THE ANALYSIS FOR CASE-CONTROL STUDY IN INFLUENZA

by

Hongyue Wang

Submitted in partial fulfillment of the requirements
for the degree of Master of Science

at

Dalhousie University
Halifax, Nova Scotia
April 2016

To my baby:

Shangyi Ouyang

Table of Contents

List of Tables

List of Figures

Abstract

This thesis is a simulation study for Serious Outcomes Surveillance (SOS) in influenza. We introduce some statistical methods appropriate for matched and unmatched data and describe an algorithm to simulate from a point process, and reports the results of some simulation studies which examining the performance of matched and unmatched analysis methods to assess the presence of a vaccine effect. We report further simulations which address the ability of matched and unmatched methods to accommodate additional predictor variables, and also investigates the effect of missing data in the form of missing matched controls.

Acknowledgements

I would like to thank my supervisor Dr.Bruce Smith, for introducing me to Statistics and supporting my Master study and career development. His generous guidance helped me in all the time of research and writing of this thesis. He is the best mentor and advisor not only for my Master study but also for my life.

Chapter 1

Introduction

Every year, 10-20% of the Canadian population becomes infected with influenza, and many will end up requiring medical attention as well as treatment for complications associated with the viral infection. On top of the economic costs associated with this illness there is also an added burden on and cost to the healthcare system due to the increased number of physician visits, hospitalizations, and emergency room trips. Fortunately, there is a publicly funded influenza vaccine that is offered annually and considered to be the best form of protection from the circulating strains of the virus.

While vaccination has been shown to reduce the prevalence of influenza, absences from work and school, and physician visits, the overall effectiveness of the influenza vaccine is quite variable. Older populations for instance have been shown to be at greater risk for vaccine failure due to immunosenescence. In Canada, sentinel surveillance for influenza amongst hospitalized Canadian adults results in estimates of vaccine effectiveness (VE) in the prevention of hospitalization in older adults of one-third [4] to one-half [5] of influenza-related hospitalizations in the elderly. This clearly represents an important contribution to the health of Canadians and justifies the investment in publicly funded influenza immunization programs.

1.1 Methods of Serious Outcomes Surveillance (SOS) Network

In 2009, the Public Health Agency of Canada/CIHR Influenza Research Network (PCIRN) Serious Outcomes Surveillance Network was established to conduct active surveillance for influenza amongst Canadian adults admitted to participating hospitals. The primary goals of this collaborative research network are to better define the burden of disease caused by influenza, and to monitor the effectiveness of the seasonal vaccine against serious outcomes such as hospitalization and death. While the Network began with nine hospitals in five provinces in 2009, today it has grown to consist of 40 hospitals in seven provinces, which encompasses about 17,000 hospital beds.

Once specific goal of the network was to address the missing gap in previous studies on vaccine effectiveness by investigating whether influenza vaccination attenuates disease severity or improves outcomes in those who are admitted to hospital despite vaccination, as well as to identify the host factors that may be predictive of vaccine failure and severe illness. Impact of prior season vaccination on estimates of VE was also explored. Such knowledge is crucial, as it may help to inform health policies and guidelines surrounding the use of new influenza vaccines, and more importantly it has the potential to change our understanding of the utility of the vaccine and improve vaccine uptake.

Nurses conducted active surveillance for flu in admitted adults from October to May (flu reason). All patients admitted with respiratory infection had a Nasopharyngeal(NP) swab for flu polymerase chain reaction (PCR). Test negative results were controls for calculation of Vaccine effectiveness(VE). Measuring VE was a challenge due to the timing of the immunization campaign and the bulk of illness occurrence.

1.2 Study Design and Data Collection

This is a prospective, test-negative case-control study that was conducted at the participating sentinel hospitals located throughout Canada. For the purposes of assessment of vaccine effectiveness, cases were defined as adult patients admitted to an SOS hospital as a result of influenza or complication of the virus, and who had tested positive for influenza. For each case enrolled, one or more controls were selected from amongst patients admitted with an acute respiratory illness to the same SOS hospital of the case, within 2 weeks of the admission date of the case, but who had a negative test for influenza on admission. The test-negative case-control design is not a case-control design in the usual sense, in that controls may be collected prospectively, or retrospectively over a small time window.

The SOS network collects data during the flu season, which normally begins in September and ends in April or May depending on the level of virus recycling. The cases are defined as adult patients with a positive test for influenza whose admission is attributable to influenza or a complication of influenza. The potential control is defined as a consenting adult patient at same site, with the same age stratum as a case, which is either greater than 65 years old or less than 65 year old; and admission date within 14 days of case; NP swab obtained within seven days of onset of symptoms, and test negative for influenza and diagnosis compatible with influenza. The study was planned such that each case would be matched with two controls, but due to difficulties in identifying sufficient numbers of controls, the planned matching was reduced to single controls in year two, with a number of cases being unmatched.

1.3 Object of interest

The object of interest for SOS study is vaccine efficacy. The primary predictor variable of interest is vaccination status (Y/N), with the goal being to assess whether the vaccination rates differ in cases and controls. For example, if the vaccination rate is higher in controls than in cases, that would suggest that vaccination is effective in reducing hospitalization with influenze like illness.

By definition, [11] [12], vaccine efficacy, also known as vaccine effectiveness, VE, is estimated as one minus some measure of relative risk, RR .

$$VE = 1 - RR \tag{1.1}$$

The SOS study focused on evaluating protective effects of influenza vaccination, the relative risk measure being the ratio of odds of vaccination in cases vs controls. [7]

$$VE = 1 - OR \tag{1.2}$$

In an unmatched study, a reasonable assumption is that the times at which vaccinated and unvaccinated individuals are admitted to hospital follow independent Poisson processs, with hazard rates $\lambda_V(t)$ and $\lambda_U(t)$ respectively, in which case the time varying vaccine efficacy is given by one minus the hazard rate ratio [11]

$$VE = 1 - \frac{\lambda_V(t)}{\lambda_U(t)} \tag{1.3}$$

In the matched design of the SOS study, where matched controls are identified within a specified time window of cases, the control times are not independent of case times. In the limit, where the interval of the matching window tends to 0, there is only one arrival process, that for cases, with no additional randomness in the control times. In this matched case, we will see that methods of survival analysis can be used to estimate vaccine efficacy.

1.4 Why match?

Why is matching useful for this study?

The quantity of circulating influenza virus in the population, and therefore the probability of influenza infection, depends on the time of year and location. By matching according to time of year and location, the unmeasured level of circulating virus can be controlled for.

An individual presenting at hospital who tests positive for influenza must have been exposed to the virus. Let $P(V)$ denote the probability that an individual was vaccinated in the current flu season, and $P(U) = 1 - P(V)$ be the probability of an unvaccinated individual. Let $P(E)$ be the probability that an individual was exposed to the influenza virus, and let $P(F|V,E)$ and $P(F|U,E)$ be the probabilities that vaccinated and unvaccinated individuals who were exposed to the virus present to hospital with the flu virus. Let $P(O)$ be the probability that an individual presents to the hospital with an other respiratory infection. It is assumed that the probability of other respiratory infections is unaffected by the influenza vaccination status, or contact with flu infected individuals.

Consider the population of individuals presenting to hospital with influenza like illness. Individuals will be of one of the following types: (Other, Vaccinated), (Other, Unvaccinated), (Flu, Vaccinated, Exposed), or (Flu, Unvaccinated, Exposed). Assuming that influenza vaccination does not change the likelihood of another respiratory infection:

- $P($ Other,Vaccinated$) = P(O)P(V)$
- $P($ Other,Unvaccinated$)= P(O)P(U)$
- $P($ Flu,Vaccinated,Exposed$)= P(Flu|V,E)P(V)P(E)$, and

- $P($ Flu,Unvaccinated,Exposed$)= P(Flu|U,E)P(U)P(E)$
- The true vaccine efficacy is defined as one minus the relative odds,

$$VE = 1 - \frac{P(Flu|V,E)}{P(Flu|U,E)}$$

Exposue is a latent variable which is not measured, but it is known that $P(E)$ changes throughout the influenza season. Consider the following table of vaccination status by case/control status, for those individuals presenting to hospital with influenza like illness. The probabilities in the table allow that exposure might differ for vaccinated and unvaccinated individuals, with probabilities $P(E|V)$ and $P(E|U)$ respectively.

	Case	Control
Vaccinated	$P(\mathrm{F}lu|V,E)P(V)P(E|V)$	$P(O)P(V)$
Unvaccinated	$P(\mathrm{F}lu|U,E)P(U)P(E|U)$	$P(O)P(U)$

The odds of vaccination for a case is $\frac{P(Flu|V,E)P(V)P(E|V)}{P(Flu|U,E)P(U)P(E|U)}$, while the odds of vaccinaton for a control is $\frac{P(V)}{P(U)}$, and the resulting odds ratio is $\frac{P(Flu|V,E)P(E|V)}{P(Flu|U,E)P(E|U)}$. Unless $P(E|V) = P(E|U)$, the odds ratio being calculated is not the odds ratio of interest $\frac{P(Flu|V,E)}{P(Flu|U,E)}$. One way to assure that $P(E|V) = P(E|U)$ is to match by time and location, and any other factors which might be related to the quantity of circulating influenza virus. Based on this argument, the SOS study was planned as a matched design.

The experience in carrying out the SOS study has been that it is often difficult to obtain matched controls. As noted above, in year one the goal was to use 2:1 matching. In year two this was reduced to a goal of 1:1 matching, and for many cases, the implementation team was unable to find any matched controls.

To carry out a matched analysis, only matched cases are included, and the loss of unmatched cases may result in reduced power. This leads to questions concerning the method of analysis. Although the design is matched, is it always best to use a matched analysis, or can an unmatched analysis provide greater power than a matched analysis, given that unmatched analysis methods can use all data, matched or otherwise. The primary goal of this thesis is to address this question.

1.5 Role of other variables

The elderly are known to be at greater risk for vaccine failure due to immunosensence, and a secondary objective for the SOS study is to assess the role of patient age as a determinant of vaccine efficacy. The SOS study recorded gender and a number of other variables, and while age was controlled for through matching, other variables were controlled for by inclusion in regression models.

In this thesis, we carry out an exploratory simulation based comparison of matched and unmatched analyses using only three variables - vaccination status, case/control status, and gender, with the latter variable included as a regressor in both matched and unmatched logistic regression models, and in addition, including vaccination status as a regressor in an unmatched logistic regression model with case/control status as the outcome.

The remainder of the thesis is organized as follows. In chapter 2, we introduce some statistical methods appropriate for matched and unmatched data. Chapter 3 describes an algorithm to simulate from a point process, and reports the results of some simulation studies which examining the peformance of matched and unmatched analysis methods to assess the presence of a vaccine effect. Chapter 4 reports further simulations which address the ability of matched and unmatched methods to accommodate additional predictor variables, and also investigates the effect of missing data in the form of missing matched controls. Chapter 5 summarizes the advantages and limitations of the methods used, and suggests areas for further research.

Chapter 2

Likelihood and Theoretical Inference

A full understanding of how the data from a case-control study permit estimation of the relative risk requires careful description of how cases and controls are sampled from the population.

2.1 Data and Model

For an unmatched design, the underlying data of interest consist of the times of admission to hostpital with influenza like illness. Let a, b, c, and d denote the numbers of admissions with influenza like illness in time (0,T], for vaccinated cases, unvaccinated cases, vaccinated controls, and unvaccinated controls, respectively.

Let $\{t_{j,i}, i = 1, ..., N_j, j \in (1, 2, 3, 4)\}$ denote the times of admission, where N_j is the number of admissions of the j'th type in (0,T]. It is assumed that the four admission processes are independent non-homogeneous Possion processes, with the j'th process having rate $\lambda_j(t)$.

2.2 The relative risks for matched data

We shall start by considering two dichotomous variables, one of which we shall regard as the vaccination status, the other a case-control variable. Suppose we had obtained, when cross-tabulating disease status against vaccination status, the following results based on pooling the data over levels of any covariates.

Table 2.1: Pooled data

	Vaccinated	Not vaccinated	
case	a	b	n_1
control	c	d	n_0
	m_1	m_0	N

The risk ratio associated with exposure to vaccination status can be approximated by the odds ratio in the above table.

$$OR = ad/bc \tag{2.1}$$

2.3 Matched analysis using McNemar's test

Where ψ denote the population odds ratio, McNemar's test is specifically designed to the hypothesis $H_0 : \psi = 1$. Under this hypothesis of no association, the probabilities of the two discordant types are equal. With two sided alternative, the null hypothesis $H_0 : \psi = 1$ may be tested by calculating the exact tail probabilities of the biomial distribution with probability equal to 1/2. Alternatively, a a continuity corrected version of the chi-square statistic is based on the standardized value of b. In carrying out McNemar's test, the data are arranged as in table 2.

Table 2.2: Data for McNemar's test

	Cases	
Controls	vaccinated	unvaccinated
vaccinated	A	B
unvaccinated	C	D

Here B is the number of matched pairs where the Control is vaccinated and the Case is unvaccinated, and so on. Known as McNemar's test [1] for the equality proportions in matched samples, the test statistic is often expressed as

$$\chi^2 = \frac{(B - C)^2}{(B + C)} \tag{2.2}$$

which has a χ^2_1 distribution under the null hypothesis.

2.4 Poisson Likelihood for Unmatched Data

In this section we develop the likelihood for a Poisson process where the intensity function is of the Cox proportional hazards form. In the context of section 2.1, we are developing the likelihood for one of the four processes. The presentation is general in that each individual may be under observation over a different period of time, and may experience 0 or more events. The development follows Lawless [9].

Suppose that there are independent observations on m individuals. Individual i is observed over the time interval (S_i,T_i), during which she experiences N_i events at times $t_{i1} < ... < t_{iN_i}$.

Without loss of generality, assume that all of the S_i's equal to 0, and let $\Lambda(T) = \int_0^T \lambda(t)dt$ denote the integrated intensity function. For example,

If individual i has covariate vector x_i, the proportional hazards model assumes

$$\lambda_{x_i}(t;\theta,\beta) = \lambda_0(t;\theta)g(x_i;\beta) \tag{2.3}$$

where $\lambda_0(t;\theta)$ is a baseline intensity function depending on parameter θ, and $g(x_i;\beta)$ is a postitive-valued function of x_i and a vector of parameters β. For the model with

$$g(x;\beta) = exp(x'\beta) \tag{2.4}$$

and $\lambda_0 = \lambda_0(t;\theta)$, the likelihood function [9] is

$$L(\theta,\beta) = \prod_{i=1}^{m} \left\{ \prod_{j=1}^{n_i} \lambda_{x_i}(t_{ij};\theta,\beta) \right\} exp\left\{-\Lambda_{x_i}(T_i;\theta,\beta)\right\} \tag{2.5}$$

which can be decomposed as

$$L(\theta,\beta) = \left\{ \prod_{i=1}^{m}\prod_{j=1}^{n_i} \frac{\lambda_0(t_{ij};\theta)}{\Lambda_0(T_i;\theta)} \right\} \times \prod_{i=1}^{m} exp[-\Lambda_0(T_i;\theta)e^{x_i'\beta}][\Lambda_0(T_i;\theta)e^{x_i'\beta}]^{n_i} \tag{2.6}$$

$$= L_1(\theta)L_2(\theta;\beta) \tag{2.7}$$

For the SOS study, there is one event per subject so that $n_i = 1$ for $i = 1, 2, ...m$, where m is the number of subjects. Therefore the likelihood for this study is

$$L = \prod_{i=1}^{m} \lambda_{x_i}(t_i)exp\left\{-\Lambda_{x_i}(T_i)\right\} \tag{2.8}$$

where

$$\Lambda_{x_i}(T_i) = \int_0^{T_i} \lambda_{x_i}(t)dt \tag{2.9}$$

In the simplest case of independent homogeneous Poisson processes with rate functions $\lambda_1(t) = \lambda_1$ for cases, and $\lambda_0(t) = \lambda_0$ for controls, with N_1 cases and N_0 controls, the likelihood is

$$L(\lambda_j) = \prod_{i=1}^{N_j} \lambda_j e^{-\int_0^T \lambda_j dt}$$

giving log likelihood

$$l = N_j log\lambda_j - T\lambda_j$$

Differenting,

$$\frac{\partial l}{\partial \lambda_j} = \frac{N_j}{\lambda_j} - T = 0$$

which implies that the MLE is $\lambda_j = \frac{N_j}{T}$.

Alternatively, the intensity function can be written as $log(\lambda) = \beta_0 + \beta_1 X$, where X is an indicator variable being 1 for cases and 0 for controls. The MLE's for β_0 and β_1, and the information matrix, can be calculated in the usual fashion, with the usual tests and confidence intervals for β_0 and β_1 being based on the asymptotic normality of the MLE. For the point process model with this parameterization, vaccine efficacy is defined as $VE = 1 - exp(\beta_1)$, for which confidence intervals and tests follow from the asymptotic distribution of $\hat{\beta}_1$. Covariates are accommodated in a straightforward manner by inclusion in the intensity function. For example, $log(\lambda) = \beta_0 + \beta_1 X + \beta_2 G$, where $G = 1$ for males, and $G = 0$ for females.

2.5 Conditional Likelihood for a Matched Design

The developement of the conditional likelihood follows Cox [2]. The original motivation was the comparison of survival functions, and contained the proposal for the Cox proportional hazards model. Where $x = (x_1, ..., x_p)$ is a vector of covariates and $\beta = (\beta_1, ..., \beta_p)$ a vector of parameters, Cox proposed that the hazard function for a subject with covariate vector β be

$$\lambda(t, x) = \lambda_0(t) exp(x'\beta) \tag{2.10}$$

where $\lambda_0(t)$ is an unspecified baseline hazard function. The multiplicative factor $exp(x'\beta)$, gives the risk of failure with covariate x relative to that at $x = \mathbf{0}$.

In the original context, the time points of interest were general event or failure times, with the possibility of several events per subject. In this thesis, it is assumed that each subject experiences exactly one event, an admission to hospital with influenza like illness.

Cox argued that if $\lambda_0(t)$ is arbitrary, then no information can be contributed about β by time intervals in which no events occurred because the component $\lambda_0(t)$ might conceivably be identically zero in such intervals. He therefore argued conditionally on the set $\{t_i\}$ of instants at which failures occur.

Let $R(t_i)$ denote the risk set at observed failure time t_i, which consists of those individuals at risk of failure just prior to t_i. In the absence of tied failure times, then t_i, the probability that the failure is on the individual as observed, conditional on the risk set, is

$$exp(x_i'\beta) / \sum_{j \in R(t_i)} exp(x_j'\beta)$$

.

The conditional, or partial, likelihood for β is the product of such terms over observed failure times.

Downton [3] showed how Cox's conditional likelihood analysis can be applied to matched pairs. In particular, for the i'th matched pair, the probability that the failure occurred to the actual individual observed is $exp(x\beta)/(1 + exp(\beta))$, where $x = 1$ for a case, and $x = 0$ for a control.

Assume that out of n discordant matched pairs, there are r for which the case was vaccinated and the control unvaccinated, and $n - r$ for which the control was vaccinated and the case was unvaccinated. The resulting conditional, or partial, log likelihood is

$$l(\beta) = r\beta - nlog(1 + exp(\beta))$$

Differentiating the above log likelihood with respect to β

$$\frac{\partial l}{\partial \beta} = r - \frac{ne^{\beta}}{(1 + e^{\beta})} = 0 \tag{2.11}$$

$$\implies r = \frac{ne^{\beta}}{(1 + e^{\beta})} \tag{2.12}$$

$$\implies r + re^{\beta} = ne^{\beta} \tag{2.13}$$

$$\implies \frac{r}{n - r} = e^{\beta} \tag{2.14}$$

$$\tag{2.15}$$

$$\implies \widehat{\beta} = \log \frac{r}{n - r} \tag{2.16}$$

Differentiating a second time,

$$\frac{\partial^2 l(\beta)}{\partial \beta^2} = -\frac{ne^{\beta}}{(1 + e^{\beta})^2} \tag{2.17}$$

The asymptotic variance is

$$VAR(\widehat{\beta}) = -E[\frac{\partial^2 l(\beta)}{\partial \beta^2}] = \frac{(1 + e^{\beta})^2}{ne^{\beta}}$$

for which the plugin estimate equals $\frac{n}{r(n-r)}$.

The regularity conditions underlying the asympotic normality of the maximum likelihood estimator are satisfied in this case, so that

$$\widehat{\beta} \sim AN\left(\beta, \frac{(1 + e^{\beta})^2}{ne^{\beta}}\right) \tag{2.18}$$

where the notation "AN" denotes asymptotic normality.

Thus to test the hypothesis $\beta = 0$ we have the test statistic

$$\chi^2 = z^2 = \frac{\widehat{\beta}^2}{var(\widehat{\beta})} = (\log \frac{r}{n-r})^2 / \frac{n}{r(n-r)} \tag{2.19}$$

Using application of Taylor's theorem, it can be shown that for any $x > 0$, $\log x \approx 2(\frac{x-1}{x+1})$ leading to the test statistic

$$\chi^2 = 4(r - n/2)^2/n \tag{2.20}$$

whose distribution when $\beta = 0$ is, asymptotically, χ^2 with one degree of freedom.

Letting the number of discordant pairs be $n = b + c$ and taking $r = b$ as in table (2.1), then $4(r - n/2)^2/n = (b - c)^2/(b + c)$, and we see that the test based on the conditional likelihood is the same as the McNemar's test.

2.6 Power and Bias

Three inference procedures have been described - the unconditional and conditional methods to compare the rates of underlying point processes assuming a Cox proportional hazards model, and McNemar's test. We are interested in estimating the bias and mean squared error of the unconditional and conditional estimators, and in comparing the power of each of the three methods when testing the hypothesis that VE=0, or equivalently, the hypothesis that relative risk equals 1.

2.6.1 Bias

The bias of an estimator is the difference between this estimator's expected value and the true value of the parameter being estimated. That is,

$$Bias[\widehat{\beta}] = E_\beta[\widehat{\beta}] - \beta \tag{2.21}$$

An estimator or decision rule with zero bias is called unbiased. Otherwise the estimator is said to be biased.

2.6.2 Mean squared error

While bias quantifies the average difference to be expected between an estimator and an underlying parameter, an estimator based on a finite sample can additionally be expected to differ from the parameter due to the randomness in the sample.

The MSE is one measure which is used to try to reflect both mean and variability. It is given by

$$MSE(\hat{\beta}) = E\big[(\hat{\beta} - \beta)^2\big] \tag{2.22}$$

and it is well known that $MSE(\hat{\beta}) = V(\hat{\beta}) + Bias[\widehat{\beta}]^2$.

2.6.3 Power

The power of a binary hypothesis test is the probability that the test correctly rejects the null hypothesis H_0 when the alternative hypothesis (H_1) is true. That is Power $= P(rejectH_0 \mid H_1$ is true).

$$\text{If under} \quad H_0 : \hat{\beta} \sim N\left(\beta_0, \sigma_0^2\right), \quad i.e. \frac{\hat{\beta} - \beta_0}{\sigma_0} \sim N(0, 1) \tag{2.23}$$

$$\text{and under} \quad H_1 : \hat{\beta} \sim N\left(\beta_1, \sigma_1^2\right), \quad i.e. \frac{\hat{\beta} - \beta_1}{\sigma_1} \sim N(0, 1) \tag{2.24}$$

then when testing $H_0 : \beta = 0$ vs $H_1 : \beta > 0$ using the test statistic

$$T = \frac{\widehat{\beta} - \beta_0}{\sigma_0} \tag{2.25}$$

one rejects H_0 if $T \geq Z_{1-\alpha} \iff \widehat{\beta} \geq \beta_0 + Z_{1-\alpha}\sigma_0$. This gives power

$$\begin{aligned}
Power(\gamma) &= P_{H_1}(\text{Reject } H_0) \qquad (2.26)\\
&= P_{\beta_1}(T \geq Z_{1-\alpha}) = P_{\beta_1}(\widehat{\beta} \geq \beta_0 + Z_{1-\alpha}\sigma_0) \qquad (2.27)\\
&= P_{\beta_1}\left(\frac{\widehat{\beta}-\beta_1}{\sigma_1} \geq \frac{\beta_0 + Z_{1-\alpha}\sigma_0 - \beta_1}{\sigma_1}\right) \qquad (2.28)\\
&= P\left(Z \leq \frac{(\beta_1-\beta_0) - Z_{1-\alpha}\sigma_0}{\sigma_1}\right) \qquad (2.29)\\
&= P\left(Z \leq Z_\gamma\right) \qquad (2.30)
\end{aligned}$$

In the above, α is referred to as the significance level, the power is γ, and Z_γ and $Z_{1-\alpha}$ are quantiles of the standard normal distribution. These calculations assume that the variances σ_0^2 and σ_1^2 are known, together with β_0 and β_1, with power calculation typically carried out prior to data collection.

Where exact formulas are not available, simulation can be used to estimate the power of a statistical procedure, and also to estimate the bias and the mean squared error of different estimators. The next chapter reports the results of a simulation study to estimate the power of the unconditional and conditional models to test the hypothesis that VE=0, and to estimate the bias and MSE of conditional and unconditional estimators.

Chapter 3

Simulation

A simulation study was conducted to examine the performance of matched and unmatched estimation procedures, with the focus being on the power of test in the binary situation.

The study population being modeled is Canadian adults admitted to hospital for influenza like illness during flu season. It is assumed that patients are admitted to hospital according to Poisson processes with rate $\lambda(t|\beta)$ [6].

As in Chapter 2, we assume that the intensity function of the admission process is:

$$\lambda_x(t) = \lambda_0(t) exp(x'\beta) \tag{3.1}$$

To model a seasonal cycle for the influenza season, we add a periodic term, as

$$log(\lambda_x(t)) \sim \beta_0 + \beta_1 x + \beta_2 cos(\frac{2\pi}{T}t) + \beta_3 sin(\frac{2\pi}{T}t) \tag{3.2}$$

where x is the indicator vaccination status, which is either 1 (vaccinated) or 0 (unvaccinated). t is the time at which either case or control is admitted to hospital, and $T = 1$ year is the period of the influenza cycle.

3.1 Algorithm to simulate from a point process

We now introduce the Lewis and Shedler(1976) algorithm [10] to simulate a point process. A point process is determined by its conditional intensity function $\lambda(t|H_t)$, where H_t is the history of the process on (0,t]. For simulation, we require knowledge of a finite bound M such that $\lambda(t|H_t) \leq M$ for all possible past histories. The process is to be simulated over the finite interval (0,A), given some initial history H_0.

The thinning algorithm is the following [13],

1. Simulate $t_1, ..., t_i$ according to a stationary Poisson process with rate M (by simulating successive interval lengths as i.i.d. exponential r.v.s. of mean 1/M), stopping when $t_i > A$
2. Simulate $y_1, ..., y_i$ as a set of i.i.d uniform (0,1) random variables.
3. Set $K = 1, j = 1$
4. If $t_k > A$, terminate, otherwise evaluate $\lambda(t_k|H_{t_k})$.
5. If $y_k \leq \lambda(t_k|H_{t_k})/M$, set $z_j = t_k$, update the history H to $H \cup \{z_j\}$ and advance j to $j + 1$
6. Advance k to $k + 1$ and return to step 4
7. The output consists of the lists $\{j; z_1, ..., z_j\}$

If the intensity function does not depend on the past history, then the process is a Poisson process, and the Lewis and Shedler algorithm is simple rejection sampling.

3.2 Simulation of data

Unmatched data was generated by simulating times from two independent Poisson processes for cases and controls. Matched data was generated by simulating case admission times from a single Poisson process, and randomly allocating successive observations to be either (case,control) or (control,case) pairs.

All simulations assumed admission times over a period of 3 years, so a sampling interval of 1095 days, over which the number of admissions is a Poisson random variable, whose mean is the integral of the intensity function from 0 to 1095.

3.2.1 Simulation of matched data

The simulation of matched data was based on the SOS study design, with a three year interval of interest, a yearly cycle, and one matched control for each simulated case time. The bound M for the Shedler-Lewis algorithm was $e^{\sum \beta}$, where the parameters were fixed as in Table 3.1. In practice, when simulating for a Poisson process, for which the history is not taken into account, rather than looping according to the Shedler-Lewis algorithm it is sufficent to generate iid exponential interarrival times with rate M, stopping when the associated event times exceed $A = 1095$, and then using rejection sampling with acceptance probability $\lambda(t|x)/M$.

Table 3.1: Regression parameters used to generate matched data

β_0	β_1	β_2	β_3
1	0	1	0

This generates times according to a Poisson process with log intensity function $log(\lambda_x(t)) \sim 1+cos(\frac{2\pi}{T}t)$, so the intensity function has an annual cycle with minimum and maximum rates of one and exp(2).

After generating the sequence of times $t_1, t_2, \ldots, t_N$, successive observations (t_{2i-1}, t_{2i}) were randomly permuted, with the first member of the pair taken as a case time, and the second member as the matched control time.

3.2.2 Simulation of unmatched data

Unmatched data were generated from independent Poisson processes for cases and controls using the parameter values in table 3.2.

Table 3.2: Regression parameters used to generate unmatched data

Status	β_0	β_1	β_2	β_3
Case	0	1	1	0
Control	0	0	1	0

The intensity function use for cases was the same as that used for cases with matched data, while the intensity function for unmatched controls was $log(\lambda_x(t)) \sim cos(\frac{2\pi}{T}t)$, so that the controls intensity function has a yearly period with minimum and maximum rates $exp(-1)$ and $exp(1)$.

3.3 Analysis of matched data

Seasonal influenza vaccine is usually effective in reducing the probability of hospitalization for influenza, and in the test negative case/control study, vaccine effectiveness is related to the relative probabilities of vaccination in cases, P(V | case), and controls, P(V | control).

The null hypothesis of 0 vaccine efficacy implies that the probability of vaccination is the same in each of the case and control groups, $P(V \mid \text{case}) = P(V \mid \text{control})$.

Having generated a sequence of matched times as described above, indicators of vaccination status were generated as a sequence of Bernoulli random variables with probabilities $P(V \mid \text{case})$ and $P(V \mid \text{control})$. In order to reduce the number of parameters, $P(V \mid \text{case})$ was fixed at .5 throughout, corresponding to a 50% control vaccination rate. $P(V \mid \text{control})$ was then varied, and the power of different statistical procedures to reject the null hypothesis was assessed.

3.3.1 Matched analysis of matched data

Two matched methods of analysis were evaluated - McNemar's test and conditional logistic regression.

The null and alternative hypotheses are:

$$H_0 : P(V \mid case) = P(V \mid control) \tag{3.3}$$

$$H_1 : P(V \mid case) \neq \quad P(V \mid control) \tag{3.4}$$

where $P(V \mid case) = .5$ throughout. All hypothesis tests, here and elsewhere, were carried out at level .05.

For each simulated dataset the p-value for McNemar's test was evaluated using the R function “mcnemar.test”.

For conditional logistic regresson the model fit was

$$logit(P(case)) = \beta_0 + \beta_1 X$$

where $P(case)$ is the probability that the associated admission time is for a case. The conditional logistic regression analysis was carried out using the R function “clogit” in the library survival, using the model statement

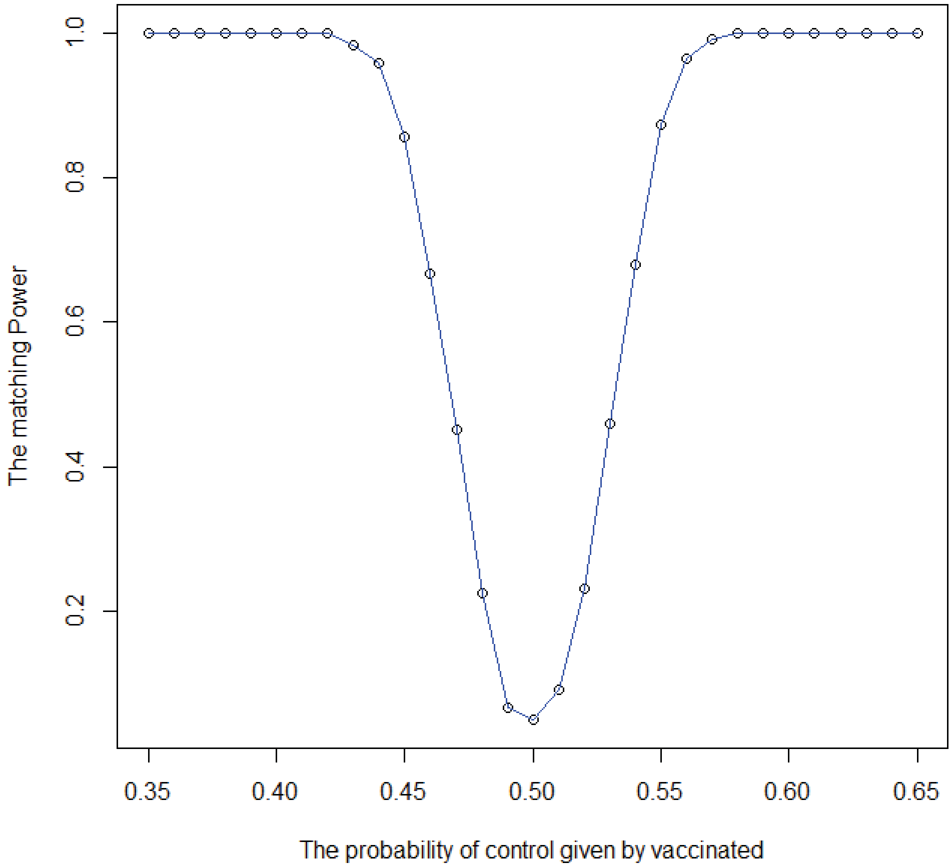
$$case/control \sim \; X + strata(pair)$$

where pair identifies the matched pair, X is the indicator of vaccination status, and case/control is an indicator identifying case vs control observation. The null hypothesis was rejected if the 95% confidence interval for β_1 did not contain 0. The power of each procedure was estimated as the proportion of the simulation batches for which the null hypothesis was rejected.

Figure 3.1 shows the simulated power of McNemar's test. The figure shows that at $P(V \mid control) = 0.5$, the power is close to the nominal level $\alpha = .05$, while the power quickly approaches 1 under the alternative, with power exceeding 80% as soon as the difference between control and vaccination probabilities differs by at least .05.

We conclude that with the length of time interval that we're looking at and the associated rate constants, there are sufficient number of observations that we could detect a difference in vaccination probability of 0.05 with high probability.

Figure 3.1: The matched power of McNemar's test



Conditional logistic regression provides another inference procedure for matched data, and we now study its power in comparison to McNemar's test.

Conditional logistic regression offers a conceptual advantage over unconditional logistic regression for case-control studies [1] in that it depends only on the relative risk parameters of interest and thus allows for construction of exact tests and estimates using matched data.

In contrast to McNemar's test, conditional logistic regression has the advantage that it allows for the inclusion of additional predictor variables. In this chapter no additional predictor variables are included, while in chapter 4, the performance of conditional logistic regression will be assessed where there are additional covariates.

The conditional approach is the best restricted to matched case-control designs, or to similar situations involving very fine stratification, where its use is in fact essential in order to avoid biased estimates of relative risk.

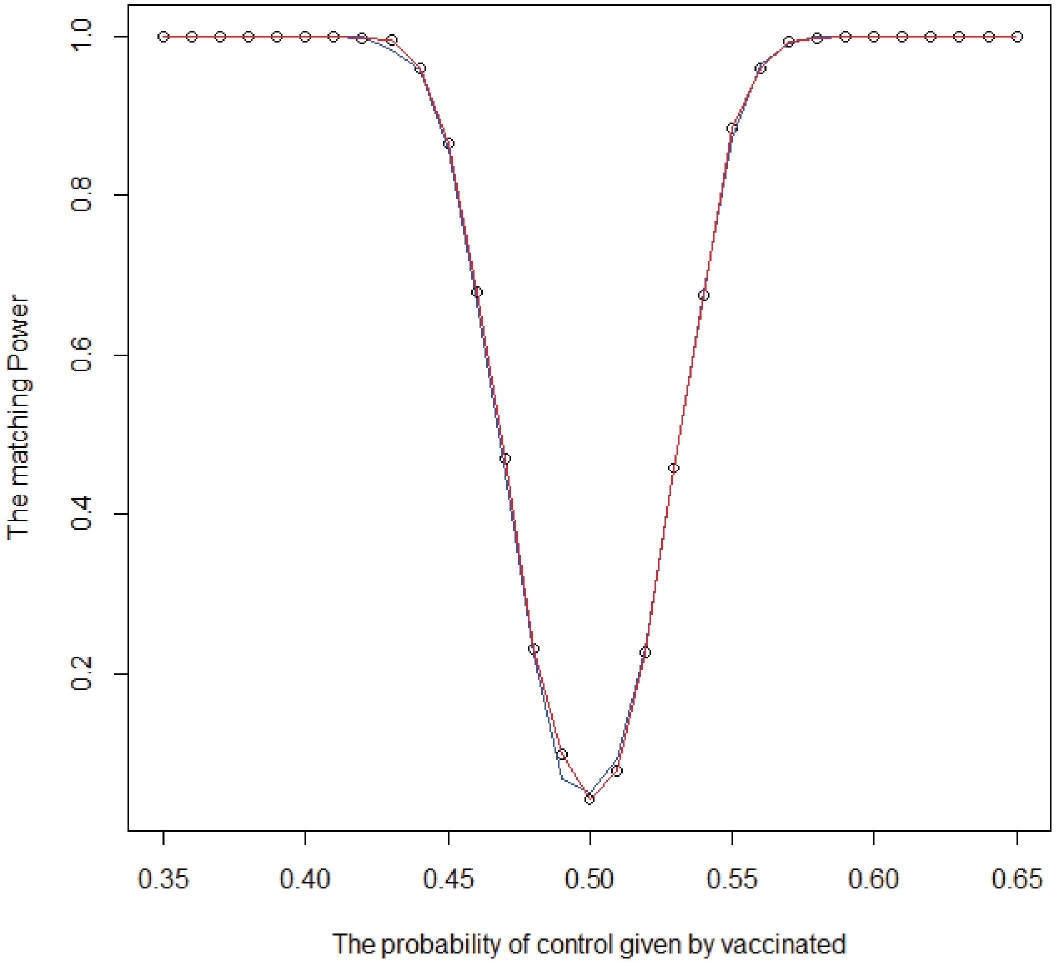
Table 3.3 and figure 3.2 show the estimated powers calculated by McNemar's test and conditional logistic regression in the neighborhood of the hull hypothesis, using 1000 simulation batches, again fixing $P(V \mid case) = .5$. Where the true power is p, the standard variation of the estimated power is $\sqrt{(p(1-p)/1000}$, which has a maximum value of about .016 when $p = .5$.

Table 3.3: Simulated power of McNemar's test and conditional logistic regression with $P(V \mid control)$ from 0.45 to 0.55 and sample size 1000

$P(V \mid control)$	McNemar's	Conditional Logistic Regression
0.45	85.7%	86.6%
0.46	66.8%	68.0%
0.47	45.1%	47.2%
0.48	22.5%	23.2%
0.49	6.9%	9.9%
0.50	5.1%	4.3%
0.51	9.4%	8.0%
0.52	23.2%	22.8%
0.53	45.9%	45.9%
0.54	67.9%	67.5%
0.55	87.2%	88.4%

These results suggest that the power of conditional logistic regression is very close to that of McNemar's test. Conditional logistic regression has the advantage over McNemar's test that additional covariates can be included in the model, and therefore it will be the method of choice for conditional analysis in chapter 4.

Figure 3.2: The power of McNemar's test and Conditional logistic regression with $P(V \mid control)$ from 0.35 to 0.65



3.3.2 Unmatched analysis of matched data

With matched data, it is possible to use an unmatched method of analysis. This entails ignoring the matching status, and using the case/control status as the outcome variable. Specifically we fit a generalized additive model using the "gam" procedure in the R library mgcv, with the probability of a case modeled as

$$logitP(case) = \beta_0 + \beta_1 X + s(X, t)$$

Here X is the indicator of vaccination, and different smooth functions of time are allowed for vaccinated and unvaccinated individuals. By including the smooth functions of time, no knowledge of the underlying periodic structure of the intensity function is required.

Table 3.4 shows the simulated power to detect a vaccine effect using both matched and unmatched methods of analysis, in the neighborhood of the null hypothesis. Where ρ is the true power, the standard deviation of $\hat{\rho}$ based on 1000 batches is $\sqrt{(\rho(1-\rho)/1000}$, which is about .007 when $\rho = .05$, and .16 when $\rho = .5$, so there appears to be little difference between the power of the matched and unmatched analyses.

Table 3.4: Power of matched analysis and unmatched analysis with $P(V \mid control)$ from 0.45 to 0.55 and sample size 1000

$P(V \mid control)$	Matched Analysis	Unmatched Analysis
0.45	85.7%	86.8%
0.46	66.8%	68.4%
0.47	45.1%	45.2%
0.48	22.5%	23.8%
0.49	6.9%	9.2%
0.50	5.1%	5.8%
0.51	9.4%	10.0%
0.52	23.2%	21.8%
0.53	45.9%	47.2%
0.54	67.9%	71.2%
0.55	87.2%	86.3%

In a matched study for which there is a problem in obtaining matched controls, unmatched cases will be unused in a matched analysis. The results in table 3.4 suggest that little power is lost in moving from a matched to an unmatched method of analysis. This suggests that when unmatched cases are common, it may be advantageous to use an unmatched analysis, which can use all data, including unmatched cases.

3.4 Analysis of unmatched data

A simulation was carried out to examine estimator properties with fully unmatched data, in which case matched analysis methods are not appropriate.

In this simulaton we are not generating cases and controls, but rather generating a two sequences of admission times where the intensity of the admission process differs for vaccinated and unvaccinated individuals, and also depends smoothly on time. Where $x = 1$ denotes vaccinated individuals and $x = 0$ denotes unvaccinated individuals, we are simulating times according to the intensity function

$$log(\lambda_x(t)) \sim \beta_0 + \beta_1 x + \beta_2 cos(\frac{2\pi}{T}t) + \beta_3 sin(\frac{2\pi}{T}t) \tag{3.5}$$

The cyclic terms represent a seasonally varying vaccination rate. We used a period of one year, and assumed observation over a 3 year window. The average number of admission events in the 3 year period depends on the value of the parameters, and if $\beta_1 > 0$, will be higher for vaccinated individuals. The goal is to assess the ability to estimate β_1 in the presence of cyclic variation.

In the simulation, β_0 was set at 0, and $\beta_1 = 1$. For unvaccinated individuals, $(\beta_2, \beta_3) = (1, 0)$ and for vaccinateds, $(\beta_2, \beta_3) = (0, .25)$. This choice of parameters entails a 90^o phase shift for event times in vaccinated individuals, as compared to unvaccinated, and also, an amplitude change in the cyclic term. The associated intensity functions are shown in the top plots of figures 3.3 and 3.4, where the left hand panels are for unvaccinated, and the right hand panels are for vaccinated. The ordinate is time in days. Admission times for vaccinated and unvaccinated were generated independently of one another.

It is straightfoward to write down the likelihood for a Poisson process with parameterized intensity function, and maximize to obtain maximum likelihood estimators.

However, in the SOS study, admission times are recorded to the nearest day, in discrete time, and so we chose to use a discrete time approximation in the simulation. The event times were transformed to a binary time series $Y(t)$ by discretizing the time interval [0,3 years] into 100000 equal length subintervals. This discretization is sufficiently fine that each subinterval contains at most 1 point. The middle panels in figures 3.3 and 3.4 show a windowed average of the resulting binary time series for one replicate, and appears to capture the cyclic nature of the intensity function.

To the associated sequence of Bernoulli random variables $Y_1, Y_2, \ldots, Y_{10000}$ we fit the binomial generalized additive model

$$logit(P(Y=1)) = \beta_0 + \beta_1 X + s(X,t)$$

This model allows for separate smooth functions of time in vaccinateds and unvaccinateds, together with a fixed effect of vaccination status. The estimated smooth functions of time are shown in the lower panels of figures 3.3 and 3.4. These plots are illustrative only, being based on just two simulated data sets. Figure 3 differs from figure 4 in that the amplitude parameter of the sinusoidal variation is reduced from .25 in figure 3.3, to .10 in figure 3.4.

Without having specified the sinusoidal nature of the underlying smooth term, the estimates from the gam fit appear to capture the underlying smooth structure.

Figure 3.3: Estimated smooth component of intensity function. Top: true inensity function. Middle: running mean of binary time series: Bottom: estimated smooth component of intensity from gam fit. Left panels: $(\beta_0, \beta_1, \beta_2, \beta_3) = (0, 1, 1, 0)$ Right panels: $(\beta_0, \beta_1, \beta_2, \beta_3) = (0, 0, 0, .25)$.

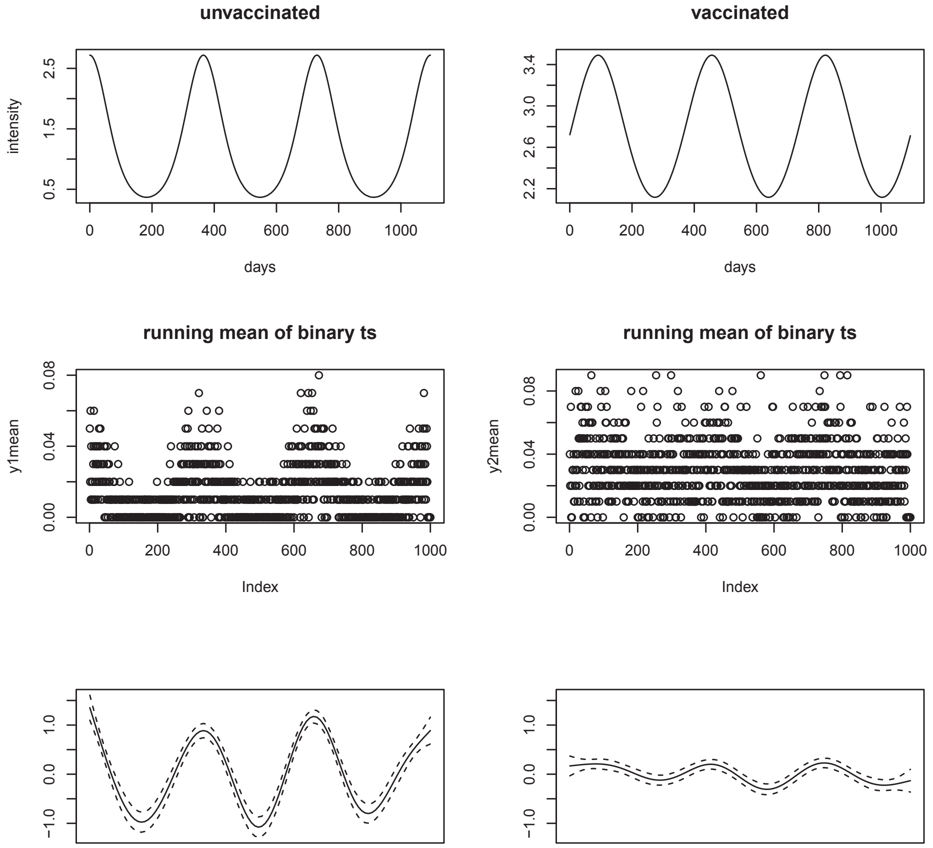
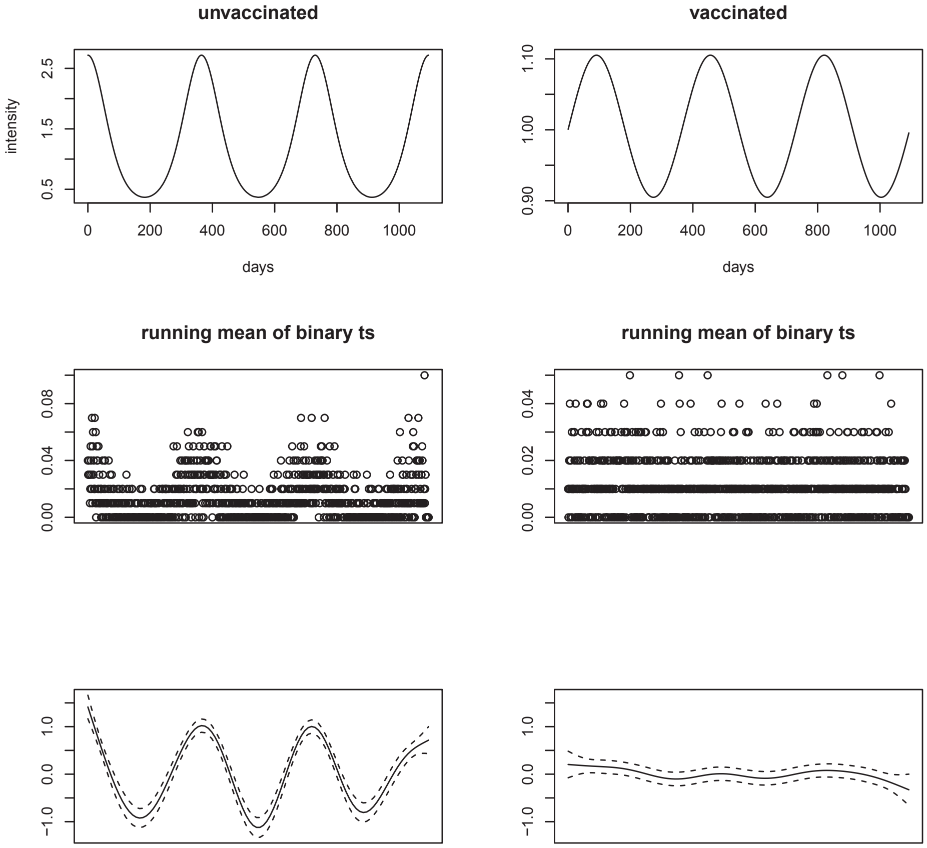


Figure 3.4: Estimated smooth component of intensity function. Top: true inensity function. Middle: running mean of binary time series: Bottom: estimated smooth component of intensity from gam fit. Left panels: $(\beta_0, \beta_1, \beta_2, \beta_3) = (0, 1, 1, 0)$ Right panels: $(\beta_0, \beta_1, \beta_2, \beta_3) = (0, 0, 0, .10)$



It is of primary interest to assess the ability of the *gam* fit to provide an accurate estimate of β_1.

By transforming the point process to a binary time series, with discretization interval sufficiently small that each interval contains at most one point, the resulting time series is very long, and this can lead to substantial computational cost. The question arises as to whether a coarser discretization can be used.

It is well known that a Poisson process has independent increments, and that the number of events in an interval (a,b] will have a Poisson distribution.

We broke the interval (0,3 years] into 100 subintervals of equal length, and counted the number of events in each subinterval, leading to a collection of 100 Poisson distributed observations, where the mean μ_j for the j'th observation is a function of X and t_j. We then fit a Poisson generalized addive model with mean parameterized as

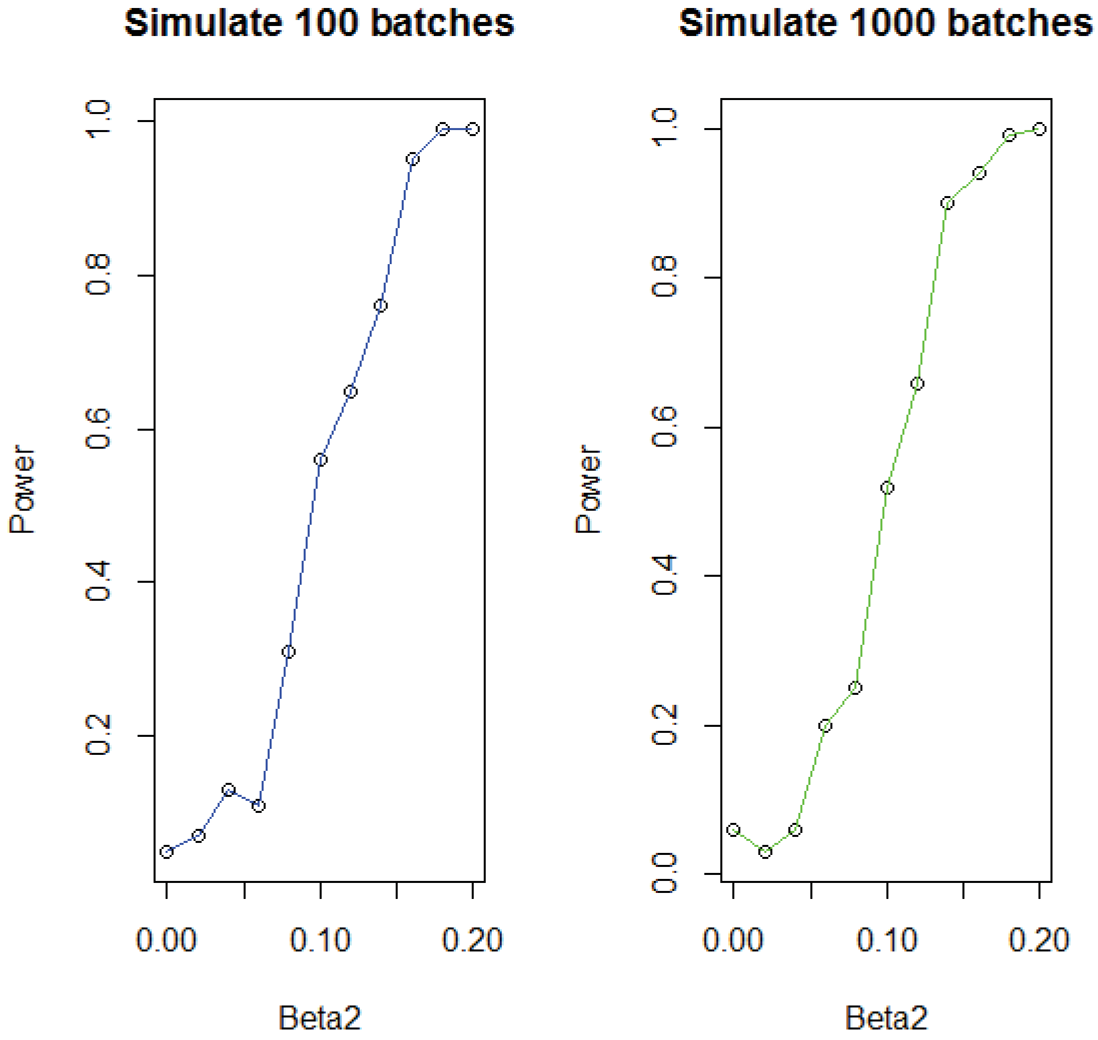
$$log(\mu_j) = \beta_0 + \beta_1 X + s(X, j)$$

in which real time has been replaced by the discrete index j.

The results of a simulation study are presented in figure 3.5, which shows the estimated power to reject $H_0 : \beta_1 = 0$ when testing against the two sided alternative at level .05. Under the alternative we used values of β_1 ranging from 0 (which represents the null hypothesis) to .2, beyond which the power was essentially 1. The left hand panel shows the estimated power using 100 simulation batches, and the right hand panel uses 1000 simulation batches, which provides a more precise power estimate.

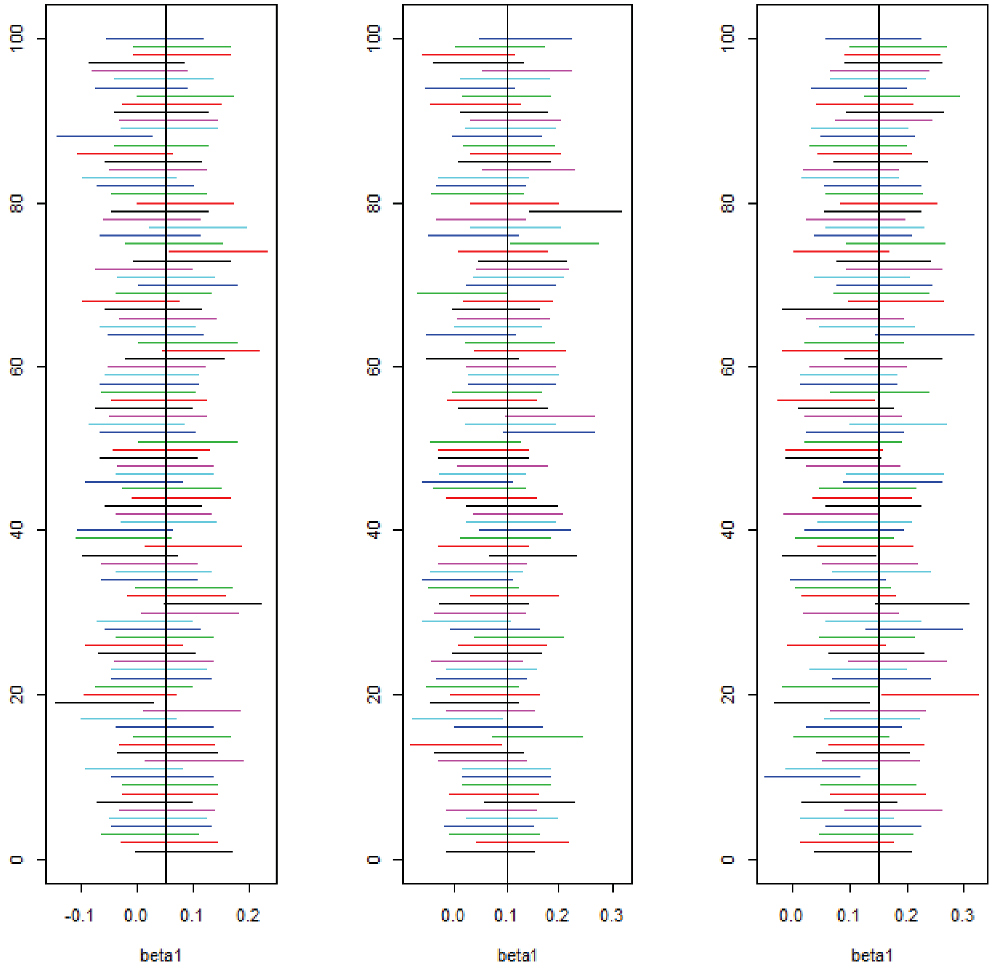
The simulation suggests that a Poisson regression model with a nonparametric smooth term can be an effective method for estimating vaccine effect in the presence of time variation.

Figure 3.5: Estimated power to detect $\beta_1 > 0$. Left hand panel is based on 100 simulation batches, right hand panel used 1000 batches.



95% confidence intervals for β_1 are shown in figure 3.6. Each panel shows 100 intervals, each interval being based on 1 simulation batch. The true values of β_1 are shown as vertical lines, and are .05, .1 and .15 respectively, for the left, middle and right hand panels. The estimated coverage probabilties based on 1000 simulation batches were 95.1%, 92.8% and 94.3%, indicating that the empirical coverage is close to the nominal. The 92.8% coverage appears significantly lower than the nominal .95, in that it leads to an observed Z statistic of -3.19, which has one sided p-value .0007.

Figure 3.6: 95% confidence intervals for β_1, where $\beta_1 = .05$ (left), .10 (centre) and .15 (right).



Chapter 4

Method appropriate to the SOS study

The SOS study was designed as a matched study with matching on time, location, subject age, and including a number of nonmatching predictor variables. The proposed method of analysis was conditional logistic regression. In the first year of the study, the goal was to match two controls to each case. Based on difficulties in finding sufficient numbers of matched controls, the matching goal was reduced to 1:1 in year 2, and exprience has been that many cases remain unmatched.

For a matched analysis, any unmatched cases are unused. This represents a possibly inefficient use of data, and the question arises as to whether there might be some advantage to using an unmatched analysis method, which would include unmatched cases. This is explored in the present chapter. Data are generated according to a matched design, after which a random subset of controls are chosen to be missing. For the matched analysis the cases associated with the missing controls are also deleted prior to analysis. For unmatched analysis, the missing controls are excluded, but the associated cases are included in the analysis. We are interested in potential bias that this might generate, and also whether, for a sufficiently large proportion of missing controls, the unmatched method might provide a more precise estimate of vaccine efficacy than the matched method, and/or increased power to detect a nonzero VE. We explored the effects of including a non-matching covariate, and of adding a time varying vaccination status. As regards the latter, it is well known that the influenza vaccination rate is cyclic, and mirrors the quantity of circulating virus to a certain degree. People generally get vaccinated when there is flu circulating in the community. There is concern that if the time variation in vaccination rate differs in cases and controls, this may reduce the ability to accurately assess the average vaccine efficacy.

Admission times were randomly generated from a Poisson process with rate $\lambda(t)$, where $log(\lambda(t)) = 1 + cos(2\pi t/T)$, such that $\lambda(t)$ has a period of 1 year, with three years of observation. Where the number of observations was odd, the last observation was removed. The resulting sequence of times $t_1, t_2, \ldots$ is divided into a collection of cases and matched controls by randomly permuting the order of each pair t_{2i-1}, t_{2i}, assigning the first permuted observation as a case, and the second as a control.

Vaccination status was then randomly generated according to probabiltities $p_1 = P(Vaccinated|Case)$ and $p_2 = P(Vaccinated|Control)$, in which case the odds ratio for vaccination, which equals the odds ratio for being an influenza case, is $OR = p_1(1-p_2)/(p2(1-p_1))$, with vaccine efficacy $1 - OR$.

In some simulations another variable was considered, referred to here as gender and generated as Bernoulli variables with probabilities $P(Male|Case)$ and $P(Male|Control)$.

Where V denotes the indicator of vaccination, and Y is the indicator of case ($Y = 1$)/ control ($Y = 0$) status, and gender was included, the unmatched models fit were Bernoulli generalized additive models of the form

$$logit(P(Case)) = \beta_0 + \beta_1 V + \beta_2 Gender + s(t|V)$$

with $s(t|V)$ being smooth functions of time, which are allowed to be different for vaccinated and unvaccinated individuals.

The matched analysis fits a conditional logistic regression model of the form

$$logit(P(case)) = \beta_0 + \beta_1 V + \beta_2 Gender$$

with stratification (matching) by time.

Section 4.1 examines the effect of adding the additional nonmatching covariate gender, and section 4.2 examines the influence of missing data on both matched and unmatched analyses. In section 4.4 we explore the effect of time varying vaccinaton rates for cases $P(V|t, Case)$, and controls $P(V|t, Control)$. In this situation the vaccine efficacy may be time varying, and an overall estimate, calculated, for example, by collapsing to a 2×2 table of vaccine vs case/control status will be estimating some type of time average, at best.

4.1 Adding covariates

There are two kinds of covariates which can be added - those involved in the matching design, and those not associated with matching. Inclusion of matching variables as covariates in a matched analysis is not appropriate, and so we restrict to covariates not used in matching controls to cases.

We considered the performance of the matched analysis after including one additional covariate, a Bernoulli variable which we have referred to as gender, with $G = 1$ representing males, and $G = 0$ females.

A simulation was carried out fixing $P(male|case) = .5$, and varying $P(male|control) = .45, .46, \ldots, .55$. We fixed $P(V|control) = .45$ and $P(V|case) = .50$ and studied the power to detect the vaccine effect in the presence of a Gender effect. The following table gives the estimated power of the conditional likelihood method based on 1000 simulation batches.

Table 4.1: Estimated power to detect a vaccine effect. $P(Male|Case) = .5$

$P(Male|Control)$	Power
0.45	84.7%
0.46	87.1%
0.47	86.1%
0.48	86.7%
0.49	85.0%
0.50	84.9%
0.51	85.9%
0.52	86.8%
0.53	85.5%
0.54	87.6%
0.55	88.2%

The results in the table suggest that the power is not dramatically altered by inclusion of a covariate whose values differ for cases and controls.

In this and other simulation results reported, there is Monte Carlo sampling variation. In particular, if the true power p is estimated as $\hat{p}$ - the proportion of N independent replicates which either reject or do not reject the null hypothesis - then a 95% confidence interval for p will have half width of about $2\sqrt{p(1-p)/N}$, which for, say $p = .85$ and $N = 1000$, is about .023. Based on this, there is some weak evidence that there is a small variation in power as the gender ratios vary between cases and controls.

4.2 Comparison of unmatched and matched analyses when the design was matched

In this section we examine the performance of matched and unmatched analyses for estimating vaccine efficacy when data are matched. For the unmatched analyses, this address the question of the effect of breaking the match.

4.2.1 Estimation

As mentioned in chapter 2, the odds ratio is

$$OR = \frac{\frac{P(Vaccinated|case)}{P(NotVaccinated|Case)}}{\frac{P(Vaccinated|Control)}{P(NotVaccinated|Control)}} \tag{4.1}$$

with $VE = 1 - OR$. Where $p_1 = P(V|case)$ and $p_2 = P(V|control)$,

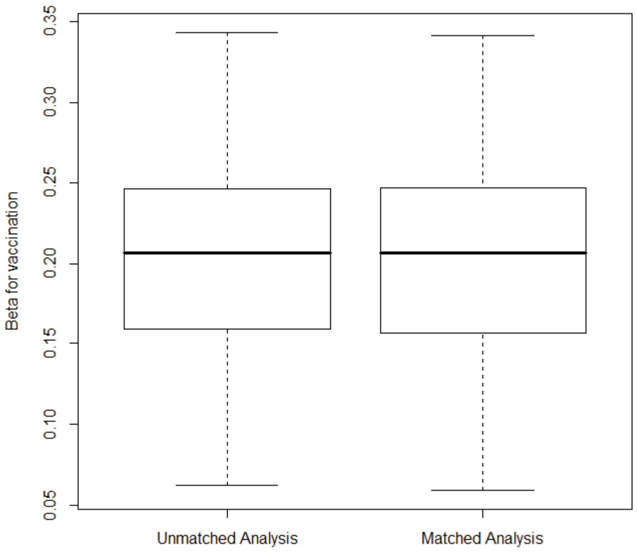
$$e^{\beta_1} = \frac{\frac{p_1}{1-p_1}}{\frac{p_2}{1-p_2}} \tag{4.2}$$

$\Longrightarrow$

$$\beta_1 = ln(\frac{\frac{p_1}{1-p_1}}{\frac{p_2}{1-p_2}}) \tag{4.3}$$

We carried out a simulation with $p_1 = 0.5$ and $p_2 = 0.45$, and also fixing $P(Male|Case) = .5$ and $P(Male|Control) = .45$. In this case, $\beta_1 = 0.2006707$. The estimated values of β_1 using matched and unmatched analyses are shown in figure 4.1 for 100 simulation batches. There appears to be little difference in the distributions of the estimated values from unmatched and matched analysis, with the boxplots centred near to the true β_1.

Figure 4.1: Boxplot of 100 simulated values of $\hat{\beta}_1$ calculated using matched and unmatched analyses



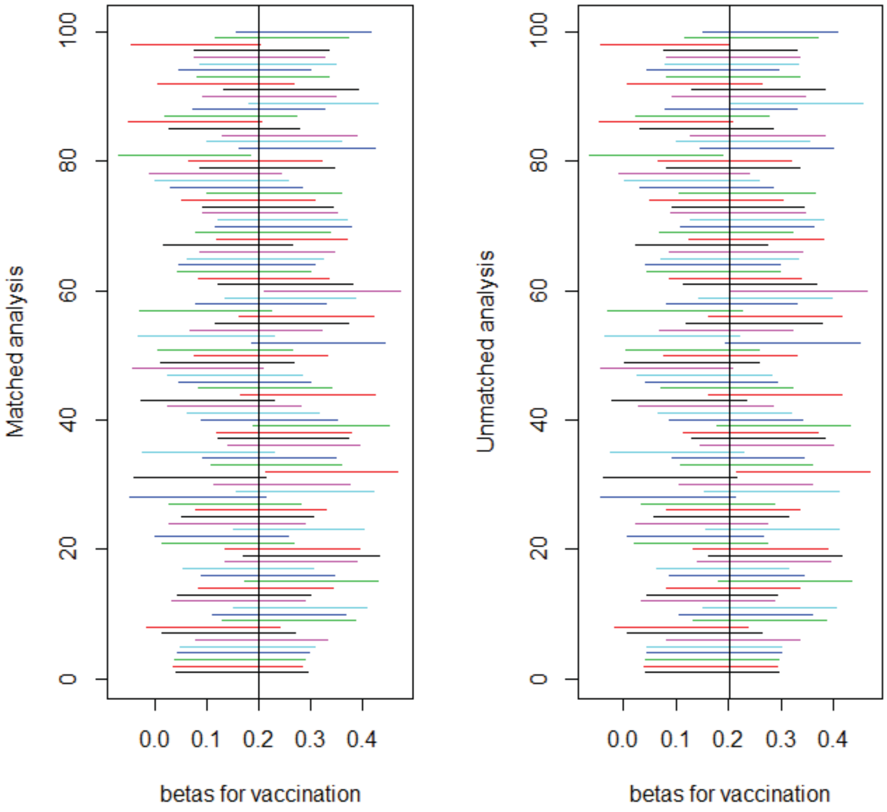
Where $Bias[\widehat{\beta_1}] = E_{\beta_1}[\widehat{\beta_1}] - \beta_1$ and $MSE(\hat{\beta}_1) = E_{\beta_1}\left[(\hat{\beta}_1 - \beta_1)^2\right]$, table 4.2 shows the estimated bias and MSE of the matched and unmatched estimators, together with the power to detect a non-zero value of β_1 when testing against the two sided alternative at level .05.

Table 4.2: Estimated bias, MSE and power of matched and unmatched analyses based on 100 simulation batches

	Matched analysis	Unmatched analysis
$\widehat{Bias}$	$1.36 * 10^{-3}$	$8.84 * 10^{-4}$
$\widehat{MSE}$	$4.09 * 10^{-3}$	$4.04 * 10^{-3}$
$\widehat{Power}$	97%	96%

The estimated MSE is very close for the two methods, while the bias is about double for the matched analysis. Figure 4.2 shows the 95% confidence intervals from the 100 simulation batches. The empirical coverage was 97% for the matched analysis and 96% for the unmatched analysis. Where the nominal coverage is .95, the estimated coverage based on 100 simulation batches would have a standard error of $\sqrt{.95(.05)/100} \approx .02$, suggesting that the empirical coverage is compatible with the nominal.

Figure 4.2: 95% confidence intervals for β_1 calculated using matched and unmatched analysis, for 100 simulation batches



4.3 Missing controls

In the SOS study, controls are matched by time and location to account for the quantity of circulating virus, and age, to control for the known age effect on the immune response. In reality, it is often impossible to identify matched controls, and in order to calculate the VE for the SOS study, the Principal Investigator decided to use only matched pairs to estimate VE. The resulting estimate was sufficiently small that the lower end of the 95% CI for VE was negative, compatible with decision in favour of the null hypothesis $H_0 : VE = 0$. This led to a discussion as to whether incorporating unmatched cases might increase precision of the VE estimate.

Jewell [8] notes that gains in precision of matched designs are most evident when the matching factors are strongly associated with the outcome of interest (in the SOS study hospitalization) and with the relevant exposure variable (here vaccination). He notes that in cohort studies, breaking the match and pooling yields a valid estimate of the odds ratio, but that accounting for the matching variables in the analysis is usually necessary to produce an appropriate assessment of variability. He argues further that for matched case-control data it is essential to control for matching factors in an unmatched analysis, in order to obtain a valid estimate of the odds ratio. In simulations examining the impact of breaking the match, we included the matching variable time as a predictor in the unmatched analysis, in which case we expect valid inferences.

A first simulation estimates the power of the matched and unmatched procedures when unmatched cases are deleted from each analysis. Table 4.3 shows the estimated power to reject the hypothesis $H_0 : VE = 0$ against the two sided alternative. For each simulation batch the odds ratio was estimated using both matched and unmatched analyses, together with 95% confidence interval. The null hypothesis was rejected if the CI for the odds ratio did not contain 1. The values reported are the proportion of 1000 simulation batches for which the null hypothesis was rejected.

As expected, estimated power decreases as the proporton of nonmissing data decreases. The unmatched analysis appears to have higher power than the matched analysis even though matched data was being used. This might be due to the matched analysis using a conditional likelihood, whereas the unmatched analysis uses a full likelihood, albeit for a generalized linear model incorporating smooth functions of

Table 4.3: Percentage of missing matched pairs, and estimated power for matched and unmatched analysis, based on 1000 simulation batches

Percentage of nonmissing matched pairs	Estimated power of unmatched analysis	Estimated power of matched analysis
20%	30.2%	28.7%
25%	31.5%	32.1%
30%	38.1%	39.2%
35%	42.0%	41.9%
40%	48.8%	50.1%
45%	54.9%	52.2%
50%	58.3%	57.0%
55%	60.6%	62.3%
60%	67.4%	67.2%
65%	70.0%	69.9%
70%	73.3%	74.4%
75%	75.7%	76.1%
80%	79.0%	78.1%
85%	81.4%	80.8%
90%	83.0%	83.4%
100%	86.6%	85.7%

time. Another explanation might be that there is no need for matching, as the simulation used constant vaccination rates for cases or controls. On the other hand, the estimated powers of the two tests are within two standard errors based on the Monte Carlo sampling variation, indicating no real difference in power.

4.3.1 Missing controls only

Another simulation was carried out in which unmatched cases were discarded for the matched analysis, but retained for the unmatched analysis. Table 4.4 shows the estimated power of the unmatched analysis, together with the percentage of non-missing controls. These values can be compared to the estimates in table 4.3, where unmatched cases were removed.

Comparison of tables 4.3 and 4.4 shows that there can be a considerable increase in power of an unmatched analysis, as compared to a matched analysis, when cases are retained for the unmatched analysis and when there is a moderately large proportion of missing controls.

Table 4.4: Estimated power of unmatched analysis based on 1000 simulation batches.

Proportion of nonmissing controls	Estimated power of unmatched analysis
100%	85.1%
90%	83.2%
80%	81.7%
70%	75.7%
60%	73.2%
50%	70.3%

Figure 4.3 illustrates the increased power of the unmatched analysis when unmatched cases are retained, as opposed to when they are dropped from analysis. Blue triangles show estimated power when unmatched cases are retained, and green circles show estimated power when they are deleted.

Figure 4.3: Estimated power of unmatched analysis vs percentage of missing controls. Blue - unmatched cases are retained in the analysis. Green - unmatched cases are deleted

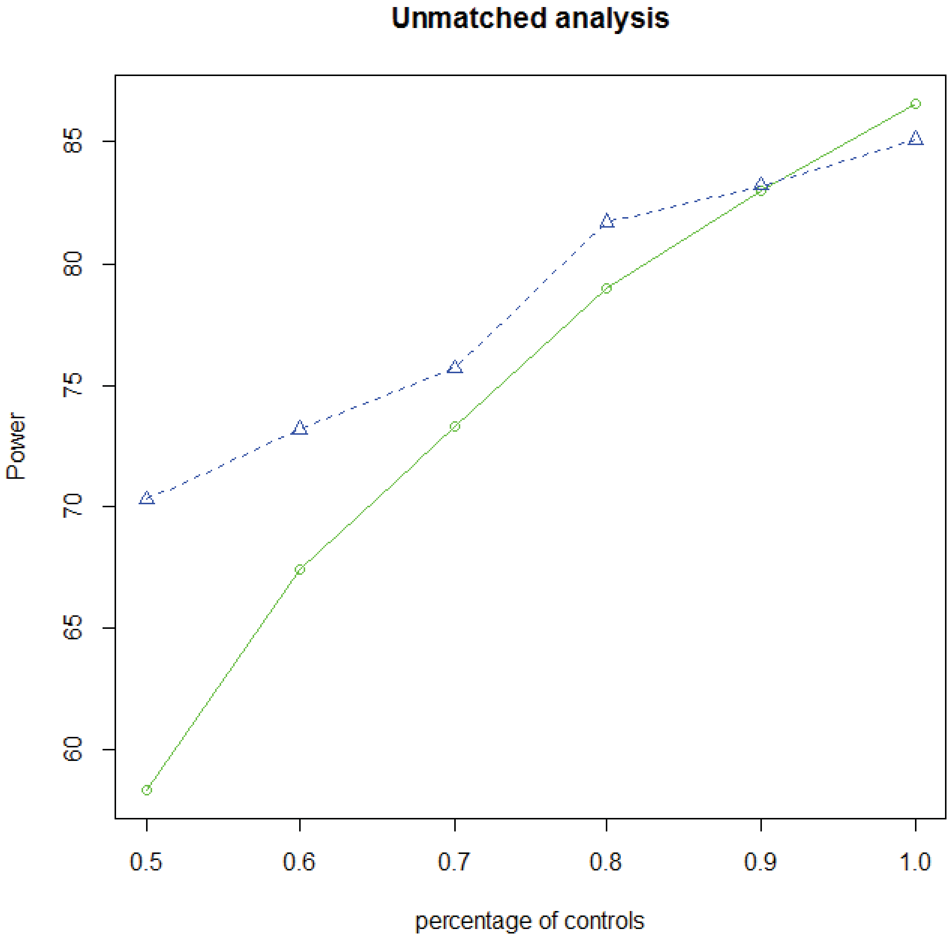


Table 4.5 provides further information in the form of 95% confidence intervals for the true power. The table reports confidence intervals on the true power based on 1000 batches, when unmatched controls are both deleted and included.

Table 4.5: 95% confidence intervals on the power of the unmatched analysis based on fully missing pairs and missing controls

Proportion of nonmissing controls	CI on power with unmatched cases removed	CI on power with unmatched cases included
100%	(84.5% , 88.7%)	(82.9% , 87.3%)
90%	(80.7% ,85.3%)	(80.9% , 85.5%)
80%	(76.5% , 81.5%)	(79.3% ,84.1%)
70%	(70.6% , 76.0%)	(73.0% , 78.4%)
60%	(64.5% , 70.3%)	(70.5% , 75.9%)
50%	(55.2% , 61.4%)	(67.5% , 73.1%)

4.4 Time varying vaccination rate

In general, we know the vacination rate is not a constant through the year, with flu vaccination rate increasing in the flu season. In setting up a simulation with time varying vaccination rate, we assumed that vaccination rate has period of one year.

We assumed that $P_1(t)$ and $P_2(t)$, the time varying probabilities of vaccination for cases and controls, respectively, are given by

$$logit(P_1(t)) \sim \alpha_{0c} + \alpha_{1c}cos(\frac{2\pi}{T}t) + \alpha_{2c}sin(\frac{2\pi}{T}t) \tag{4.4}$$

and

$$logit(P_2(t)) \sim \alpha_{0o} + \alpha_{1o}cos(\frac{2\pi}{T}t) + \alpha_{2o}sin(\frac{2\pi}{T}t) \tag{4.5}$$

for which the log odds ratio is

$$logOR(t) = logitP_t(V|C) - logitP_t(V|O) \tag{4.6}$$

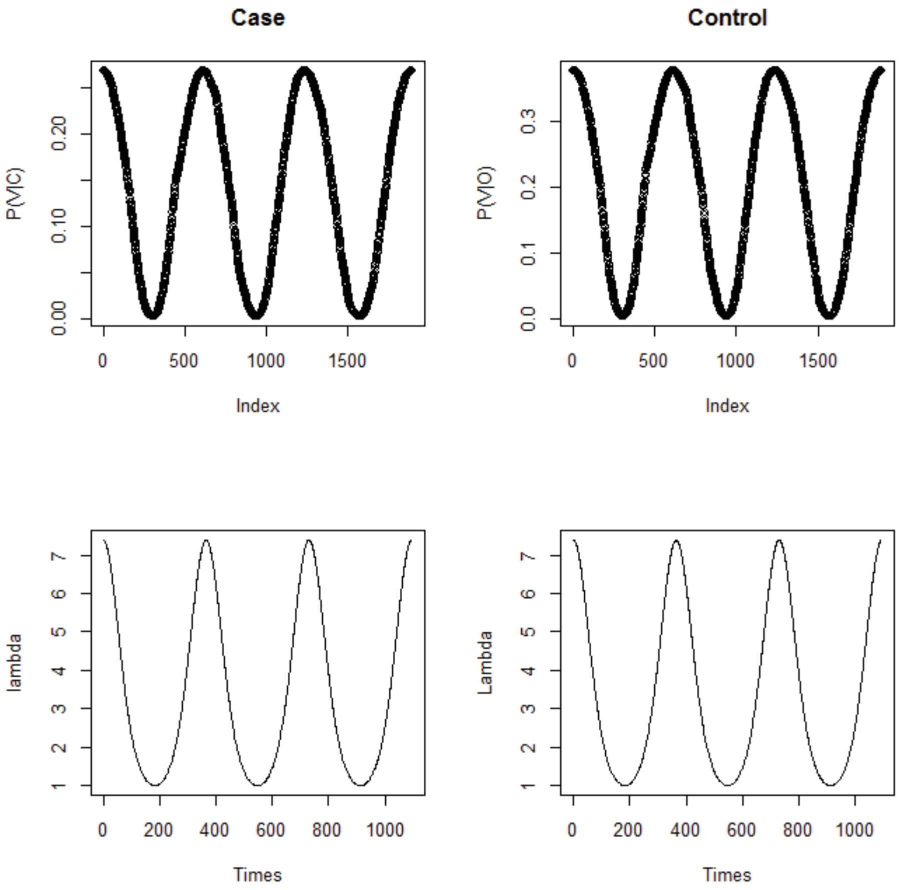
or

$$logOR(t) = (\alpha_{0c} - \alpha_{0o}) + (\alpha_{1c} - \alpha_{1o})cos(\frac{2\pi}{T}t) + (\alpha_{2c} - \alpha_{2o})sin(\frac{2\pi}{T}t) \tag{4.7}$$

Note that unless $(\alpha_{1c} = \alpha_{1o})$ and $(\alpha_{2c} = \alpha_{2o})$, the odds ratio will be time varying.

Figure 4.4 shows the time varying probabilities of vaccination for cases and controls, together with the intensity function $\lambda(t)$.

Figure 4.4: The probability function and lambda function in terms of times with cases and controls



4.4.1 Estimating β

A small simulation study was carried out using the following six models with time varying vaccination rate.

Table 4.6: Coefficients of time varying vaccination rate

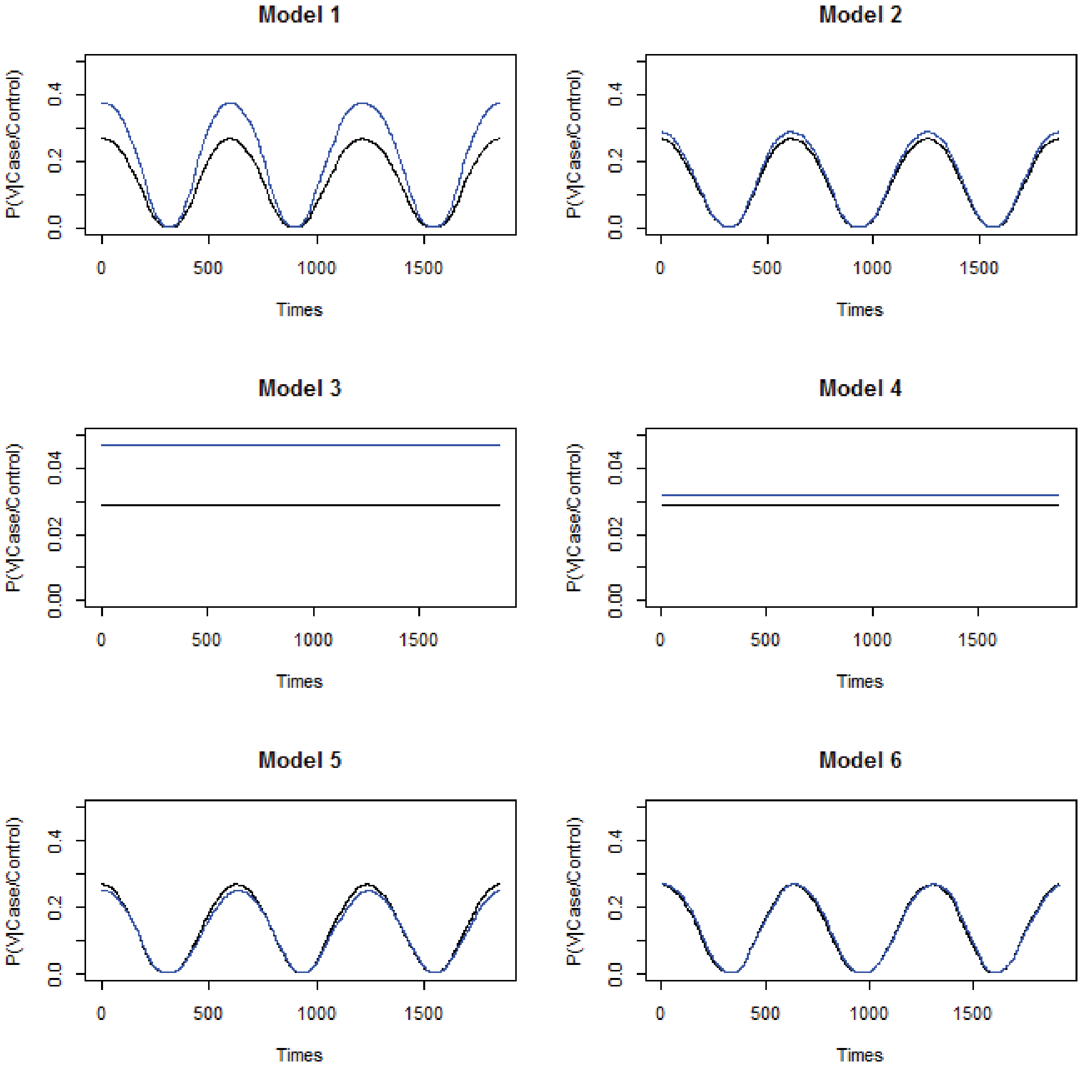
Model	Status	α_0	α_1	α_2
1	Case	-3.5	2.5	0
	Control	-3	2.5	0
2	Case	-3.5	2.5	0
	Control	-3.4	2.5	0
3	Case	-3.5	0	0
	Control	-3	0	0
4	Case	-3.5	0	0
	Control	-3.4	0	0
5	Case	-3.5	2.5	0
	Control	-3.5	2.4	0.1
6	Case	-3.5	2.5	0
	Control	-3.4	2.4	0.1

The odds ratios for models 1-4 are constant, because the time varying structure of the vaccination rate is the same for cases and controls. Models 1 and 2 differ in the magnitude of the odds ratio. For models 3 and 4, the vaccination rates are constant, as are the odds ratios, and a comparison of those models with 1 and 2 shows the effect, if any, of a time varying vaccination rate.

Models 5 and 6 have time varying odds ratio with a small (model 6) or absent (model 5) constant term in the log odds ratio. These models were not considered further at this time, but are discussed in Chapter 5.

Figure 4.5 shows the probability of vaccination for the six models for controls (blue) and cases (black). It is expected that the ability to detect a vaccine effect will be greatest in those models where the ratio of the vaccination probabilities is largest.

Figure 4.5: Probability of vaccination for cases (black) and controls (blue) for models 1-6



An unmatched model was fit, in which the rate of the case/control binary series is given by

$$log(\lambda_x(t)) \sim \beta_0 + \beta_1 x + s(t|x) \tag{4.8}$$

where x is the indicator of vaccination status, and the goal was to examine the behaviour of the estimate of β_1 using matched and unmatched methods.

In the log odds ratio calculations above, the sine and cosine terms cancel for models 1 through 4. The difference of intercept terms $\alpha_{0c} - \alpha_{0o}$ remains, and that is the true β_1 which is being estimated, so that for model 1 and model 3, $\beta_1 = -0.5$ and for model 2 and 4, $\beta_1 = -0.1$. Models 5 and 6, which have different time varying vaccination rates for cases and controls, so a model with constant value of β_1 is not appropriate.

For each of 100 simulation batches, β_1 was estimated using both matched and unmatched analyses for models 1-4 . The boxplots of the estimated values $\hat{\beta}_1$ are shown in figure 4.6 All $\hat{\beta}_1$ in those 8 boxplots are very close to the true β_1. The median of the matched estimator is closer to β_1 than the median of the unmatched estimator for model 1, while for models 2 through 4, the distribution of the estimates from the matched and unmatched analyses are similar.

The bias and MSE of the estimates was also calculated for models I-IV, for both matched and unmatched analyses.

4.4.2 Bias and MSE with no missing data

The results shown in table 4.7 are for the case that there are no unmatched cases. The estimated bias is greater for the unmatched analysis than for matched analysis for these 4 models, while the estimated MSE's are similar. For the matched analysis, the bias in model 3 is twice bigger than the bias in model 1; the bias in model 4 is also twice bigger than the bias in model 2; the MSE in model 3 is much bigger than the MSE in model 1; the same as the MSE in model 4 is much bigger than the MSE in model 2. For the unmatched analysis, the bias in model 3 is the same as the bias in model 1; the bias in model 4 is the same as the bias in model 2 as well. The MSE in model 3 is much bigger than the MSE in model 1; the same as the MSE in model 4 is much bigger than the MSE in model 2.

Figure 4.6: The boxplot of estimated β_1 using matched and unmatched analysis in model 1 through model 4

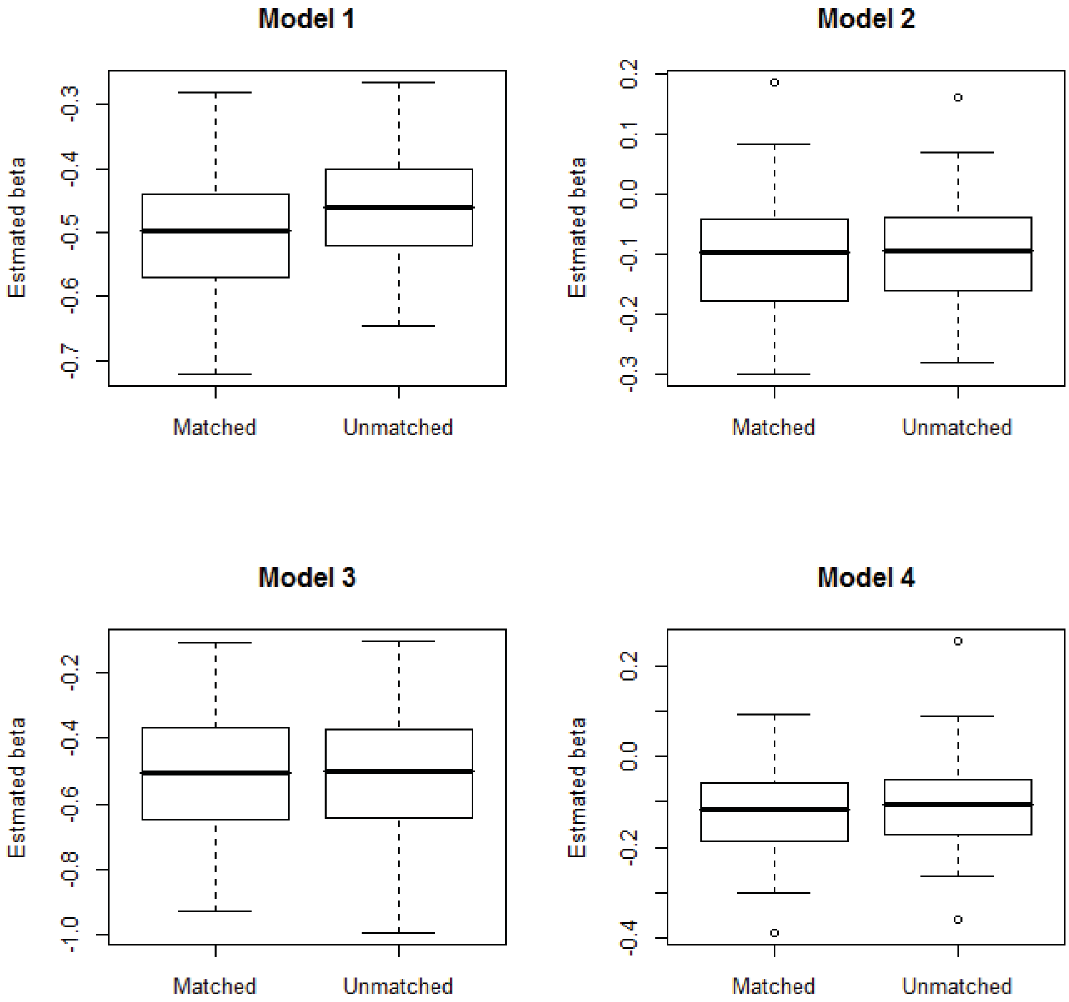


Table 4.7: Estimated Bias and MSE of matched and unmatched analysis with time varying vaccinated rate-models with 100% of matching data

Models		Matched analysis	Unmatched analysis
Model 1	Bias	$-1.1 * 10^{-2}$	$3.87 * 10^{-2}$
	MSE	$9.03 * 10^{-3}$	$8.65 * 10^{-3}$
Model 2	Bias	$2.13 * 10^{-3}$	$1.04 * 10^{-2}$
	MSE	$6.90 * 10^{-3}$	$5.87 * 10^{-3}$
Model 3	Bias	$-2.55 * 10^{-2}$	$-3.24 * 10^{-2}$
	MSE	$2.77 * 10^{-2}$	$2.94 * 10^{-2}$
Model 4	Bias	$-4.67 * 10^{-2}$	$-4.96 * 10^{-2}$
	MSE	$2.08 * 10^{-1}$	$2.16 * 10^{-1}$

4.4.3 Bias and MSE with missing data

Figure 4.7 shows the estimated bias and MSE of $\hat{\beta}_1$ for matched and unmatched methods, as a function of the proportion of missing data. For matched analysis, the proportion of matched data runs from 100% to 50%. For unmatched analysis, this is the proportion of controls used, while both matched and unmatched cases are included.

There is more bias in the unmatched analysis than in the matched analysis in the first 2 models. In model 3 and 4, there is more bias in the unmatched analysis with complete and with 90% of matched data. For matched analysis, the bias in model 3 is bigger than the bias in model 1; the bias in model 4 is also bigger than the bias in model 2.

The MSE started getting bigger in matched analysis with the percentage of missing data increasing in model 1 to model 3. The MSE in model 4 is much bigger than the MSE in model 2. The MSE using matched analysis is the same as the MSE using unmatched analysis in model 4.

So the vaccination function with time varying which are model 1 and model 2 are closer than real models. The bias from model 1 and model 2 are smaller than model 3 and 4. As long as keeping 80% of matched pairs in the models, we can get a good estimate.

Figure 4.7: Estimated bias as a function of proportion of nonmissing controls, for both matched and unmatched analyses, based on 100 simulation batches

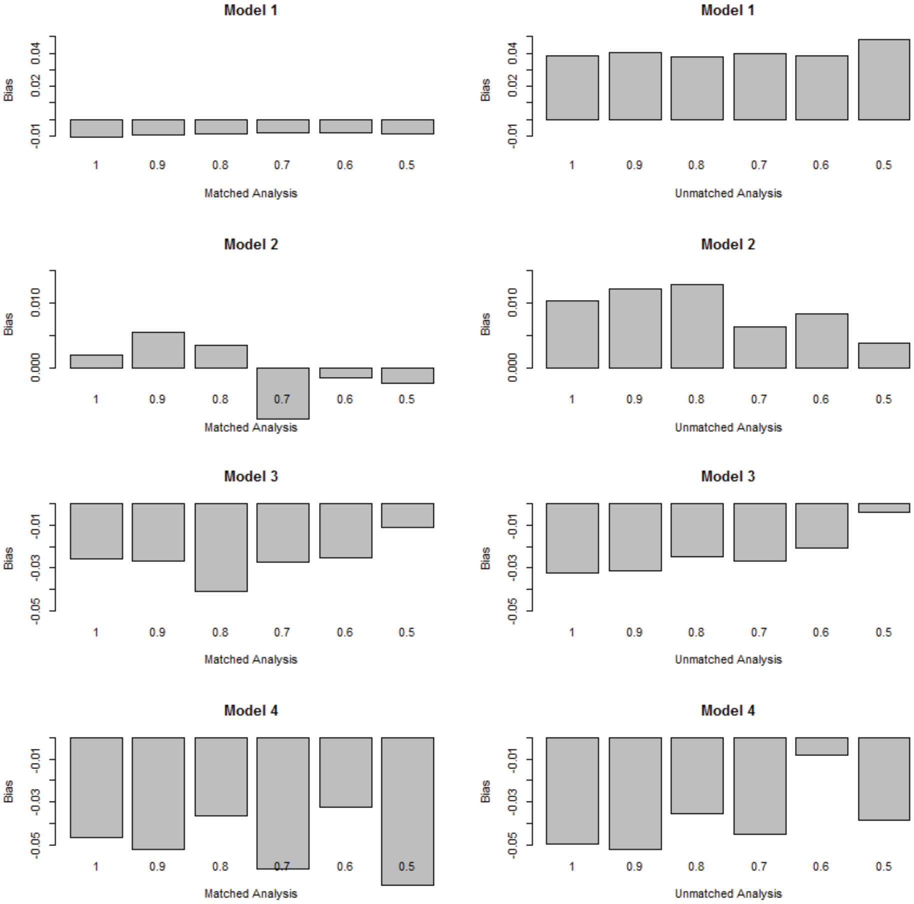
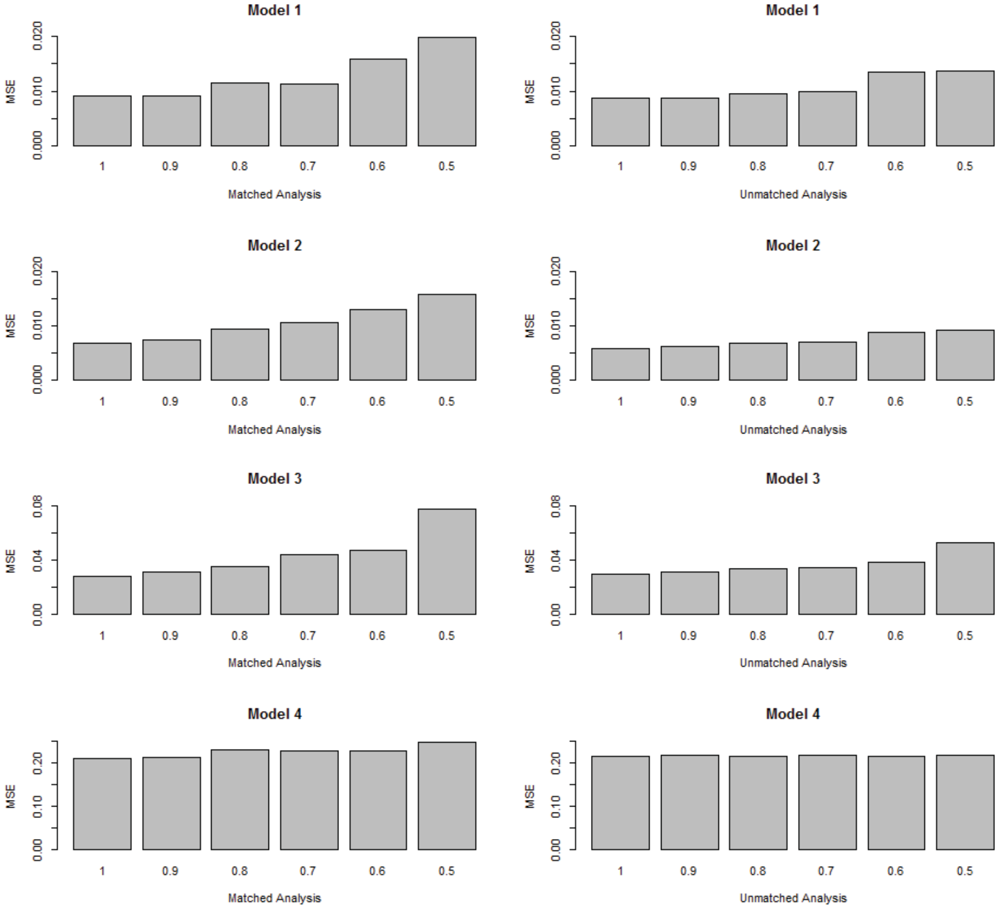


Figure 4.8: Estimated MSE as a function of proportion of nonmissing controls, for both matched and unmatched analyses, based on 100 simulation batches



Chapter 5

Conclusions and Further Work

In this thesis, we conducted a simulation study to investigate the performance of matched and unmatched methods of analysis for estimation of vaccine efficacy with a case negative case control design. A number of issues were addessed, in particular, the influence of unmatched controls, which reduces power, and the consequences of breaking the match, whereby unmatched cases can be included with unmatched inference methods.

In the case that there are no variables measured in addition to vaccine status and case/control status, we showed that a hypothesis test based on the partial likelihood for Cox's proportional hazards model is equivalent to McNemar's test, which is the usual approach to testing for paired binary observations.

We then carried out a simulation study comparing unconditional and conditional likelihood based methods. Data were simulated from separate non-homogeneous Poisson processes for cases and controls, including additional predictor variables in some simulations. The results of the simulation suggest that there is little reduction in power or increase in bias or MSE in moving from a matched to an unmatched method of analysis, even when data were generated in a matched fashion, and that there is a potential gain in power for unmatched procedures when the proportion of unmatched cases is moderately large.

In the simulations carried out, additional predictor variables were not confounded with vaccination status in determining the outcome, case vs control. Further research will address the performance of the methods when one or more additional predictor variables are confounded with vaccine status such that they change the probability of case/control status.

In addition, futher work will be carried out to modify the algorithm used for data simulation. In the present thesis, matched data were generated by simulating from one Poisson process, and then randomly selecting one observation from each contiguous pair as a case, with the other set as the time matched control. A better approach would be to simulate from independent processes for vaccinated cases, vaccinated controls, unvaccinated cases and unvaccinated controls, and then process these data to select matched pairs.

Bibliography

[1] N. E. BRESLOW and N. E. DAY. *Statistical Methods in Cancer research.* International Agency Fro Research on Cancer Lyon, 1980.

[2] D.R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B*, 34:187–220, 1972.

[3] F. Downton. Discussion onn professor cox's paper. *Journal of the Royal Statistical Society. Series B*, 34:202–205, 1972.

[4] McNeil S. et al. *Effectiveness of 2011/12 seasonal influenza vaccines in the prevention of influenza-related hospitalization in Canadian adults: A PCIRN Serious Outcomes Surveillance (SOS) Network Study.* Presented at Options for the Control of Influenza VIII, 2013.

[5] McNeil S. et al. *Effectiveness of 2012/13 seasonal influenza vaccines in the prevention of influenza-related hospitalization in Canadian adults: A Public Health Agency of Canada/ Canadian Institutes of Health Research (PCIRN)Serious Outcomes Surveillance Network Study.* Presented at Options for the Control of Influenza VIII, 2013.

[6] Peter Guttorp. *Stochastic Modeling of Scientific Data.* Chapman and Hall, 1995.

[7] M. Elizabeth Halloran and Michael G Hudgens. Causal inference for vaccine effects on infectiousness. *The international journal of biostatistics*, 8(2), 2012.

[8] Nicholas P. Jewell. *Statistics for Epidemiology.* Chapman and Hall/CRC, 2004.

[9] J.F.Lawless. *Regression Method for Poisson Process Data.* Taylor and Francis, 2015.

[10] P.A.W. Lewis and G.S. Shedler. Simulation of nonhomogeneous poisson process with log linear rate function. *Biometrika*, 63:501–505, 1976.

[11] Jr.Claudio J.Struchiner M. Elizabeth Halloran, Ira M.Longini. *Design and Analysis of Vaccine Studies.* Springer, 2010.

[12] Marie-Pierre Preziosi M. Elizabeth Halloran and Haitao Chu. Estimating vaccine efficacy form secondary attack rates. *American Statistical Association*, 98:38–46, 2003.

[13] Jin Yue and Bruce Smith. *A Note on the Relationship between Point Process and Binary Time Series.* Research India Publications, 2011.