# Bayesian Inference for Two-Phase Studies with Categorical Covariates

Michelle Ross and Jon Wakefield,

Department of Biostatistics, University of Washington, Seattle, WA 98195.

Working Paper no. 123
Center for Statistics and the Social Sciences
University of Washington

Abstract

In this paper, we consider two-phase sampling in the situation in which all covariates are categorical. Two-phase designs are appealing from an efficiency perspective since, if carefully implemented, they allow sampling to be concentrated in informative cells. A number of likelihood-based methods have been developed for the analysis of two-phase data, but we describe a Bayesian approach which has previously been unavailable. The methods are first compared with existing approaches via a simulation study, and are then applied to data collected on Wilms tumour. The benefits of a Bayesian approach include relaxation of the reliance on asymptotic inference, particularly in sparse data situations, and the potential to model data with complex dependencies, for example, via the introduction of random effects. The sparse data situation is illustrated via a simulated example.

Keywords: Contingency tables; Efficiency; Markov chain Monte Carlo; Outcome-dependent sampling.

# 1 Introduction

Two-phase studies have a long history beginning with Neyman (1938) who described estimation methods for what he termed double sampling. In this paper, we adopt the notation of epidemiology, though this distinction is for descriptive purposes only and the methods proposed will be relevant to a two-phase design in any discipline. Hence, the binary outcome $y$ will represent disease status. At phase I, a sample is taken from a population either via simple random or case-control sampling. These phase I data are then cross-classified with respect to the binary outcome variable and to strata of *confounder* variables that we label $z$. At phase II, individuals are sampled within the cells of the cross-classified data, with additional data collected on *exposure* variables $x$. Clearly such a design requires specialized methods of analysis to acknowledge the non-random (outcome-dependent) sampling scheme. The benefit of the two-phase design is that large efficiency gains are possible by judicial choice of the phase I confounder variables and the phase II sample sizes. Table 1 defines the notation that we adopt with $N_{yxz}$ being the number of individuals in the population with disease outcome $y$, exposure level $x$ and confounder level $z$ ($y = 0, 1; x = 1, ..., m_x; z = 1, ..., m_z$). In measurement error situations, what we call confounder variables may correspond to surrogate exposure variables (Breslow and Chatterjee, 1999).

Table 2 displays data, described by Breslow and Chatterjee (1999), that we will use to demonstrate the Bayesian methodology that we develop in Section 2. The population here consists of 4,088 children diagnosed with Wilms tumour, a rare cancer of the kidney, enrolled in two clinical trials. The response is treatment outcome (non-relapsed/relapsed). The full data are in Table 2, and analysis of these complete data provide a benchmark with which analyses based on subsets of data may be compared. Following Breslow and Chatterjee (1999), we take the phase I data as the bottom line of Table 2, so that the confounder is the binary (favourable/unfavourable) institutional histologic (IH) diagnosis which is measured by the pathologist on duty at the time of treatment. This variable is available on all children and is a surrogate for true clinical histology (denoted central histology). At phase II, we assume that additional data (cancer stage and central histology) are collected on 316 non-cases and 415 cases with a favourable IH, and 255 non-cases and 156 cases with an unfavourable IH. The phase II data are listed in the caption of Table 2 and reveal that we have only sampled approximately 10% of the largest favourable IH non-case group. The aim is to estimate the association between relapse and the two covariates cancer stage and central histology. We return to these data in Section 4.

The development of two-phase studies within the context of epidemiology began with the papers of Walker (1982) and White (1982). Subsequently, the methodology has become more advanced. A number of approaches have their origin in survey sampling. In this framework, weighted design-based estimators are frequently used. In a two-phase study, the reciprocal weights for each individual are a product of the phase I and phase II inclusion probabilities. The phase II weights are computationally complex to calculate because they depend on the data sampled at phase I (and are not just a simple product of phase I and phase II sampling probabilities). This is because, with respect to Table 1, the $N_{y.z}$ are random and therefore the phase II $n_{y.z}$ are random also. Hence, to calculate the weights in phase II, we would need

to average over all possible phase I samples. To avoid this computationally expensive step, a number of simplifications have been proposed. The weighted likelihood approach (Flanders and Greenland, 1991) fits a weighted logistic regression model to the phase two data, with sampling weights $N_{y.z}/n_{y.z}$ applied to observations $n_{yxz}$ at phase II (Table 1), thereby adjusting for the outcome-dependent sampling scheme. The result is a Horvitz-Thompson estimator. A more general estimating equation technique was introduced by Reilly and Pepe (1995) but results in the same estimating equation as Flanders and Greenland (1991) in the two-phase setting. More sophisticated survey sampling techniques have been suggested and are described in detail in Chapter 9 of Särndal et al. (1992). These approaches assign the weights in a more careful fashion and are implemented in the `survey` package within `R`, an early version of which was described in Lumley (2004). Further discussion, including a description of maximum likelihood estimation in a survey sampling context, is provided by Whittemore (1997), providing results on estimating functions that are special cases of those contained in Breckling et al. (1994). The latter make connections with the missing data literature, as do Lawless et al. (1999); in the two-phase design, the missingness mechanism depends on the outcome. Many variants of likelihood, other than weighted likelihood, have been proposed in the two–phase context with pseudo-likelihood discussed by Breslow and Cain (1988), Scott and Wild (1991) and Schill et al. (1993) and nonparametric maximum likelihood (NPML) by Breslow and Holubkov (1997a) and Breslow and Holubkov (1997b). A full likelihood approach, which we describe in Section 2, would specify a joint probability model for $(y, x, z)$ and requires estimation of parameters in the model for $(x, z)$. Leaving the $(x, z)$ margin unspecified leads to a problem in semiparametric inference that was solved for random sampling at phase I by Scott and Wild (1991, 1997) and for case-control sampling by Breslow and Holubkov (1997a). In fact, the latter demonstrated that asymptotic inference is identical under the two sampling schemes. The NPML estimate places mass on the observed values of $x$ and does not require a distributional form to be assumed for $x$.

In this paper, we describe a fully Bayesian approach. Beyond pure academic interest, a Bayesian approach offers the possibility of improved small sample properties, since the asymptotic arguments of likelihood-based approaches are not required, and the possibility of modeling complex dependencies in the data through the introduction of random effects. With respect to small-sample inference, the NPML method critically depends on the sampling fractions being bounded away from zero. The method therefore breaks down completely with any zero count in the phase II data, due to infinite or indeterminate values for some of the sampling fractions (see Breslow and Chatterjee, 1999, p. 466). An example of this phenomenon is presented in Section 3. This section also contains a simulation study.

# 2 Bayesian Model

In this section we develop the Bayesian model and begin by giving a description of the sampling scheme.

Table 1: Two-phase study design: the observed data are $N_{y\cdot z}$ at phase I, and $n_{y\cdot z}$ and $n_{yxz}$ at phase II; the $N_{yxz}$ entries are unobserved in a two-phase study

Phase I Data

| | $Z=1$ | | $Z=2$ | | ... | $Z=m_z$ | |
|---|---|---|---|---|---|---|---|
| | $Y=0$ | $Y=1$ | $Y=0$ | $Y=1$ | ... | $Y=0$ | $Y=1$ |
| $X=1$ | $N_{011}$ | $N_{111}$ | $N_{012}$ | $N_{112}$ | ... | $N_{01m_z}$ | $N_{11m_z}$ |
| $X=2$ | $N_{021}$ | $N_{121}$ | $N_{022}$ | $N_{122}$ | ... | $N_{02m_z}$ | $N_{12m_z}$ |
| ... | ... | ... | ... | ... | ... | ... | ... |
| $X=m_x$ | $N_{0m_x1}$ | $N_{1m_x1}$ | $N_{0m_x2}$ | $N_{1m_x2}$ | ... | $N_{0m_xm_z}$ | $N_{1m_xm_z}$ |
| | $N_{0\cdot1}$ | $N_{1\cdot1}$ | $N_{0\cdot2}$ | $N_{1\cdot2}$ | ... | $N_{0\cdot m_z}$ | $N_{1\cdot m_z}$ |

Phase II Data

| | $Z=1$ | | $Z=2$ | | ... | $Z=m_z$ | |
|---|---|---|---|---|---|---|---|
| | $Y=0$ | $Y=1$ | $Y=0$ | $Y=1$ | ... | $Y=0$ | $Y=1$ |
| $X=1$ | $n_{011}$ | $n_{111}$ | $n_{012}$ | $n_{112}$ | ... | $n_{01m_z}$ | $n_{11m_z}$ |
| $X=2$ | $n_{021}$ | $n_{121}$ | $n_{022}$ | $n_{122}$ | ... | $n_{02m_z}$ | $n_{12m_z}$ |
| ... | ... | ... | ... | ... | ... | ... | ... |
| $X=m_x$ | $n_{0m_x1}$ | $n_{1m_x0}$ | $n_{0m_x1}$ | $n_{1m_x1}$ | ... | $n_{0m_xm_z}$ | $n_{1m_xm_z}$ |
| | $n_{0\cdot1}$ | $n_{1\cdot1}$ | $n_{0\cdot2}$ | $n_{1\cdot2}$ | ... | $n_{0\cdot m_z}$ | $n_{1\cdot m_z}$ |

Table 2: Number of Wilms tumour non-cases and cases by institutional histology (IH), central histology (CH) and stage of disease: full data. The two-phase design we analyze takes all of the unfavourable IH non-cases and cases and all of the favourable IH cases, but takes a sample of 316 of the available 3262 favourable IH non-cases. This sample produced the vector of responses (which replace the first numeric column of the table): 115, 86, 79, 27, 2, 3, 3, 1.

| | | Favourable IH | | Unfavourable IH | |
|---|---|---|---|---|---|
| | | Non-cases | Cases | Non-cases | Cases |
| | Stage I | 1363 | 91 | 15 | 1 |
| Favourable | Stage II | 816 | 117 | 11 | 2 |
| CH | Stage III | 693 | 100 | 18 | 3 |
| | Stage IV | 307 | 60 | 23 | 3 |
| | Stage I | 37 | 9 | 64 | 16 |
| Unfavourable | Stage II | 25 | 14 | 51 | 33 |
| CH | Stage III | 14 | 15 | 60 | 57 |
| | Stage IV | 7 | 9 | 13 | 41 |
| Total | | 3262 | 415 | 255 | 156 |

## 2.1 The Likelihood

The data at phase I are commonly available through one of two mechanisms, simple random or case-control sampling. To cover both situations we specify a joint log-linear model for the binary random variable $y$ and the univariate discrete random variables $x$ and $z$.

Consider $\mu_{yxz} = E[N_{yxz} \mid \boldsymbol{\lambda}]$, the mean of the disease-exposure-confounder count in cell $(y, x, z)$ of the $2 \times m_x \times m_z$ contingency table, $y = 0, 1; x = 1, ..., m_x; z = 1, ..., m_z$. We begin by specifying a saturated model but this is not always desirable, a point to which we return in Section 5. The saturated log-linear model is:

$$\log(\mu_{yxz}) = \mu + \lambda_y^Y + \lambda_x^X + \lambda_z^Z + \lambda_{yx}^{YX} + \lambda_{yz}^{YZ} + \lambda_{xz}^{XZ} + \lambda_{yxz}^{YXZ} \tag{1}$$

with identifiability gained through the sum-to-zero constraints

$$\begin{aligned}
\sum_y \lambda_y^Y = \sum_x \lambda_x^X = \sum_z \lambda_z^Z &= 0, \\
\sum_y \lambda_{yx}^{YX} = \sum_x \lambda_{yx}^{YX} &= 0, \\
\sum_y \lambda_{yz}^{YZ} = \sum_z \lambda_{yz}^{YZ} &= 0, \\
\sum_x \lambda_{xz}^{XZ} = \sum_z \lambda_{xz}^{XZ} &= 0, \\
\sum_y \lambda_{yxz}^{YXZ} = \sum_x \lambda_{yxz}^{YXZ} = \sum_z \lambda_{yxz}^{YXZ} &= 0.
\end{aligned}$$

We first consider the distribution of the phase I data. Under simple random sampling at phase I, the joint probabilities are given by $r_{yxz} = \text{pr}(Y = y, X = x, Z = z \mid \boldsymbol{\lambda})$, while for case-control sampling we have $r_{xz|y} = \text{pr}(X = x, Z = z \mid Y = y, \boldsymbol{\lambda})$. Under simple random sampling, the intercept $\mu$ cancels when the probabilities

$$\begin{aligned}
r_{yxz} &= \frac{\mu_{yxz}}{\sum_u \sum_v \sum_w \mu_{uvw}} \\
&= \frac{\exp(\lambda_y^Y + \lambda_x^X + \lambda_z^Z + \lambda_{yx}^{YX} + \lambda_{yz}^{YZ} + \lambda_{xz}^{XZ} + \lambda_{yxz}^{YXZ})}{\sum_u \sum_v \sum_w \exp(\lambda_u^Y + \lambda_v^X + \lambda_w^Z + \lambda_{uv}^{YX} + \lambda_{uw}^{YZ} + \lambda_{vw}^{XZ} + \lambda_{uvw}^{YXZ})}
\end{aligned} \tag{2}$$

are calculated. Moreover, under case-control sampling, the intercept $\mu$ and main effect term for $Y$, $\lambda_y^Y$, cancel when the probabilities

$$\begin{aligned}
r_{xz|y} &= \frac{\mu_{yxz}}{\sum_v \sum_w \mu_{yvw}} \\
&= \frac{\exp(\lambda_x^X + \lambda_z^Z + \lambda_{yx}^{YX} + \lambda_{yz}^{YZ} + \lambda_{xz}^{XZ} + \lambda_{yxz}^{YXZ})}{\sum_v \sum_w \exp(\lambda_v^X + \lambda_w^Z + \lambda_{yv}^{YX} + \lambda_{yw}^{YZ} + \lambda_{vw}^{XZ} + \lambda_{yvw}^{YXZ})}
\end{aligned}$$

are calculated.

The observed data consist of three vectors of counts: the response-confounder margin observed at phase I, $\boldsymbol{N}^{y\cdot z} = \{N_{y\cdot z}, y = 0, 1; z = 1, ..., m_z\}$, the phase II sample sizes

$\boldsymbol{n}^{y\cdot z} = \{n_{y\cdot z}, y = 0, 1; z = 1, ..., m_z\}$ and the phase II outcomes $\boldsymbol{n}^{yxz} = \{n_{yxz}, y = 0, 1; x = 1, ..., m_x; z = 1, ..., m_z\}$. Throughout the paper, variables super-scripted with $x, y, z$ will represent vectors. The phase II sample sizes are determined by the investigator, though $n_{y\cdot z} \leq N_{y\cdot z}$ and the latter are random so that technically the $n_{y\cdot z}$ are random variables also (Schill et al., 1993). However, since this term does not depend on $\boldsymbol{\lambda}$, it need not be considered. Under simple random sampling at phase I, we condition on the grand total $N_{\cdot\cdot\cdot}$ only and the likelihood is

$$\begin{aligned} \text{pr}(\boldsymbol{N}^{y\cdot z}, \boldsymbol{n}^{y\cdot z}, \boldsymbol{n}^{yxz} | N_{\cdot\cdot\cdot}, \boldsymbol{\lambda}) &\propto \prod_{y=0}^{1} \prod_{z=1}^{m_z} \text{pr}(Y = y, Z = z)^{N_{y\cdot z}} \\ &\quad \times \prod_{y=0}^{1} \prod_{z=1}^{m_z} \prod_{x=1}^{m_x} \text{pr}(X = x | Z = z, Y = y)^{n_{yxz}} \\ &= \text{pr}(\boldsymbol{N}^{y\cdot z} | N_{\cdot\cdot\cdot}, \boldsymbol{\lambda}) \times \text{pr}(\boldsymbol{n}^{yxz} | \boldsymbol{n}^{y\cdot z}, \boldsymbol{\lambda}), \end{aligned} \tag{3}$$

using analogous arguments from proposition 1 of Holubkov (1995). The first term in (3) is

$$\boldsymbol{N}^{y\cdot z} | N_{\cdot\cdot\cdot}, \boldsymbol{\lambda} \sim \text{Multinomial}\left(N_{\cdot\cdot\cdot}, \{\text{pr}(Y = y, Z = z)\}_{y=0,1;z=1,...,m_z}\right),$$

where $\text{pr}(Y = y, Z = z) = \sum_{x=1}^{m_x} r_{yxz}$. The second term of (3) is

$$\boldsymbol{n}^{yxz} | \boldsymbol{n}^{y\cdot z}, \boldsymbol{\lambda} \sim \prod_{y=0}^{1} \prod_{z=1}^{m_z} \text{Multinomial}\left(n_{y\cdot z}, \{\text{pr}(X = x | Z = z, Y = y)\}_{x=1,...,m_x}\right),$$

where $\text{pr}(X = x | Z = z, Y = y) = \frac{r_{yxz}}{\sum_{v=1}^{m_x} r_{yvz}}$.

Under case-control sampling at phase I we condition on $N_{0\cdot\cdot}$ and $N_{1\cdot\cdot}$ and the likelihood is

$$\begin{aligned} \text{pr}(\boldsymbol{N}^{y\cdot z}, \boldsymbol{n}^{y\cdot z}, \boldsymbol{n}^{yxz} | N_{0\cdot\cdot}, N_{1\cdot\cdot}, \boldsymbol{\lambda}) &\propto \prod_{y=0}^{1} \prod_{z=1}^{m_z} \text{pr}(Z = z | Y = y)^{N_{y\cdot z}} \\ &\quad \times \prod_{y=0}^{1} \prod_{z=1}^{m_z} \prod_{x=1}^{m_x} \text{pr}(X = x | Z = z, Y = y)^{n_{yxz}} \\ &= \text{pr}(\boldsymbol{N}^{y\cdot z} | N_{1\cdot\cdot}, N_{0\cdot\cdot}, \boldsymbol{\lambda}) \times \text{pr}(\boldsymbol{n}^{yxz} | \boldsymbol{n}^{y\cdot z}, \boldsymbol{\lambda}), \end{aligned} \tag{4}$$

using proposition 1 of Holubkov (1995). The first term of (4) is

$$\boldsymbol{N}^{y\cdot z} | N_{1\cdot\cdot}, N_{0\cdot\cdot}, \boldsymbol{\lambda} \sim \prod_{y=0}^{1} \text{Multinomial}\left(N_{y\cdot\cdot}, \{\text{pr}(Z = z | Y = y)\}_{z=1,...,m_z}\right),$$

where $\text{pr}(Z = z | Y = y) = \sum_{x=1}^{m_x} r_{xz|y}$. The second term of (4) is

$$\boldsymbol{n}^{yxz} | \boldsymbol{n}^{y\cdot z}, \boldsymbol{\lambda} \sim \prod_{y=0}^{1} \prod_{z=1}^{m_z} \text{Multinomial}\left(n_{y\cdot z}, \{\text{pr}(X = x | Z = z, Y = y)\}_{x=1,...,m_x}\right),$$

where $\text{pr}(X = x | Z = z, Y = y) = \frac{r_{xz|y}}{\sum_{v=1}^{m_x} r_{vz|y}}$. In some studies, the phase I sample may be the result of a matched case-control study; the above derivation follows through in this situation with the obvious conditioning on the case and control sample sizes within each matching set.

In many situations, we are interested in the interpretations of the coefficients from fitting the linear logistic disease model

$$\log\left[\frac{\mathrm{pr}(Y=1 \mid X=x, Z=z)}{\mathrm{pr}(Y=0 \mid X=x, Z=z)}\right] = \boldsymbol{W}\boldsymbol{\beta}. \tag{5}$$

Depending on the context, the design matrix $\boldsymbol{W}$ may contain terms for both main effects and interactions. Note that if interactions are not desired then $\lambda^{YXZ}_{yxz}$ is not required in the log-linear model. There is a direct correspondence between $\boldsymbol{\lambda}$ and $\boldsymbol{\beta}$:

$$\begin{aligned}
\beta_0 &= -2(\lambda^Y_0 + \lambda^{YX}_{01} + \lambda^{YZ}_{01} + \lambda^{YXZ}_{011}) \\
\beta_x &= -2(\lambda^{YX}_{0x} - \lambda^{YX}_{01}) - 2(\lambda^{YXZ}_{0x1} - \lambda^{YXZ}_{011}) \\
\beta_z &= -2(\lambda^{YZ}_{0z} - \lambda^{YZ}_{01}) - 2(\lambda^{YXZ}_{01z} - \lambda^{YXZ}_{011}) \\
\beta_{xz} &= -2(\lambda^{YXZ}_{0xz} - \lambda^{YXZ}_{0x1} - \lambda^{YXZ}_{01z} + \lambda^{YXZ}_{011}),
\end{aligned}$$

for $x = 2, \ldots, m_x$ and $z = 2, \ldots, m_z$. We perform computation for the $\boldsymbol{\lambda}$ collection and then transform to the $\boldsymbol{\beta}$ set using the above relationships. Note that $\lambda^X_x$, $\lambda^Z_z$ and $\lambda^{XZ}_{xz}$ describe the relationship between $x$ and $z$ in the $y = 0$ group. In the phase I case-control situation, we can no longer estimate the intercept $\beta_0$ since $\lambda^Y_y$ cancels in the calculation of $r_{xz|y}$. In the frequentist development, an intercept is present but is not interpretable unless information is available on the sampling frequencies of the cases and controls.

## 2.2 Prior Specification

Under a log-linear specification, a convenient choice of prior for $\boldsymbol{\lambda}$ is the Dirichlet, but this choice is restrictive since there are insufficient free parameters. Specifically, there is a single parameter only to control the precision for all elements. We suppose there is simple random sampling at phase I; the case-control alternative follows in an obvious fashion. With respect to (1), let $\boldsymbol{\lambda}^Y = \lambda^Y_0$, $\boldsymbol{\lambda}^X = (\lambda^X_1, ..., \lambda^X_{m_x-1})$, $\boldsymbol{\lambda}^Z = (\lambda^Z_1, ..., \lambda^Z_{m_z-1})$, $\boldsymbol{\lambda}^{YX} = (\lambda^{YX}_{01}, ..., \lambda^{YX}_{0,m_x-1})$, $\boldsymbol{\lambda}^{YZ} = (\lambda^{YZ}_{01}, ..., \lambda^{YZ}_{0,m_z-1})$, $\boldsymbol{\lambda}^{XZ} = (\lambda^{XZ}_{11}, ..., \lambda^{XZ}_{m_x-1,m_z-1})$ and $\boldsymbol{\lambda}^{YXZ} = (\lambda^{YXZ}_{011}, ..., \lambda^{YXZ}_{0,m_x-1,m_z-1})$ represent the reduced sets of parameters for which we specify prior distributions. The prior we adopt assumes

$$\begin{aligned}
\pi(\boldsymbol{\lambda}^Y, \boldsymbol{\lambda}^X, \boldsymbol{\lambda}^Z, \boldsymbol{\lambda}^{YX}, \boldsymbol{\lambda}^{YZ}, \boldsymbol{\lambda}^{XZ}, \boldsymbol{\lambda}^{YXZ}) = \\
\pi(\boldsymbol{\lambda}^Y)\pi(\boldsymbol{\lambda}^X)\pi(\boldsymbol{\lambda}^Z)\pi(\boldsymbol{\lambda}^{YX})\pi(\boldsymbol{\lambda}^{YZ})\pi(\boldsymbol{\lambda}^{XZ})\pi(\boldsymbol{\lambda}^{YXZ}).
\end{aligned}$$

Since $\boldsymbol{\beta}$ depends on $\boldsymbol{\lambda}^Y$, $\boldsymbol{\lambda}^{YX}$, $\boldsymbol{\lambda}^{YZ}$ and $\boldsymbol{\lambda}^{YXZ}$ these parameters have special status as parameters of interest. Hence, we place independent multivariate normal priors on these parameters, with means and variances chosen based on the context. For example, if we wish to have inference based on the data alone, we may assign zero means and large variances for these sets of parameters. In other situations, for example when we have sparse data, more informative choices will be desirable. We propose a slightly different approach for $\boldsymbol{\lambda}^X$, $\boldsymbol{\lambda}^Z$ and $\boldsymbol{\lambda}^{XZ}$ which will usually be nuisance parameters. For these parameters we use the specification proposed by Knuiman and Speed (1988), and refined by Dellaportas and Forster (1999), in

which independent multivariate normal priors are placed on the collections of main effects and interactions. For $a \in \{X, Z, XZ\}$ let $\boldsymbol{\lambda}^a \sim N(\mathbf{0}_{d_a}, \alpha_a^2 \boldsymbol{V}_a)$ where $d_a = |\boldsymbol{\lambda}_a|$, $\mathbf{0}_{d_a}$ is a vector of zeros of length $d_a$, and

$$\boldsymbol{V}_a = \frac{1}{m_x m_z} \prod_{\gamma \in a} |I_\gamma| \otimes \left( \boldsymbol{I}_{|I_\gamma|-1} - \frac{1}{|I_\gamma|} \boldsymbol{J}_{|I_\gamma|-1} \right) \tag{6}$$

where $I_\gamma$ is the set of levels of factor $\gamma$, $\boldsymbol{I}_{|I_\gamma|}$ is the $|I_\gamma| \times |I_\gamma|$ identity matrix and $\boldsymbol{J}_{|I_\gamma|}$ is the $|I_\gamma| \times |I_\gamma|$ matrix containing all ones. Dellaportas and Forster (1999) discuss the choice of $\alpha_a^2$, with an invariance argument suggesting $\alpha_a^2 = k \times m_x m_z$. We adopt a value of $k = 1$, corresponding closely to a unit information prior (Kass and Wasserman, 1995) see Section 3.4 of Dellaportas and Forster (1999).

## 2.3 Computation

The likelihoods given by (3) and (4) are complicated functions of $\boldsymbol{\lambda}$ with no standard forms. As detailed in Web Appendix A, we have experimented with an auxiliary variable scheme in which the full data $N_{yxz}$ are introduced into this model. This leads to simplified conditional forms within a Markov chain Monte Carlo scheme, but at the expense of requiring the simulation of additional variables. As an alternative, we may use a Metropolis-Hastings algorithm for $\boldsymbol{\lambda}$ only. For the examples of the paper, the $\boldsymbol{\lambda}$ only schemed proved more efficient and so that is the method we describe here. To obtain a Markov chain with good mixing properties, we sample the complete $\boldsymbol{\lambda}$ vector via a single Metropolis-Hastings step via a multivariate random walk based on a normal proposal.

Sampling the *complete* $\boldsymbol{\lambda}$ vector via a multivariate normal proposal requires a covariance matrix that mimics well the dependence within the posterior distribution. Again we suppose we have simple random sampling at phase I and a saturated log-linear model. Case-control sampling and/or non-saturated models follow in an obvious fashion. The phase II data $\boldsymbol{n}_{yxz}$ provide information on the $y \times x \times z$ table, but these data are biased due to the phase I sampling. However, we can correct for the bias by creating expected counts $N^\star_{yxz} = E[N_{yxz}] = N_{\cdots} \times \widehat{\text{pr}}(Y = y, X = x, Z = z)$, with

$$\begin{aligned} \widehat{\text{pr}}(Y = y, X = x, Z = z) &= \widehat{\text{pr}}(X = x \mid Y = y, Z = z) \times \widehat{\text{pr}}(Y = y, Z = z) \\ &= \frac{n_{yxz}}{n_{y\cdot z}} \times \frac{N_{y\cdot z}}{N_{\cdots}} \end{aligned}$$

to give $N^\star_{yxz} = n_{yxz} N_{y\cdot z} / n_{y\cdot z}$. A saturated log-linear model is fitted to the $N^\star_{yxz}$ data using maximum likelihood estimation, and the variance-covariance matrix of the proposal for $\boldsymbol{\lambda}$ is taken as a constant $c$ times the asymptotic variance-covariance of the maximum likelihood estimate. The constant is chosen to achieve acceptance rates of approximately 25–30% (Roberts et al., 1997). The above scheme captures the dependence in the posterior and has proved to be well-behaved Markov chains in the examples we have worked on. Evidence for this is presented in Web Appendices C–F in which we report trace plots for the two examples of the paper (as described in Sections 3 and 4) and two additional examples, one involving case-control sampling at phase I and another a table of large $(2 \times 2 \times 2 \times 3 \times 7)$ dimension.

# 3 Simulated Data

## 3.1 A Simulation Study

In this section, we compare our proposed method with NPML using simulated data. There are clearly many possible scenarios to investigate and we choose to consider binary $x$ and $z$, with a fixed phase I sample size. Specifically, the sample sizes are 4039 ($Z = 0, Y = 0$), 477 ($Z = 0, Y = 1$), 306 ($Z = 1, Y = 0$), 178 ($Z = 1, Y = 1$). Web Appendix B contains the full data from which the phase II data are sampled. We report three simulations, with varying phase II sample sizes, denoted by "small", "medium" and "large". For the "small" scenario, we fix the phase II sample sizes as 125 each of favorable and unfavorable IH non-cases, and 50 each of favorable and unfavorable IH cases for a total of 350 individuals. For the "medium" scenario, we fix the phase II sample sizes as 250 each of favorable and unfavorable IH non-cases, and 100 each of favorable and unfavorable IH cases totaling 700 individuals, whereas for the "large" scenario, we fix the phase II sample sizes as 500 favorable IH non-cases, 200 favorable IH cases, and all unfavorable IH individuals for a total of $1,184$ individuals.

In each simulation, individuals were randomly sampled from the population to obtain the phase II data, where a total of 500 data sets were generated. For each sampled data set, we analysed the data using WL, PL, NPML, as well as the Bayesian two-phase method. In the event of a zero count in the phase II data, due to the unavailability of NPML in such cases, only WL, PL, and the Bayesian method were fit to the data.

Two different sets of prior distributions were used for $\boldsymbol{\beta}$. The first used a normal prior distribution with a large variance for $\beta_0$, and more informative but realistic priors for $\beta_x, \beta_z$ and $\beta_{xz}$. Specifically, we assigned independent normal prior distributions with variance $\log(5)/1.96$, which corresponds to a 95% prior interval of $(-5, 5)$ for each of the log odds ratios. The second set of priors were normal distributions with large variance for each of $\beta_0, \beta_x, \beta_z$ and $\beta_{xz}$ ("flat priors"). We ran the chains for a total of 450,000 iterations, where the first 50,000 were used for tuning.

The bias, variance and mean squared error (MSE) were calculated for all five of the analyses across the 500 simulated data sets (tables of results shown in Web Appendix B). Figure 1 compares the results from NPML and the two Bayesian two-phase analyses using small, medium and large phase II sample sizes. While the Bayesian analyses produced results which were slightly more biased than those of NPML, the variances were much smaller, resulting in smaller MSEs. As the phase II sample sizes got larger, both the bias and the variance decreased, as expected, with the performance of NPML and the Bayesian method being nearly identical in the case of large phase II sample sizes. In the case of small phase II sample sizes, 31 phase II data sets were generated which contained a zero count. When these data sets were included in the bias, variance and MSE calculations for the Bayesian method, the Bayesian and NPML methods performed comparably. However, when these data sets were excluded, the Bayesian method generally had smaller bias and variance, and hence smaller MSEs, than the NPML method. Hence, the Bayesian method generally resulted in estimates with smaller MSE than NPML, and as we will see in Section 3.2, Bayesian

Table 3: Number of Wilms tumour cases and non-cases by institutional histology (IH) and central histology (CH) in the simulated data: full data in the left columns and phase II data in the right columns (indicated in bold type)

| | Favourable IH | | | | Unfavourable IH | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Non-cases | | Cases | | Non-cases | | Cases | | | |
| | Full | Phase II | Full | Phase II | Full | Phase II | Full | Phase II | | |
| Favourable CH | 1171 | **295** | 128 | **128** | 25 | **11** | 3 | **0** | | |
| Unfavourable CH | 24 | **5** | 19 | **19** | 75 | **39** | 55 | **50** | | |
| Total | 1195 | **300** | 147 | **147** | 100 | **50** | 58 | **50** | 1500 | **547** |

inference is reliable in sparse data situations, in particular in the presence of zero cell counts where NPML is unavailable.

## 3.2 A Sparse Data Example

To illustrate the potential advantages of the Bayesian approach, we present an analysis of a single dataset for which the NPML approach fails. We generated the data in the same spirit as the Wilms tumour data set, but for simplicity ignored stage information while retaining central ($x$) and institutional ($z$) histology. The observed proportions of individuals with unfavourable central and institutional histology, as well as the odds ratio characterizing the dependence between $x$ and $z$, were used to simulate the data. The parameters used in the simulation were: $\beta_0 = -2.16, \beta_x = 1.59, \beta_z = 0.15, \beta_{xz} = 0.17$, $\text{pr}(X = 1) = 0.11, \text{pr}(Z = 1) = 0.1$, with an $xz$ odds ratio of 120. The total number of subjects in the phase I population for the simulation is 1500. Table 3 displays the unobserved population data (in the left columns), along with the phase II data (in the right columns). All favourable IH cases and samples of the non-cases and favourable IH cases were used as the phase II sample and we picked a realization of the data in which a zero count resulted in the phase II data. Due to the zero count in the upper right corner of this table, the NPML estimate is not available. We attempted to use the weighted and pseudo likelihood approaches for these data but the presence of the zero produced spurious results (which are reported in Web Appendix C).

A normal prior distribution with a large variance was used for $\beta_0$, while more informative but realistic priors were used for $\beta_x, \beta_z$ and $\beta_{xz}$. Specifically, we assigned independent normal prior distributions with variance $\log(5)/1.96$, which corresponds to a 95% prior interval of $(-5, 5)$ for each of the log odds ratios. We ran the chain for a total of 200,000 iterations, where the first 50,000 were used to tune the Markov chain, with a constant of $c = 0.97$ giving an acceptance rate of approximately 30%. This analysis took 9 minutes to run on a 2 GHz Intel Core Duo processor with 2 GB of 667 MHz DDR2 SDRAM.

We compare the results of the Bayesian two-phase analysis to those with fitting a logistic

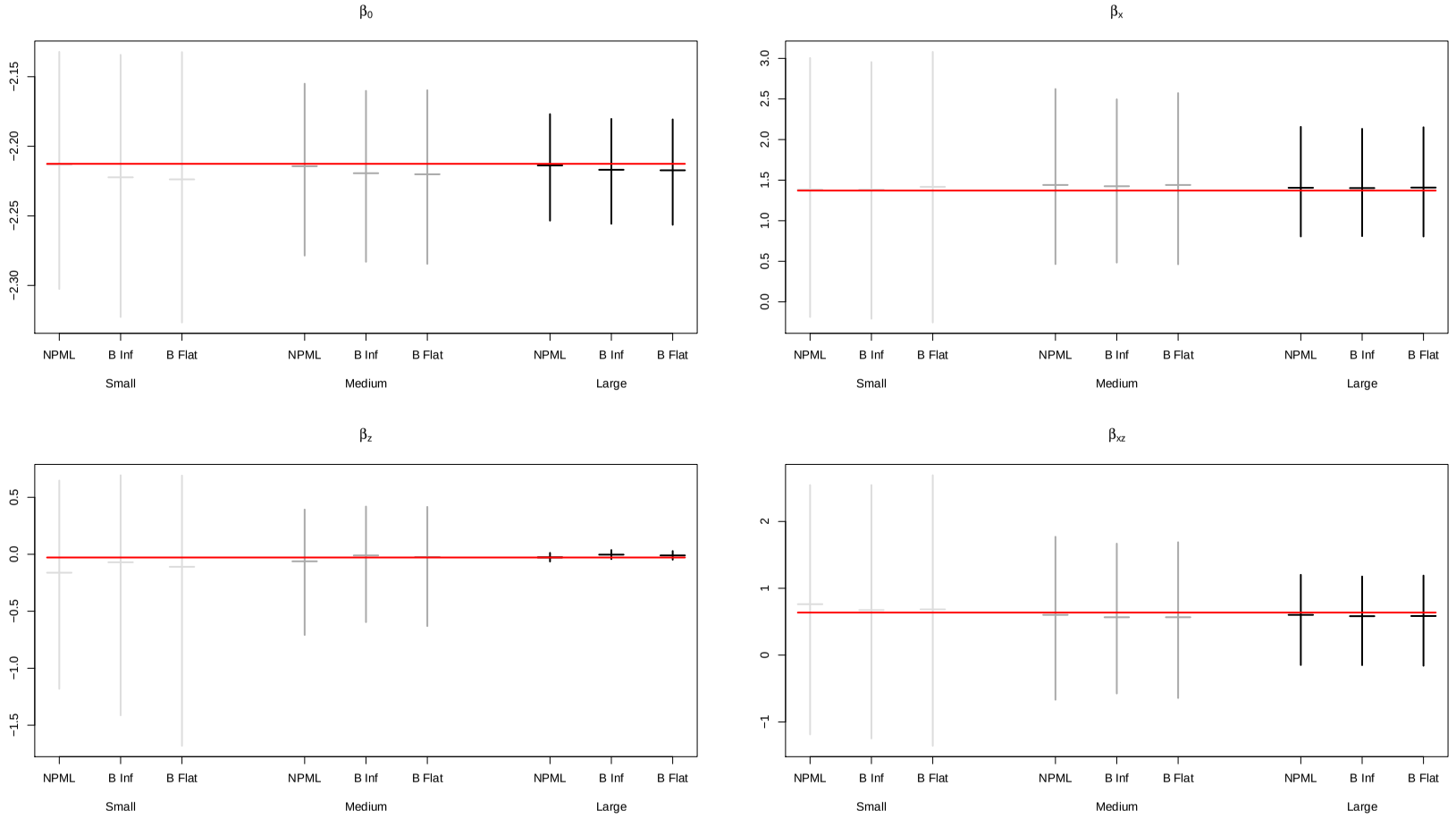


Figure 1: Comparing the results from NPML and two Bayesian two-phase analyses using small, medium and large phase II sample sizes with informative (B Inf) and flat priors (B Flat). The red lines indicate the true value of the parameters.

disease model to the complete data. The posterior means and 95% credible intervals obtained from the Bayesian analysis for the log odds parameters $\beta$ and log linear parameters $\lambda$ are displayed in Figure 2, along with summaries from fitting a logistic model to the full phase I data. The Bayesian credible intervals are wide but contain the true values of each parameter. Hence, Bayesian inference is reliable in this sparse data situation.

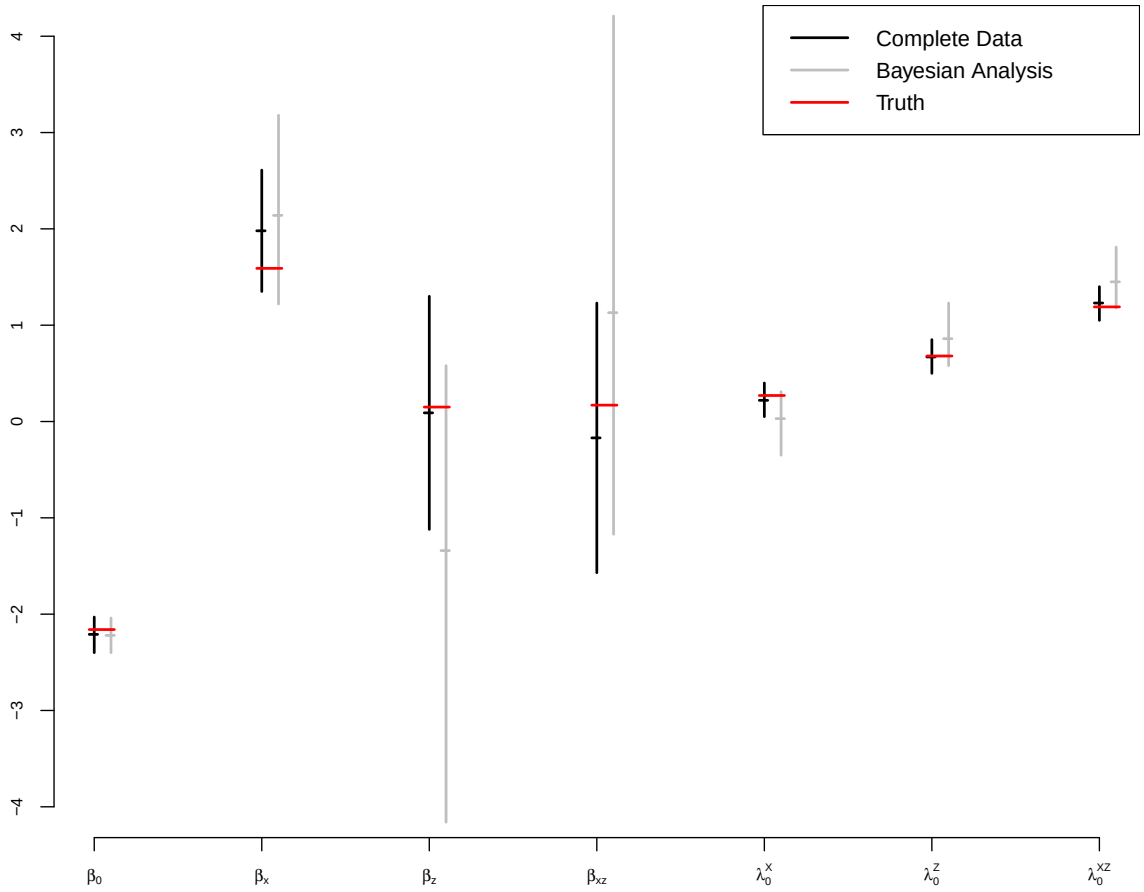


Figure 2: Comparison of log odds ratios $\beta$ and log-linear parameters $\lambda$, for the simulated data with a zero cell count.

# 4 Wilms Tumour Example

We return to the Wilms tumour data introduced in Section 1 in which simple random sampling is assumed at phase I. The phase II data was obtained by selecting all the Wilms

tumour cases, all unfavourable institutional histology cases and approximately 10% of the favourable institutional histology controls. These data are listed in the caption of Table 2. Using a disease model that includes main effects for stage and unfavourable central histology, as well as their interaction, we compare the NPML approach with the Bayesian two-phase approach. We also report results from a survey sampling approach as described by Särndal et al. (1992) and implemented in the `survey` package, for details see Chapter 8 of Lumley (2010). The Bayesian approach used zero mean normal distributions with large variances on each of the $\boldsymbol{\beta}$ parameters and the approach outlined in Section 2.2 for the remaining $\boldsymbol{\lambda}$ parameters. The first 100,000 iterations were used to tune the Markov chain, with a constant of $c = 0.26$ giving an acceptance rate of approximately 30%. We then ran the chain for a further 500,000 iterations (though convergence was achieved at around 200,000 iterations). The 600,000 iterations took 77 minutes to run on a 2 GHz Intel Core Duo processor with 2 GB of 667 MHz DDR2 SDRAM. Web Appendix D contains trace plots of the $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$ parameters.

Results from the complete data given in Table 2 are shown in the second column of Table 4. The parameter estimates and 95% intervals for the NPML and Bayesian approaches are displayed in columns 3 and 4 of Table 4. The NPML and Bayesian two-phase results agree reasonably well, with the Bayesian credible intervals tending to be slightly narrower. Compared to the results using the full data, the NPML and Bayesian two-phase approaches slightly underestimate the main effects, while slightly overestimating the interaction terms. Analysis of the phase II data only (column 2) results in a large loss of precision, as we would expect. The results of a survey sampling weighted estimator are given in column 5 and produce point estimates that are quite different to the NPML/Bayes results. In this example, the NPML and Bayes approach give essentially identical inference.

# 5 Discussion

The Bayesian approach we have outlined requires a far greater level of model specification than the NPML approach, namely an exposure-confounder model. This is a major drawback and in situations in which the data are numerous and there are no dependencies that need to be modeled (via the use of random effects, for example), we would recommend the NPML approach. For the situation in which the covariates are all discrete and the data are sparse, the Bayesian approach is beneficial, however. In situations in which $x$ and/or $z$ take on many values, it may be beneficial to avoid the saturated log-linear model (1), which will contain many parameters, in favour of a simpler form. Web Appendix F illustrates and provides an example in which the dimensionality of the covariate space is much larger than in the Wilms tumour example.

Our approach also only considers the case when the covariates are all discrete. The extension to continuous or mixed continuous and discrete covariates situations is not straightforward, though the approaches of Sinha et al. (2004) and Dunson and Xing (2009) offer possible avenues for addressing this deficiency.

A motivation for the full probability modeling approach described here is that random effects

may be introduced to account for dependencies in the data. For example, Wakefield and Haneuse (2008) describe the use of two phase studies in a spatial epidemiological context, in which one may wish to acknowledge confounding by location using spatial random effects. In this case, a natural approach would be to extend (5) to

$$\log\left[\frac{\text{pr}(Y=1 \mid X=x, Z=z)}{\text{pr}(Y=0 \mid X=x, Z=z)}\right] = \boldsymbol{W}\boldsymbol{\beta} + V_z + U_z$$

where $\boldsymbol{W}$ is the design matrix and we have asssumed that the confounding variable $z$ is area, and that $U_z$ and $V_z$ are random effects with and without spatial structure, respectively. For example, the convolution model of Besag et al. (1991) could be adopted.

# Supplementary Materials

Web Appendices, Tables, and Figures referenced in Sections 5 are available with this paper at the Biometrics website on Wiley Online Library.

# Acknowledgement

This research was supported by grants R01 CA095994 and R01 CA125081 from the National Institutes of Health. The authors would like to thank the editor, associate editor and referees for constructive comments that greatly helped in the presentation of the material.

# References

Besag, J., York, J., and Mollié, A. (1991). Bayesian image restoration with two applications in spatial statistics. *Annals of the Institute of Statistics and Mathematics* **43,** 1–59.

Breckling, J., Chambers, R., Dorfman, A., Tam, A., and Welsh, A. (1994). Maximum likelihood inference from sample survey data. *International Statistical Review* **62,** 349–363.

Breslow, N. and Chatterjee, N. (1999). Design and analysis of two-phase studies with binary outcome applied to wilms tumour prognosis. *Applied Statistics* **48,** 457–468.

Breslow, N. and Holubkov, R. (1997a). Maximum likelihood estimation of logistic regression parameters under two-phase, outcome-dependent sampling. *Journal of the Royal Statistical Society, Series B* **59,** 447–461.

Breslow, N. and Holubkov, R. (1997b). Weighted likelihood, pseudo likelihood and maximum likelhood methods for logistic regression analysis of two-stage data. *Statistics in Medicine* **16,** 103–116.

Breslow, N. E. and Cain, K. C. (1988). Logistic regression for two-stage case-control data. *Biometrika* **75,** 11–20.

Dellaportas, P. and Forster, J. (1999). Markov chain Monte Carlo model determination for hierarchical and graphical log-linear models. *Biometrika* **86,** 615–633.

Dunson, D. and Xing, C. (2009). Nonparametric bayes modeling of multivariate categorical data. *Journal of the American Statistical Association* **104,** 1042–1051.

Flanders, W. and Greenland, S. (1991). Analytic methods for two-stage case-control studies and other stratified designs. *Statistics in Medicine* **10,** 739–747.

Holubkov, R. (1995). *Maximum Likelihood Estimation in Two-Stage Case-Control Studies*. PhD thesis, University of Washington.

Kass, E. and Wasserman, L. (1995). A reference Bayesian test for nested hypothesis and its relationship to the Schwarz criterion. *Journal of the American Statistical Association* **90,** 928–934.

Knuiman, M. and Speed, T. (1988). Incorporating prior knowledge into the analysis of contingency tables. *Biometrics* **44,** 1061–1071.

Lawless, J., Kalbfleisch, J., and Wild, C. (1999). Semiparametric methods for response-selective and missing data problems in regression. *Journal of the Royal Statistical Society, Series B* **61,** 413–438.

Lumley, T. (2004). Analysis of complex survey samples. *Journal of Statistical Software* **9,**.

Lumley, T. (2010). *Complex Surveys: A Guide to Analysis using R*. John Wiley and Sons, Hoboken, Jersey.

Neyman, J. (1938). Contribution to the theory of sampling human populations. *Journal of the American Statistical Association* **33,** 101–116.

Reilly, M. and Pepe, M. (1995). A mean score method for missing and auxiliary covariate data in regression models. *Biometrika* **82,** 299–314.

Roberts, G., Gelman, A., and Gilks, W. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *The Annals of Applied Probability* **7,** 110–120.

Särndal, C.-E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer, New York.

Schill, J., Jockel, J. H., Drescher, K., and Timm, J. (1993). Logistic analysis in case-control studies under validation sampling. *Biometrika* **84,** 57–71.

Scott, A. and Wild, C. (1991). Fitting logistic regression in stratified case-control studies. *Biometrics* **47,** 497–510.

Scott, A. and Wild, C. (1997). Fitting regression models to case-control data by maximum likelihood. *Biometrika* **51,** 54–71.

Sinha, S., Mukherjee, B., and Ghosh, M. (2004). Bayesian semiparametric modeling for matched case-control studies with multiple disease states. *Biometrics* **60,** 41–49.

Wakefield, J. and Haneuse, S. (2008). Overcoming eological bias using the two-phase study design. *American Journal of Epidemiology* **167,** 908–916.

Walker, A. (1982). Anamorphic analysis: sampling and estimation for covariate effects when both exposure and disease are known. *Biometrics* **38,** 1025–1032.

White, E. (1982). A two stage design for the study of the relationship between a rare exposure and a rare disease. *American Journal of Epidemiology* **115,** 119–128.

Whittemore, A. (1997). Multistage sampling designs and estimating equations. *Journal of the Royal Statistical Society, Series B* **59,** 589–602.

Table 4: Disease model point and interval estimates for the Wilms tumour relapse data for various methods

| | Complete Data Estimate (95% CI) | Phase II Sample Estimate (95% CI) | NPML Estimate (95% CI) | Bayesian Two-Phase Estimate (95% CI) | Survey Sampling Estimate (95% CI) |
|---|---|---|---|---|---|
| $\beta_0$ | -2.71 (-2.92, -2.50) | -0.35 (-0.61, -0.08) | -2.58 (-2.83, -2.32) | -2.58 (-2.83, -2.33) | -2.57 (-2.82, -2.32) |
| Stage II | 0.77 (0.48, 1.05) | 0.55 (0.17, 0.93) | 0.55 (0.17, 0.94) | 0.56 (0.18, 0.95) | 0.55 (0.16, 0.94) |
| Stage III | 0.77 (0.48, 1.07) | 0.41 (0.02, 0.79) | 0.49 (0.09, 0.88) | 0.49 (0.09, 0.89) | 0.48 (0.08, 0.88) |
| Stage IV | 1.05 (0.71, 1.39) | 0.58 (0.12, 1.03) | 0.96 (0.47, 1.46) | 0.97 (0.49, 1.46) | 1.00 (0.50, 1.51) |
| UH* | 1.31 (0.82, 1.80) | -0.63 (-1.16, -0.09) | 1.26 (0.70, 1.82) | 1.25 (0.68, 1.80) | 1.35 (0.74, 1.96) |
| Stage II:UH | 0.15 (-0.49, 0.78) | 0.28 (-0.43, 0.99) | 0.29 (-0.47, 1.05) | 0.30 (-0.46, 1.07) | 0.12 (-0.75, 0.98) |
| Stage III:UH | 0.59 (-0.03, 1.21) | 0.70 (0.01, 1.39) | 0.73 (0.01, 1.46) | 0.75 (0.03, 1.49) | 0.51 (-0.32, 1.34) |
| Stage IV:UH | 1.26 (0.50, 2.02) | 1.67 (0.79, 2.55) | 1.43 (0.51, 2.36) | 1.43 (0.51, 2.37) | 0.98 (-0.24, 2.20) |

*: Unfavourable central histology