^* = [\tau I + J^T I_E(\hat{\theta}) J]^{-1}. \tag{3.19}$$

Incorporation of the data, as done here, allows for the elements of the covariance matrix to adapt to the sample size. This approach could be extended once more to allow for the incorporation

of prior information as follows

$$\Sigma^* = \left[\tau I + \left(J^T I_E(\hat{\theta}) J + \frac{d^2 p(\theta)}{d\theta d\theta^T}\right)\right]^{-1}, \tag{3.20}$$

where $p(\theta)$ represents the prior distribution on θ. Note that (3.20) is equivalent to $\Sigma^* = \left[\tau I + J^T \frac{d^2 p(\theta|\eta,x)}{d\theta d\theta^T} J\right]^{-1}$, where $p(\theta|\eta, x)$ represents the posterior distribution for θ given η and x which we recall are the observed probabilities and total number of observations respectively for our leptospirosis example.

A potential disadvantage to specifying the proposal distribution in this way is the increased computational burden. Regardless of the choice of Σ^* above we are required to re-calculate Σ^* for each iteration of the MCMC algorithm. Therefore, even if the method explores the posterior more rapidly than other methods it may perform poorly in terms of efficiency. Additionally, we are required to specify two different tuning parameters. Through experience we have found that selecting the tuning parameter c, when using Σ^* defined by (3.18) can be particularly difficult. A very small value of c is necessary to achieve acceptance between 20-50%. Incorporation of the Fishers information as in (3.19) and prior information as in (3.20) improves the ability to tune the method.

3.2.4 Convergence and efficiency diagnostics for MCMC chains

Assessing whether a chain has converged to the stationary distribution or not can sometimes be problematic. Many methods currently exist for assessing convergence and efficiency of MCMC samplers. Trace plots provide a simple visual assessment of the state of the Markov chain over time and allow one to determine if the chain is mixing well. An ideal trace plot will show little autocorrelation and explore the distribution in areas of both low and high densities. Although trace plots can give a good indication of the behavior of the chain, one should not rely solely on them to assess convergence. In some situations it is possible for a trace plot to appear to have converged, but in actuality the chain has not entirely explored the parameter space.

Evaluation of m multiple chains in parallel with different starting values, rather than a single long chain, can provide greater insight into whether the chain has actually explored the entirety of the parameter space and converged (Gelman and Rubin, 1992). Gelman and Rubin (1992) assess m parallel chains for convergence by examining a "potential scale reduction factor". Before the scale reduction factor can be monitiored, the ratio of the within and between chain variations must be calculated. For each scalar estimate θ we label the draws from m parallel sequences of length n as θ_{ij} where $i = 1, \ldots, n$ and $j = 1, \ldots, m$. The within chain variation can be computed as $W = \frac{1}{m(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{m} (\theta_{ij} - \bar{\theta}_{.j})^2$, and between variation as $B = \frac{n}{m-1} \sum_{j=1}^{m} (\theta_{.}j - \bar{\theta}_{..})^2$, where $\bar{\theta}_{.j} = \frac{1}{n} \sum_{i=1}^{n} \theta_{ij}$ and $\bar{\theta}_{..} = \frac{1}{m} \sum_{j=1}^{m} \bar{\theta}_{.j}$. The marginal posterior variance of the estimates can then be calculated by $\text{var}(\theta|x) = \frac{n-1}{n} W + \frac{1}{n} B$, which is a weighted average of the within and between variation. The potential scale reduction factor can then be calculated as $R = \sqrt{\text{var}(\theta|x)/W}$, which estimates the factor by which the scale of the current distribution for θ might be reduced as the number of iterations in the simulation tends to infinity. Values of R close to 1 suggest that the chain has converged to the stationary distribution. Brooks and Gelman (1998) provide a graphical approach to visualize the change in R over time, `gelman.plot`, which has now been included as a part of the R software package `CODA` (Plummer et al., 2006).

Theoretically if we take infinitely many iterations our Markov chain will converge, under fairly mild conditions (Rosenthal, 2006), to the desired stationary distribution. In practice however, we only perform a finite number of iterations until we deem the chain to have achieved stationarity. Ideally we wish for each of the draws in our chain to be iid. Unfortunately, it is likely that there will be some serial correlation amongst the θ^t's, recalling that θ^t is the value of θ for time step t, resulting in a sampler which is inefficient. In order to compare the loss in efficiency, due to the use of a finite chain, for different MCMC samplers we can look at the effective sample size (ESS). The ESS is defined as $n/\kappa(h)$, where n is the number of iterations of the Markov chain and $\kappa(h) = 1 + 2 \sum_{t=1}^{\infty} \text{corr}(h(\theta^0), h(\theta^t))$, the autocorrelation associated with the sequence of $h(\theta^t)$ for some function $h(.)$ (Robert and Casella, 2010). The autocorrelation

$\kappa(h)$, can be estimated using the R function `spectrum0`, although, the ESS can be found directly by using the R function `effectiveSize` (both functions can be found in the `CODA` package). The ESS tells us the number of iid samples equivalent to a Markov chain of length n. Thus, the larger the ESS the more efficient the sampler. Although a large value for the ESS implies that the sampler is efficient at producing an iid sample, it is not useful if that sampler requires an extremely large amount of computation time. To consider both of these criteria in a single measure we can simply estimate the efficiency in terms of the number of iid samples taken per second (i.e ESS/run time). In order for the efficiency of each sampler to be comparable each chain must be run for the same number of iterations.

3.2.5 Monte-Carlo importance sampling for under-identified models

A novel approach to determine the posterior distribution for under-identified models was recently proposed by Gustafson (2015). His approach makes use of the general strategy of importance sampling. The idea behind importance sampling is to draw samples from the "wrong", but convenient, joint posterior distribution and then to correct for having choosen from the "wrong" distribution by multiplying by an appropriate weighted factor (Kahn, 1955). When the model is under-identified first an appropriate transparent reparameterisation must be found, where the distribution for the data depends only on Φ_I and not Φ_N, such that the joint prior density, $p(\Phi_I, \Phi_N)$, can be evaluated. A convenient prior density, $p^*(\Phi_I, \Phi_N)$, can be selected by specifying a marginal density for $p^*(\Phi_I)$ that makes sampling easy, and then specifying the conditional density $p^*(\Phi_N|\Phi_I)$. A Monte Carlo sample of size n, denoted $(\Phi_I^i,\ \Phi_N^i)$ for $i = 1, \ldots, n$, can then be drawn from the posterior distribution arising from the convenience prior, $p^*(\Phi_I, \Phi_N) = p^*(\Phi_I)p^*(\Phi_N|\Phi_I)$. Adjusting $p^*(\Phi_I, \Phi_N)$ by applying the weights:

$$w_i \propto \frac{p(\Phi_I^i, \Phi_N^i)}{p^*(\Phi_I^i)p^*(\Phi_N^i|\Phi_I^i)} \tag{3.21}$$

scaled such that $\sum_{i=1}^{N} w_i = 1$, will represent the desired posterior distribution $p(\Phi_I, \Phi_N|x)$. See section 3.2.6.6 for implementation details. If $p^*(\Phi_I)$ is chosen to be a conjugate prior for the "embedded" identifiable model and $p^*(\Phi_N|\Phi_I)$ a standard distribution, then direct iid sampling from $p^*(\Phi_I, \Phi_N|x)$ is possible (Gustafson, 2015). This simplification will result in much shorter computational run times, and in turn efficiency, over MCMC methods.

3.2.6 Simulation study: measurement error example

The ultimate aim of this simulation study is to be able to propose a sampling method (or methods) which will provide accurate posterior distributions for the parameters (p, q, e, Se and Sp) from the under-identified model (3.4) specified on the leptospirosis data, so that the PAR, PAF and their credible intervals can be estimated. Each of the sampling methods described over the previous sections were applied for samples of size 380, 3800 and 38000 (i.e. the leptospirosis data where each entry in Table 2.1 is multiplied by 1, 10 or 100 respectively), since it has been reported that sample size can affect the convergence of a Markov chain when the model is under-identified (Johnson et al., 2001). This effect due to increasing sample size is a result of the narrowing of the posterior ridge as it tends towards the LPD. BGR analysis was performed to assess the convergence of each method and a total of $100,000$ iterations, including burn in, for each sampler was carried out. A tuning period was implement pre-simulation for each method, for every sample size, in the hope of there being no disadvantage due to a poor choice in initial conditions. The tuning parameters selected (excluding HMC) are presented in Table 3.2. Recall from Section 3.2.2 that the prior distributions assigned to the parameters of interest for the leptospirosis data are:

$$p, q \sim \text{Beta}(1, 1) \qquad e \sim \text{Beta}(2, 2) \tag{3.22}$$

$$Se \sim \text{Beta}(25, 3) \qquad Sp \sim \text{Beta}(30, 1.5). \tag{3.23}$$

$n = 380$	τ	c
MH-random walk	NA	2.15
MALA	NA	0.2
MH-$\Sigma^* = c[\tau I + (J^T DJ + p''(\theta))]^{-1}$	0.1	0.5
MH-$\Sigma^* = c[\tau I + J^T DJ]^{-1}$	10	0.1
MH-$\Sigma^* = c[\tau I + J^T J]^{-1}$	0.2	0.00075
$n = 3,800$		
MH-random walk	NA	2.15
MALA	NA	0.00675
MH-$\Sigma^* = c[\tau I + (J^T DJ + p''(\theta))]^{-1}$	0.1	0.5
MH-$\Sigma^* = c[\tau I + J^T DJ]^{-1}$	20	0.1
MH-$\Sigma^* = c[\tau I + J^T J]^{-1}$	0.1	0.00009
$n = 38,000$		
MH-random walk	NA	2.15
MALA	NA	0.001
MH-$\Sigma^* = c[\tau I + (J^T DJ + p''(\theta))]^{-1}$	0.1	0.3
MH-$\Sigma^* = c[\tau I + J^T DJ]^{-1}$	40	0.1
MH-$\Sigma^* = c[\tau I + J^T J]^{-1}$	0.005	0.000005

Table 3.2: Tuning parameters for MCMC approaches (excluding HMC), where MH-$\Sigma^* = c[\tau I + (J^T I_E(\hat{\theta})J + p''(\theta))]^{-1}$ represents the adjusted random walk sampler with Σ^* given by (3.20), MH-$\Sigma^* = c[\tau I + J^T I_E(\hat{\theta})J]^{-1}$ by (3.19) and MH-$\Sigma^* = c[\tau I + J^T J]^{-1}$ by (3.19). Note the random walk approach was implemented component-wise where the value for c remained the same for each of the 5 parameters in θ for both the the standard sampler and independence sampler. Additionally, $D = I_E(\hat{\theta})$ in Σ^* for the adapted MH-random walk approaches.

Before discussing the results of the simulation study we first provide additional details for the implementation of the different sampling approaches outlined in Sections 3.2.3 and 3.2.5 specific to the leptospirosis example.

3.2.6.1 Metropolis-Hastings random walk sampler and Independence sampler

In order to implement the random walk algorithm initial conditions for the parameters of interest $\theta = (p, q, e, Se, Sp)$ must be specified. To initalise Se and Sp we simply draw these from their prior distributions (3.23). Once values for Se and Sp have been drawn and η calculated based on the observed data, the system of equations (3.1) can be solved for π (using LU decomposition) and thus p, q and e according to (1.10). Note that if all four equations in the system (3.1) are

used when solving for π, as opposed to only the three independent equations, then it needs to be checked that $\sum_{i=1}^{2}\sum_{j=1}^{2}\pi_{ij} = 1$, as it is possible for $\sum_{i=1}^{2}\sum_{j=1}^{2}\pi_{ij} > 1$. This situation does not occur when only three independent equations are used as π_{22} is calculated according to $1 - \pi_{11} - \pi_{12} - \pi_{22}$.

For this example we implement the random walk approach component wise with a proposal distribution $Q(\theta^{\dagger}|\theta^{t}) \sim \text{Normal}(\theta^{t}, c\sigma^{*})$, where σ^{*} and c were is assigned according to Christensen et al. (2010). An initial pre-simulation period was run where arbitrary values for σ^2 and c were assigned for each parameter. After 1,000 iterations of the chain had been carried out the standard deviation for those iterations was calculated and set equal to σ^{*}. The tuning parameter c was then adjusted to achieve and acceptance rate around 20-50%. The tuning parameters used are provided in Table 3.2. As σ varied for each of the simulations of differing sample size (i.e. $n = \{380, 3800, 38000\}$) the value of $c = 2.15$ provided an acceptance rate around 20-50% in each case.

We carry out two different methods for the Metropolis-Hastings random walk approach on the leptospirosis data. The first is a random walk making use of the priors proposed in (3.22 - 3.23) for θ. Whereas, in the second method priors for θ are set equal to the proposal distribution (i.e. independence sampling).

3.2.6.2 MALA: Deriving $\nabla \log(p(\theta))$

Initialisation of θ for the MALA is performed following the exact same procedure as that carried out for the random walk approach (see Section 3.2.6.1). In order to perform an iteration of the MALA we are required to define the gradient function for the target distribution. Given the priors (3.22-3.23) on p, q, e, Se and Sp along with the log-likelihood for the model (3.2.2)

$$l(\eta|x) = \sum_{i=1}^{2}\sum_{j=1}^{2} x_{ij} \log \eta_{ij} + K, \tag{3.24}$$

where K represents a normalising constant, the joint posterior distribution can be described by

$$p(\theta) = \eta_{11}^{x_{11}}\eta_{12}^{x_{12}}\eta_{21}^{x_{21}}\eta_{22}^{x_{22}}(1-Se)^{24}Sp^{29}(1-Sp)^{0.5}e(1-e), \tag{3.25}$$

and the log joint posterior by

$$\begin{aligned}\log(p(\theta)) = x_{11}\log(\eta_{11}) + x_{12}\log(\eta_{12}) + x_{21}\log(\eta_{21}) + x_{22}\log(\eta_{22}) + 24\log(Se) + \\ 2\log(1-Se) + 29\log(Sp) + 0.5\log(1-Sp) + log(e) + log(1-e).\end{aligned} \tag{3.26}$$

The gradient vector is then simply the 5×1 matrix made up of the partial derivatives of (3.26) with respect to each element of θ:

$$\frac{\partial \log(p(\theta))}{\partial p} = \frac{eSex_{11}}{\eta_{11}} - \frac{Sex_{12}e}{\eta_{12}} + \frac{(1-Se)ex_{21}}{\eta_{21}} + \frac{e(Se-1)x_{22}}{\eta_{22}}, \tag{3.27}$$

$$\frac{\partial \log(p(\theta))}{\partial q} = \frac{(1-Sp)(1-e)x_{11}}{\eta_{11}} + \frac{(1-Sp)(e-1)x_{12}}{\eta_{12}} + \frac{Sp(1-e)x_{21}}{\eta_{21}} + \frac{Sp(e-1)x_{22}}{\eta_{22}}, \tag{3.28}$$

$$\begin{aligned}\frac{\partial \log(p(\theta))}{\partial e} = &\frac{[pSe + q(Sp-1)]x_{11}}{\eta_{11}} + \frac{[Se(1-p) + (1-Sp)(q-1)]x_{12}}{\eta_{12}} + \\ &\frac{[(1-Se)p - qSp]x_{21}}{\eta_{21}} + \frac{[(1-Se)(1-p) + Sp(q-1)]x_{22}}{\eta_{22}} + \frac{1}{e} - \frac{1}{1-e},\end{aligned} \tag{3.29}$$

$$\frac{\partial \log(p(\theta))}{\partial Se} = \frac{pex_{11}}{\eta_{11}} + \frac{(1-p)ex_{12}}{\eta_{12}} - \frac{pex_{21}}{\eta_{21}} + \frac{e(p-1)x_{22}}{\eta_{22}} + \frac{24}{Se} - \frac{2}{(1-Se)}, \tag{3.30}$$

$$\begin{aligned}\frac{\partial \log(p(\theta))}{\partial Sp} = &\frac{q(e-1)x_{11}}{\eta_{11}} + \frac{(1-q)(e-1)x_{12}}{\eta_{12}} + \frac{(1-e)qx_{21}}{\eta_{21}} + \frac{(1-q)(1-e)x_{22}}{\eta_{22}} + \\ &\frac{29}{Sp} - \frac{1}{2(1-Sp)}.\end{aligned} \tag{3.31}$$

3.2.6.3 HMC: Defining $\nabla U(q)$, ϵ and L

Similarly to the MALA, to implement the HMC algorithm we need to incorporate gradient information for the target distribution, $\nabla U(q)$. Given that the potential energy $U(q)$ can be described by minus the log-likelihood of the posterior distribution for the parameters of interest

θ, $\nabla U(q)$ simply becomes the 5×1 matrix of negated partial derivatives given by (3.27-3.31). Additionally, the initialisation of the position vector q for the HMC algorithm is carried out following the exact same procedure as that of the MALA and random walk.

Tuning the HMC algorithm for the leptospirosis model (3.4) was particularly difficult. In order to achieve an acceptance rate around 20-50% a very small step size, ϵ, was needed. As the sample size increased the ϵ required to achieve the desired acceptance became smaller still. This resulted in the chains for samples of size $n = 3800$ and $n = 38000$, being unable to converge within the $100,000$ iterations performed. For $n = 3800$ an $\epsilon = 1.0 \times 10^{-5}$ and $L = 5000$ achieved an acceptance rate of 33.5%, whereas for $n = 38000$ an $\epsilon = 3.0 \times 10^{-7}$ and $L = 5,000$ achieved and acceptance rate of 49.7%. Although, the desired acceptance rate is achieved the small ϵ means that a very large number of iterations would be required to explore the entirety of the parameter space and for the chain to converge. For $n = 380$ a much larger step size ($\epsilon = 0.02$) and smaller number of leapfrog steps ($L = 50$) could be adopted. This resulted in an acceptance rate of 63.8% and a chain which appeared to converge based on visual inspection of trace plots. Due to the difficulty of tuning the HMC algorithm for large sample sizes, the leptospirosis model analysis was also carried out using the R interface to the Stan software (Stan Development Team, 2017), which uses the adaptive HMC algorithm that attempts to automatically tune the HMC algorithm. Interestingly, the Stan implementation of HMC also failed to tune for our particular model, even for $n = 380$.

3.2.6.4 Gibbs sampler: the full joint conditional distributions

For our particular example in order to determine the full conditionals for each parameter from the model (3.4), we follow the approach of Joseph et al. (1995) by introducing latent variables. A latent variable is one that cannot be observed itself, but may be inferred from other variables. Let Y_{ij} and Z_{ij}, where $i, j \in \{1, 2\}$ (as seen in Tables 3.4 and 3.3), be latent variables which represent the number of subjects that are correctly and incorrectly classified. Additionally, it

must hold that:

$$X_{11} = Z_{21} + Y_{11} \qquad X_{12} = Z_{22} + Y_{12} \tag{3.32}$$

$$X_{21} = Z_{11} + Y_{21} \qquad X_{22} = Z_{12} + Y_{22}, \tag{3.33}$$

recalling that X_{ij} for $i, j \in \{1, 2\}$ is what was actually observed. We can now express the likelihood for our model (3.4) in terms of the latent variables $Y = (Y_{11}, Y_{12}, Y_{21}, Y_{22})$ and $Z = (Z_{11}, Z_{12}, Z_{21}, Z_{22})$ as follows:

$$\begin{aligned} L(X|Y, \pi, Se, Sp) \propto (\pi_{11}Se)^{Y_{11}}(\pi_{12}Se)^{Y_{12}}(\pi_{21}Sp)^{Y_{21}}(\pi_{22}Sp)^{Y_{22}} \times \\ [(1-Sp)\pi_{12}]^{Z_{11}}\,[(1-Sp)\pi_{22}]^{Z_{12}}[(1-Se)\pi_{11}]^{Z_{21}}\,[(1-Se)\pi_{12}]^{Z_{22}}, \end{aligned} \tag{3.34}$$

where $Z_{21} = X_{11} - Y_{11}$, $Z_{22} = X_{12} - Y_{12}$, $Z_{11} = X_{21} - Y_{21}$ and $Z_{12} = X_{22} - Y_{22}$. Given we apply a Dirichlet$(1, 1, 1, 1)$ prior on π, it follows that the conditional posterior for the true probabilities, π, in the table is:

$$p(\pi|X, Y, Se, Sp) \sim \text{Dirichlet}(Y_{11} + Z_{11} + 1, Y_{12} + Z_{12} + 1, Y_{21} + Z_{21} + 1, Y_{22} + Z_{22} + 1).$$

Given the beta priors (3.23) on Se and Sp, the conditional posterior for Se and Sp are also beta:

$$p(Se|X, Y, \pi, Sp) \sim \text{Beta}(Y_{11} + Y_{12} + \alpha_e, Z_{11} + Z_{12} + \beta_e)$$

$$p(Sp|X, Y, \pi, Se) \sim \text{Beta}(Y_{21} + Y_{22} + \alpha_p, Z_{21} + Z_{22} + \beta_p),$$

where $\alpha_e = 25$ and $\beta_e = 3$ which represent the shape parameters from the beta prior on Se and similarly $\alpha_p = 30$ and $\beta_p = 1.5$ which represent the shape parameters from the beta prior on Sp. Finally for the latent variables Y_{ij} the conditional posterior distributions are binomial:

$$P(Y_{11}|\pi, X, Se, Sp) \sim \text{Binomial}\left(X_{11}, \frac{\pi_{11}Se}{\pi_{11}Se + (1-Sp)\pi_{21}}\right)$$

	Diseased	
Exposure	D^+	D^-
E^+	Y_{11}	Y_{12}
E^-	Y_{21}	Y_{22}

Table 3.3: Subjects correctly classified

	Diseased	
Exposure	D^+	D^-
E^+	Z_{11}	Z_{12}
E^-	Z_{21}	Z_{22}

Table 3.4: Subjects incorrectly classified

$$
\begin{aligned}
P(Y_{12}|\pi, X, Se, Sp) &\sim \text{Binomial}\left(X_{12}, \frac{\pi_{12}Se}{\pi_{12}Se + (1-Sp)\pi_{22}}\right) \\
P(Y_{21}|\pi, X, Se, Sp) &\sim \text{Binomial}\left(X_{21}, \frac{\pi_{21}Sp}{(1-Se)\pi_{12} + \pi_{21}Sp}\right) \\
P(Y_{22}|\pi, X, Se, Sp) &\sim \text{Binomial}\left(X_{22}, \frac{\pi_{21}Sp}{(1-Se)\pi_{12} + \pi_{21}Sp}\right).
\end{aligned}
$$

The conditional posterior for the latent variables Z_{ij} are also binomial, but in practice it is more efficient to determine Z_{ij} using the relationships (3.32-3.33). Initialisation of the Gibbs sampling procedure followed the same approach as that of the Metropolis-Hastings random walk.

3.2.6.5 Adapted Metropolis-Hastings random walk: deriving Σ^*

In Section 3.2.3.5 we introduced three different alternatives for the covariance matrix, Σ^*, to be incorporated into the proposal distribution. To derive Σ^* given by (3.18) we first determine the Jacobian matrix for the model (3.1), which is given by

$$
\begin{bmatrix}
eSe & (Sp-1)(e-1) & pSe + q(Sp-1) & ep & q(e-1) \\
-eSe & -(Sp-1)(e-1) & -Se(p-1) - (Sp-1)(q-1) & -e(p-1) & -(e-1)(q-1) \\
-e(Se-1) & -Sp(e-1) & -qSp - p(Se-1) & -ep & -q(e-1) \\
e(Se-1) & Sp(e-1) & (1-Se)(1-p) + Sp(q-1) & e(p-1) & (1-q)(1-e)
\end{bmatrix}
\tag{3.35}
$$

where the rows represent the observed probabilities η_{11}, η_{12}, η_{21} and η_{22} and columns the parameters in the following order: p, q, e, Se and Sp. Secondly, recall that the log-likelihood for

the transparent reparameterised model (3.4) is given by (3.24). Taking the derivative of (3.24) with respect to the parameter vector θ using the chain rule provides

$$\frac{\partial\, l(\eta|x)}{\partial\theta} = \frac{\partial\, l(\eta|x)}{\partial\eta}\frac{\partial\eta}{\partial\theta}, \tag{3.36}$$

where $\frac{\partial\, l(\eta|x)}{\partial\eta}$ is a 1×4 matrix made up of x_{ij}/η_{ij} for $i,j\in\{1,2\}$, and $\frac{\partial\eta}{\partial\theta}=J$, the 4×5 Jacobian matrix given by (3.35). Taking the second derivative of (3.24) using the chain rule again we get

$$\frac{\partial^2\, l(\eta|x)}{\partial\theta\partial\theta^t} = J^T\frac{\partial^2\, l(\eta|x)}{\partial\eta^2}J, \tag{3.37}$$

where $\frac{\partial^2\, l(\eta|x)}{\partial\eta^2}$ represents the 4×4 diagonal matrix $I_E(\hat{\theta})$, which for this particular example has diagonal elements (n^2/x_{ij}) for $i,j\in\{1,2\}$. To derive Σ^* given by (3.20) we simply add the 5×5 diagonal matrix $\frac{d^2p(\theta)}{d\theta d\theta^T}$ to (3.37). Given the priors for this example are described by the beta distributions the diagonal components of $\frac{d^2p(\theta)}{d\theta d\theta^T}$ are expressed as $-[\alpha_i/\theta_i^2]-[\beta_i/(1-\theta_i)^2]$.

Similarly to the standard random walk approach, initialisation of the parameters θ are preformed following the same procedure. Furthermore, tuning of each of these approaches was conducted by carrying out an initial simulation to determine which tuning parameters provided an acceptance rate between 20-50%.

3.2.6.6 Monte-Carlo importance sampling

As discussed in Section 3.2 the model (1.10) for the leptospirosis data expressed in terms of the original parameter vector $\theta=(\pi_{11},\pi_{12},\pi_{21},Se,Sp)$ can be described by the transparent parameterisation to $\Phi=(\eta_{11},\eta_{12},\eta_{21},Se,Sp)$ according to (3.2) and modeled via (3.4). To calculate the importance sampling weights, w_i, we propose making an adaptation to (3.21) proposed by Gustafson (2015), which uses the product of the marginal prior for the identifiable part and the conditional prior for the non-identifiable part. The problem with implementing this is that the constraints on the conditional prior depend on the values of the identifiable part, so the

normalizing constant in the conditional prior is a complex function of η. By replacing Gustafson's $p^*(\Phi_I^i)p^*(\phi_N^i|\Phi_I^i)$ by an overall prior $p^*(\Phi)$ we can avoid this problem as the normalising constant is now fixed. Thus w_i becomes

$$w_i \propto \frac{p(\Phi)}{p^*(\Phi)}, \tag{3.38}$$

where $p(\Phi)$ is the prior distribution induced on Φ by the actual prior on θ and $p^*(\Phi)$ the convenience prior specified on Φ. The prior for Φ must be restricted to the set of values of Φ, say A, for which $\pi_{ij} \in [0,1]$. We take the convenience prior for Φ as

$$p^*(\Phi) \propto Se^{24}(1-Se)^2 Sp^{29}(1-Sp)^{0.5}\mathbb{1}(A), \tag{3.39}$$

where $\mathbb{1}$ represents the indicator function which is 1 when $\pi_{ij} \in [0,1]$ and 0 otherwise. By specifying the convenience prior on Φ by (3.39), we find that the full posterior density is

$$p^*(\Phi|x) \propto \eta_{11}^{x_{11}}\eta_{12}^{x_{12}}\eta_{21}^{x_{21}}(1-\eta_{11}-\eta_{12}-\eta_{21})^{x_{21}} Se^{24}(1-Se)^2 Sp^{29}(1-Sp)^{0.5}\mathbb{1}(A). \tag{3.40}$$

We can sample from this posterior by drawing η from a Dirichlet$(x_{11}+1,\ldots,x_{22}+1)$, sampling Se and Sp from their beta priors (3.23), and rejecting any parameter sets that fail to satisfy the constraints $\pi_{ij} \in [0,1]$. The normalizing constant is now marginalized so can be ignored in the importance weights. The prior induced on Φ by the actual prior on θ is given by

$$p(\Phi) \propto Se^{24}(1-Se)^2 Sp^{29}(1-Sp)^{0.5}|J|^{-1}\mathbb{1}(A), \tag{3.41}$$

where the Jacobian, $|J| = |\partial\Phi/\partial\theta|$, is given by

$$\begin{vmatrix} Se & 0 & (1-Sp) & \pi_{11} & -\pi_{21} \\ -(1-Sp) & Se+Sp-1 & -(1-Sp) & \pi_{12} & -(1-\pi_{11}-\pi_{12}-\pi_{21}) \\ (1-Se) & 0 & Sp & -\pi_{11} & \pi_{21} \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{vmatrix}$$

$$= (Se + Sp - 1)^2.$$

Note that the restriction of Φ to A are still enforced by the restriction $\pi_{ij} \in [0, 1]$. The importance weights in (3.21) are thus $(Se+Sp-1)^{-2}$. The LPD can also be derived through this importance sampling procedure by fixing η to be the observed probabilities according to Table 3.1.

3.2.7 Simulation results

The acceptance rates for each of the sampling methods performed on the sample sizes $n = 380$, $n = 3800$ and $n = 38000$ are given in Table 3.5. As expected the acceptance rates for the independence sampler are particularly low for some parameters (i.e. q and e). The acceptance rates are so low in fact that when the sample size reaches $n = 38000$ the low acceptance impedes the convergence of the Markov chain. The acceptance rate less than 100% for the importance sampling approach, proposed by Gustafson (2015), suggests that approximately 12% of the time a solution for $\pi_{ij} \notin [0, 1]$. Tuning the approaches with proposal variances involving J presented difficulties, especially as the sample size increased. Specifically when $\Sigma^* = c[\tau I + J^T J]^{-1}$ both τ and c need to be very small in order for the proposed $\theta^\dagger$ to be accepted at all. The HMC algorithm also posed difficulties when it came to tuning for the sample sizes $n = 3800$ and $n = 38000$. In an attempt to avoid the problem of tuning, our example was implemented on the R platform of the Stan software, which aims to automatically tune the HMC algorithm using adaptive methods. Unfortunately though our example was unable to be tuned in the Stan

implementation of the HMC algorithm for all sample sizes.

Table 3.6 provides a comparison of the ESS per $1,000$ iterations for each of the different samplers. What is overwhelmingly clear from Table 3.6 is that the importance sampling based method vastly out performs the MCMC methods. Even as the sample size becomes large (i.e. $n = 38000$) the importance sampling approach provides a similar ESS to that seen when the sample size is $n = 380$. In terms of computational efficiency the importance sampling approach is greatly superior, as can be seen in Table 3.7. The downside of this approach in general however, is the need for a transparent re-parameterization. If this re-parameterization is too difficult to determine (e.g. due to high dimensionality) and an MCMC approach adopted, then the choice of sampler should be based on the sample size.

It can be seen that when $n = 380$ the HMC sampler performs better than the other MCMC algorithms in terms of ESS for almost every parameter. However, Table 3.7 shows that this superior ESS comes at the cost of increased computational effort, in comparison to the random walk and Gibbs sampling approaches. The random walk and Gibbs sampler perform less well than HMC in terms of ESS for $n = 380$, but better than the other MCMC methods investigated. Given their superior efficiency at $n = 380$ and the ease with which they can be implemented, the random walk or Gibbs sampling approaches may be a viable option if a transparent parameterisation can not be found for implementation of Gustafson's approach or tuning an HMC algorithm presents difficulties. As the sample size increases the performance of the random walk and Gibbs sampler diminishes dramatically. This dramatic reduction in performance (in terms of ESS) for the random walk and Gibbs sampler occurs because as $n \to \infty$ the posterior ridge becomes narrower as it tends to the LPD. Figure 3.2 shows how the posterior distribution tends towards the LPD for the PAR and PAF, for selected samplers, as the sample size increases from $n = 380$ to $n = 38000$. The relatively wider posterior distribution we get when the sample size is small allows for larger steps in any direction to be taken without moving off the ridge. Moreover, Table 3.8 provides the posterior estimates of the mean and median PAR and PAF with their corresponding credible intervals for each sampler.

$n = 380$	p	q	e	Se	Sp	PAR	PAF
MC importance sampling	87.2	87.2	87.2	87.2	87.2	87.2	87.2
MH-independent proposal	31.7	7.6	3.4	82.2	46.3	39.1	39.1
MH-random walk	43.1	45.2	30.4	42.3	30.3	78.3	78.3
MALA	60.0	60.0	60.0	60.0	60.0	60.0	60.0
HMC	63.8	63.8	63.8	63.8	63.8	63.8	63.8
MH-$\Sigma^* = c[\tau I + (J^T DJ + p''(\theta))]^{-1}$	28.2	28.2	28.2	28.2	28.2	28.2	28.2
MH-$\Sigma^* = c[\tau I + J^T DJ]^{-1}$	25.7	25.7	25.7	25.7	25.7	25.7	25.7
MH-$\Sigma^* = c[\tau I + J^T J]^{-1}$	21.0	21.0	21.0	21.0	21.0	21.0	21.0
$n = 3,800$							
MC importance sampling	87.5	87.5	87.5	87.5	87.5	87.5	87.5
MH-independent proposal	10.7	2.0	1.0	56.3	18.3	13.7	13.7
MH-random walk	16.8	42.9	15.2	33.7	15.2	59.8	59.8
MALA	57.1	57.1	57.1	57.1	57.1	57.1	57.1
HMC							
MH-$\Sigma^* = c[\tau I + (J^T DJ + p''(\theta))]^{-1}$	23.7	23.7	23.7	23.7	23.7	23.7	23.7
MH-$\Sigma^* = c[\tau I + J^T DJ]^{-1}$	28.0	28.0	28.0	28.0	28.0	28.0	28.0
MH-$\Sigma^* = c[\tau I + J^T J]^{-1}$	24.2	24.2	24.2	24.2	24.2	24.2	24.2
$n = 38,000$							
MC importance sampling	87.4	87.4	87.4	87.4	87.4	87.4	87.4
MH-independent proposal							
MH-random walk	32.9	40.4	27.9	26.6	32.6	71.7	71.7
MALA							
HMC							
MH-$\Sigma^* = c[\tau I + (J^T DJ + p''(\theta))]^{-1}$	22.9	22.9	22.9	22.9	22.9	22.9	22.9
MH-$\Sigma^* = c[\tau I + J^T DJ]^{-1}$	23.8	23.8	23.8	23.8	23.8	23.8	23.8
MH-$\Sigma^* = c[\tau I + J^T J]^{-1}$	28.2	28.2	28.2	28.2	28.2	28.2	28.2

Table 3.5: Percent acceptance rates for each method with a chain of length $100,000$. Acceptance rates are removed from the table when the method does not converge within the $100,000$ iterations for all parameters according to the BGR diagnostic, for the specified sample size, or when the method could not be tuned. Note that the Gibbs sampler is not included here as the acceptance probability is 1 and $D = I_E(\hat{\theta})$ in Σ^* for the adapted MH-random walk approaches. Additionally, for methods where block-wise updating has been adopted the acceptance rate will be the same for all parameters.

$n = 380$	p	q	e	Se	Sp	PAR	PAF
MC importance sampling	872.5	875.6	872.5	881.3	872.5	861.2	872.5
MH-independent proposal	17.5	47.6	7.4	94.7	10.5	67.7	59.1
MH-random walk	50.9	210.7	36.5	144.4	33.8	186.2	177.4
Gibbs sampler	61.3	704.7	50.4	204.3	43.8	359.6	354.4
MALA	3.4	114.4	12.5	14.3	10.8	15.1	17.5
HMC	225.3	165.2	127.9	335.2	133.0	144.6	163.3
MH-$\Sigma^* = c[\tau I + (J^T DJ + p''(\theta))]^{-1}$	28.2	29.3	26.7	29.7	26.4	30.4	29.8
MH-$\Sigma^* = c[\tau I + J^T DJ]^{-1}$	9.6	6.8	12.4	62.4	22.1	6.0	6.0
MH-$\Sigma^* = c[\tau I + J^T J]^{-1}$	9.4	33.4	23.3	34.3	19.2	30.0	30.1
$n = 3,800$							
MC importance sampling	874.6	874.6	874.6	874.6	874.6	864.4	866.3
MH-independent proposal	0.9	12.9	0.4	12.7	0.8	13.5	11.7
MH-random walk	2.7	106.4	3.0	34.7	2.7	46.5	46.6
Gibbs sampler	4.9	154.6	5.1	27.8	4.6	57.5	57.1
MALA	0.7	41.0	1.6	1.8	1.5	8.9	9.0
HMC							
MH-$\Sigma^* = c[\tau I + (J^T DJ + p''(\theta))]^{-1}$	8.7	24.5	10.0	22.7	9.5	24.2	24.1
MH-$\Sigma^* = c[\tau I + J^T DJ]^{-1}$	5.9	8.0	7.9	39.0	7.5	10.2	10.3
MH-$\Sigma^* = c[\tau I + J^T J]^{-1}$	3.5	40.2	5.5	11.4	4.7	24.0	23.5
$n = 38,000$							
MC importance sampling	886.6	874.0	874.0	865.1	884.7	876.7	876.8
MH-independent proposal							
MH-random walk	0.5	7.8	0.6	5.2	0.6	5.2	5.2
Gibbs sampler	0.5	5.5	0.6	3.2	0.5	3.6	3.6
MALA							
HMC							
MH-$\Sigma^* = c[\tau I + (J^T DJ + p''(\theta))]^{-1}$	1.8	13.2	2.7	12.1	2.4	12.2	12.2
MH-$\Sigma^* = c[\tau I + J^T DJ]^{-1}$	1.9	8.7	2.8	14.2	2.8	11.6	10.8
MH-$\Sigma^* = c[\tau I + J^T J]^{-1}$	3.6	23.3	3.5	16.0	5.1	18.7	18.4

Table 3.6: Effective sample size (ESS) per 1,000 iterations. ESS values are removed from the table when the method does not converge within $100,000$ iterations according to the BGR diagnostic, for the specified sample size, or when the method could not be tuned. Note that $D = I_E(\hat{\theta})$ in Σ^* for the adapted MH-random walk approaches.

$n = 380$	p	q	e	Se	Sp	PAR	PAF
MC importance sampling	300.0	301.1	300.0	303.1	300.0	296.3	300.0
MH-independent proposal	9.7	26.5	4.1	52.6	5.8	37.6	32.8
MH-random walk	42.4	175.6	30.4	120.3	28.2	155.1	147.8
Gibbs sampler	36.7	422.0	30.2	122.3	26.2	215.3	212.2
MALA	5.5	184.5	20.2	23.1	17.4	24.4	28.3
HMC	41.2	30.2	23.4	61.3	24.3	26.4	29.9
MH-$\Sigma^* = c[\tau I + (J^T DJ + p''(\theta))]^{-1}$	23.7	24.7	22.4	25.0	22.2	25.5	25.0
MH-$\Sigma^* = c[\tau I + J^T DJ]^{-1}$	8.5	6.0	10.9	54.8	19.4	5.3	5.3
MH-$\Sigma^* = c[\tau I + J^T J]^{-1}$	7.0	24.9	17.4	25.6	26.4	22.4	22.5
$n = 3,800$							
MC importance sampling	321.3	321.3	321.3	321.3	321.3	317.6	317.2
MH-independent proposal	2.9	41.6	1.3	41.0	2.4	43.7	37.7
MH-random walk	2.4	97.6	2.8	31.9	2.5	42.7	42.7
Gibbs sampler	1.5	48.9	1.6	8.8	1.5	18.2	18.0
MALA	1.1	62.1	2.4	2.8	2.3	13.4	13.6
HMC							
MH-$\Sigma^* = c[\tau I + (J^T DJ + p''(\theta))]^{-1}$	8.5	24.0	9.9	22.2	9.3	23.7	23.6
MH-$\Sigma^* = c[\tau I + J^T DJ]^{-1}$	4.3	5.8	5.8	28.4	5.5	7.5	7.6
MH-$\Sigma^* = c[\tau I + J^T J]^{-1}$	3.1	35.9	4.9	10.2	4.2	21.4	17.2
$n = 38,000$							
MC importance sampling	336.9	332.1	332.1	328.7	336.0	333.1	333.2
MH-independent proposal							
MH-random walk	0.1	2.3	0.2	1.5	0.2	1.5	1.5
Gibbs sampler	0.1	1.1	0.1	0.6	0.1	0.7	0.7
MALA							
HMC							
MH-$\Sigma^* = c[\tau I + (J^T DJ + p''(\theta))]^{-1}$	3.0	19.2	2.9	13.2	4.2	10.2	10.2
MH-$\Sigma^* = c[\tau I + J^T DJ]^{-1}$	1.6	7.2	2.3	11.8	2.3	10.1	9.5
MH-$\Sigma^* = c[\tau I + J^T J]^{-1}$	1.7	12.8	2.6	11.8	2.3	14.0	13.8

Table 3.7: Effective samples performed per second (i.e. method efficiency). Efficiency values are removed from the table when the method does not converge within the $100,000$ iterations according to the BGR diagnostic, for the specified sample size, or when the method could not be tuned. Note that $D = I_E(\hat{\theta})$ in Σ^* for the adapted MH-random walk approaches.

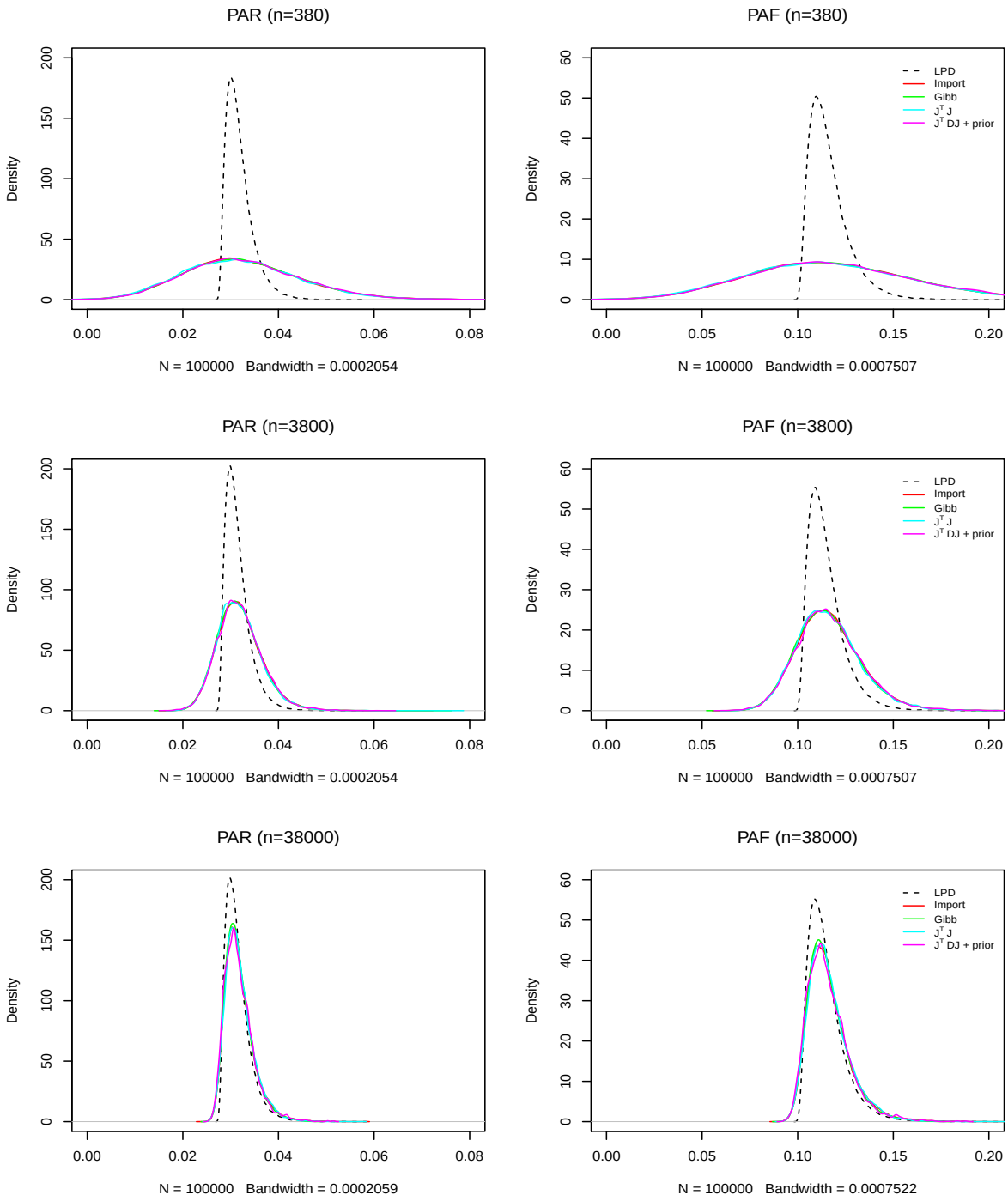

Figure 3.2: Density plots for PAR and PAF comparing selected samplers with the LPD for differing sample sizes. Note that Import = Gustafson (2015) importance sampler, Gibb = Gibbs sampler, J^TJ = the adapted MH random walk sampler with $\Sigma^* = c[\sigma I + J^TJ]$ and J^TDJ + prior = the adapted MH random walk sampler with $\Sigma^* = c[\tau I + (J^T I_E(\hat{\theta})J + p''(\theta))]^{-1}$ and LPD = limiting posterior distribution.

The MALA algorithm performs poorly in all situations and may not be a wise choice of sampler for analysing an under-identified model. It is particularly useful at reaching a point of high likelihood on the posterior ridge. However, as the ridge is a set of values of equal likelihood the algorithm gets stuck and can not explore the entirety of the ridge. This phenomenon becomes particularly apparently when $n = 38000$ where the Markov chain was unable to converge within the $100,000$ iterations.

For the largest sample size, $n = 38000$, the adapted random walk approach with proposal covariance matrix $\Sigma^* = c[\sigma I + J^T J]$ is the preferred MCMC option, performing slightly better in terms of ESS than all other MCMC based approaches. The elliptical shape of the proposal distribution appears to help the chain with exploring along the posterior ridge, although the low ESS suggests there is still a large amount of autocorrelation in the chain. The adapted random walk methods with proposal covariance $\Sigma^* = c[\tau I + (J^T I_E(\hat{\theta})J + p''(\theta))]^{-1}$ and $\Sigma^* = c[\tau I + J^T I_E(\hat{\theta})J]^{-1}$ performed only slightly poorer, in terms of the ESS, than when $\Sigma^* = c[\sigma I + J^T J]$. The adapted random walk methods with covariance $\Sigma^* = c[\tau I + (J^T I_E(\hat{\theta})J + p''(\theta))]^{-1}$ and $\Sigma^* = c[\tau I + J^T I_E(\hat{\theta})J]^{-1}$ however, are simpler to tune than when $\Sigma^* = c[\sigma I + J^T J]$ due to adapting to the sample size. The slightly improved performance of the adapted approach with $\Sigma^* = c[\sigma I_n + J^T J]$ at $n = 38000$ is likely a result of the smaller step size being taken. Computationally however, this approach can be quite intensive due to the matrix inversion required to provide the proposal variance, which is carried out for every iteration of the algorithm. For $n = 38000$ though the effective samples generated per second for this method out-perform all other Metropolis-Hastings based methods.

3.3 Case-control study example

As previously mentioned (see Section 1.2.2), case-control studies are conducted by selecting pre-assigned numbers of cases and controls. This means that the prevalence of disease can not be estimated from data arising from this type of study. If prior information is available on

	PAR				PAF			
$n = 380$	Mean	Median	95% CI	CI length	Mean	Median	95% CI	CI length
MC importance sampling	0.0325	0.0317	(0.011, 0.059)	0.048	0.118	0.115	(0.038, 0.213)	0.175
MH-independent proposal	0.0325	0.0316	(0.012, 0.060)	0.048	0.118	0.115	(0.039, 0.214)	0.175
MH-random walk	0.0325	0.0316	(0.011, 0.059)	0.048	0.118	0.115	(0.039, 0.214)	0.175
Gibbs sampler	0.0326	0.0318	(0.011, 0.059)	0.048	0.118	0.116	(0.039, 0.214)	0.175
MALA	0.0323	0.0316	(0.010, 0.059)	0.049	0.117	0.115	(0.037, 0.211)	0.174
HMC	0.0325	0.0317	(0.010, 0.059)	0.049	0.118	0.115	(0.039, 0.213)	0.174
MH-$\Sigma^* = c[\tau I + (J^T DJ + p''(\theta))]^{-1}$	0.0327	0.0318	(0.011, 0.059)	0.048	0.119	0.116	(0.039, 0.214)	0.175
MH-$\Sigma^* = c[\tau I + J^T DJ]^{-1}$	0.0325	0.0316	(0.010, 0.059)	0.049	0.118	0.115	(0.038, 0.210)	0.172
MH-$\Sigma^* = c[\tau I + J^T J]^{-1}$	0.0322	0.0315	(0.010, 0.059)	0.049	0.117	0.114	(0.037, 0.211)	0.174
$n = 3,800$								
MC importance sampling	0.0319	0.0315	(0.024, 0.042)	0.018	0.117	0.115	(0.087, 0.154)	0.067
MH-independent proposal	0.0315	0.0312	(0.024, 0.042)	0.018	0.115	0.114	(0.085, 0.152)	0.067
MH-random walk	0.0319	0.0315	(0.024, 0.043)	0.019	0.117	0.115	(0.087, 0.154)	0.067
Gibbs sampler	0.0318	0.0314	(0.024, 0.042)	0.018	0.116	0.115	(0.087, 0.154)	0.067
MALA	0.0316	0.0314	(0.024, 0.041)	0.017	0.116	0.115	(0.087, 0.155)	0.068
HMC								
MH-$\Sigma^* = c[\tau I + (J^T DJ + p''(\theta))]^{-1}$	0.0316	0.0316	(0.024, 0.042)	0.018	0.117	0.115	(0.086, 0.154)	0.068
MH-$\Sigma^* = c[\tau I + J^T DJ]^{-1}$	0.0323	0.0318	(0.024, 0.043)	0.019	0.118	0.116	(0.088, 0.156)	0.068
MH-$\Sigma^* = c[\tau I + J^T J]^{-1}$	0.0314	0.0314	(0.024, 0.042)	0.018	0.116	0.115	(0.087, 0.154)	0.067
$n = 38,000$								
MC importance sampling	0.0318	0.0313	(0.027, 0.039)	0.012	0.116	0.114	(0.100, 0.144)	0.044
MH-independent proposal								
MH-random walk	0.0316	0.0311	(0.027, 0.039)	0.012	0.115	0.114	(0.100, 0.142)	0.042
Gibbs sampler	0.0317	0.0312	(0.027, 0.039)	0.012	0.116	0.114	(0.100, 0.142)	0.042
MALA								
HMC								
MH-$\Sigma^* = c[\tau I + (J^T DJ + p''(\theta))]^{-1}$	0.0318	0.0312	(0.027, 0.039)	0.012	0.116	0.114	(0.099, 0.144)	0.045
MH-$\Sigma^* = c[\tau I + J^T DJ]^{-1}$	0.0318	0.0313	(0.027, 0.039)	0.012	0.116	0.114	(0.100, 0.144)	0.044
MH-$\Sigma^* = c[\tau I + J^T J]^{-1}$	0.0319	0.0313	(0.027, 0.039)	0.012	0.116	0.114	(0.100, 0.143)	0.043

Table 3.8: Mean and median estimates for the PAR and PAF with the corresponding 95% credible interval and its length. Estimates are omitted from the table when the sampler did not converge within 100,000 iterations or could not be tuned. Note that $D = I_E(\hat{\theta})$ in Σ^* for the adapted MH-random walk approaches.

the prevalence of disease, in the form of a prior distribution, then a Bayesian approach can be adopted. Alternatively, due to the relationship between exposure and disease a prior could be specified for the $P(E^+)$, which would then induce a prior on the prevalence of disease. Over the following sections we describe how each of these approaches could be implemented in order to derive the PAR, PAF and their credible intervals. As a numerical example we simply use the 2×2 Table 2.1 (i.e. leptospirosis data) and assume that the data had instead been collected using a case-control design.

3.3.1 Using prior information for $P(D^+)$

Recall from Section 1.2.2 that $\phi_1 = P(E^+|D^+)$ and $\phi_2 = P(E^+|D^-)$. Let $\phi_3 = P(D^+)$ and $\boldsymbol{\phi} = (\phi_1, \phi_2, \phi_3)$. Assuming that the leptospirosis data was collected in a case-control type manner, we assign the priors $\phi_1, \phi_2 \sim \text{Beta}(1,1)$ and $\phi_3 \sim \text{Beta}(1, 1000)$, where the prior on ϕ_3 makes the assumption that the $P(D^+)$ is rare. Given the underlying model (1.12) and the fact that beta and binomial distributions are conjugates, the joint posterior distribution for ϕ, calculated by multiplying the likelihood for the model by the priors, is as follows:

$$p(\boldsymbol{\phi}|n_1, n_2, X) \propto \phi_3^{\alpha_3 - 1}(1-\phi_3)^{\beta_3 - 1} \prod_{i=1}^{2} \phi_i^{x_{1i}+\alpha_i-1}(1-\phi_i)^{\beta_i + n_i - x_{1i} - 1}, \tag{3.42}$$

where $\alpha_3 = 1$ and $\beta_3 = 1000$. As no information can be gained about ϕ_3 from the data, the prior for ϕ_3 simply becomes the posterior. Note that the marginal posterior distributions for ϕ_1 and ϕ_2 are given by:

$$p(\phi_1|n_1, X) \sim \text{Beta}(\alpha_1 + x_{11}, \beta_1 + n_1 - x_{11}) \quad \text{and} \quad p(\phi_2|n_2, X) \sim \text{Beta}(\alpha_2 + x_{12}, \beta_1 + n_2 - x_{12}), \tag{3.43}$$

where $\alpha_1 = \alpha_2 = \beta_1 = \beta_2 = 1$, $n_1 = 104$ and $n_2 = 276$ (see Table 2.1) in this example.

In order to calculate the PAR and PAF we require estimates of $p = P(D^+|E^+)$, $q =$

$P(D^+|E^-)$ and $e = P(E^+)$. We can specify these parameters in terms of $\boldsymbol{\phi}$ as follows:

$$p = \frac{\phi_1\phi_3}{\phi_1\phi_3 + \phi_2(1-\phi_3)} \tag{3.44}$$

$$q = \frac{(1-\phi_1)\phi_3}{(1-\phi_1)\phi_3 + (1-\phi_2)(1-\phi_3)} \tag{3.45}$$

$$e = \phi_1\phi_3 + \phi_2(1-\phi_3). \tag{3.46}$$

The PAR and PAF are then given by:

$$PAR = e(p-q) = \phi_1\phi_3 - \frac{(1-\phi_1)\phi_3[\phi_1\phi_3 + \phi_2(1-\phi_3)]}{(1-\phi_1)\phi_3 + (1-\phi_2)(1-\phi_3)} \tag{3.47}$$

$$PAF = \frac{PAR}{\phi_3}. \tag{3.48}$$

Re-sampling ϕ from their posterior distribution above (3.42) the PAR and PAF can then be estimated. Drawing 10,000 iterations from the posterior with the `rbeta` function in R and using the mean as the point estimate we get $PAR = 0.0014$ (95% CI: $2.56 \times 10^{-6}, 5.52 \times 10^{-4}$) and $PAF = 0.14$ $(0.05, 0.23)$. Note that the PAR calculated here is different from that calculated in Chapter 2, where the underlying study design was assumed to be cross-sectional, due to the effect of the informative prior. Additionally, if the disease is rare the PAR will inevitably be small as seen here.

3.3.2 Using prior information for $P(E^+)$

In order to estimate p, q, PAR and PAF using equations (3.44-3.48), ϕ_3 must first be expressed in terms of ϕ_1, ϕ_2 and e as follows:

$$\phi_3 = \frac{e - \phi_2}{\phi_1 - \phi_2}. \tag{3.49}$$

Since ϕ_3 represents a probability, it must be constrained to the interval $[0, 1]$. This introduces the additional constraints that: $\phi_2 < e \leq \phi_1$ or $\phi_1 < e \leq \phi_2$ and $\phi_1 \neq \phi_2$. Note that if $\phi_1 > \phi_2$, then for $\phi_3 \geq 0$ we need $e \geq \phi_2$. For $\phi_3 \leq 1$ we need $\phi_1 - \phi_2 \geq e - \phi_2$, and therefore $\phi_1 \geq e$.

Furthermore, if $\phi_1 < \phi_2$, then for $\phi_3 \geq 0$ we need $e \leq \phi_2$. For $\phi_3 \leq 1$ we need $e - \phi_2 \geq \phi_1 - \phi_2$, and therefore $e \geq \phi_1$.

To account for these constraints let $A = [\min(\phi_1, \phi_2), \max(\phi_1, \phi_2)]$. When assigning priors for ϕ_1, ϕ_2 and e one way we can go about this is by specifying each prior independently then constraining these such that $e \in A$. For our particular example we assign Beta$(1, 1)$ priors on ϕ_1 and ϕ_2 as before, and suppose there is prior information about the exposure rate specifying low exposure, $e \sim$ Beta$(1, 10)$. The joint posterior distribution for ϕ_1, ϕ_2 and e can then be represented by:

$$p(\phi_1, \phi_2, e|n_1, n_2, X) \propto \begin{cases} e^{\alpha_4 - 1}(1-e)^{\beta_4 - 1} \prod_{i=1}^{2} \phi_i^{(x_{1i} + \alpha_i) - 1}(1-\phi_i)^{(\beta_i + n_i - x_{1i}) - 1}, & \text{if } e \in A. \\ 0, & \text{if } e \notin A, \end{cases} \tag{3.50}$$

where in this example $\alpha_4 = 1$ and $\beta_4 = 10$. As the full conditional posterior distribution can be identified for ϕ and e we can adopt the Gibbs sampling procedure (discussed in more detail in Section 3.2.3.4) for updating our parameters. For example the full conditional posterior distribution for $e|\phi$ is the Beta(α_4, β_4) distribution truncated to $[\phi_1, \phi_2]$ when $\phi_1 < \phi_2$ and $[\phi_2, \phi_1]$ when $\phi_2 < \phi_1$. Sampling from this truncated beta distribution can be carried out by taking draws from Beta(α_4, β_4) then simply rejecting those values which do not fall inside the appropriate interval.

The parameter space for this example is split into the space where $\phi_1 > \phi_2$ and $\phi_2 > \phi_1$. To explore the entirety of this parameter space we use a Metropolis-Hastings step to allow for a switch between the two spaces. This is done by comparing the likelihood for ϕ_1^t and ϕ_2^t, that is ϕ_1 and ϕ_2 at time t for the sampler, with the likelihood where $\phi_1^\dagger = \phi_2^t$ and $\phi_2^\dagger = \phi_1^t$. In other words, the acceptance probability α compares the likelihood of ϕ_1^t and ϕ_2^t with the likelihood where ϕ_1^t and ϕ_2^t have being switched. For example, this means that if $\phi_1^t > \phi_2^t$ the proposed values would be $\phi_2^\dagger > \phi_1^\dagger$. Note that the simplified acceptance probability, $\alpha = L(\phi_2, \phi_1, e)/L(\phi_1, \phi_2, e)$, is used in this example as Beta$(1, 1)$ flat prior distributions are adopted for both ϕ_1 and ϕ_2 and a

symmetric proposal distribution (which is always to switch) used. A comprehensive outline of the sampling procedure is as follows:

1. Specify initial values ϕ_1^0 and ϕ_2^0 such that $\phi_1^0 \neq \phi_2^0$, according to (3.43), calculate $A^0 = [min(\phi_1^0, \phi_2^0), max(\phi_1^0, \phi_2^0)]$ and initialise an iteration counter at $t = 0$.

2. As ϕ_1^t and ϕ_2^t will be sampled conditional on e^t set a secondary counter s, used to monitor switching, which if $\phi_1^t > \phi_2^t$ then $s = 1$ and if $\phi_1^t < \phi_2^t$ then $s = -1$.

3. Draw a value for e^t from the Beta$(1, 10)$ distribution. Calculate A^t and if $e^t \notin A^t$ then re-draw until $e^t \in A^t$.

4. Select a ϕ_1^t from (3.43) and update A^t. If $e^t \notin A^t$, then re-draw until $e^t \in A^t$.

5. Select a ϕ_2^t from (3.43) and update A^t. If $e^t \notin A^t$, then re-draw until $e^t \in A^t$.

6. Let $\phi_1^\dagger = \phi_2^t$ and $\phi_2^\dagger = \phi_1^t$, where $\phi_1^\dagger$ and $\phi_2^\dagger$ represent candidate values and ϕ_1^t and ϕ_2^t the values for current time step t.

7. Compute the probability of acceptance for switching

$$\alpha = \min \left\{ 1, \frac{L(\phi_2^\dagger, \phi_1^\dagger, e)}{L(\phi_1^t, \phi_2^t, e)} \right\}. \tag{3.51}$$

8. Take a random draw, r, from a Uniform$(0, 1)$ distribution. If $r \leq \alpha$ set $\phi_1^{t+1} = \phi_1^\dagger$, $\phi_2^{t+1} = \phi_2^\dagger$ and multiply s by -1, otherwise set $\phi_1^{t+1} = \phi_1^t$ and $\phi_2^{t+1} = \phi_2^t$.

9. Set the iteration counter from t to $t + 1$.

10. Repeat steps 2-9 until desired number of iterations is complete.

Performing 1,000 iterations following the above procedure and applying the formulae (3.44-3.48) to estimate the PAR and PAF for the leptospirosis data, resulted in a mean estimate for the PAR of 0.024 (95% CI: $0.0015, 0.054$) and for the PAF of 0.097 (95% CI: $0.0067, 0.21$). As $\hat{\phi}_1$

and $\hat{\phi_2}$ estimated from the leptospirosis data are quite different ($\hat{\phi_1} = 0.21$ and $\hat{\phi_2} = 0.091$) no candidate values proposed by switching ϕ_1^t and ϕ_2^t were accepted. If ϕ_1 and ϕ_2 are more similar then it would be expected that more switching between the two parts of the parameter space would occur. However, if ϕ_1 and ϕ_2 are more similar the sampler proposed takes a very long time to find values of $e \in A$. The addition of a more informative prior for e may help reduce the computational burden, but choosing such a prior would be difficult in practice.

3.4 Cohort study example

As previously discussed (see Section 1.2.3) cohort studies involve following a group of individuals who share a similar characteristic, such as being exposed or not exposed to a certain risk factor, over a period of time. The numbers of individuals in the cohort which are exposed or not exposed to the risk factor of interest are fixed in advanced. This means that the probability of exposure cannot be estimated from the data. Under the Bayesian framework to overcome this problem we can simply apply a prior distribution to the probability of exposure, e. Alternatively, we could specify a prior on the prevalence of disease, ϕ_3, which induces a prior distribution on e.

3.4.1 Using prior information for $P(E^+)$

Applying a prior to e and deriving the posterior distributions for PAR and PAF can be done in a somewhat similar fashion to the case-control example where a prior was applied to $\phi_3 = P(D^+)$ (see Section 3.3.1). Let the priors on e, p and q be the $\sim$ Beta(α_4, β_4), Beta(α_5, β_5) and Beta(α_6, β_6) respectively. Given the underlying model (1.13) the joint posterior distribution for p, q and e calculated by multiplying the likelihood for the model and the priors can be derived:

$$p(p, q, e|m_1, m_2, X) \propto e^{\alpha_4 - 1}(1-e)^{\beta_4 - 1} p^{x_{11} + \alpha_5 - 1}(1-p)^{\beta_5 + m_1 - x_{11} - 1} q^{x_{21} + \alpha_6 - 1}(1-q)^{\beta_6 + m_2 - x_{21} - 1}, \tag{3.52}$$

recalling that $m_1 = x_{11} + x_{12}$ and $m_2 = x_{21} + x_{22}$ as in Table 1.1. Note that the marginal posterior distribution for p and q can be given analytically by

$$p(p|m_1, X) \sim \text{Beta}(\alpha_5 + x_{11}, \beta_5 + m_1 - x_{11}) \quad \text{and} \quad p(q|m_2, X) \sim \text{Beta}(\alpha_6 + x_{21}, \beta_6 + m_2 - x_{21}). \tag{3.53}$$

After taking random draws from the posterior for p, q and e the posterior for PAR can then be estimated using (1.8). The posterior for the PAF can then be generated by dividing the posterior draws for the PAR by $\phi_3 = pe + q(1 - e)$.

3.4.2 Using prior information for $P(D^+)$

If we wish to specify a prior on ϕ_3 (as opposed to e), say $\phi_3 \sim \text{Beta}(\alpha_3, \beta_3)$, then in order to estimate the PAR and PAF using equations (3.47-3.48), e must first be expressed in terms of p, q and ϕ_3 as follows:

$$e = \frac{\phi_3 - q}{p - q} \tag{3.54}$$

Since e represents a probability it must be constrained to the interval $[0, 1]$. This introduces the additional constraints that: $q \leq \phi_3 \leq p$ or $p \leq \phi_3 \leq q$. Similarly to the case-control example where a prior was placed on e (see Section 3.3.2) we need to account for these constraints by truncating our joint posterior distribution. If we let $B = [\min(p, q), \max(p, q)]$, then the joint posterior distribution for (p, q, ϕ_3) is

$$p(p, q, \phi_3|m_1, m_2, X) \propto \begin{cases} \phi_3^{\alpha_3 - 1}(1 - \phi_3)^{\beta_3 - 1} p^{x_{11} + \alpha_5 - 1}(1 - p)^{\beta_5 + m_1 - x_{11} - 1} \times \\ \qquad\qquad q^{x_{21} + \alpha_6 - 1}(1 - q)^{\beta_6 + m_2 - x_{21} - 1}, & \text{if } \phi_3 \in B. \\ 0, & \text{if } \phi_3 \notin B. \end{cases} \tag{3.55}$$

As the full conditional posterior distribution can be identified as truncated beta distributions, similar to those derived under the case-control example (Section 3.3.2), we can adopt the Gibbs sampling procedure for updating our parameters. The parameter space for this example is split

into the space where $p > q$ and $q > p$. In practice it unlikely for $q > p$ unless E^+ represents a protective exposure such as vaccination. To explore the entirety of this parameter space we can adopt a Metropolis-Hastings step to allow for a switch between the two spaces. This approach described is analogous to that outlined for the case-control study where a prior is applied to e, therefore we do not repeat it here (see Section 3.3.2).

3.5 Concluding remarks

The overwhelming majority of statistical models proposed in the literature are identifiable (Gustafson, 2015). When a model is under-identified, as is the case when estimating the PAR and PAF from case-control, cohort, or measurement error data in a 2×2 table, a Bayesian approach must be adopted to take into consideration all sources of uncertainty. Here we show that a Bayesian approach for estimating the PAR and PAF from a case-control or cohort study is very straightforward if beta priors are applied to ϕ_3 or e respectively. This is because the joint posterior distribution in these cases can be specified analytically. If a prior distribution were specified on e rather than ϕ_3 for case-control data, then the induced posterior distribution on ϕ_3 results in values such that $\phi_3 \notin [0, 1]$. To account for this problem we develop an MCMC sampler that constrains e to lie between ϕ_1 and ϕ_2, such that $\phi_3 \in [0, 1]$. Imposing this constraint splits the parameter space into two parts, that where $\phi_1 > \phi_2$ and where $\phi_1 < \phi_2$. It was found that exploring this parameter space fully was difficult. Switching between the part of the parameter space where $\phi_1 > \phi_2$ and $\phi_1 < \phi_2$, occurs very rarely when ϕ_1 and ϕ_2 have a large difference. However, when the difference between ϕ_1 and ϕ_2 is small, finding an e which falls between these values is difficult and results in increased computation.

The leptospirosis study, where an imperfect diagnostic test was used to assess exposure status, gave rise to another example of an under-identified model. In this case the posterior distribution for the parameters of interest could not be derived analytically. Standard MCMC samplers (i.e. Metropolis-Hastings and Gibbs) have also been reported to perform poorly when

the model lacks of identification (Gustafson, 2015). We provide a comparison of the performance of several different MCMC samplers and develop a sampler which aims to effectively explore the posterior ridge of an under-identified model, by taking into consideration the shape of the ridge. Comparison of effective sample size shows that the importance sampling approach proposed by Gustafson (2015) is by far superior to all MCMC methods. If a transparent parameterisation is difficult to find or work with, MCMC simulation may be preferred. The choice of sampler in this situation should be based on the sample size of the data. When the sample size is small the HMC algorithm provided a greater number of effective samples per 1,000 iterations than the other MCMC samplers examined. Tuning the HMC algorithm though can be a difficult task. If the HMC algorithm can not be tuned then the Gibbs sampler provides the next best performance. As the sample size increases an adapted random walk approach which takes into consideration the shape of the likelihood is suggested. Specifically, the adapted random walk approach with covariance matrix given by $\Sigma^* = c[\sigma I + J^T J]$ provides the greatest effective sample size. Given that this approach performs well for large samples further investigation could be conducted to determine whether improved performance can be attained via a combination of this sampler and the MALA or HMC algorithms.

Chapter 4

Bayesian Inference for Adjusted Population Attributable Measures from More Complex Designs

Until now we have focused solely on estimating the PAR, PAF and their credible intervals for data with only a single risk factor for disease (i.e. data in the form of a 2×2 table). In practice however, many analyses consider multiple risk factors for disease, both numerical and categorical, to estimate the true effect of exposure. Taking into account additional variables when calculating the PAR and PAF requires the following:

1. A model of the outcome, y, as a function of exposure to the risk factor being considered for removal, E, and other covariates, $x = (x_1, \ldots, x_k)$.

2. An estimate of the joint distribution for the risk factor E and other covariates x in the population.

In this chapter we explore two datasets which incorporate multiple risk factors for disease. The first explores the risk factors associated with low birth weights of newborn babies (see

Section 4.1). This example is a simple extension of the 2×2 table which incorporates a third categorical variable, with three levels, resulting in a $2 \times 2 \times 3$ table (see Table 4.1). Newson (2013) uses this dataset to demonstrate his method of calculating the PAR, PAF and their confidence intervals via logistic regression. Implementing the logistic regression approach under the Bayesian framework, we show through simulation that when the PAR is large, accounting for variation in the joint distribution of (E, x) results in improved coverage of the intervals, for only a slight increase in length.

The second dataset results from an experiment which evaluates the effectiveness of a treatment for bovine mastitis (see Section 4.2). This example provides three further extensions including: many covariates both numeric and categoric, "person-time" type data and data arising from an experiment rather than an observational study. "Person-time" data is that where the observations are counts of occurences of an event within a specified time interval. With the mastitis example the number of occurrences of disease may differ for each cow over the period of time at risk. Moreover, a cow is only at risk of mastitis during calving, which could differ in length of time for each cow.

Since the modeled response for the mastitis data is based on a rate it no longer makes sense to interpret the PAR as a probability in the interval $[0, 1]$. Instead we adopt the population attributable rate (PARate), which describes the change in the rate of disease that could occur if a risk factor for that disease were removed from the population, and the population attributable rate fraction (PARF), the proportion of the rate of disease which can be attributed to the risk factor being considered for removal (see Section 4.4). In order to calculate the PARate and PARF for this dataset we specify a Bayesian Poisson regression model on λ_i, the rate of cases of mastitis per cow per day, as a function of the cow's age, previous history of mastitis and exposure to the protective treatment. The model is later extended to incorporate random effects so that the hierarchical nature of the data (i.e. that cows are clustered into herds) can be taken into consideration.

Several methods for estimating the population distribution of the covariates have also been

	White		Black		Other		
	S	NS	S	NS	S	NS	Total
Low	19	4	6	5	5	20	59
Not low	33	40	4	11	7	35	130
Total	52	44	10	16	12	55	189

Table 4.1: Cross-tabulation of the low birth weight data. Birth weight $< 2500g$ is considered as low whereas birth weight $\geq 2500g$ is not low. Smoking status is indicated by S for smoker and NS for non-smoker.

implemented. These include using fixed estimates based on the sample, allowing for uncertainty by adopting Bayesian bootstrap re-sampling as proposed by Rubin (1981), and finally, extending the Bayesian bootstrap to accommodate cluster sampling. Through simulation we show that when the PARate is large, taking into consideration all possible sources of uncertainty results in improved coverage of the credible intervals, in terms of their Frequentist properties, whilst remaining of reasonable interval length.

4.1 Low birth weight data

Infant mortality and birth defect rates are higher in babies of low birth weight (Hosmer et al., 2013). Hosmer et al. (1988), and later Hosmer et al. (2013), describe a dataset collected at the Baystate Medical Center in Springfield, Massachusetts, investigating potential risk factors associated with low birth weights. The low birth weight dataset consists of information on 189 births from women seen in the obstetrics clinic. Of the 189 births 59 were classified as having low birth weight, whilst the remaining 130 were classified as being of normal birth weight.

The dataset consisted of eight potential risk factors for low birth weights in newborns. However, Newson (2013) who uses this data to assess his own method of PAR and PAF estimation, only considers the joint effect of race and smoking status on birth weight, claiming this relationship to be of most interest. All three of these variables are categorical with race having the three levels white, black and other, smoking status the levels smoker and non-smoker, and

	White		Black		Other	
	S	NS	S	NS	S	NS
Low	e_1p_1	e_2q_1	e_3p_2	e_4q_2	e_5p_3	e_6q_3
Not low	$e_1(1-p_1)$	$e_2(1-q_1)$	$e_3(1-p_2)$	$e_4(1-q_2)$	$e_5(1-p_3)$	$e_6(1-q_3)$
Total	e_1	e_2	e_3	e_4	e_5	e_6

Table 4.2: Probability model for the low birth weight data, where $e = (e_1, \dots e_6)$ gives the joint probability distribution of exposure and race, p_i and q_i give the probability of low birth weight in racial group i for exposed and unexposed respectively. Smoking status is indicated by S for smoker and NS for non-smoker.

birth weight the levels less than 2500g and greater than or equal to 2500g. Since all these variables are categorical they can be presented in a $2 \times 2 \times 3$ contingency table, such as Table 4.1. This data can be described by the distribution

$$Y \sim \text{Multinomial}(n = 189, \pi = P(D^+|E, x)) \tag{4.1}$$

where Y represents the response variable low birth weight, x the factor race and the $P(D^+|E, x)$ is as specified in Table 4.2.

The PAR calculated by Newson (2013) was 0.084 (95% CI: 0.03, 0.14) and the PAF 0.27 (95% CI: 0.09, 0.41). These estimates were achieved using a logistic regression model. However, from the paper it is unclear exactly how the joint distribution of exposure and race has been treated. Through comparison with our methods it appears that Newson (2013) has used fixed parameters for the joint distribution of (E, x) (i.e. not accounting for their sampling uncertainty, see Section 4.7).

4.2 Bovine mastitis data

Mastitis is a reaction to infection which causes inflammation in the mammary glands of cows (Petrovski, 2007). It is also one of the most prevalent diseases affecting the dairy industry (Petrovski, 2007). Dairy cows can have up to four cases of mastitis, one for each quarter of

the udder, at any one time. The disease can resolve on its own or be treated, meaning that it is possible for a cow to contract mastitis several times over its lifespan. The dataset we explore was collected by Heuer et al. (2001) to examine the effect of monensin treatment on milk production, health and reproduction. The dataset consists of 601 observations collected from 8 different research farms, where the monensin treatment at each farm was assigned according to a randomised block design. Cows with similar parity, milk yield and time of calving were assigned to a block which contained between two to six cows. Within each of these blocks the treatment was then randomly assigned.

The total number of days that each cow was at risk of mastitis, due to calving, was recorded. The number of cases of mastitis per cow, Y_i, can be represented by:

$$Y_i \sim \text{Poisson}(\lambda_i \, DAR_i), \tag{4.2}$$

where λ_i is the daily rate of disease estimated by $\hat{\lambda}_i = \text{Cases}_i / DAR_i$ for $i = 1, \ldots, 601$ (i.e. the rate of mastitis cases per cow per day) and DAR_i is the number of days at risk for cow i. Data on several other variables considered to be risk factors for the disease were also collected and can be seen in Table 4.3. When calculating the PARrate and PARF for this example the variable we consider for removal is the monensin treatment (Tx), which is protective against mastitis, as shown by Heuer et al. (2001). Note that we decide to assign those cows on the treatment as unexposed ($E-$) so that the PARate and PARF are positive. Doing this allows for easier interpretation.

4.3 Population attributable risk and population attributable fraction for multiple covariates

As pointed out by Hanley (2001) estimation formulae for the PAF (i.e. $\widehat{PAF}$) are often all that is provided in the literature, due to formal definitions not being directly applicable in

Cases	Number of cases of mastitis during the entire duration of lactation
Tx	Treatment (control=0, treatment=1)
DAR	Days at risk of getting mastitis
Herd	Herd ID
D START	Days after calving when Tx started
MF	Milk Fever - a disease caused by calcium deficiency around calving (no=0, yes=1)
PARCAT	Number of lactations (parities) as an approximation for age
Premast	Mastitis before D START (no=0, yes=1)

Table 4.3: List of variables in the bovine mastitis dataset

application. Here we formally define and provide estimates for the PAR and PAF when there is a vector x of other risk factors to be considered. We suppose that the probability of disease is modeled as a function of exposure to the risk factor of interest, E, and the covariate vector x, say $P(D^+ \mid E, x)$. Note that the probability of exposure may also be related to the other risk factors. The reduction in risk for an individual with exposure E and covariates x is given by

$$P(D^+ \mid E, x) - P(D^+ \mid E^-, x) = \begin{cases} P(D^+ \mid E^+, x) - P(D^+ \mid E^-, x) & \text{if } E = E^+ \\ 0 & \text{if } E = E^-. \end{cases} \tag{4.3}$$

Integrating over the joint distribution of E and x, noting that there is no contribution when $E = E-$, gives

$$PAR = \int_{\mathcal{X}} [P(D^+ \mid E^+, x) - P(D^+ \mid E^-, x)] P(E^+ \mid x)\, dFx, \tag{4.4}$$

which is a natural extension of (1.8), where Fx represents the cumulative distribution function for the distribution of x and $\mathcal{X}$ denotes the domain of x. To obtain the PAF, the PAR is simply divided by

$$P(D^+) = \int_{\mathcal{X}} \sum_{E} P(D^+ \mid E, x) P(E \mid x)\, dFx, \tag{4.5}$$

where $E \in \{E^-, E^+\}$. Estimation of the PAR and PAF now requires a model which provides an estimate of $P(D^+ \mid E, x)$, for example by logistic regression, together with an estimate of the joint distribution of E and x.

In the low birth weight example, x is a single factor with three levels, say $\{X_1, X_2, X_3\}$, so

$$PAR = \sum_{i=1}^{3} [P(D^+ \mid E^+, X_i) - P(D^+ \mid E^-, X_i)] P(E^+ \mid X_i) P(X_i), \tag{4.6}$$

and

$$PAF = \frac{PAR}{\sum_{i=1}^{3} \sum_{E} P(D^+ \mid E, X_i) P(E \mid X_i) P(X_i)}. \tag{4.7}$$

The term in square brackets in PAR (4.6) is estimated from a logistic regression model, whilst the other probabilities are estimated from marginal tables (i.e. Table 4.1).

4.4 Population attributable rate and rate fraction

In the case where the response variable is a rate, as in the mastitis dataset, the definitions of PAR in terms of probabilities for single (1.8) and multiple covariates (4.6) are no longer appropriate. We now build on the concept of a population attributable measure by extending (1.8), which considers only the risk factor being selected for removal, to account for rates. We propose the population attributable rate (PARate) defined as:

$$PARate = P(E^+)(\lambda_{E^+} - \lambda_{E^-}), \tag{4.8}$$

where λ_{E^+} is the rate of disease in the exposed population and λ_{E^-} the rate of disease in the non-exposed population. This parameterisation is equivalent to that proposed by MacMahon

and Trichopoulos (1996) given by (1.9). The PARate can be estimated by

$$\widehat{PARate} = \widehat{P(E^+)}(\hat{\lambda}_{E^+} - \hat{\lambda}_{E^-}), \tag{4.9}$$

where $\hat{\lambda}_{E^+}$ and $\hat{\lambda}_{E^-}$ are estimated from a Poisson regression model and $\widehat{P(E^+)}$ from the sample. Note that λ_{E^+} and λ_{E^-} must be defined over the same unit interval (e.g. cases per day, cases per year, etc.). The PARate is interpreted as the change in the rate of disease that could occur if a risk factor for that disease were removed from the population. Unlike the PAR, the PARate can take on any value in the interval $(-\infty, \infty)$.

Similarly the PAF defined by (1.2), which considers only the risk factor selected for removal, can be extended to account for rates. We now define the population attributable rate fraction (PARF) as

$$PARF = \frac{P(E^+)(\lambda_{E^+} - \lambda_{E^-})}{P(E^+)\lambda_{E^+} + (1 - P(E^+))\lambda_{E^-}}, \tag{4.10}$$

which can be estimated by

$$\widehat{PARF} = \frac{\widehat{P(E^+)}(\hat{\lambda}_{E^+} - \hat{\lambda}_{E^-})}{\widehat{P(E^+)}\hat{\lambda}_{E^+} + (1 - \widehat{P(E^+)})\hat{\lambda}_{E^-}}. \tag{4.11}$$

Similarly to the PARate this formulation for the PARF is equivalent to that proposed by MacMahon and Trichopoulos (1996). The PARF is interpreted as the proportion of the rate of disease which can be attributed to the risk factor being considered for removal and can take on any value in the interval $(-\infty, 1]$.

To extend the PAR (4.4) to account for rates when there is a vector x of other risk factors, we let $\lambda(E, x)$ denote the rate for an individual with exposure E and covariate vector x. Then as with PAR

$$PARate = \int_{\mathcal{X}} [\lambda(E^+, x) - \lambda(E^-, x)]P(E^+ \mid x)dFx. \tag{4.12}$$

and

$$PARF = \frac{PARate}{\int_{\mathcal{X}} \lambda(E,x)P(E \mid x)dFx}. \tag{4.13}$$

An estimate of $\lambda(E,x)$ can be found using Poisson regression, but the joint distribution of (E,x) must be estimated empirically. Given a large finite population of N individuals with exposures E_i and covariates x_i, we could define PARate as

$$PARate = \frac{1}{N}\sum_{i=1}^{N}[\lambda(E_i,x_i) - \lambda(E^-,x_i)] \tag{4.14}$$

which can be estimated from sample data of size n as

$$\widehat{PARate} = \frac{1}{n}\sum_{E_i=E^+}[\hat{\lambda}(E_i,x_i) - \hat{\lambda}(E^-,x_i)], \tag{4.15}$$

which requires a model for $\lambda(E,x)$ which is fitted to the sample data. Note that we need only sum over $E_i = E^+$, since when $E_i = E^-$ we have $\lambda(E_i,x_i) - \lambda(E^-,x_i) = 0$. Similarly the PARF for a large finite population can be calculated according to

$$PARF = \frac{PARate}{\frac{1}{N}\sum_{i=1}^{N}\lambda(E_i,x_i)}, \tag{4.16}$$

which can be estimate from sample data as

$$\widehat{PARF} = \frac{\widehat{PARate}}{\frac{1}{n}\sum_{i=1}^{n}\hat{\lambda}(E_i,x_i)}. \tag{4.17}$$

For experimental data like the mastitis example in which the exposure is controlled, therefore independent of x, and the population is currently exposed (where E^+ = no treatment and E^- = treatment) we can use

$$\widehat{PARate} = \frac{1}{n}\sum_{i=1}^{n}[\hat{\lambda}(E^+,x_i) - \hat{\lambda}(E^-,x_i)] \tag{4.18}$$

and

$$\widehat{PARF} = \frac{\widehat{PARate}}{\frac{1}{n}\sum_{i=1}^{n}\hat{\lambda}(E^+, x_i)}. \tag{4.19}$$

To incorporate uncertainty in the estimation of the distribution of (E, x), we introduce weights as follows for

$$\widehat{PARate}_w = \sum_{i=1}^{n} w_i[\hat{\lambda}(E^+, x_i) - \hat{\lambda}(E^-, x_i)] \tag{4.20}$$

and

$$\widehat{PARF}_w = \frac{\widehat{PARate}_w}{\sum_{i=1}^{n} w_i\hat{\lambda}(E^+, x_i)} \tag{4.21}$$

where the $\sum w_i = 1$ and w can be generated using the Bayesian bootstrap.

4.5 Bayesian bootstrap

Rubin (1981) was the first to introduce a Bayesian analogue of the bootstrap approach proposed by Efron (1979) (see Section 2.2.3). Rubin's paper not only describes implementation details for the Bayesian bootstrap, but also discusses that although these approaches allow one to literally "pull themselves up by the their bootstraps", both methods draw inference based on model assumptions that may not be reasonable. For example, he questions whether it is sensible to use a model which assumes that all possible distinct values of X have been observed. Moreover, having assumed that all distinct values of X have been observed he questions whether it is reasonable to assume *a priori* independent weights, which are constrained to sum to 1, for these values of X. Nevertheless, his examples suggest that the Bayesian bootstrap outperforms the bootstrap.

The Bayesian bootstrap is not what one may consider a fully Bayesian approach: a pos-

terior distribution is generated but no prior information is required. Implementation of a fully Bayesian approach would require specification of a prior on the joint distribution of the covariates and exposure, which would be very difficult both in theory and practically. Use of the Bayesian bootstrap incorporates uncertainty in this joint distribution without the need to specify this joint distribution explicitly.

Before describing the implementation details for the Bayesian bootstrap first recall that if we have a sample of size n, say $x_1, \ldots, x_n$ which is viewed as n iid realisations of the random variable X, then a bootstrap replicate can be derived by taking a simple random sample, of size n with replacement from $x_1, \ldots, x_n$. Taking a simple random sample with replacement is equivalent to assigning a fixed weight $w_i \in \{0, 1/n, 2/n, \ldots, 1\}$, where $\sum w_i = 1$, to each x_i. The bootstrap distribution for a statistic, say $\hat{\phi}$, can then be generated by considering all possible bootstrap replicates (or in practice a large random sample) of $\hat{\phi}$. Now a Bayesian bootstrap replicate begins by assigning a posterior probability, $w_i \in [0, 1]$ where $\sum w_i = 1$, for each x_i, where the values of X that are not observed are assigned a posterior probability of zero. The posterior probability for all X that are observed can be generated by drawing $(n-1)$ values, $u_1, \ldots, u_{n-1}$, from a Uniform$(0, 1)$ distribution, ordering them and calculating the gaps between, $w_i = u_i - u_{i-1}$, for $i = 1, \ldots, n$ where $u_0 = 0$ and $u_n = 1$. The vector of probabilities $w = (w_1, \ldots, w_n)$ is then attached to each data value $x_1, \ldots, x_n$ for that Bayesian bootstrap replicate. Since the $n - 1$ variables in each replication are iid Uniform$(0, 1)$, w_i follows the n-dimensional Dirichlet$(1, \ldots, 1)$ distribution (Wilks, 1962).

The Bayesian bootstrap approach was adopted by Lehnert-Batar et al. (2006) in order to randomly generate a vector of weights which was then attached to the data to calculate the adjusted PAF and its confidence interval. If there is no obvious hierarchical structure to the data then generating weights using the Bayesian bootstrap as described maybe appropriate. However, in the mastitis example we have cows which are clustered into herds. Since cows in the same herd are likely to be more similar to one another than those cows from a different herd, we want to generate the weights in a way which reflects this structure. Therefore, we extend

Rubin's Bayesian bootstrap to take into consideration clustered data.

4.5.1 Cluster Bayesian bootstrap

Implementation of the cluster Bayesian bootstrap is similar to that of the Bayesian bootstrap described above. Except rather than generating a posterior probability, w_i, for each individual from the n-dimensional Dirichlet$(1\ldots,1)$ distribution, we instead draw weights, w_c for $c = 1,\ldots,C$, for the cluster level from the C-dimensional Dirichlet$(1\ldots,1)$ distribution, where C is the number of clusters. The weight, w_c^i, for each individual in cluster c can then be calculated by simply dividing w_c by the sample size for the cluster as follows:

$$w_c^i = \frac{w_c}{n_c}, \tag{4.22}$$

where $\{w_c\}_{c=1}^C \sim$ Dirichlet$(1,\ldots,1)$ and n_c is the sample size for cluster c. Generating individuals weights in this way means that individuals in the same cluster will have the same weight assigned to them. This gives a Bayesian equivalent of the ordinary cluster bootstrap, first suggested by Gross (1980) for variance estimation of the weighted median, in which clusters rather than individuals are re-sampled.

4.6 Generalised linear models

Linear models provide a means of modeling the effect of a given set of covariates (or explanatory variables) $x_1,\ldots,x_k$ on the response variable of interest y. A linear model with k covariates is given by the equation

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \epsilon_i, \tag{4.23}$$

where x_{ij} is the value of the jth covariate, $j = 1,\ldots,k$, for the ith observation, $i = 1,\ldots,n$, and ϵ_i iid distributed errors, such that $E(\epsilon_i) = 0$ and Var$(\epsilon_i) = \sigma^2$. The simplest form of the linear model is where the errors are assumed to be normally distributed, $\epsilon_i \sim Normal(0,\sigma^2)$. In

this case the response variable is conditionally independent and normally distributed given the covariates, $y_i \sim Normal(\mu_i, \sigma^2)$, with $\mu_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}$. This particular model is known as the classical linear regression model (Fahrmeir et al., 2013). If the response variable is continuous then it may be normally distributed. However, if the response is binary, a count, categorical or even a continuous variable whose distribution is considerably skewed, then this assumption will not hold. It should be noted though that in the latter case a transformation may be applied to achieve a more symmetric distribution.

Generalised linear models (glm) are a class of models that allows for the analysis of data where the assumption of normal variation is not appropriate. To specify a glm the following are required (Gelman et al., 2004):

- The linear predictor, $\eta = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} = X\beta$,
- A link function $g(.)$ that relates the linear predictor to the mean of the outcome variable $E(y|X) = g^{-1}(\eta)$,
- The random component specifying the distribution of the outcome variable y with mean $E(y|X)$.

In the case where the data is binary like the low birth weight data, the appropriate model is the logistic regression (or logit) model. When there is count data like the mastitis dataset, then the Poisson regression model is most appropriate. In the following sections we outline the basis of both the logistic and Poisson regression models.

4.6.1 Logistic regression

Logistic regression models are commonly used for describing data with a binary response variable, similar to the low birth weight dataset. A binary response variable y_i can be described by the Bernoulli(π) distribution, where $\pi_i = P(y_i = 1)$, the probability of a success and $(1 - \pi_i) = P(y_i = 0)$, the probability of a failure. For multiple binary observations with the

same covariate vector these Bernoulli trials can be combined into a Binomial(n, π) distribution, where n represents the total number of Bernoulli trials and π the vector of probabilities for the success of each trial. We wish to specify a model that describes the relationship between the probability of success (or failure) and the covariates $x_1, \ldots x_k$. Expressing π_i as a simple linear model $\pi_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}$ violates the definition of probability, since we get $\pi_i \in (-\infty, \infty)$. Therefore, we must choose an appropriate link function which restricts π_i to the interval $[0, 1]$. A commonly used link function in practice is the logit or logistic function

$$g(\pi) = \ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}. \tag{4.24}$$

This link function provides a linear model for the log-odds, which if exponentiated provides $\pi/(1-\pi) = \exp(\beta_0)\exp(\beta_1 x_1)\cdot\ldots\cdot\exp(\beta_k x_k)$, showing the multiplicative effect of the covariates on the odds. The inverse of the logit link function is

$$\pi = \frac{\exp(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik})}, \tag{4.25}$$

which can be used to describe the response probability given the covariates. Other link functions which have been used for binary response variables include the inverse normal (or probit) function, $g(\pi) = \Phi^{-1}(\pi)$, the complementary log-log function, $g(\pi) = \log\{-\log(1-\pi)\}$ and the log-log function, $g(\pi) = -\log\{-\log(\pi)\}$ (McCullagh and Nelder, 1989; Fahrmeir et al., 2013).

Implementation of a Bayesian logistic regression model means applying a prior distribution on the parameter vector β (i.e. $p(\beta)$). Eliciting prior information from a subject matter expert on regression parameters, however, may not be a simple task. A flat or weakly informative prior is often applied instead. Assigning the improper flat prior $p(\beta) \propto 1$ usually has little effect on the posterior distribution (Christensen et al., 2010). Weakly informative priors based on the t-distribution family have been proposed by Gelman et al. (2008). He suggests a Cauchy(0, 2.5) prior on the β coefficients after standardising all non-binary variables to have a mean of 0 and standard deviation 0.5. He states that in the simplest setting, the Cauchy(0, 2.5) prior

is a longer-tailed version of the distribution attained by assuming the addition of one-half a success and one-half a failure in a logistic regression. Gelman et al. (2008) also indicates that depending on the context it may make more sense to use a weaker prior distribution on the intercept term. He selects a Cauchy(0, 10) prior, which would imply that the probability of a case is between 10^{-11} and $(1-10^{-11})$. A Normal$(0, \sigma^2)$ where σ^2 is large has also been suggested as a weakly informative prior for the elements of the parameter vector β (Christensen et al., 2010). This prior induces a prior on the effect of each component of β on π that puts half the probability near 0 and the other half near 1. Christensen et al. (2010) states "as silly as these priors seem, they will usually not affect the posterior very much". They also suggest that σ^2 could be adjusted such that the distributions on θ_i, for $i = 1, \ldots, n$, are approximately uniform. Once a prior is selected for the parameter vector β these coefficients can then be updated using standard MCMC approaches (e.g. Metropolis-Hastings) as described for the low birth weight data in Section 4.7.1.

4.6.2 Poisson regression

Poisson regression is a commonly used model for describing count data, like the mastitis data set. It assumes that counts, y_i, observed in a time interval of length T_i, can be described by the distribution $y_i \sim \text{Poisson}(\lambda_i)$, where

$$\lambda_i = E(y_i|x, T_i) = T_i \times \exp(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}) \tag{4.26}$$

or alternatively

$$\log(\lambda_i) = \log(T_i) + \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}. \tag{4.27}$$

The addition of T_i in (4.26) or $\log(T_i)$ in (4.27) is called an offset term, which has its coefficient set equal to 1.

Implementation of a Poisson regression model under a Bayesian framework is similar to

that described for the logistic regression. A prior distribution must be specified on the parameter vector β and once assigned the coefficients can then be updated using MCMC sampling, as discussed for the mastitis example in Section 4.8. The weakly informative Cauchy priors proposed by Gelman et al. (2008) may still be a reasonable default choice of prior in this situation.

4.6.3 Generalised linear mixed effects models

Mixed effects models arise from the need to accurately describe grouped or nested data. When data is grouped or nested it is likely that observations within the same group will be correlated. For the mastitis dataset cows in the same herd are likely to be more similar to one another than to cows in different herds, due to influences occurring at the herd level. Therefore, assuming independence of these observations as is done with standard linear regression models is not appropriate. A simple way to account for the grouping structure is to introduce an extra parameter into the model which is common to all observations in a particular group. For example:

$$Y_{ij} = x_{ij}^T \beta + u_i + \epsilon_{ij}, \tag{4.28}$$

where Y_{ij} denotes the jth observation in group i, u_i the intercept for each group with the corresponding error structure $u_i \sim \text{Normal}(0, \sigma_u^2)$, where σ_u^2 denotes the group level variation, and ϵ_{ij} the error associated with the jth observation in the ith group following the $\epsilon_{ij} \sim \text{Normal}(0, \sigma^2)$ distribution. This type of model is referred to as a random intercept model where u_i is often called a random effect and $x_{ij}^T \beta$ fixed effects. As pointed out by Gelman (2005) a large number of definitions, sometimes conflicting, exist for describing fixed and random effects. To avoid confusion Gelman (2005) suggests simply describing mixed models in terms of whether there is a random intercept, slope, or both.

In the class of generalised linear mixed effects models the error terms at the individual level need not be normally distributed. As already established in the previous section a Poisson model is most appropriate for the mastitis dataset. In order to carry out a Poisson mixed

effects model in R the function `glmer` from the `lme4` package (Bates et al., 2015) can be used. Implementation of this model under the Bayesian framework is similar to that of a generalised linear model, except with the addition of having to update the random effect u_i and its the hyper-parameter σ_u^2 (or as often done $\tau_u = 1/\sigma_u^2$, where τ_u is referred to as the precision). Due to conjugacy if a gamma distributed prior is applied to τ_u then the posterior distribution for τ_u is

$$\tau_u \sim \text{Gamma}\left(a + \frac{n}{2}, b + \frac{\sum_{i=1}^{n}(y_i - \mu)^2}{2}\right), \tag{4.29}$$

where a and b are the shape parameters of the gamma prior on τ_u. More details on the specifics of the MCMC sampler for this model using the mastitis dataset as an example are provided in Section 4.8.3.1.

4.7 Simulation study: Low birth weight data

To assess the performance of our method for estimating the PAR, PAF and their credible intervals we implement a simulation study based on the low birth weight data. We begin by selecting fixed known parameters for $P(E|x)$ where x represents the factor race, $P(x)$ and sample size N. Using the chosen parameters we can simulate the number of those exposed and not exposed for each race according to

$$E, x \sim \text{Multinomial}(n = N, \pi = P(E|x)P(x)), \tag{4.30}$$

where $P(E|x)P(x)$ can be described by the vector e as in Table 4.2. In order to simulate a complete dataset information is also required about $P(D^+|E, x)$, or alternatively $P(D^-|E, x)$. This information is obtained by applying to the original data set the logistic regression model

$$\log\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \beta_1 E_i + \beta_2 X_i, \tag{4.31}$$

where π represents the components of $P(D^+|E,x)$, and can be extracted from the maximum likelihood estimates of the β's via back transformation using (4.25). The number of diseased individuals can then be determined according to

$$D^+|E,x \sim \text{Binomial}(n = N_{(E,x)}, \pi = P(D^+|E,x)). \tag{4.32}$$

Once the number of diseased for each exposure and race combination is known, the number of non-diseased for each exposure and race combination can be calculated by simple subtraction.

Notice that the above procedure has been separated into two parts. That is, defining the joint distribution on exposure and race followed by specifying a logistic regression model for $P(D^+|E,x)$. After simulating a $2 \times 2 \times 3$ table of data in this way estimation of the posterior distributions can be separated in a similar fashion. The posterior distribution on $P(E,x)$ can be derived analytically if the appropriate conjugate prior is selected, whereas for updating the logistic regression parameters the random walk sampler can be adopted (more details are given in Section 4.7.1). Having specified posterior distributions on $P(D|E,x)$ and $P(E,x)$ the PAR and PAF can be calculated according to equations (4.6) and (4.7) respectively. To determine whether Newson (2013) used fixed constants on the joint distribution of exposure given race when calculating confidence intervals for the PAR, we also calculate the PAR and PAF using $P(E,x)$ and $P(x)$ as calculated directly from the data. Note that this means not allowing for uncertainty in the joint distribution (E,x).

For each simulation described in the following sections 1,000 tables were generated from which the PAR, PAF and their 95% credible intervals were calculated. The performance of the methods, based on the percent coverage, is evaluated whilst taking note of the interval length. Before discussing the simulations carried out we first describe the details of the MCMC sampler in this case.

4.7.1 MCMC sampler for a $2 \times 2 \times 3$ table

When running any MCMC sampler initial conditions stating where the sampler should begin for each parameter need to be specified. For the low birth weight example these parameters include $P(E, x)$ and the β coefficients of the logistic regression model (4.31). To initialise the β coefficients a logistic regression model is applied to the simulated dataset and the resulting coefficients used as initial conditions of the MCMC chain. If any components of the vector $P(D^+|E, x)$, used to simulate the $2 \times 2 \times 3$ table, are close to zero then it is likely that zeros will be present in the table. The existence of zeros, although possible in practice, poses problems when calculating the log-likelihood for the data corresponding to (4.1). This results in the `glm` function in R, used to get starting values from a logistic regression model, failing. To circumvent this problem, when zeros are present they are replaced by a fraction of an observation (we use 0.1). Having initalised β, a starting value can be calculated for $P(D^+|E, x)$ by multiplying the vector β by the model matrix then back-transforming using (4.25).

Recall that (E, x) can be described by the multinomial distribution (4.30). If a Dirichlet prior is chosen for $P(E, x)$ then due to the conjugacy relationship for these distributions we could specify the posterior analytically. Therefore, let the prior distribution on $P(E, x)$ be a Dirichlet$(1, \ldots, 1)$ distribution. It then follows that the posterior distribution on $P(E, x)$ is a Dirichlet$(n_1+1, \ldots, n_6+1)$ distribution where $\sum_{i=1}^{6} n_i = N$, and n_i is the total number of those exposed and non-exposed for each race.

Unlike $P(E, x)$, the posterior distribution for β and thus $P(D^+|E, x)$, PAR and PAF, can not be defined analytically. Therefore, we make use of the Metropolis-Hastings random walk sampler (see Section 3.2.3.1) to generate the posterior distributions for β. To update the random walk sampler a prior and proposal distribution for β needs to be specified. We implement this sampler using block-wise updating with the proposal distribution $Q(\beta^\dagger|\beta^t)$ being the multivariate normal distribution $MVN(\beta^t, c\Sigma)$, where $\beta^\dagger$ represents the proposed value of β, β^t the current value of β, c a fixed constant tuning parameter which provides an acceptance

rate of approximately 24% (Roberts et al., 1997) and Σ the covariance matrix of the β's from the initial fit of the logistic regression model (4.31). Proposals for β are obtained using eigenvalue decomposition (i.e. $\sqrt{\Sigma} = MVM^T$, where M is the matrix of eigenvectors of Σ, V the diagonal matrix of the square root of the eigenvalues of Σ and M^T the transpose of the matrix M) and setting $\beta^\dagger = \beta^t + c\sqrt{\Sigma}z$ where z is a vector of independent Normal$(0, 1)$ variables. In terms of choice of prior for β we try two different options: first a flat prior where the contribution to the acceptance probability (α) is simply set equal to 1, and then the weakly informative Cauchy prior, Cauchy$(0, 2.5)$, suggested by Gelman et al. (2008).

To calculate the acceptance probability the contribution to the likelihood function also needs to be estimated. Recall that the distribution of $(D^+|E, x)$ is given by the binomial distribution (4.32). That is, each combination of exposure and race can be be represented by its own binomial distribution. Thus the log-likelihood function for (4.32) is the product binomial

$$l(D^+|E, X) = \log\left(\prod_{i=1}^{6} \frac{N_i!}{(N_i - y_i)!\, y_i!}\right) + \sum_{i=1}^{6} y_i \log(\pi_i) + (N_i - y_i)\log(1 - \pi_i), \tag{4.33}$$

where π_i is $P(D^+|E, x)$. Note that it is common practice to work on the log scale to decrease error introduced by rounding. If any of the probabilities in the vector π are 0 or 1, then the log-likelihood is undefined. To avoid this problem a zero probability is assigned to the term of the log-likelihood corresponding to any element of $P(D|E, x)$ equal to 0 or 1.

We now have all the necessary information to outline the MCMC sampler used within the simulation of the low birth weight data. Note the similarity to the general structure outlined in Chapter 3 (Section 3.2.3.1).

1. Define initial conditions for $P(E, x)$, β, and $P(D^+|E, x)$ and initialize an iteration counter at $t = 0$.

2. Generate a posterior value for $P(E, x)$ from the Dirichlet$(n_1 + 1, \ldots, n_6 + 1)$ distribution where $\sum_{i=1}^{6} n_i = N$ and represent the total number of those exposed and non-exposed in

the data set for each race.

3. Generate a candidate value, $\beta^{\dagger}$, from the proposal distribution $Q(\beta^{\dagger}|\beta^{t}) \sim \text{MVN}(\beta^{t}, c\Sigma)$, where β^{t} is the value of the Markov chain at iteration t, c a fixed tuning parameter and Σ the covariance matrix for the estimated β's in the logistic regression model (4.31).

4. Calculate the $P(D^{+}|E,x)$ by multiplying $\beta^{\dagger}$ by the model matrix, then back-transform using (4.25) to calculate π_i in (4.33).

5. Compute the probability of acceptance

$$\alpha = \min\left\{1, \frac{p(\beta^{\dagger})L(P(D^{+}|E,x)^{\dagger})Q(\beta^{t}|\beta^{\dagger})}{p(\beta^{t})L(P(D^{+}|E,x)^{t})Q(\beta^{\dagger}|\beta^{t})}\right\}, \tag{4.34}$$

   where $p(\beta)$ represents the prior for β and $L(P(D^{+}|E,x))$ the likelihood given by exponentiating (4.33).

6. Take a random draw, U, from a Uniform$(0,1)$ distribution. If $U \leq \alpha$ set $\beta^{t+1} = \beta^{\dagger}$, otherwise set $\beta^{t+1} = \beta^{t}$.

7. Set iterate counter from t to $t+1$.

8. Repeat steps 2-7 until desired number of iterations is reached.

After obtaining the posterior distributions for $P(E,x)$, $P(D^{+}|E,x)$ and β following the above algorithm, the posterior distribution for the PAR and PAF can be estimated by applying the formulae (4.4) and (4.5) to each iteration. To calculate the PAR and PAF using fixed estimates for comparison with Newson's approach, $P(E,x) = P(E^{+}|X_i)P(X_i)$ in (4.4) and (4.5) is replaced the probabilities calculated directly from the data, rather than re-sampled from the posterior distribution.

4.7.2 Simulation 1: Original data

This simulation is based on the low birth weight data provided in Table 4.1, using a fixed sample size of $N = 189$. The vector of input parameters for the $P(E|x) = \{P(E^+|X_1), P(E^-|X_1), P(E^+|X_2), P(E^-|X_2), P(E^+|X_3), P(E^-|X_3)\}$ where the levels of x (i.e. $X_1 \ldots, X_3$) are white, black and other respectively is $\{0.6, 0.4, 0.4, 0.6, 0.2, 0.8\}$ and $P(x)$ is $\{0.5, 0.2, 0.3\}$. These parameter inputs result in a true PAR and PAF of 0.095 and 0.292 respectively.

Before running the simulation a tuning period was applied to the original dataset for the two different choices of prior on β (i.e. flat or weakly informative Cauchy). The tuning parameter was initially set to 1 then adjusted to achieve an acceptance rate of approximately 24% as suggested by Roberts et al. (1997). The tuning parameters which achieved closest to optimal acceptance were 1.5 when using the flat prior (acceptance rate: 21.9%) and 1.2 when using the weakly informative Cauchy prior (acceptance rate: 21.8%). After assigning the tuning parameters a BGR analysis was run, using the original data and three Markov chains with differing initial conditions, to assess convergence and determine if a burnin period is required. A burnin of 8,000 iterations when using a flat prior and 6,000 iterations when using a Cauchy prior was chosen. The chains appeared to converge in at least half the number of iterations than the burnin specified. However, when performing the simulation different tables to the one we have performed the BGR analysis on will be generated. Although, these tables will be similar the time for the chain to converge may differ. Due to the MCMC algorithm being reasonably computationally efficient (i.e. 10,000 iterations takes approximately 20s), a conservative burnin period was adopted. A total of 10,000 iterations were run for each of the 1,000 datasets generated from the input parameters with no thinning applied.

The results of the simulation study, provided in Table 4.4, show that there is little difference in the percent coverage of the flat prior compared to the Cauchy prior, with both resulting in slightly less than nominal 95% coverage for the PAR and PAF. Interestingly, the mean and median estimates of the PAR and PAF when using the Cauchy prior have been shrunk towards

N=189	Percent coverage		Mean PAR		Median PAR		Mean CI length	
	Cauchy	Flat	Cauchy	Flat	Cauchy	Flat	Cauchy	Flat
PAR								
Fixed $P(E\|x)P(x)$	92.7	93.1	0.087	0.093	0.086	0.092	0.114	0.111
Variable $P(E\|x)P(x)$	93.6	94.4	0.088	0.093	0.087	0.094	0.117	0.117
PAF								
Fixed $P(E\|x)P(x)$	92.9	93.4	0.271	0.292	0.271	0.292	0.343	0.343
Variable $P(E\|x)P(x)$	93.1	93.8	0.271	0.292	0.271	0.293	0.350	0.351

Table 4.4: The percent coverage, mean, median and interval length achieved for the PAR and PAF using both fixed and variable estimates for $P(E|x)$, from Simulation 1. The true PAR and PAF for this simulation are 0.095 and 0.292 respectively.

0, whilst estimates where the flat prior is used are closer to the true PAR (0.095) and PAF (0.292). Fixing $P(E|x)P(x)$ when calculating the PAR and PAF resulted in similar but slightly lower coverage than when $P(E|x)P(x)$ was allowed to vary. This is likely a result of the slightly larger interval length, when $P(E|x)P(x)$ is allowed to vary, due to accounting for a greater amount of uncertainty. The difference in interval length for these two approaches of calculating the PAR and PAF will hopefully allow for us to determine which approach Newson (2013) took in calculating his intervals. For this example the length of Newson's confidence interval for the PAR was 0.103 (PAR = 0.084 [0.032,0.135]) and for the PAF was 0.324 (PAF = 0.268 [0.088,0.412]). Both of these interval lengths are closer to our results for when $P(E|x)P(x)$ is fixed. Due to the similarity in the performance for the differing calculations of the PAR and PAF in this situation, where the PAR is small, the use of fixed estimates on $P(E|x)P(x)$ may not be too problematic.

4.7.3 Simulation 2: Large PAR

The true PAR for simulation 1 based on the low birth weight data set was very small (PAR=0.095). Since differences in performance become more obvious when the PAR is large (see Chapter 2), we adjust the value of $P(E|x)$ given in simulation 1 to create a larger effect due to exposure and thus a larger value for PAR. The vector of input parameters for $P(E|x)$ is now

	White		Black		Other		
	S	NS	S	NS	S	NS	Total
Low	49	1	17	2	30	1	100
Not low	1	9	3	3	2	11	29
Total	50	10	20	5	32	12	129

Table 4.5: Pseudo low birth weight data set created to more accurately reflect the input parameters for Simulation 2 and aid in initialisation of the MCMC sampler. Birth weight $< 2500g$ is considered as low whereas birth weight $\geq 2500g$ not low. Smoking status is indicated by S for smoker and NS for non-smoker.

$\{0.8, 0.2, 0.6, 0.4, 0.7, 0.3\}$ and for $P(X)$ it is $\{0.5, 0.2, 0.3\}$ similarly to simulation 1. These parameter inputs result in a true PAR of 0.569 and true PAF of 0.781. A range of sample sizes, $N = \{80, 129, 200, 500\}$, are also considered in this simulation to observe their effect on the performance.

For the sole purpose of initialising the MCMC algorithm the low birth weight data set has been adjusted (see Table 4.5) to more closely reflect the input parameters. Applying the `regpar` function in Stata, developed by Newson (2013), to the data provided in Table 4.5 the PAR was estimated as 0.619 (95% CI: 0.487, 0.724; interval length=0.237). When applying our Bayesian approach to this dataset the PAR, when estimates of $P(E|x)P(x)$ were fixed, was 0.619 (95% CI: 0.480, 0.717, interval length=0.237) and when allowed to vary was 0.61 (95% CI: 0.464, 0.724, interval length=0.26). This result suggests that Newson (2013) did in fact use fixed estimates for $P(E|x)P(x)$, which does not take into consideration the uncertainty surrounding these parameters. Similarly to simulation 1 a tuning period was performed on the data in Table 4.5 along with BGR to assess convergence. The same tuning parameters and burnin period as previously used appear to be appropriate for this situation as well.

The results of this simulation, provided in Tables 4.6-4.9, show that the percent coverage achieved by allowing for uncertainty in the estimate of $P(E|x)P(x)$ is superior to fixing $P(E|x)P(x)$ in every situation explored. This difference in percent coverage is more prominent for PAR than it is for PAF. The posterior median for the PAR and PAF is closer to the true

N=80	Percent coverage		Mean estimate		Median estimate		Mean CI length	
	Cauchy	Flat	Cauchy	Flat	Cauchy	Flat	Cauchy	Flat
PAR								
Fixed $P(E\|x)P(x)$	83.7	82.3	0.545	0.555	0.545	0.568	0.246	0.243
Variable $P(E\|x)P(x)$	91.9	89.4	0.533	0.543	0.534	0.556	0.287	0.290
PAF								
Fixed $P(E\|x)P(x)$	89.5	85.9	0.746	0.708	0.747	0.784	0.385	0.428
Variable $P(E\|x)P(x)$	90.4	86.7	0.742	0.705	0.744	0.780	0.394	0.494

Table 4.6: The percent coverage, mean, median and interval length achieved for the PAR and PAF using both fixed and variable estimates for $P(E|x)$, from Simulation 2 where the sample size is $N = 80$.

N=129	Percent coverage		Mean estimate		Median estimate		Mean CI length	
	Cauchy	Flat	Cauchy	Flat	Cauchy	Flat	Cauchy	Flat
PAR								
Fixed $P(E\|x)P(x)$	87.7	86.7	0.556	0.572	0.557	0.574	0.196	0.190
Variable $P(E\|x)P(x)$	93.5	93.9	0.548	0.564	0.549	0.566	0.228	0.225
PAF								
Fixed $P(E\|x)P(x)$	93.1	90.3	0.765	0.789	0.770	0.794	0.306	0.301
Variable $P(E\|x)P(x)$	93.3	90.7	0.763	0.787	0.767	0.793	0.311	0.306

Table 4.7: The percent coverage, mean, median and interval length achieved for the PAR and PAF using both fixed and variable estimates for $P(E|x)$, from Simulation 2 where the sample size is $N = 129$.

N=200	Percent coverage		Mean estimate		Median estimate		Mean CI length	
	Cauchy	Flat	Cauchy	Flat	Cauchy	Flat	Cauchy	Flat
PAR								
Fixed $P(E\|x)P(x)$	88.6	88.5	0.559	0.569	0.560	0.570	0.160	0.160
Variable $P(E\|x)P(x)$	93.5	93.9	0.554	0.564	0.555	0.565	0.185	0.186
PAF								
Fixed $P(E\|x)P(x)$	94.7	94.1	0.769	0.780	0.770	0.787	0.255	0.253
Variable $P(E\|x)P(x)$	94.8	94.4	0.767	0.779	0.769	0.785	0.26	0.257

Table 4.8: The percent coverage, mean, median and interval length achieved for the PAR and PAF using both fixed and variable estimates for $P(E|x)$, from Simulation 2 where the sample size is $N = 200$.

N=500	Percent coverage		Mean estimate		Median estimate		Mean CI length	
	Cauchy	Flat	Cauchy	Flat	Cauchy	Flat	Cauchy	Flat
PAR								
Fixed $P(E\|x)P(x)$	90.2	90.0	0.566	0.570	0.566	0.570	0.102	0.103
Variable $P(E\|x)P(x)$	94.4	94.3	0.564	0.568	0.564	0.568	0.119	0.119
PAF								
Fixed $P(E\|x)P(x)$	93.9	93.6	0.777	0.783	0.777	0.784	0.164	0.164
Variable $P(E\|x)P(x)$	94.1	93.8	0.776	0.782	0.776	0.783	0.165	0.166

Table 4.9: The percent coverage, mean, median and interval length achieved for the PAR and PAF using both fixed and variable estimates for $P(E|x)$, from Simulation 2 where the sample size is $N = 500$.

value for the population in every case. As expected the interval length is slightly larger when allowing for variation in $P(E|x)P(x)$, but not so much that the estimates for PAR and PAF are no longer useful. It can also be seen that when the sample size is small (i.e. $N = \{80, 129\}$), the coverage for both PAR and PAF is less than the nominal 95%. For these smaller sample sizes the weakly informative Cauchy prior provides slightly greater coverage than using the flat prior. This is due to the shrinkage effect of the prior, introducing bias and a reduction of the variance following the usual bias-variance trade-off. The magnitude of this shrinkage effect can be seen for $N = \{80, 129\}$ in Table 4.10, which provides the bias, variance and MSE for the posterior median when a flat or Cauchy prior is used. Increasing the sample size to $N = 500$ results in coverage that is closer to nominal and similar regardless of the prior used.

For the simulations investigated, taking into consideration the uncertainty in the joint distribution of (E, x) when calculating the PAR and PAF resulted in similar or improved coverage of the credible intervals in terms of their Frequentist properties. When the PAR is close to 0, using fixed estimates on $P(E|x)P(x)$ calculated directly from the data appears to provide similar coverage to when $P(E, x)P(x)$ is allowed to vary. Therefore, the use of the methodology developed by Newson (2013) may be adequate in most practical situations. However, we still recommend the Bayesian approach be applied to account for the additional uncertainty. When the PAR is large, allowing the estimates of $P(E|x)P(x)$ to vary results in improved coverage of

the intervals in most cases. This improved coverage is particularly noticeable with the PAR, as opposed to the PAF, and especially as the sample size decreases. The choice of prior applied to the β coefficients of the logistic regression model becomes more important when the PAR is large. The weakly informative Cauchy prior results in improved coverage, particularly when the sample size decreases. However, this is at the cost of increased bias, which for very extreme values of PAR (e.g. 0.9-1) could result in a decrease in coverage.

4.8 Simulation Study: Bovine mastitis data

The simulations carried out in this section aim to assess the performance of our Bayesian methods for estimating the PARate, PARF and their credible intervals using the bovine mastitis dataset as an example. Three different simulation studies have been incrementally carried out to take into consideration additional sources of uncertainty. Each simulation study is begun by generating a large population of cows, where their covariates, x, are based on specified distributions and their rate of mastitis per day is determined via a Poisson regression model. From this population a separate random sample can be drawn for each iteration of the simulation study. For the first simulation the population is generated without taking into consideration the fact that cows are clustered into herds (see Section 4.8.1). When generating the population for the second simulation, the covariates for each cow were derived under the assumption that cows from the same herd are more likely to be similar in terms of their covariates than cows from different herds (see Section 4.8.2). That is, clustering has only been considered in the covariates and not in the rate of disease. The final simulation aimed to take into consideration clustering in both the covariates and the rate of disease by adopting a generalised linear mixed effects model, where herd is treated as a random effect (see Section 4.8.3). Before simulations were carried out however, model selection was performed on the original mastitis dataset. A model incorporating each of the variables listed in Table 4.3 was considered. Any variables that were not significant at the 5% level were sequentially removed from the model. Models were compared using the

	Median	Bias (10^{-2})	Variance(10^{-3})	MSE (10^{-3})
Cauchy prior				
PAR (N=80)	0.534	-3.5	5.3	6.56
PAF (N=80)	0.744	-3.7	11.5	12.88
PAR (N=129)	0.549	-2	3.4	3.83
PAF (N=129)	0.767	-1.4	7.1	7.25
Flat prior				
PAR (N=80)	0.556	-1.3	15.2	15.24
PAF (N=80)	0.780	-0.1	258.2	258.2
PAR (N=129)	0.567	-0.24	3.8	3.85
PAF (N=129)	0.793	1.2	8.7	8.82

Table 4.10: Comparison of the posterior median estimates of the PAR and PAF, using the Cauchy and flat priors, when the $P(E|x)P(x)$ is allowed to vary. The true PAR and PAF for this case are 0.569 and 0.781 respectively.

likelihood ratio test, with the most favored model being:

$$\log(\lambda_i) = -7.1706 - 0.4476\, Tx + 0.2338\, Age - 0.3013\, Age^2 + 0.7717\, Premast, \qquad (4.35)$$

where age has being standardised according to $[PARCAT - mean(PARCAT)]/sd(PARCAT)$, and $PARCAT$ is treated as numeric, rather than categorical. The quadratic term for age was considered as it is not unreasonable to assume that the risk of mastitis in younger cows, calving for the first time, and older cows which are not calved as often, will be less than those of parity 2 and 3 which are calved intensively. It should also be noted that herd was found not to be significant at the 5% level. One may think this means herd need not be considered further. However, the distribution of the other covariates in the model could vary between herds. Such clustering could significantly affect the precision of estimation of the joint distribution of (E, x).

When calculating the PARate and PARF an estimate for the joint distribution of exposure and the other covariates is also required. For the mastitis dataset where we have many covariates, both numeric and categoric, we provide estimates for (E, x) from a discrete distribution which places point masses (i.e. weights), w, on the observed data. If equal fixed weights were assigned to each individual cow in the dataset (i.e. $w_i = 1/n$, so $\sum_{i=1}^{n} w_i = 1$), this would correspond to

the empirical distribution of (E, x) in the sample, and the uncertainty in the estimates of the joint distribution (E, x) would not be taken into consideration. To allow for this uncertainty the Bayesian bootstrap approach (see Section 4.5), where weights are drawn from an n-dimensional Dirichlet$(1, \ldots, 1)$ distribution, is adopted. This results in different weights, w_i, being assigned to each individual cow from a distribution, where again $\sum_{i=1}^{n} w_i = 1$. Since cows are clustered into herds it maybe more appropriate to assign weights using a clustered version of the Bayesian bootstrap (see Section 4.5.1), so that each cow in the same herd is assigned the same weight. These methods of assigning weights to estimate the joint distribution of (E, x) are compared over the simulations. Furthermore, the PARate and PARF must be defined over a pre-specified interval of time. For the mastitis dataset the PARate and PARF are estimated as the rate of mastitis per year. Once the PARate, PARF and their credible intervals have been estimated for the desired number of iterations, the overall percent coverage is calculated taking note of the interval length.

4.8.1 Simulation 1: No clustering considered

A finite population of one million cows was generated for this simulation, where the effect of clustering has not being taken into consideration in the covariates or the response variable. To generate the population, distributions were specified for each of the covariates in the model (4.35) based on those seen in the mastitis dataset. Note that it may not be unreasonable to assume that the DAR for a particular cow will be dependent on its age. However, no obvious difference could be seen in the distribution of DAR for each parity group.

Once each of the covariates has been specified, $\lambda(E, x)$ can be estimated using a Poisson regression model. Since previous simulations based on the PAR showed that larger PAR values highlight differences, as well as investigating the model (4.35) two other models are proposed. With the additional models the β coefficients from (4.35) are changed to provide a larger effect due to E (i.e. the treatment) and thus a large value of PARate. Creating a larger effect with respect to treatment can be achieved by either changing the coefficient β_1 for treatment to a large

negative value, which implies the treatment provides a larger protective effect, or by changing the intercept term β_0 which affects all the rates. The two additional models investigated are:

$$\log(\lambda_i) = -4 - 0.8Tx + 0.2338Age - 0.3013Age^2 + Premast, \tag{4.36}$$

and

$$\log(\lambda_i) = -2 - 0.8Tx + 0.2338Age - 0.3013Age^2 + Premast. \tag{4.37}$$

A population was generated under each of these models using the following procedure:

1. Draw an age for each cow from a Poisson($\lambda = 1$) + 1 distribution.
2. Select the cow's premast status from a Binomial($n = 1, p = P$[premast|age]) distribution, based on logistic regression of the original data.
3. Assign each cow to either the treatment or control group according to a Binomial($n = 1, p = 0.5$) distribution.
4. Determine the length of time in days that the cow is at risk of mastitis from a Normal($\mu = 320, \sigma = 30$) distribution.
5. Estimate the rate of disease, λ_i, from the Poisson regression model (4.35), (4.36) or (4.37).
6. Determine the number of cases of mastitis for each cow according to the Poisson(λ_i) distribution.

From each population a total of 1,000 datasets, of size 601, were generated using simple random sampling. The true PARate per year for the simulated populations under models (4.35), (4.36) and (4.37) was calculated as 0.085, 3.22 and 23.81 respectively based on equations (4.14) and (4.16). The true PARF for the simulated populations generated under these models was 0.359, 0.550 and 0.551 respectively. A reduction in the rate of mastitis by 23 cases is unrealistic for this example. However, in the context of quality control a reduction in the rate of defects in a

large lot by 23 may not be unreasonable. Therefore, we make use of the model (4.37) to simply further explore the parameter space.

4.8.1.1 MCMC sampler

To calculate the PARate and PARF for each simulated dataset from the population we require estimates for $\lambda(E^+, x)$ and $\lambda(E^-, x)$ where E^+ represents the control and E^- the treatment. To estimate these parameters we implement a Bayesian Poisson regression model, where each of the β coefficents are updated using the Metropolis-Hastings random walk sampler (see Section 3.2.3.1). To initalise the random walk sampler we specify each of the βs according to their maximum likelihood values for the Poisson regression model run on each simulated dataset. To update the random walk sampler, prior and proposal distributions need to be specified. Similarly to the low birth weight sampler, we choose to perform block-wise updating using the multivariate normal proposal distribution $Q(\beta^\dagger|\beta^t) \sim \text{MVN}(\beta^t, c\Sigma)$, where $\beta^\dagger$ represents the proposed value of β, β^t the current value of β, c a fixed constant tuning parameter which provides an acceptance rate of approximately 24% (Roberts et al., 1997) and Σ the covariance matrix for the βs from the Poisson regression model. For choice of prior we assign a Cauchy$(0, 10)$ to the intercept term and a Cauchy$(0, 2.5)$ to all other coefficients as suggested by Gelman et al. (2008). Note that the variable age, which is the only non-binary variable in the models (4.35-4.37), was standardised as required for use of these priors.

To calculate the acceptance probability the contribution to the likelihood function also needs to be estimated. Recall that the mastitis data can be described by the Poisson distribution (4.2), which has the log-likelihood function

$$l(\lambda(E, x)) = \sum_{i=1}^{n} y_i \log(\lambda_i DAR_i) - \lambda_i \, DAR_i, - \log(\prod_{i=1}^{n} y_i!), \tag{4.38}$$

where DAR is the number of days at risk. We now have all the necessary information to outline the MCMC sampling procedure which is used within the simulation.

1. Define the initial conditions for β and initialize an iteration counter at $t = 0$.

2. Generate n weights to describe the joint distribution of (E, x), either by assigning equal probability to each individual (i.e. $1/n$), or by drawing from an n-dimensional Dirichlet$(1, \ldots, 1)$ distribution (i.e. Bayesian bootstrap).

3. Generate a candidate value, $\beta^\dagger$, from the proposal distribution $Q(\beta^\dagger|\beta^t) \sim \text{MVN}(\beta^t, c\Sigma)$, where β^t is the value of the Markov chain at iteration t, c a fixed tuning parameter and Σ the covariance matrix of the β's from the Poisson regression model.

4. Calculate $\lambda(E^+, x)$ by multiplying β by the control model matrix X_0 where $E = E^+$, and similarly $\lambda(E^-, x)$ by multiplying β by the treatment model matrix X_1 where $E = E^-$, then back-transform by exponentiating.

5. Compute the probability of acceptance

$$\alpha = \min\left\{1, \frac{p(\beta^\dagger)L(\lambda^\dagger)Q(\beta^t|\beta^\dagger)}{p(\beta^t)L(\lambda^t)Q(\beta^\dagger|\beta^t)}\right\}, \tag{4.39}$$

   where $p(\beta)$ represents the prior for β and $L(\lambda(E, x))$ the likelihood given by exponentiating (4.38).

6. Take a random draw, U, from a Uniform$(0, 1)$ distribution. If $U \leq \alpha$ set $\beta^{t+1} = \beta^\dagger$, otherwise set $\beta^{t+1} = \beta^t$.

7. Set the iterate counter from t to $t + 1$.

8. Repeat steps 2-7 until desired number of iterations is reached.

Having obtained weights and the posterior distributions for β and $\lambda(E, x)$ following the above procedure, the posterior distribution for the PARate and PARF can be estimated by applying the formulae (4.20) and (4.21) for a year long duration to each iteration. Before running this MCMC algorithm though, a tuning period should be carried out on a simulated

dataset from the population to ensure an appropriate acceptance rate. For each population generated under the models (4.35-4.37) a tuning parameter of 1 provided an acceptance rate of approximately 24%. Convergence should also be inspected. BGR analysis suggested the sampler converged very quickly ($< 1,000$ iterations) on simulated data from each population, so a burnin of 1,000 iterations was applied in this case, with a total of 10,000 iterations for each chain run after burnin.

4.8.1.2 Simulation results

Table 4.11 shows the percent coverage, posterior mean and median, and the mean interval length for the PARate and PARF from the population generated under each of the models (4.35)-(4.37). It can be seen that when the PARate is small, as is the case under under model (4.35), there is no difference in the percent coverage when using fixed weights of equal value, compared to allowing these values to vary by re-sampling using the Bayesian bootstrap. Moreover, nominal 95% coverage is achieved in this case. When the PARate is large, as is the case under models (4.36-4.37), re-sampling the weights using the Bayesian bootstrap results in superior coverage that is close to nominal. As expected the mean interval length is larger when re-sampling the weights as additional uncertainty in the joint distribution of (E, x) is being taken into consideration. However, the mean interval length is not much larger than that provided when fixed weights of equal value are used.

The PARF had close to nominal coverage in all cases. Interestingly though, for PARF it was found that regardless of whether fixed or re-sampled weights were used the coverage, intervals and estimates remain the same. This occurred because having estimated $\lambda(E^+, x)$ and $\lambda(E^-, x)$ for the same model, $\lambda(E^+, x)$ is simply a scalar multiple, k, of $\lambda(E^-, x)$. That is, $\lambda(E^-, x)$ and $\lambda(E^+, x)$ are linearly dependent. This means that the PARF simplifies to:

$$PARF = 1 - \frac{w_1\lambda(E^-, x_1) + \ldots + w_n\lambda(E^-, x_n)}{w_1\lambda(E^+, x_1) + \ldots + w_n\lambda(E^+, x_n)} \tag{4.40}$$

Model		Coverage (%)	Mean	Median	CI length
PARate					
(4.35)	Fixed $w = 1/n$	95.1	0.085	0.085	0.15
	Bayesian bootstrap	95.1	0.085	0.086	0.15
(4.36)	Fixed $w = 1/n$	93.7	3.21	3.22	0.71
	Bayesian bootstrap	95.7	3.21	3.22	0.75
(4.37)	Fixed $w = 1/n$	85.0	23.80	23.79	1.92
	Bayesian bootstrap	94.6	23.80	23.79	2.74
PARF					
(4.35)	Fixed $w = 1/n$	94.5	0.34	0.35	0.53
	Bayesian bootstrap	94.5	0.34	0.35	0.53
(4.36)	Fixed $w = 1/n$	95.7	0.55	0.55	0.08
	Bayesian bootstrap	95.7	0.55	0.55	0.08
(4.37)	Fixed $w = 1/n$	94.7	0.55	0.55	0.03
	Bayesian bootstrap	94.7	0.55	0.55	0.03

Table 4.11: Simulation 1 results for the PAR and PARF. The true PARate for model 1, 2 and 3 are 0.085, 3.22 and 23.81 respectively and true PARF 0.36, 0.55 and 0.55 respectively. Where each chain of the MCMC algorithm was run for 10,000 iterations after burnin with a tuning parameter of 1. Note that the CI length represents the mean CI length for each of the 100,000 generated datasets.

$$= 1 - \frac{w_1 k\lambda(E^+, x_1) + \ldots + w_n k\lambda(E^+, x_n)}{w_1\lambda(E^+, x_1) + \ldots + w_n\lambda(E^+, x_n)} \qquad \text{since } \lambda(E^-, x) = k\lambda(E^+, x) \tag{4.41}$$

$$= 1 - k \tag{4.42}$$

irrespective of the distribution of x. However, the above simplification only occurs in this case due to the mastitis data being based on an experiment. As the treatment is not present in the underlying population it will be independent of the covariates x.

4.8.2 Simulation 2: Clustering in x

In this simulation we wish to take into consideration that the distribution of the covariates, excluding the randomly assigned treatment, could vary between herds. This is achieved by generating the population at herd level, rather than the cow level as was done in Simulation 1. Similarly to Simulation 1, covariates for each cow in the population distributions are generated

based on those seen in the data at the herd level. As the largest discrepancy in percent coverage for PARate occurred for the population generated under model (4.37) in the previous simulation, this model is used again here. A population of 1,000 herds was generated for this simulation in the following way:

1. Select a herd size from a Poisson($\lambda = 75$) distribution.
2. Select the age for each cow from a Poisson(λ_h = Average herd age) + 1 where $\lambda_h \sim$ log normal($\mu = 1, \sigma = 0.3$).
3. Select the premast status for each cow from a Binomial($n = 1, p = P(\text{premast}|\text{age})$) distribution.
4. Assign each cow to the treatment or control group following the Binomial($n = 1, p = 0.5$) distribution.
5. Determine the length of time in days that the cow is at risk of disease from a Normal($\mu = 320, \sigma = 30$) distribution.
6. Estimate the rate of disease, λ_i, according to the Poisson regression model (4.37).
7. Determine the number of cases of disease each cow has according to the distribution Poisson(λ_i).

From this population 1,000 simulated datasets were generated using cluster sampling of h herds, where an h of 8 and 25 were investigated. The true PARate and PARF for this population, calculated according to (4.14) and (4.16), are 15.72 and 0.55 respectively. To estimate the posterior distributions for β, and by extension the PARate and PARF, the random walk sampler described in Section 4.8.1.1 is applied with the addition that weights describing the joint distribution of (E, x) are generated according to the cluster Bayesian bootstrap. The tuning period of the sampler on simulated data, with $h = 8$ and $h = 25$ herds, indicated that the tuning parameter of 1.25 provided closest to optimal acceptance in both cases. BGR analysis showed

that a longer burnin period than Simulation 1 was required for the sampler to converge in this case. A total of 8,000 iterations were run, after a burnin of 3,000 iterations, on each simulated dataset.

4.8.2.1 Simulation results

The results provided in Table 4.12 clearly show that the method used for assigning w when estimating the PARate has a huge effect on the percent coverage of the credible interval. As expected assigning equal weight to each individual, which disregards the use of cluster sampling in the experimental design and uncertainty on the joint distribution of (E, x), has resulted in short intervals that provide poor coverage. When uncertainty in the joint distribution of (E, x) is accounted for by allowing w to vary via Bayesian bootstrap re-sampling, but the cluster-sampling in the experimental design is not considered, the percent coverage is superior to using fixed equal weights, but still less than nominal. This is likely a result of the increased interval length. The preferred approach however, and that with the highest percent coverage, was estimation of the PARate where w was assigned using the cluster Bayesian bootstrap. Coverage was still less than nominal, but it can be seen that the coverage improved when the number of herds sampled was increased from 8 to 25. Furthermore, it appears that when the number of herds is small (i.e. $h = 8$) the improved coverage using the cluster Bayesian bootstrap comes at the cost of a much larger interval length.

Similar to what was found in Simulation 1, the PARF provides the same results regardless of how w is assigned due to the linear dependence of $\lambda(E^+, x)$ and $\lambda(E^-, x)$. Table 4.13 shows that the PARF achieves greater than nominal coverage when the herd size is small, but when the number of herds increases the coverage decreases slightly due to a decrease in interval length.

Number of herds		Coverage (%)	Mean	Median	CI length
PARate					
8	Fixed $w = 1/n$	60.0	15.76	15.75	1.56
	Bayesian bootstrap	80.7	15.76	15.75	2.37
	Cluster Bayesian bootstrap	90.3	15.75	15.75	3.24
25	Fixed $w = 1/n$	59.6	15.75	15.72	0.88
	Bayesian bootstrap	77.6	15.75	15.72	1.34
	Cluster Bayesian bootstrap	92.5	15.75	15.72	1.99
PARF					
8	Fixed $w = 1/n$	96.2	0.55	0.55	0.037
	Bayesian bootstrap	96.2	0.55	0.55	0.037
	Cluster Bayesian bootstrap	96.2	0.55	0.55	0.037
25	Fixed $w = 1/n$	94.8	0.55	0.55	0.021
	Bayesian bootstrap	94.8	0.55	0.55	0.021
	Cluster Bayesian bootstrap	94.8	0.55	0.55	0.021

Table 4.12: Simulation 2 results for the PARate and PARF. The true PARate and PARF for this model are 15.72 and 0.55 respectively. Each chain of the MCMC algorithm was run for 8,000 iterations after burnin with a tuning parameter of 1.25. Note that the CI length represents the mean CI length for each of the 100,000 generated datasets.

4.8.3 Simulation 3: Clustering in x and y

Cows that are in the same herd may be more similar in disease status than those in other herds, even after allowing for covariate effects. This may especially be true if the disease is infectious, but could also be due to the effect of unobserved covariates. The most appropriate model to use, which takes into consideration the hierarchical structure of cows being clustered into herds, is a generalised linear mixed effects model where herd is specified as a random effect. Specifically the model used to generate the population for this simulation is an extension of model (4.37) where the random effect of herd is described by $u_i \sim \text{Normal}(\mu = 0, \sigma = 0.3)$. As the original mastitis data did not show any significant effect due to herd the standard deviation of 0.3 for the random effect was chosen to provide a larger herd effect. Similarly to Simulation 2, to generate the covariates for each cow in the population, distributions for each covariate are selected based on those seen in the dataset at the herd level. A population of 1,000 herds was generated for this simulation in the following way:

1. Select the herd size from a Poisson($\lambda = 75$) distribution.

2. Select the age for each cow from a Poisson(λ_h = Average herd age) + 1 where $\lambda_h \sim$ log normal($\mu = 1, sd = 0.3$).

3. Determine the premast status for each cow from a Binomial($n = 1, p = P(\text{premast}|\text{age})$) distribution.

4. Assign each cow to the treatment or control group following the Binomial($n = 1, p = 0.5$) distribution.

5. Determine the length of time in days that each cow is at risk of disease from the Normal($\mu = 320, \sigma = 30$) distribution.

6. Determine the random effect term for each herd from a Normal($\mu = 0, \sigma = 0.3$) distribution, and assign each cow the random effect value corresponding to their herd.

7. Estimate the rate of disease according to $\lambda_i = \exp(x_{ij}^T\beta + u_i)$, where $x_{ij}^T\beta$ is described by (4.37).

8. Determine the number of cases of disease each cow has according to the distribution Poisson(λ_i).

From this population 1,000 simulated datasets were generated using cluster sampling for herds of size $h = 8$ and $h = 25$. The true PARate and PARF for this population, calculated according to (4.14) and (4.16), are 16.43 and 0.55 respectively. To generate the posterior distributions of PARate and PARF the MCMC sampler now needs to be extended so that u_i along with its corresponding precision τ_u are also updated. An outline of the details of the MCMC sampler is given in the following section.

4.8.3.1 MCMC sampler

The MCMC sampler for this simulation is made up of four distinct processes. First is the updating of w according to the cluster bootstrap re-sampling procedure. Second is the updating of the random effect u_i using an independence sampler (i.e. a random walk sampler where the prior distribution is equal to the proposal distribution). Third is the updating of τ_u and by extension σ_u^2 the hyper-parameters of the distribution of u_i. Then finally comes the updating of β using the random walk sampler similarly to that described in Simulation 1 (steps 3-6 Section 4.8.1.1). Before we can update the parameters β, u_i and τ_u they must first be initalised. A Poisson mixed effects model is run for each simulated dataset, from which the βs are initalised by taking the maximum likelihood estimate for each of the fixed effects. The u_is are initalised by taking the maximum likelihood estimate for each of the random effects, and σ_u^2 (and thus τ_u) by taking the models variance associated with the random effect.

To update u_i we choose to adopt a component-wise random walk approach with proposal distribution $Q(u_i^\dagger|u_i^t) \sim \text{Normal}(0, \sigma_u^2)$. This proposal distribution is also used as the prior distribution for u_i. By doing this the calculation of the acceptance probability (α_1) in the random walk approach is simplified due to cancellation of proposal and prior contributions. This type of sampler is known as an independence sampler, where the acceptance probability is simply a ratio of the likelihoods. The log-likelihood in this example is given by (4.38). To update the variance component, σ_u^2, for the random effect it is often easier to work with the precision $\tau_u = 1/\sigma_u^2$. If a gamma prior distribution, say $\tau_u \sim \text{Gamma}(a = 0.001, b = 0.001)$ which implies a large variation between herds, is specified then due to the conjugacy relationship with the normal distribution of known mean, Gibbs sampling can be used because the posterior distribution for τ_u can be specified analytically as given by (4.29). Incorporating these additional extensions to account for using a mixed effects model the MCMC procedure becomes:

1. Define the initial conditions for β, u_i and σ_u^2 and initialise an iteration counter to $t = 0$.

2. Generate n weights to describe the joint distribution of (E, x) via cluster bootstrap re-

sampling.

3. Begin implementation of the independence sampler for the random effect by generating a candidate value $u_i^\dagger$ from a Normal($\mu = 0, \sigma = \sqrt{1/\tau_u^t}$) distribution.

4. Estimate $\lambda^\dagger$ according to $\lambda^\dagger = \exp(x_{ij}^T\beta + u_i^\dagger)$.

5. Compute the probability of acceptance

$$\alpha_1 = \min\left\{1, \frac{L(\lambda^\dagger)}{L(\lambda^t)}\right\}. \tag{4.43}$$

6. Take a random draw, r_1, from a Uniform$(0, 1)$ distribution. If $r_1 \leq \alpha_1$ set $u_i^{t+1} = u_i^\dagger$, otherwise set $u_i^{t+1} = u_i^t$.

7. Repeat steps 3-6 until the entire vector of u_i has been updated.

8. Update the precision τ_u for the random effect by drawing from the gamma posterior distribution (4.29).

9. Begin implementation of the random walk sampler for the βs by generating a candidate value, $\beta^\dagger$, from the proposal distribution $Q(\beta^\dagger|\beta^t) \sim \text{MVN}(\beta^t, c\Sigma)$, where β^t is the value of the Markov chain at iteration t, c a fixed tuning parameter and Σ the covariance matrix of the βs from the Poisson regression model.

10. Estimate $\lambda(E^+, x)$ by multiplying β by the model matrix X_0 where $E = E^+$ (i.e. control), and similarly $\lambda(E^-, x)$ by multiplying β by the model matrix X_1 where $E = E^-$ (i.e. treatment), then back-transform by exponentiating.

11. Compute the probability of acceptance

$$\alpha_2 = \min\left\{1, \frac{p(\beta^\dagger)L(\lambda^\dagger)Q(\beta^t|\beta^\dagger)}{p(\beta^t)L(\lambda^t)Q(\beta^\dagger|\beta^t)}\right\}, \tag{4.44}$$

where $p(\beta)$ represents the prior for β and $L(\lambda(E,x))$ the likelihood given by exponentiating (4.38).

12. Take a random draw, r_2, from a Uniform$(0,1)$ distribution. If $r_2 \leq \alpha_2$ set $\beta^{t+1} = \beta^{\dagger}$, otherwise set $\beta^{t+1} = \beta^t$.

13. Set the iteration counter from t to $t+1$.

14. Repeat steps 2-13 until desired number of iterations is reached.

After obtaining weights and the posterior distribution for $\lambda(E,x)$ through the above procedure, the posterior distribution for the PARate and PARF can be estimated by applying the formulae (4.20) and (4.21) to each iterate. Before carrying out this MCMC algorithm a tuning period should be carried out to ensure an appropriate acceptance rate. For the population generated for this simulation a tuning parameter of 0.5 provided an acceptance rate of approximately 24%. Convergence should also be inspected. BGR analysis on simulated data for this MCMC sampler suggested a burnin period of 3,000 iterations would be sufficient. A total of 5,000 iterations after burnin were carried out for each of the 1,000 simulated datasets.

4.8.3.2 Simulation results

The results provided in Table 4.13 show that the PARF has closer to nominal coverage than the PARate. Coverage for the PARate was particularly low when the number of herds was small. An increase in the number of herds from 8 to 25 resulted in an increase in coverage, but this was still less than nominal. Mean and median estimates of the PARate and PARF were both similar to that of the population. The PARF achieved close to nominal coverage regardless of the total number of herds in the sample. However, it should be noted that when the number of herds was small (i.e. $h = 8$) this close to nominal coverage has come at the cost of a larger interval length.

Taking into consideration the simulations performed it can be seen that not accounting

Number of herds		Coverage (%)	Mean	Median	CI length
PARate					
8	Cluster Bayesian bootstrap	86.7	16.39	16.25	6.68
25	Cluster Bayesian bootstrap	91.6	16.45	16.41	4.25
PARF					
8	Cluster Bayesian bootstrap	94.6	0.55	0.55	0.037
25	Cluster Bayesian bootstrap	94.4	0.55	0.55	0.021

Table 4.13: Simulation 3 results for the population attributable rate. The true PARate and PARF for this population were 16.43 and 0.550 respectively. Each chain of the MCMC algorithm was run for 5,000 iterations after burnin with a tuning parameter of 0.5. Note that the CI length represents the mean CI length for each of the 100,000 generated datasets.

for the appropriate sampling design, and its effect on the covariates, can result in poor coverage of the credible intervals when PARate is large. Similarly to the PAR and PAF, it was confirmed that accounting for the uncertainty in the joint distribution of (E, x) resulted in superior coverage of the intervals for the PARate and PARF.

4.9 Concluding remarks

Clear definitions for estimating the PAR and PARate when confounding variables are present are not currently found in the literature. Here we provide formal definitions for both these measures, along with the PAF and PARF, and illustrate how these and their corresponding uncertainty can be estimated from real-world epidemiological data. We confirm that Newson (2013) used fixed estimates for the joint distribution of (E, x) when calculating confidence intervals for the PAR and PAF from the low birth weight data. Newson's approach provides approximately nominal coverage when PAR is small, however, this is not the case when PAR is large. Using the Bayesian approach where (E, x) is described by a Dirichlet$(n_1 + 1, \ldots, n_6 + 1)$ distribution and the response modeled using Bayesian logistic regression, provides superior coverage for all cases examined. This is in part due to the increased interval length, as a larger amount of uncertainty has being taken into consideration, but these intervals are not so much larger that they are no longer useful.

In practice data collected may often have a hierarchical structure, which needs to be considered in both the joint distribution of (E, x) as well as the modeled rate (or probability) of disease. Through sequential simulations we examine the performance of different model choices and methods for specifying the joint distribution of (E, x), when the underlying population has no clustering, clustering in x but not y and clustering in both x and y. When no clustering is present in the population and y modeled using Bayesian Poisson regression, allowing (E, x) to varying via Bayesian bootstrap re-sampling, in comparison to using fixed estimates, resulted in superior coverage when the PARate was large. When clustering had an effect on x but not y in the population, re-sampling (E, x) using the cluster Bayesian bootstrap, as opposed to the standard Bayesian bootstrap or fixed estimates, provided improved (but still less than nominal) coverage for the PARate. Sampling a larger number of clusters resulted in closer to nominal coverage when the cluster Bayesian bootstrap was used to specify (E, x). When clustering had an effect on both x and y in the population, modeling y using a Poisson mixed effects model and specifying (E, x) using the cluster Bayesian bootstrap provided less than nominal coverage when the cluster size was 8 and 25.

As we are measuring the performance of each these Bayesian methods through their Frequentist properties, this does not necessarily guarantee that asymptotically the coverage achieved will be exact. With clustered data, Frequentist asymptotics rely on the number of clusters which are sampled, as well as the total number of individuals. Therefore, better coverage might be obtained with a much larger sample of clusters, although in practice only a moderate number of clusters are likely to be sampled.

Chapter 5

Conclusions and Future Research

The population attributable risk (or rate) provides practitioners with a means of describing the effect of an intervention in advance. Chapter 1 shows the many different definitions that have been proposed in the literature under the name population attributable risk, which have resulted in much confusion. We suggest that this ambiguity in definition surrounding population attributable measures has resulted in the gap in our knowledge for estimating the uncertainty for the PAR expressed by (1.7). This thesis begins to address this knowledge gap through three main objectives, originally outlined in Chapter 1:

1. To develop and compare Frequentist and Bayesian methodologies for estimating the PAR from a cross-sectional study with a single dichotomous risk factor.

2. To develop Bayesian methodologies to estimate the PAR and PAF for situations where the statistical model is under-identified.

3. To extend the Bayesian methodologies developed to account for more complex designs and attributable rates.

In this chapter we summarise the statistical advances made throughout this thesis in exploring these three objectives and provide direction for future avenues of research.

5.1 Summary

Chapter 2 was devoted to achieving the first of our thesis objectives. For cross-sectional studies we adopted standard Frequentist approaches for variance estimation and confidence interval construction for the PAR which were previously adopted by many authors (Walter, 1975, 1976; Benichou and Gail, 1990; Kooperberg and Petitti, 1991; Lehnert-Batar et al., 2006). These Frequentist approaches (including the delta, jackknife and bootstrap methods) were then compared, in terms of their Frequentist properties, to a Bayesian approach through simulation. We saw adopting a Bayesian approach as necessary due to it providing a flexible platform for extension to situations where the probability of disease or exposure may not be accurately estimated from the data (e.g. case-control and cohort studies). The Bayesian approach with the standard reference prior Dirichlet$(1, 1, 1, 1)$ resulted in nominal coverage for almost all practically realistic values of PAR and is recommended especially when there are small or moderate observed counts in the table. In extreme situations such as mass outbreak of disease, where the PAR is close to +1 (or alternatively close to -1 in the case of vaccination) the use of the Dirichlet$(1, 1, 1, 1)$ prior resulted in inadequate coverage. For these extreme cases, we found the use of the Dirichlet$(1, 1, \epsilon, \epsilon)$ prior provided a flat distribution over the parameter space of PAR as well as close to nominal coverage.

Chapter 3 addresses the second objective of this thesis by extending the work of Chapter 2 to account for the situation where the model is under-identified. We demonstrate that a Bayesian approach for estimating the PAR, PAF and their credible intervals for case-control and cohort studies is straightforward, if a conjugate beta prior is applied to either the $P(D^+)$ or $P(E^+)$ respectively. If prior information were only available for the $P(E^+)$ for a case-control study, or $P(D^+)$ for a cohort study, then we must resort to MCMC approaches. We describe a new MCMC sampler which proposes switching between two different portions of the parameter space and discuss its limitations. Additionally, we explore the more complex situation where the uncertainty associated with an imperfect diagnostic test is incorporated into the model,

specifically for data arising from a cross-sectional study. As there are more parameters to estimate than independent observations in this case a Bayesian approach is the only option. Since standard MCMC samplers (e.g. Metropolis-Hastings and Gibbs) are reported to perform poorly when the model lacks of identification (Gustafson, 2015) we develop a novel MCMC sampler which aims to effectively explore the posterior ridge of the under-identified model, by taking into consideration the shape of the ridge. Although our approach does not perform as well as Gustafson's importance sampling approach, it does outperform standard MCMC samplers in terms of effective sample size when the sample size is large. Gustafson's method may perform poorly in circumstances where few of the sampled values satisfy the necessary constraints, due to the acceptance rate being particularly low. Under such circumstances there is the potential for our model to out perform the importance sampling approach. Theoretical justifications provided within this chapter also demonstrate that if perfect test sensitivity can be achieved then the PAR becomes identifiable.

Our third thesis objective is explored in Chapter 4 by extending the PAR and PARate (as well as PAF and PARF) to take into consideration multiple covariates in addition to the risk factor being considered for removal. Estimating the PAR and PAF (and PARate/PARF) whilst incorporating these additional confounders requires not only a model for the response, y, in terms of the covariates, x, but also an estimate of the joint distribution of the exposure and the covariates, (E, x). We confirm that Newson (2013) used fixed estimates for the joint distribution of (E, x) when calculating confidence intervals for the PAR and PAF. We then show that, using a Bayesian approach where x is the categorical variable race, (E, x) described by a Dirichlet$(n_1 + 1, \ldots, n_6 + 1)$ distribution and y modeled using Bayesian logistic regression, provides superior coverage and reasonable interval length for all cases examined. Within Chapter 4 we also make extensions to the developed methodology to account for clustered data and numerical covariates. For the situation where clustering has an effect on x but not y, we use a Bayesian Poisson regression model to describe the rate of disease and cluster Bayesian bootstrap re-sampling to allow for variability in joint distribution of (E, x). This approach results in superior coverage,

in comparison to simply using fixed values or standard Bayesian bootstrap re-sampling, when the PARate was large. Extending to a Bayesian Poisson mixed effects model to account for the case where clustering has an effect on both x and y resulted in less than nominal coverage for PARate. However, when the number of clusters sampled was increased the coverage for the PARate improved.

5.2 Future research

This thesis has highlighted the need for standard approaches for estimating the uncertainty surrounding the population attributable risk and population attributable rate. Our research begins to fill the current knowledge gap by providing methods for estimating credible intervals for the PAR, PAF, PARate and PARF for commonly used epidemiological study designs. However, many other complex study and experimental designs can be carried out by practitioners which require thorough investigation.

In Chapter 4 we investigated a moderately sized experiment where cluster sampling had been carried out. It was noted that using a Bayesian Poisson mixed effects model to describe the rate of disease along with the cluster Bayesian bootstrap to describe the joint distribution (E, x), resulted in slightly less than nominal coverage, and that a large number of clusters may improve this result. In particular we could extend this Bayesian approach to allow for estimation of the PARate and PARF (or PAR and PAF) when the data come from a complex sample survey with stratification as well as clustering. This would involve extending the model used to describe the rate of disease, to one which takes into consideration the complex survey design. Additionally, the cluster Bayesian bootstrap can be used to account for clustering in the survey data when specifying the joint distribution (E, x). Stratification could then be accounted for by performing the cluster bootstrap separately within each stratum, ensuring that the total weight in each stratum is kept constant.

In Chapter 3 we described a new MCMC sampler for under-identified models arising from

cross-sectional studies where an imperfect diagnostic test was adopted. This provided improved effective sample sizes over standard MCMC samplers (i.e. Metropolis-Hastings random walk and Gibbs) when the sample size was large. This is due to more efficient exploration of the posterior ridge for this sampler as a result of incorporating information about the shape of the ridge through the Jacobian matrix. However, currently this sampler suffers from potentially spending long periods of time exploring or revisiting parts of the parameter space which are of low posterior probability, a limitation inherited from its predecessor the Metropolis-Hastings random-walk sampler. To hopefully improve the efficiency with which this sampler explores the parameter space we could try using a combination of our adapted random walk sampler with the MALA or HMC samplers. However, such an approach would require an increased number of tuning parameters which may present difficulties.

One aspect we have yet to touch on in this research is the proposal of a population attributable measure where the exposure of interest is a continuous variable. To our knowledge there is no current attributable measure that allows one to describe the reduction of risk that could be achieved if the distribution of exposure, (E, x), could be changed in some way. An example of where such a measure could be useful is in changing legal blood alcohol limits for driving. Different scenarios could be carried out to estimate the reduction in the risk, with a measure of uncertainty, for each scenario.

Regardless of the measure, statistically rigorous methods for quantifying the uncertainty are required. By successfully addressing the objectives outlined in this thesis we have presented a start to the development of such methodology for population attributable measures. We provide a deeper understanding of population attributable measures and the models used to estimate them; comparisons of method performance when more than one was available; a range of real-world applications and a collection of open source code for the implementation of our developed methods. Furthermore, this thesis also highlights plentiful opportunities for future work surrounding population attributable measures as each underlying study or experimental design warrants its own investigation.

Bibliography

Baker, S. G. (1994). The multinomial-poisson transformation. *J. Royal Stat. Soc. Series D*, 43(4):495–504.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.*, 67(1):1–48.

Benichou, J. (1991). Methods of adjustment for estimating the attributable risk in case-control studies: A review. *Stat. Med.*, 10:1753–1773.

Benichou, J. and Gail, M. H. (1990). Variance calculations and confidence intervals for estimates of the attributable risk based on logistic models. *Biometrics*, 46(4):991–1003.

Bennett, J. E., Racine-Poon, A., and Wakefield, J. C. (1996). *MCMC for nonlinear hierarchical models*. Chapman and Hall.

Besag, J., Green, P., Higdon, D., and Mengersen, K. (1995). Bayesian computation and stochastic systems (with discussion). *Stat. Sci.*, 10(1):3–41.

Besag, J. E. (1994). Comment on "Representations of knowledge in complex systems". *J. Royal Stat. Soc Series B*, 56(4):549–603.

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1975). *Discrete multivariate analysis*. The MIT Press.

Brooks, S., Gelman, A., Jones, G. L., and Meng, X. L., editors (2011). *Handbook of Markov Chain Monte Carlo*. CRC Press.

Brooks, S. P. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *J. Comput. Graph. Stat.*, 7(4):434 – 455.

Bruzzi, P., Green, S. B., Byar, D. P., Brinton, L. A., and Schairer, C. (1985). Estimating the population attributable risk for multiple risk factors using case-control data. *Am J Epidemiol*, 122(5):904–914.

Caraguel, C. G. B., Stryhn, H., Gagne, N., Dohoo, I. R., and Hammell, L. (2011). Selection of a cutoff value for real-time polymerase chain reaction results to fit a diagnostic purpose: analytical and epidemiologic approaches. *J Vet Diagn Invest*, 23:2–15.

Christensen, R., Johnson, W., Branscum, A., and Hanson, T. E. (2010). *Bayesian Ideas and Data Analysis*. CRC Press.

Cole, P. and MacMahon, B. (1971). Attributable risk percent in case-control studies. *Brit. J. prev. soc. Med.*, 25(4):242– 244.

Denman, D. W. and Schlesselman, J. J. (1983). Interval estimation of the attributable risk for multiple exposure levels in case-control studies. *Biometrics*, 39:185–192.

Diciccio, T. and Efron, B. (1991). More accurate confidence intervals in exponential families. Technical Report 368, Standord University.

Dreyfus, A., Heuer, C., Wilson, P., and Collins-Emerson, J. (2014). Risk of infection and associated influenza-like disease among abattoir workers due to two leptospira species. *Epidemiol. Infect*, 143(10):2095–2105.

Dreyfus, A., Verdugo, C., Benschop, J., Collins-Emerson, J., Wilson, P., and Heuer, C. (2011). Leptospirosis sero-prevalence and associated economic loss in New Zealand livestock. In *Proceedings of the Food Safety, Animal Welfare & Biosecurity, Epidemiology & Animal Health*

Management, and Industry branches of the NZVA, pages 3.12.1 – 3.12.10. Epidemiology & Animal Health Management Branch of the New Zealand Veterinary Association.

Duane, S., Kennedy, A. D., Pendleton, B. J., and Roweth, D. (1987). Hybrid Monte Carlo. *Phys. Lett. B*, 195:216–222.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann. Stat.*, 7(1):1–26.

Efron, B. (1980). Non-parametric estimates of standard error: The jackknife, the bootstrap and other resampling plans. Technical Report 163, Stanford University.

Efron, B. (1981). Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods. *Biometrika*, 68(3):589–599.

Efron, B. (1987). Better bootstrap confidence intervals. *JASA*, 82(397):171–185.

Efron, B. (1990). Six questions raised by the bootstrap. Technical Report 139, Stanford University.

Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Chapman & Hall.

Fahrmeir, L., Kneib, T., Lang, S., and Marx, B. (2013). *Regression models, methods and applications*. Springer.

Fleiss, J. L. (1973). *Statistical Methods for Rates and Proportions*. John Wiley and Sons.

Fleiss, J. L. (1979). Inference about population attributable risk from cross-sectional studies. *Am J Epidemiol*, 110(2):103–104.

Gamerman, D. and Lopes, H. F. (2006). *Markov Chain Monte Carlo*. CRC Press, 2nd edition.

Gart, J. J. (1962). Approximate confidence limits for the relative risk. *J. Royal Stat. Soc. Series B*, 24:454–463.

Gart, J. J. and Thomas, D. G. (1972). Numerical results on confidence limits for the odds ratio. *J. Royal Stat. Soc. Series B*, 34:441–447.

Gefeller, O. (1990). Theory and application of attributable risk estimation in cross-sectional studies. *Statistica Applicata*, 2(4):323–331.

Gelman, A. (2005). Analysis of variance why it is more important than ever. *Ann. Stat.*, 33(1):1–53.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis*. CRC Press, 3rd edition.

Gelman, A., Jakulin, A., Pittau, M. G., and Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *Ann. Appl. Stat.*, 2(4):1360–1383.

Gelman, A., Roberts, G. O., and Gilks, W. (1996). Efficient metropolis jumping rules. *Bayesian Statistics 5*, pages 599–607.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Stat. Sci.*, 7(4):457–511.

Geman, S. and Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Trans. Pattern. Anal. Mach. Intell.*, 6(6):721–741.

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61(2):215–231.

Graham, P. (2000). Bayesian inference for a generalized population attributable fraction: the impact of early vitamin a levels on chronic lung disease in very low birthweight infants. *Stat. Med.*, 19:937–956.

Greenland, S. (1987). Variance estimators for attributable fraction estimates consistent in both large strata and sparse data. *Stat. Med.*, 6:701–708.

Greenland, S. and Drescher, K. (1993). Maximum likelihood estimation of the attributable fraction from logistic models. *Biometrics*, 49:865–872.

Greenland, S. and Robins, J. M. (1988). Conceptual problems in the definition and interpretation of attributable fractions. *Am J Epidemiol*, 128(6):1185–1197.

Gross, S. T. (1980). Median estimation in sample surveys. In *Proceedings of the survey and research methods section of the American Statisistical Association*, pages 181–184. Am. Statist. Ass.

Gustafson, P. (2005). On model expansion, model contraction, identifiability and prior information: two illustrative scenarios involving mismeasured variables. *Stat. Sci.*, 20(2):111–140.

Gustafson, P. (2010). Bayesian inference for partially identified models. *Int. J. Biostat.*, 6(2):1–18.

Gustafson, P. (2015). *Bayesian Inference for Partially Identified Models: Exploring the limits of limited data*. Taylor & Francis Group, LLC.

Hanley, J. A. (2001). A heuristic approach to the formulas for population attributable fraction. *J Epidemiol Community Health*, 55:508–514.

Hastings, W. (1970). Monte Carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109.

Heuer, C., Schukken, Y., Jonker, L., Wilkinson, J., and Noordhuizen, J. (2001). Effect of monensin on blood ketone bodies, incidence and recurrence of disease and fertility in dairy cows. *J. Dairy Sci*, 84(5):1085–1097.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Hosmer, D. W., Lemeshow, S., and Klar, J. (1988). Goodness-of-fit testing for the logistic regression model when the estimated probabilities are small. *Biom. J.*, 30(8):911–924.

Hosmer, D. W., Lemeshow, S., and Sturdivant, R. X. (2013). *Applied Logistic Regression*. Wiley, 3rd edition.

Hui, S. L. . and Walter, S. D. (1980). Estimating the error rates of diagnostic tests. *Biometrics*, 36:167–171.

Jeffreys, H. (1945). An invariant form for the prior probability in estimation problems. *Proc. R. Soc. Lond. A*, 186(1007):453–461.

Johnson, W. O., Gastwirth, J. L., and Pearson, L. M. (2001). Screening without a gold standard: The Hui-Walter paradigm revisited. *Am J Epidemiol*, 153(9):921–924.

Johnson, W. O. and Hanson, T. E. (2005). Comment on "On model expansion, model contraction, identifiability and prior information: Two illustrative scenarios involving mismeasured variables". *Stat. Sci.*, 20(2):111–140.

Jones, G. and Johnson, W. O. (2014). Prior elicitation: Interactive spreadsheet graphics with sliders can be fun and informative. *Am. Stat.*, 68(1):42–51.

Jones, G., Johnson, W. O., Hanson, T. E., and Christensen, R. (2010). Identifiability of models for multiple diagnostic testing in the absence of a gold standard. *Biometrics*, 66:855–863.

Joseph, L., Gyorkos, T. W., and Coupa, L. (1995). Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. *Am J Epidemiol*, 41(3):263–272.

Kahn, H. (1955). Use of different Monte Carlo sampling techniques. Technical report, The RAND Corporation.

Kleinbaum, D. G., Kupper, L. L., and Morgenstern, H. (1982). *Epidemiologic Research*. Van Nostrand Reinhold.

Kooperberg, C. and Petitti, D. B. (1991). Using logistic regression to estimate the adjusted attributable risk of low birthweight in an unmatched case-control study. *Epidemiology*, 2(5):363–366.

Lehnert-Batar, A., Pfahlberg, A., and Gefeller, O. (2006). Comparison of confidence intervals for adjusted attributable risk estimates under multinomial sampling. *Biom. J.*, 48(5):805–819.

Leung, H. M. and Kupper, L. L. (1981). Comparisons of confidence intervals for attributable risk. *Biometrics*, 37:293–302.

Levin, M. L. (1953). The occurrence of lung cancer in man. *Acta Unio Int Contra Cancrum*, 9(3):531–41.

Levin, M. L. and Bertell, S. R. (1978). Simple estimation of population attributable risk from case-control studies. *Am J Epidemiol*, 108(1):78–79.

Leviton, A. (1973). Letter: Definitions of attributable risk. *Am J Epidemiol*, 98(3):231.

Lilienfeld, A. M. (1973). Epidemiology of infectious and non-infectious disease: some comparisons. *Am J Epidemiol*, 97(3):135–147.

Lui, K. J. (2001). Notes on interval estimation of the attributable risk in cross-sectional sampling. *Stat. Med.*, 20:1797–1809.

MacMahon, B. and Pugh, T. F. (1970). *Epidemiology Principles and Methods*. Little, Brown and Company, 1st edition.

MacMahon, B., Pugh, T. F., and Ipsen, J. (1960). *Epidemiologic methods*. Little, Brown and Company.

MacMahon, B. and Trichopoulos, D. (1996). *Epidemiology Principles and Methods*. Little, Brown and Company, 2nd edition.

Markush, R. E. (1977). Levin's attributable risk statistic for analytic studies and vital statistics. *Am J Epidemiol*, 105(5):401–406.

Markush, R. E. and Seigel, D. G. (1968). Prevalence at death. I. A new method for driving death rates for specific disease. *Am J Public Health*, 58:544–557.

Martin, A. D., Quinn, K. M., and Park, J. H. (2011). MCMCpack: Markov chain monte carlo in R. *J. Stat. Softw.*, 42(9):22.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*. Chapman & Hall, 2nd edition.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equations of state calculations by fast computing machines. *J. Chem. Phys.*, 21(6):1089–1092.

Miettinen, O. S. (1974). Proportion of disease caused or prevented by a given exposure, trait or intervention. *Am J Epidemiol*, 99(5):325–322.

Miller, R. G. (1974). The jackknife: A review. *Biometrika*, 61(1):1–15.

Morgenstern, H. and Bursic, E. S. (1982). A method for using epidemiologic data to estimate the potential impact of an intervention on the health status of a target population. *J. Community Health*, 7(4):292–309.

Muller, P. (1994). Metropolis based posterior integration schemes. Technical report, Numerical Recipes in Fortran.

Neal, R. M. (1996). *Bayesian learning for neural networks*. Springer, New York.

Neal, R. M. (2005). The short-cut metropolis method. Technical Report 0506, Department of Statistics, University of Toronto.

Neal, R. M. (2011). *MCMC using Hamiltonian Dynamics*, chapter 5. CRC Press.

Newson, R. B. (2013). Attributable and unattributable risks and fractions and other scenario comparisons. *The Stata Journal*, 3(4):672–698.

Ouellet, B. L., Romeder, J.-M., and Lance, J.-M. (1979). Premature mortality attributable to smoking and hazardous drinking in Canada. *Am J Epidemiol*, 109(4):451–463.

Petrovski, K. R. (2007). Bovine mastitis in New Zealand. Master's thesis, Massey University, Palmerston North, New Zealand.

Plummer, M., Best, N., Cowles, K., and Vines, K. (2006). Coda: Convergence diagnosis and output analysis for MCMC. *R News*, 7(1):7–11.

Politis, D. N. (1998). Computer-intensive methods in statistical analysis. *IEEE Signal. Process. Mag.*, 15(1):39–55.

Politis, D. N., Romano, J. P., and Wolf, M. (1999). *Subsampling*. Springer.

Poole, C. (2015). A history of the population attributable fraction and related measures. *Ann Epidemiol*, 25:147–154.

Quenouille, M. H. (1949). Approximate tests of correlation in time-series. *J R Stat Soc Series B*, 11(1):68–84.

Quenouille, M. H. (1956). Notes on bias in estimation. *Biometrika*, 43(3/4):353–360.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing.

Robert, C. P. and Casella, G. (2010). *Introducing Monte Carlo Methods with R*. Springer.

Roberts, G. and Rosenthal, J. (1998). Optimal scaling of discrete appoximations to Langevin diffusions. *J. Royal Stat. Soc Series B*, 60(1):255–268.

Roberts, G. O., Gelman, A., and Gilks, W. R. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *Ann. Appl. Probab.*, 7(1):110–120.

Rockhill, B., Newman, B., and Weinberg, C. (1998). Use and misuse of population attributable fractions. *Am. J. Public Health*, 88(1):15–19.

Rosenthal, J. S. (2006). *A First Look At Rigorous Probability Theory*. second edition.

Rothenberg, T. J. (1971). Identification in parametric models. *Econometrica*, 39:577–591.

Rothman, K. J. (2002). *Epidemiology an introduction*. Oxford University Press.

Rubin, D. B. (1981). The Bayesian bootstrap. *Ann. Stat.*, 9(1):130–134.

Seigel, D. G. and Markush, R. E. (1968). Prevalence at death. II. Methodological consideration for use in mortality studies. *Am. J Public Health*, 58:772–776.

Silverman, B. W. and Young, G. A. (1987). The bootstrap: to smooth or not to smooth? *Biometrika*, 74(3):469–479.

Smith, D. J. and Teo, K. L. (1989). *Linear Algebra*. New Zealand Mathematical Society.

Stan Development Team (2017). *RStan: the R interface to Stan*. R package version 2.17.3.

Stevenson, M., Nunes, T., Heuer, C., Marshall, J., Sanchez, J., Thornton, R., Reiczigel, J., Robison-Cox, J., Sebastiani, P., Solymos, P., Yoshida, K., Jones, G., Pirikahu, S., Firestone, S., Kyle, R., Popp, J., and Jay, M. (2017). *epiR: Tools for the Analysis of Epidemiological Data*. R package version 0.9-93.

Taylor, J. W. (1977). Simple estimation of population attributable risk from case-control studies. *Am J Epidemiol*, 106(4):206.

Thornley, C., Baker, M., Weinstein, P., and Maas, E. (2002). Changing epidemiology of human leptospirosis in New Zealand. *Epidemiol. Infect*, 128:29–36.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *Ann. Stat.*, 22(4):1701–1762.

Tukey, J. W. (1958). Bias and confidence in not-quite large sample. *Ann. Math. Stat.*, 29:614.

Uter, W. and Pfahlberg, A. (2001). The concept of attributable risk in epidemiological practice. *Stat. Methods Med. Res*, 10:231–237.

Wagner, B. A., Wells, S. J., and Kott, P. S. (2001). Variance estimation for population attributable risk in a complex cross-sectional animal health survey. *Prev. Vet. Med*, 48:1–13.

Walter, S. D. (1975). The distribution of Levin's measure of attributable risk. *Biometrika*, 62(2):371–374.

Walter, S. D. (1976). The estimation and interpretation of attributable risk in health research. *Biometrics*, 32:829–849.

Whittemore, A. S. (1982). Statistical methods for estimating attributable risk from retrospective data. *Stat. Med.*, 1:229–243.

Whittemore, A. S. (1983). Estimating attributable risk from case-control studies. *Am J Epidemiol*, 117(1):76–85.

Wilks, S. S. (1962). *Mathematical Statistics*. Wiley, New York.

Young, G. A. (1988). A note on bootstrapping the correlation coefficient. *Biometrika*, 75(2):370–373.

MASSEY UNIVERSITY
GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION
DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the candidate and the candidate's Primary Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

Name of candidate:	Sarah Pirikahu
Name/title of Primary Supervisor:	Professor Geoff Jones
Name of Research Output and full reference:	
Pirikahu, S., Jones, G., Hazelton, M. L., and Heuer, C. (2016). Bayesian methods of confidence interval construction for the population attributable risk from cross- sectional studies. Stat. Med., 35(18):3117-3130.	
In which Chapter is the Manuscript /Published work:	2
Please indicate:	
• The percentage of the manuscript/Published Work that was contributed by the candidate:	80
and	
• Describe the contribution that the candidate has made to the Manuscript/Published Work:	
All authors made contributions to the editing of the manuscript and interpretation of the results. Sarah Pirikahu was responsible for coding the jointly discussed methods, performing the simulations and writing the paper.	
For manuscripts intended for publication please indicate target journal:	
Candidate's Signature:	Sarah Digitally signed by Sarah Date: 2019.02.07 11:04:34 +13'00'
Date:	7/02/2019
Primary Supervisor's Signature:	Geoff Jones Digitally signed by Geoff Jones Date: 2019.02.07 13:09:29 +13'00'
Date:	07-02-2019

(This form should appear at the end of each thesis chapter/section/appendix submitted as a manuscript/ publication or collected as an appendix at the end of the thesis)