UW Biostatistics Working Paper Series

5-22-2006

Hierarchical Models for Combining Ecological and Case-control Data

Sebastien Haneuse
Group Health Cooperative, Seattle, haneuse.s@ghc.org

Jon Wakefield
University of Washington, jonno@u.washington.edu

1. Introduction

Ecological studies may be defined as designs that examine associations between *groups* of individuals, rather than between the individuals themselves, and are in wide-spread use in a variety of scientific disciplines (Achen and Shively, 1995, King, 1997, Morgenstern, 1998). Although the ecological design is controversial, its continued use may be attributed to the lack of availability of high-quality data for an individual-level analysis (perhaps due to logistic, financial or ethical constraints), and the ease and low-cost with which ecological data may often be obtained (Richardson and Monfort, 2000). Further, exposures which exhibit greater between-area variability relative to within-area variability lend themselves to studies at the group level (Prentice and Sheppard, 1995). When scientific interest lies at the level of the individual, the ecological design is susceptible to a range of methodological issues. These include problems common to all observational studies as well as a variety unique to their design (Richardson et al., 1987, Greenland, 1992, Greenland and Robins, 1994, Morgenstern, 1998, Wakefield, 2003). The collective impact of these difficulties is often referred to as *ecological* bias, and the *ecological fallacy* occurs when conclusions drawn from an ecological study are interpreted as representing individual-level associations, while an individual-level study (such as a cohort or case-control study) would have led to different conclusions.

In epidemiological settings, the fundamental difficulty is the lack of information regarding within-group exposure and confounder variation. A consequence of this is that it is not possible to uniquely identify individual-level models on the basis of ecological data alone. Additional information is required and may include imposing (generally untestable) assumptions regarding the individual-level model (e.g. King, 1997), or the collection of individual-level data (e.g. Prentice and Sheppard, 1995). In general, however, without individual-level information on both responses and confounders/exposures one cannot adequately assess any assumed individual-level model (Wakefield, 2004).

As a means to overcoming methodological issues associated with the ecological design, we supplement ecological data with a sample of carefully collected individual-level data. We refer this general class of designs as *hybrid designs* where, given an individual-level model, the individual-level data provide the basis for identifiability while the ecological data provide efficiency gains.

The remainder of the paper is as follows. In Section 2 we present a general likelihood-based framework for combining ecological and individual-level data. Section 3 provides the details of the proposed approach in the context of a study of infant mortality in North Carolina. Specific interest lies in the joint impact the of infant's race and the mother's age at the time of birth. Section 3 also outlines computational details, and Section 4 provides results. Finally, Section 5 concludes with a discussion, together with an outline of avenues for further research.

2. Framework

In the context of a rare outcome, Haneuse and Wakefield (2006) proposed to combine group-level ecological data with a sample of individual-level case-control data. They derived the exact likelihood, and corresponding score and information matrix, in two settings: (i) the unadjusted association between a binary outcome and binary exposure and (ii) adjustment for a binary confounder, where the joint outcome/confounder and exposure/confounder distributions are observed. Here we present a more general framework for likelihood development, while Section 3 provides details in a more specific setting. In the previous paper, the models contained fixed effects only, and inference was made via the asymptotic distribution of the maximum likelihood estimate. Here we focus on hierarchical models, with area-specific random effects, and follow a Bayesian approach with computation via Markov chain Monte Carlo (MCMC).

Suppose the study population may be partitioned into K mutually exclusive areas. Let $\mathbf{Y}$ and $\mathbf{X}$ denote individual-level outcome and exposure information that would be observed for all individuals in each of K study areas, had a fully individual-level analysis been performed. We refer to $\mathbf{Y}$ and $\mathbf{X}$ as the *complete outcome* and *complete exposure* data. Let $\boldsymbol{\theta}$ denote a parameter vector that indexes the model relating $\mathbf{X}$ to $\mathbf{Y}$, which is assumed to be specified on the basis of a scientific question of interest. Given $(\mathbf{Y}, \mathbf{X})$, estimation and inference regarding $\boldsymbol{\theta}$ would proceed on the basis of the *complete data* or *individual-level* likelihood

$$L(\boldsymbol{\theta}; \mathbf{Y}|\ \mathbf{X}) \tag{1}$$

For both the ecological study design and the design proposed in this research, *incomplete* data are observed. In ecological studies, such data is typically summary or aggregated individual-level data and, consequently, information provided by the incomplete data may be viewed as being contained in that provided by the complete data. A simple example of this would be the marginal totals of a 2×2 table compared to the internal counts. More generally, the extent of the incompleteness depends on the setting, and may refer to incomplete outcome data alone or both incomplete outcome and exposure data. We denote incomplete outcome data by $\mathbf{Y}^*$ and incomplete exposure data by $\mathbf{X}^*$.

Given complete exposure information but incomplete outcome information, the distribution of the data may be obtained by considering the joint distribution of the complete and incomplete outcome data given the complete exposure data:

$$\begin{aligned} P(\mathbf{Y}^*|\ \mathbf{X}, \boldsymbol{\theta}) &= \sum_{\mathbf{Y}|\ \mathbf{Y}^*, \mathbf{X}} P(\mathbf{Y}^*, \mathbf{Y}|\ \mathbf{X}, \boldsymbol{\theta}) \\ &= \sum_{\mathbf{Y}|\ \mathbf{Y}^*, \mathbf{X}} P(\mathbf{Y}|\ \mathbf{X}, \boldsymbol{\theta})\ P(\mathbf{Y}^*|\ \mathbf{Y}, \mathbf{X}, \boldsymbol{\theta}) \end{aligned}$$

In the above, the summation conditions on $\mathbf{Y}^*$ since the complete outcome data must be consistent with the incomplete data. For example, one cannot sum over values of $\mathbf{Y}$ where the number of cases is fewer than that observed in $\mathbf{Y}^*$. Estimation and inference therefore proceeds on the basis of a likelihood which is derived by averaging the individual-level likelihood (1), $P(\mathbf{Y}|\ \mathbf{X}, \boldsymbol{\theta})$, over the distribution of the observed incomplete outcome data, given the observed complete exposure and unobserved incomplete outcome data:

$$L(\boldsymbol{\theta}; \mathbf{Y}^*|\ \mathbf{X})\ =\ \sum_{\mathbf{Y}|\ \mathbf{Y}^*, \mathbf{X}} L(\boldsymbol{\theta}; \mathbf{Y}|\ \mathbf{X})\ P(\mathbf{Y}^*|\ \mathbf{Y}, \mathbf{X}, \boldsymbol{\theta}) \qquad (2)$$

Given incomplete exposure information, one is further required to average over the uncertainty in the unknown complete exposure data. This requires the specification of the conditional distribution of the complete exposure data given the incomplete exposure data, which is presumed to be indexed by the parameter vector $\boldsymbol{\phi}$. The distribution of the data may be obtained in a similar manner to that used above to give

$$\begin{aligned} P(\mathbf{Y}^*|\ \mathbf{X}^*, \boldsymbol{\theta}, \boldsymbol{\phi})\ &=\ \sum_{\mathbf{X}|\ \mathbf{Y}^*, \mathbf{X}^*} P(\mathbf{Y}^*, \mathbf{X}|\ \mathbf{X}^*, \boldsymbol{\theta}, \boldsymbol{\phi}) \\ &=\ \sum_{\mathbf{X}|\ \mathbf{Y}^*, \mathbf{X}^*} P(\mathbf{Y}^*|\ \mathbf{X}, \boldsymbol{\theta})\ P(\mathbf{X}|\ \mathbf{X}^*, \boldsymbol{\phi}) \end{aligned}$$

Given incomplete data ($\mathbf{Y}^*$, $\mathbf{X}^*$), therefore, estimation and inference proceeds for both $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ simultaneously via the likelihood

$$L(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{Y}^*|\ \mathbf{X}^*)\ =\ \sum_{\mathbf{X}|\ \mathbf{Y}^*, \mathbf{X}^*} \left\{ \sum_{\mathbf{Y}|\ \mathbf{Y}^*, \mathbf{X}} L(\boldsymbol{\theta}; \mathbf{Y}|\ \mathbf{X})\ P(\mathbf{Y}^*|\ \mathbf{Y}, \mathbf{X}, \boldsymbol{\theta}) \right\} P(\mathbf{X}|\ \mathbf{X}^*, \boldsymbol{\phi}) \qquad (3)$$

For analyses within the frequentist statistical framework, expressions for both the score vector and information matrix associated with (3) may be obtained by exploiting its mixture representation (Haneuse and Wakefield, 2006). Specifically, starting with expressions for the complete data score and information matrixes, based on (1), the corresponding expressions for (3) are obtained by consideration of the conditional distribution of the complete data given the incomplete data, $\mathbf{Y}, \mathbf{X}|\ \mathbf{Y}^*, \mathbf{X}^*$. For analyses within the Bayesian statistical framework, posterior distributions may be obtained via MCMC methods where we introduce the complete data $(\mathbf{Y}, \mathbf{X})$ as auxillary variables to be estimated simultaneously with $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$.

3. North Carolina infant mortality data

To illustrate the combination of ecological and case-control data, we consider a hypothetical study of infant ($<$ 1 years old) mortality in North Carolina. Information on vital statistics are provided by the North Carolina State Center for Health Statistics, and available for download from the Odum Institute for Research in Social Science at the University of North Carolina at Chapel Hill (`http://www.irss.nc.edu/ncvital`). For each of 100 counties, we obtained the total number births and infant deaths by race and the age of the mother, aggregated across years 2000-2004. For the purpose of illustrating the methods of this paper we consider two binary exposure variables: 'minority' defined as whether or not the babies race was non-white, and 'teen mother' defined as whether or not the mother was a teenager at the time of birth. An attractive feature of these data is they consist of (de-identified) individual-level records, so that complete joint information on outcome, race and mother's age are available. We may therefore construct a hypothetical ecological study by considering the corresponding county-specific marginal totals. Further, having individual-level information provides a basis for the direct assessment of competing methods that do not use all information.

3.1 *Data description*

Across the 100 counties, there is substantial variation in both the number of births, ranging from 221 to 61,960, and in the number of (all cause) infant deaths, ranging from 0 to 484. Figure 1 provides county-specific crude mortality rates, the percent minority and the percent teen mothers. The mortality rates vary from 0 per 1,000 births to a maximum of 17.5 per 1,000. County-specific percent minority exhibits substantial variation across North Carolina, ranging from 0.6% to 73.3%, while the county-specific percent teen mothers ranges from 5.9% to 21.0%. Figure 2 presents two ecological correlation analyses, for minority and teen status respectively. In each plot, two least square fits have been added corresponding to an ordinary (unweighted) fit and a weight fit, with the latter weighted by the county-specific number of births.

[Figure 1 about here.]

[Figure 2 about here.]

In the following we present details regarding a hybrid scheme, where the ecological data (i.e. the marginal rates in Figures 1 and 2), are supplemented with a sample of case-control data from each county.

3.2 *Individual-level model*

We emphasize that the scientific goal is inference with respect to individual-level associations, and consequently the first task is to write down an individual-level model. Let Y denote the binary outcome, $X = 0/1$ represent white/minority and $Z = 0/1$ represent non-teen/teen mother. Also, let N_{yxzk} denote the number of individuals in the $[Y{=}y,\ X{=}x,\ Z{=}z]$ outcome/minority/teen stratum in the k^{th} county, and M_{xzk} denote the corresponding population in the $[X{=}x,\ Z{=}z]$ minority/teen stratum, for y, x, $z = 0$, 1. We assume the county-specific outcome counts, N_{yxzk}, to be distributed according to a Binomial(M_{xzk}, p_{xzk}) distribution with

$$\text{logit}(p_{xzk}) = \ \beta_0 + \ \beta_{\text{X}}x + \ \beta_{\text{Z}}z + \ \beta_{\text{XZ}}xz + \ V_k, \tag{4}$$

and $\mathbf{V} = \{V_1, \ldots, V_K\}^T$ is a vector of county-specific random effects. We assume the components of $\mathbf{V}$ are independent and identically distributed according to a zero mean Normal distribution with variance $\sigma_v^2 > 0$. Let $\boldsymbol{\beta} = \{\beta_0, \beta_{\text{X}}, \beta_{\text{Z}}, \beta_{\text{XZ}}\}$ and $\boldsymbol{\theta} = \exp\{\boldsymbol{\beta}\} = \{\theta_0, \theta_{\text{X}}, \theta_{\text{Z}}, \theta_{\text{XZ}}\}$ the corresponding baseline odds and odds ratio parameters.

3.3 *Hybrid sampling scheme*

We initially present development of the hybrid likelihood for a generic area, temporarily ignoring the county-specific index, k, and random effect, V_k. Table 1 provides an overview of the notation for an area of size N; N_{yxz} and M_{xz} are presented within square brackets to emphasize that in an ecological study, and hence the scheme we propose, they are unobserved. We assume, therefore, that the marginal covariate data are available for X and Z, but not the cross-classification. Further non-cases and cases are not classified by either covariate.

Suppose we obtain a sample of n_0 controls and n_1 cases. Let n_{yxz} denote the number of individuals in the [$Y{=}y$, $X{=}x$, $Z{=}z$ outcome/minority/teen stratum of the case-control sample. This case-control sampling scheme differs from Haneuse and Wakefield (2006) in which cases and controls were gathered separately within confounder-defined strata.

[Table 1 about here.]

If the $\mathbf{N_{yxz}} = \{N_{yxz}; y, x, z, = 0, 1\}$ were observed, then conditional on the joint exposure distribution, $\mathbf{M_{xz}} = \{M_{xz}; x, z, = 0, 1\}$, the individual-level likelihood, denoted $L^I(\boldsymbol{\beta}; \mathbf{N_{yxz}} | \mathbf{M_{xz}})$, is the product of four independent Binomial distributions. Using the notation of Section 2 we have $\mathbf{Y} \equiv \mathbf{N_{yxz}}$ and $\mathbf{X} \equiv \mathbf{M_{xz}}$.

Let $\mathbf{Y}^* = (\mathbf{N_y}, \mathbf{n_{yxz}})$ denote the totality of the observed data under the hybrid sampling scheme. Given $\mathbf{M_{xz}}$, the hybrid likelihood may be derived via the introduction of the $\mathbf{N_{yxz}}$ as auxiliary variables. Specifically, we consider the joint distribution of ($\mathbf{Y}^*$, $\mathbf{N_{yxz}}$) and integrate over the $\mathbf{N_{yxz}}$ margin to obtain the following weighted average of individual-level likelihoods

$$L^H(\boldsymbol{\beta}; \mathbf{Y}^* |\ \mathbf{M_{xz}}) \ = \sum_{\mathbf{N_{yxz}} \in \mathcal{R}_N(\mathbf{Y}^*, \mathbf{M_{xz}})} W(\mathbf{n_{yxz}} | \mathbf{N_{yxz}})\ L^I(\boldsymbol{\beta}; \mathbf{N_{yxz}} | \mathbf{M_{xz}}), \qquad (5)$$

where the weights

$$W(\mathbf{n_{yxz}}|\mathbf{N_{yxz}}) \;=\; \left\{\binom{N_0}{n_0}\binom{N_1}{n_1}\right\}^{-1}\left\{\prod_{x=0}^{1}\prod_{z=0}^{1}\binom{N_{0xz}}{n_{0xz}}\binom{N_{1xz}}{n_{1xz}}\right\}, \qquad (6)$$

consist of multivariate hypergeometric terms for the cases and controls respectively. Equation (5) corresponds to (2) in the general framework of Section 2. Due to the constraints imposed by both the marginal totals and case-control data, $\mathcal{R}_N(\mathbf{Y}^*, \mathbf{M_{xz}})$, the space of admissible configurations of the $\mathbf{N_{yxz}}$, is complex. The Appendix provides one representation which is computationally convenient.

To obtain the form of the hybrid likelihood given solely marginal information regarding X and Z, denoted $\mathbf{M_{x+}} = \{M_{0+}, M_{1+}\}$ and $\mathbf{M_{+z}} = \{M_{+0}, M_{+1}\}$ respectively, we adopt the same general approach. Specifically, we introduce the $\mathbf{M_{xz}}$ as an additional set of auxiliary variables and then integrate over their distribution. The latter depends solely on the underlying odds ratio between X and Z, denoted by ϕ_{XZ}, which must be estimated. Using the notation of Section 2 we have $\mathbf{X}^* = \{\mathbf{M_{x+}}, \mathbf{M_{+z}}\}$. In the aggregate data design of Prentice and Sheppard (1995), information on within-area joint exposure distributions is obtained via supplementary survey samples. In the setting of the hybrid design, information is provided by the retrospective exposure observations in the case-control data. The hybrid likelihood is given by

$$L^H(\boldsymbol{\beta}, \phi_{\mathrm{XZ}}; \mathbf{Y}^*|\ \mathbf{X}^*) \;=\; \sum_{\mathbf{M_{xz}} \in \mathcal{R}_M(\mathbf{X}^*, \mathbf{Y}^*)} L^H(\boldsymbol{\beta}; \mathbf{Y}^*|\ \mathbf{M_{xz}})\ L(\phi_{\mathrm{XZ}}; \mathbf{M_{xz}}|\ \mathbf{X}^*), \qquad (7)$$

where

$$\mathcal{R}_M(\mathbf{Y}^*, \mathbf{X}^*) \equiv [\max(m_{11}, M_{1+} - M_{+0} + m_{00}), \min(M_{1+} - m_{10}, M_{+1} - m_{01})]$$

is the support over which the unknown exposure/confounder cross-classification $\mathbf{M_{xz}}$ is marginalized and $L(\phi_{\mathrm{XZ}}; \mathbf{M_{xz}}|\ \mathbf{X}^*)$ denotes the extended hypergeometric distribution of $\mathbf{M_{xz}}$ given $\mathbf{X}^*$ (Harkness, 1965, Johnson and Kotz, 1969). The hybrid likelihood (7) corresponds to (3) in the framework of Section 2. As in previous cases, its form has the intuitive interpretation of a weighted average of hybrid likelihoods, where we average over the uncertainty in the unknown $\mathbf{M}_{\mathbf{xz}}$. An alternative hybrid scheme could be to collect supplementary information solely on cases, with the weights in (6) consisting of the multivariate hypergeometric term for the cases. Under both the case-control and cases-only scheme, in settings where the total number of cases N_1 is small it may be that $n_1 = N_1$, so that all cases are sampled. Given the ecological margins and complete information on the cases, the hybrid likelihood reduces to the individual-level likelihood, and no further information is provided by the collection of controls.

Finally, we also consider extending the above model to allow the exposure/confounder odds ratio to vary across areas. Specifically we incorporate heterogeneity into the model by introducing area-specific odds ratio parameters, $\phi_{\mathrm{xz}k}$. We assume the area-specific log-odds ratio parameters to be independently and identically distributed according to a Normal distribution with mean $\log(\phi_{\mathrm{xz}})$, and variance $\sigma_\phi^2 > 0$.

3.4 *Estimation and Inference*

In a frequentist analysis, estimation and inference may proceed via maximization of the hybrid likelihood (7) and evaluation of the corresponding information matrix. A key difficulty however, is the computational burden of repeated evaluations of univariate and multivariate hypergeometric distributions, potentially over very large spaces $\mathcal{R}^N$ and $\mathcal{R}^M$. The introduction of random effects into the disease model, as in (4), further requires integration with respect to their distribution to obtain a marginal likelihood. For most realistic disease models this will be computationally prohibitive. One possible approach is to consider approximations to the likelihood contributions. In the scheme we propose, where case-control sampling may result in small cell counts for some areas, such approximations may be inaccurate, and an investigation of the trade-off between computational tractability and accuracy of likelihood evaluations will be the subject of future work.

As an alternative, we consider a Bayesian implementation. While computation is still an issue, implementation via MCMC offers the opportunity of fitting more flexible disease models in a relatively straightforward and structured manner. To complete the Bayesian specification, we outline a priori distributional assumptions regarding the unknown $\boldsymbol{\beta}$, σ_v^2, ϕ_{XZ}, and possibly σ_ϕ^2. In the setting of a purely ecological study, considerable care is required in the specification of priors since identifiability may be driven solely by such choices (Wakefield, 2004). Under the hybrid sampling scheme, the case-control data provide identifiability and, hence, improper priors need not necessarily be avoided and we adopt an improper flat prior for $\boldsymbol{\beta}$. For the random effects variance components, a standard approach is to assume $\tau_v = \sigma_v^{-2}$ follows a conjugate Gamma(a_v, b_v) distribution. In Section 4 we explore sensitivity to the choice of a_v and b_v. Finally, in the setting where ϕ_{XZ} is allowed to vary across areas, we assume a vague but proper prior for the mean log-odds ratio and a Gamma(a_ϕ, b_ϕ) distribution for the inverse variance.

3.5 *Auxiliary variable scheme*

Samples from the hybrid posterior are obtained via an auxiliary variable scheme (Tanner and Wong, 1987). Here we present the scheme where ϕ_{XZ} is assumed fixed across areas, although it is easily modified to accommodate heterogeneity. Consider the joint distribution of the unknown parameters, $\boldsymbol{\gamma} = \{\boldsymbol{\beta}, \mathbf{V}, \sigma_v^2, \phi_{\mathrm{XZ}}\}$ and the two sets of unknown auxiliary variables, $\mathbf{N_{1xz}}$ and $\mathbf{M_{xz}}$, given by

$$\begin{aligned} \pi^H(\mathbf{N_{1xz}}, \mathbf{M_{xz}}, \gamma | \, \mathbf{Y}^*, \mathbf{X}^*) \;=\;& \pi^H(\mathbf{N_{1xz}} | \, \mathbf{Y}^*, \mathbf{X}^*, \boldsymbol{\gamma}, \mathbf{M_{xz}}) \\ & \pi^H(\mathbf{M_{xz}} | \, \mathbf{Y}^*, \mathbf{X}^*, \boldsymbol{\gamma}) \; \pi^H(\boldsymbol{\gamma} | \, \mathbf{Y}^*, \mathbf{X}^*). \end{aligned}$$

The final component on the right-hand side, $\pi^H(\boldsymbol{\gamma}|\ \mathbf{Y}^*, \mathbf{X}^*)$, is the target posterior. We consider an MCMC scheme which alternates between the full conditionals:

$$\begin{array}{ll} (i) & \pi^H(\boldsymbol{\gamma}|\ \mathbf{Y}^*, \mathbf{X}^*, \mathbf{N_{1xz}}, \mathbf{M_{xz}}) \\ (ii) & \pi^H(\mathbf{N_{1xz}}|\ \mathbf{Y}^*, \mathbf{X}^*, \boldsymbol{\gamma}, \mathbf{M_{xz}}) \\ (iii) & \pi^H(\mathbf{M_{xz}}|\ \mathbf{Y}^*, \mathbf{X}^*, \boldsymbol{\gamma}, \mathbf{N_{1xz}}). \end{array}$$

The first set of conditionals correspond to a standard Bayesian logistic regression analysis. The second set of conditionals correspond to the conditional distribution of the unobserved $\mathbf{N_{yxz}}$ counts, given the underlying disease model, the totality of the observed data and the auxiliary variables $\mathbf{M_{xz}}$. We refer to this distribution as the multivariate supplemented extended hypergeometric distribution (see the Appendix). Sampling from this distribution follows from a scheme developed for the closely related extended hypergeometric distribution (Liao and Rosen, 2001). For the final set of conditionals, we apply a Metropolis step. Ignoring terms which act as normalising constants with respect to $\mathbf{M_{xz}}$, the full conditional may be decomposed as

$$\begin{aligned} \pi^H(\mathbf{M_{xz}}|\ \mathbf{Y}^*, \mathbf{X}^*, \mathbf{N_{1xz}}, \boldsymbol{\gamma}) \quad \propto \quad & \Pr(\mathbf{N_y}|\ \mathbf{N_{1xz}}, \mathbf{M_{xz}}, \boldsymbol{\gamma}) \\ & \times \Pr(\mathbf{n_{yxz}}|\ \mathbf{N_y}, \mathbf{n_y}, \mathbf{N_{1xz}}, \mathbf{M_{xz}}, \boldsymbol{\gamma}) \qquad (8) \\ & \times \Pr(\mathbf{N_{1xz}}|\ \mathbf{M_{xz}}, \boldsymbol{\gamma}) \times \Pr(\mathbf{M_{xz}}|\ \mathbf{X}, \boldsymbol{\gamma}) \end{aligned}$$

The first component of (8) is determined trivially, since the elements of $(\mathbf{N_{1xz}}, \mathbf{M_{xz}})$ must satisfy the constraints imposed by the marginal outcome totals $\mathbf{N_y}$. The second component is the product of two independent multivariate hypergeometric distributions. The third component is the product of four Binomial distributions and the final component is an extended hypergeometric distribution with odds ratio ϕ_{XZ}.

4. Results

Table 2 provides a summary of results based on the methods of Section 3. Care must be taken in the specification of the gamma prior for the precision of the random effects. For example, Kelsall and Wakefield (1999) point out that the choice Gamma(ϵ, ϵ), with ϵ small, leads to very little weight on small values of the standard deviation and hence may impose between-area variability. For all analysis, we consider two prior choices; Gamma(0.5, 0.001) and Gamma(0.5, 0.1). The first induces a prior for σ_v with median 0.05 and central 95% credible range of (0.01, 1.01), while the second induces a prior with median 0.66 and central 95% credible range of (0.20, 14.24). We note that a Gamma(0.1, 0.1), a common choice, induces a prior median of 12.9 with 95% credible range of approximately (0.32, 4.1×10^7).

Given the ecological data (see Figures 1 and 2), we consider two sampling designs for the collection of case-control data. The first collects $n = 20$ case-control samples from each of the 100 counties. In counties for which the total number of cases N_1 exceeds 10, we take 10 cases and 10 controls. In counties for which N_1 is less than or equal to 10 (there are 18 such counties), we take all available cases and the remaining samples are taken from the controls. This scheme yielded a total of 885 cases and 1115 controls. The second design collects $n = 100$ case-control samples form each of the 9 counties in which there are at least 100 cases, while no individual-level data are obtained from the remaining 91 counties. Consequently, under this scheme, there are a total of 450 cases and 450 controls.

[Table 2 about here.]

Table 2 outlines posterior results based on a Gamma(0.5, 0.001) prior for the precision of the random effects, assuming a common exposure/confounder odds ratio across the 100 counties. For each design, a single data set was generated and six analyses performed, the first being based on the complete individual-level data. For the purposes of comparison with the alternative analyses which rely, to various extents, on incomplete data, the individual-level analysis acts as a gold standard. For the case-control analysis the intercept is identifiable via the known totals in each area and the introduction of an appropriate offset into the regression (Breslow and Day, 1980). In this case, however, the random effects standard deviation, σ_v, is not identifiable. Further, as the case-control sample represents a biased (marginal) sample for the minority/teen mother status relationship, ϕ_{XZ} is not estimated. For the hybrid design we present four scenarios which depend on the extent of the available ecological and individual-level data. For the ecological data we consider the situation where a complete cross classification of minority status and teen mother status is available (denoted $\mathbf{M_{xz}}$), as well as the situation where only their marginal totals are observed (denoted $\{\mathbf{M_{x+}}, \mathbf{M_{+z}}\}$). Within each of these, we consider an analysis based on the cases only as well as an analysis based on all case-control samples.

Although not presented, there is little sensitivity in the results of the regression coefficients, or for the ϕ_{XZ} odds ratio, to the choice of Gamma(a_v, b_v) prior. With the exception of the case-control analysis, the posterior summaries for the random effects standard deviation, σ_v, are the same across each analysis. Based on the hybrid design which collected 20 case-control samples from each county, the posterior median (95% credible interval) for σ_v under the Gamma(0.5, 0.001) and Gamma(0.5, 0.1) priors are 0.14 (0.08, 0.20) and 0.18 (0.13, 0.23) respectively, showing moderate sensitivity to the prior and in the expected direction. We note that, a useful interpretation of σ_v is to consider the extent of residual variability in the relative risks. Based on an estimate of $\hat{\sigma}_v = 0.14$, we find 95% of the residual relative risk lies between 0.76 and 1.32.

From Table 2 we find that for both designs the analysis based on the case-control data alone performs quite poorly. For each of the odds ratio parameters, point estimates differ substantially from their individual-level counter parts. Each of the analyses based on the hybrid likelihood provide point estimates that are closer to the gold standard. The inclusion of the ecological data also provides improvements in efficiency, over the case-control analysis, as is evident from the tighter credible intervals. With the introduction of the ecological data the credible intervals tighten, indicating the utility of combining the two sources of information. Given the joint minority/teen mother status distribution, there seems to be little difference in sampling cases only in the hybrid design. However, given marginal data alone we see that there is slightly more sensitivity with the loss of information having the greatest impact on the estimation of ϕ_{XZ}.

Finally, for the model where the ϕ_{XZ} are allowed to vary the results did not differ significantly from those presented in Table 2. In particular, the posterior summaries for the regression parameters remained the same. Under a Gamma(0.5, 0.001) prior for the (inverse) variance component σ^2_ϕ, the posterior median (95% credible interval) for ϕ_{XZ} was 1.96 (1.86, 2.06). The corresponding posterior summaries for σ_ϕ were 0.21 (0.17, 0.26).

5. Discussion

The use of an ecological study design, which examines associations among *groups* of individuals, results in a disconnect between the level of the hypothesis and the level of the analysis. The disconnect arises, in part, from the inability to characterise within-group exposure/confounder variation, which results in non-identifiability of the individual-level model. While ecological studies are subject to potential biases common to all observational studies, they are further subject to biases which arise from inappropriate assumptions made to overcome the issue of non-identifiability. A fundamental difficulty in the use of ecological studies is that assumptions are required to overcome methodological issues which give rise to ecological bias. Unfortunately, however, such assumptions may not be critically assessed given ecological data alone.

The solution to the ecological inference problem, where we seek to elluciate individual-level associations from group-level data, is to collect and incorporate information on individuals. In this paper we have proposed a study design aimed at combining group-level ecological information with individual-level data for a sample of the population, which we refer to as the *hybrid study design*. While there exist methods aimed at combining ecological and individual-level data, such as the aggregate data approach of Prentice and Sheppard (1995), they generally concentrate on retrieving the within-area exposure/confounder distribution. The basis for their approach is the induced aggregate-level model, and hence the analysis is viewed as being at the level of the group. The basis for the statistical analysis for the hybrid design is the induced likelihood that corresponds to the observed data, and so is at the level of the individual. This in turn allows individual-level model checking and the assessment, for example, of the need for contextual effects (Wakefield, 2004). The latter refer to the case where an individuals risk is not only determined by their own exposure but also by that of other individuals in their shared area (via the group-level measure), and are often of interest in the social sciences. An alternative interpretation of the combination of the two sources of information is to consider supplementing a small case-control study with ecological information obtained from the same population.

It is critical to ensure the compatability of the two sources of information, and in particular, to ensure that the underlying individual-level likelihood/model is common for both sets of data. In our scheme, we assume that the case-control samples are drawn directly from the population for which the ecological data provide summary information, but in practice this requires care. Valid estimation and inference based on the hybrid likelihood will depend on the assumption of no selection bias in the case-control samples. Such a requirement may viewed as being a part of broader epidemiological issues concerned with traditional case-controls studies which include, for example, the issue of the compatability between the control and case populations from which samples are being drawn. While we concentrate on retrospective case-control sampling of individuals, motivated by efficiency gains in the setting of a rare outcome, much of the methodology follows when individuals are collected via a prospective cohort scheme.

The Bayesian framework we have adopted here provides an algorithmic basis for estimation and inference which may offer a more reasonable approach to extending the methods to more complex settings. For the Ohio data, it would be natural to incorporate an additional set of spatially structured random effects (e.g. Besag et al., 1991), which may help account for some of the residual relative risk variability. This will be the subject of further work, together with extending these methods to continuous exposures. Other advantages of the Bayesian approach are the ability to incorporate prior information and the absence of reliance on asymptotics. The latter is especially useful for the hybrid design since one may only require small sample sizes, perhaps even using case information alone, to induce identifiability of the individual-level model. Finally, further development of the hybrid design we propose will likely benefit from exploring connections with the missing data literature (e.g. Little and Rubin, 2002, Robins et al., 1994) and sample survey literature (e.g. Breckling et al., 1994).

The collection and incorporation of the case-control data into the analysis is motivated by the need to avoid making untestable assumptions which may result in ecological bias. The simulation studies of Haneuse and Wakefield (2006) indicate that only a small amount of individual-level data is required to induce identifiability of the underlying disease model. Further, their simulations suggest that in a variety of settings there is utility in jointly modeling the ecological and case-control data, rather than performing analyses based solely on the case-control data. The greatest benefit is in terms of efficiency gains, the extent of which depend on the underlying disease model and the interplay between the ecological and case-control information.

Acknowledgements

This work was supported by grant R01CA095994 from the National Institutes of Health.

References

Achen, C. and Shively, W. (1995). *Cross-Level Inference*. University of Chicago Press, Chicago.

Besag, J., York, J. and Mollie, A. (1991). Bayesian image restoration, with two applications in spatial statistics (Disc: p21-59). *Annals of the Institute of Statistical Mathematics* **43**, 1–20.

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

Breslow, N. E. and Day, N. E. (1980). *Statistical methods in cancer research (Vol. 1): The analysis of case-control studies*. World Health Organization [Distribution and Sales Service].

Greenland, S. (1992). Divergent biases in ecologic and individual-level studies. *Statistics in Medicine* **11**, 1209–1223.

Greenland, S. and Robins, J. (1994). Ecologic studies – Biases, misconceptions, and counterexamples (Disc: p761-771). *American Journal of Epidemiology* **139**, 747–760.

Guthrie, K., Sheppard, L. and Wakefield, J. (2002). A hierarchical aggregate data model with spatially correlated disease rates. *Biometrics* **58**, 898–905.

Haneuse, S. and Wakefield, J. (2006). The combination of ecological and case-control data. *Journal of the Royal Statistical Society, Series B* In revision.

Harkness, W. L. (1965). Properties of the extended hypergeometric distribution. *The Annals of Mathematical Statistics* **36**, 938–945.

Johnson, N. and Kotz, S. (1969). *Distributions in Statistics: Discrete Distributions*. John

Wiley and Sons, New York.

Kelsall, J. and Wakefield, J. (1999). Contribution to: "Bayesian models for spatially correlated disease and exposure data", by Best, N.G., Waller, L.A., Thomas, A., Conlon, E.M., and Arnold, R. In Bernardo, J., Berger, J., Dawid, A. and Smith, A., editors, *Bayesian Statistics 6, Proceedings of the Sixth Valencia International Meeting.* Oxford University Press.

King, G. (1997). *A Solution to the Ecological Inference Problem.* Princeton University Press, Princeton, New Jersey.

Lasserre, V., Guihenneuc-Jouyaux, C. and Richardson, S. (2000). Biases in ecological studies: Utility of including within-area distribution of confounders. *Statistics in Medicine* **19**, 45–59.

Liao, J. G. and Rosen, O. (2001). Fast and stable algorithms for computing and sampling from the noncentral hypergeometric distribution. *The American Statistician* **55**, 366–369.

Little, R. and Rubin, D. (2002). *Statistical analysis of missing data.* John Wiley and Sons, New Jersey, second edition.

Morgenstern, H. (1998). Ecological studies. In Rothman, K. and Greenland, S., editors, *Modern Epidemiology*, pages 459–480. Lipincott-Raven, second edition.

Prentice, R. L. and Sheppard, L. (1995). Aggregate data studies of disease risk factors. *Biometrika* **82**, 113–125.

Richardson, S. and Monfort, C. (2000). Ecological correlation studies. In Elliott, P., Wakefield, J., Best, N. and Briggs, D., editors, *Spatial Epidemiology: Methods and Applications*, pages 205–220. Oxford University Press, Oxford.

Richardson, S., Stuecker, I. and Hemon, D. (1987). Comparison of relative risks obtained in ecological and individual studies: Some methodological considerations. *International*

Journal of Epidemiology **16**, 111–120.

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* **89**, 846–866.

Tanner, M. and Wong, W. (1987). The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association* **82**, 528–550.

Wakefield, J. (2003). Sensitivity analyses for ecological regression. *Biometrics* **59**, 9–17.

Wakefield, J. (2004). Ecological inference for 2×2 tables (with discussion). *Journal of the Royal Statistical Society, Series A* **167**, 385–445.

Appendix A

Multivariate supplemented extended hypergeometric distribution

The auxiliary variable MCMC scheme of Section 3 requires the ability to sample from the conditional distribution of the unobserved $\mathbf{N_{yxz}}$ counts, given the remaining components of the model. Given $\mathbf{M_{xz}}$, and in the absence of case-control data, this distribution is the multivariate analogue of the well-known extended hypergeometric distribution (Harkness, 1965, Johnson and Kotz, 1969), which we denote as MXHG($\mathbf{N_{yxz}}|\ \mathbf{M_{xz}}$). The probability mass function for this latter distribution is

$$P(\mathbf{N_{yxz}}|\ \mathbf{M_{xz}}, \theta) \ = \ \frac{\prod_{x=0}^{1} \prod_{z=0}^{1} \binom{M_{xz}}{N_{1xz}} \xi_{xz}^{N_{1xz}}}{\sum_{\mathbf{u} \in \mathcal{R}_N(\mathbf{M_{xz}})} \prod_{x=0}^{1} \prod_{z=0}^{1} \binom{M_{xz}}{u_{xz}} \xi_{xz}^{u_{xz}}}, \qquad \text{(A.1)}$$

where $\xi_{00} = 1$ and ξ_{xz} denotes the odds ratio comparing exposure level $X/Z = x/z$ to $X/Z = 0/0$. Under (4), we have $\xi_{10} = \theta_{\mathrm{X}}$, $\xi_{01} = \theta_{\mathrm{Z}}$ and $\xi_{11} = \theta_{\mathrm{X}}\theta_{\mathrm{Z}}\theta_{\mathrm{XZ}}$.

The space $\mathcal{R}_N(\mathbf{M_{xz}})$ denotes the range of possible configurations for the unknown

$\mathbf{N_{yxz}}$, given $\mathbf{M_{xz}}$. In the setting of Section 3, $\mathcal{R}_N(\mathbf{M_{xz}})$ is a complex three-dimensional space; given the marginal totals, only three of the eight components are required to completely specify $\mathbf{N_{yxz}}$. A computationally convenient ordering of $\mathcal{R}_N(\mathbf{M_{xz}})$ may be obtained by successive reductions of the 4×2 table via a series of 2×2 tables. Initially, consider collapsing all exposure groups in Table 1, with the exception of the $X/Z = 1/1$ level. This results in a 2×2 table, where the marginal outcome totals are (N_0, N_1) and the marginal exposure totals are ($M_{00} + M_{10} + M_{01}$, M_{11}). The cell corresponding to the case total in the $X/Z = 1/1$ exposure group takes on values in the range

$$N_{111} \in [\max(0, N_1 - (M_{00} + M_{10} + M_{01})),\ \min(N_1, M_{11})]. \tag{A.2}$$

For a given value of N_{111} in this range (and $N_{011} = M_{11} - N_{111}$), there are three remaining exposure levels. Collapsing the first two of these results in a 2×2 table where the marginal outcome totals are ($N_0 - N_{011}$, $N_1 - N_{111}$) and the marginal outcome totals are ($M_{00} + M_{10}$, M_{01}). The cell corresponding to the case total in the $X/Z = 0/1$ exposure group takes on values in the range

$$N_{101}|N_{111} \in [\max(0, (N_1 - N_{111}) - (M_{00} + M_{10})),\ \min(N_1 - N_{111}, M_{01})]. \tag{A.3}$$

For a given value of N_{011} in this range (and $N_{001} = M_{01} - N_{101}$), there are two remaining exposure levels. The cell corresponding to the case total in the $X/Z = 1/0$ exposure group, in the corresponding 2×2 table, takes on values in the range

$$N_{110}|N_{111}, N_{101} \in [\max(0, (N_1 - N_{111} - N_{101}) - M_{00}),\ \min(N_1 - N_{111} - N_{101}, M_{10})]. \tag{A.4}$$

Finally, the space $\mathcal{R}_N(\mathbf{M_{xz}})$ is taken to be the recursive product of these three ranges.

Haneuse and Wakefield (2006) introduce the univariate supplemented extended hypergeometric distribution, where a single 2×2 is supplemented with case-control data under a hybrid sampling scheme. In the setting of Section 3, the multivariate analogue has probability mass function

$$P(\mathbf{N_{yxz}}|\ \mathbf{M_{xz}}, \theta) \ = \ \frac{W(\mathbf{n_{yxz}}|\mathbf{N_{yxz}})\mathrm{MXHG}(\mathbf{N_{yxz}}|\ \mathbf{M_{xz}})}{\sum\limits_{\mathbf{u} \in \mathcal{R}_N(\mathbf{Y}^*, \mathbf{M_{xz}})} W(\mathbf{n_{yxz}}|\mathbf{u})\mathrm{MXHG}(\mathbf{u}|\ \mathbf{M_{xz}})} \tag{A.5}$$

where the weights are given by (6). An expression for the space $\mathcal{R}_N(\mathbf{Y}^*, \mathbf{M_{xz}})$ may be obtained in the same recursive way as above, with the addition of the case-control data modifying each of the components.

Sampling from this distribution follows from an approach developed for the extended hypergeometric distribution in the 2×2 case (Liao and Rosen, 2001). When supplemental case-control data are available, only minor modifications are required. Sampling from the multivariate version follows from the above recursive partitioning of the 4×2 table into a series of 2×2 tables, and applying the methods of Liao and Rosen.

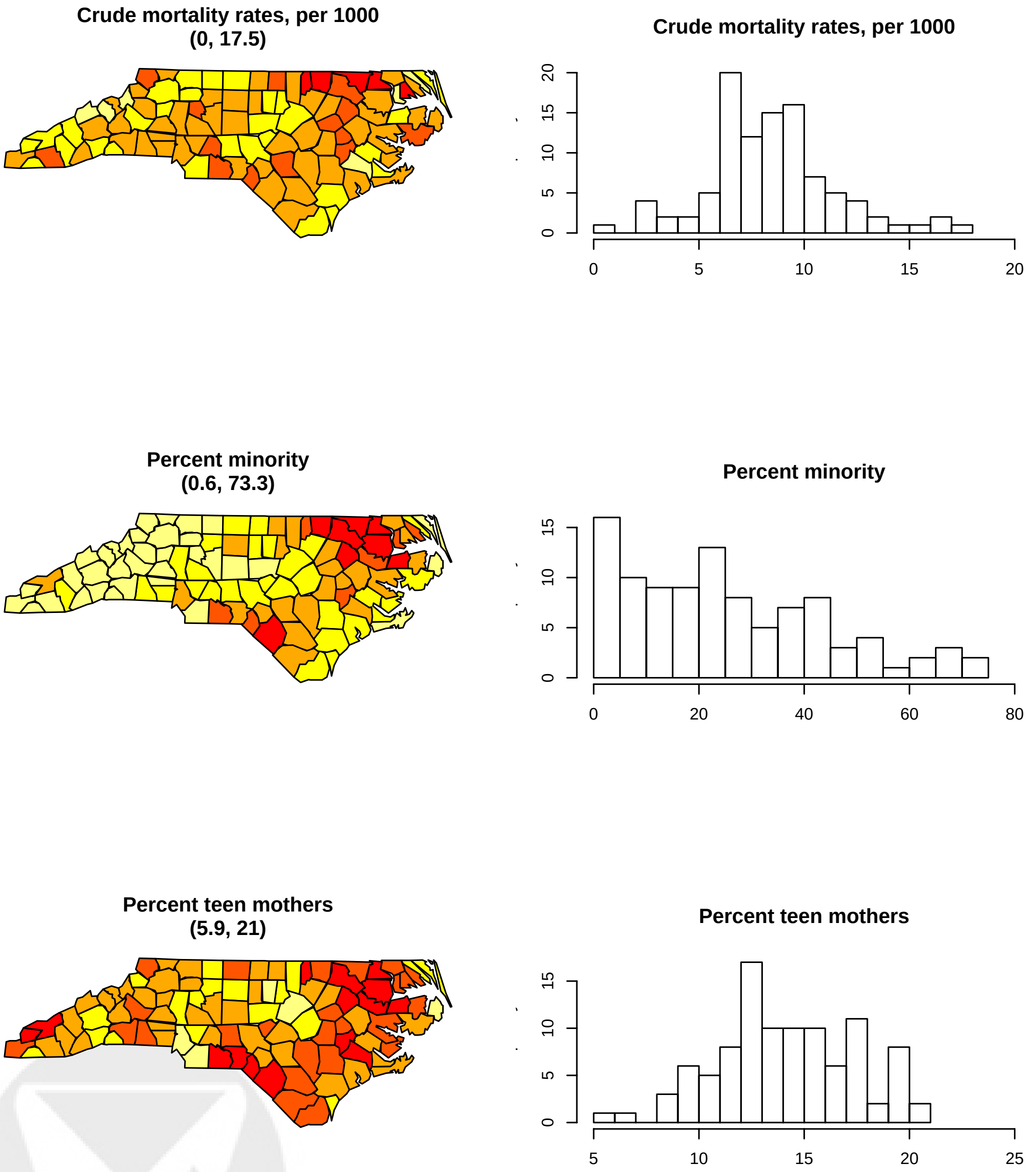

Figure 1. Crude infant mortality death rates (×1,000), percent minority and percent teen mother for 100 counties in the state of North Carolina.

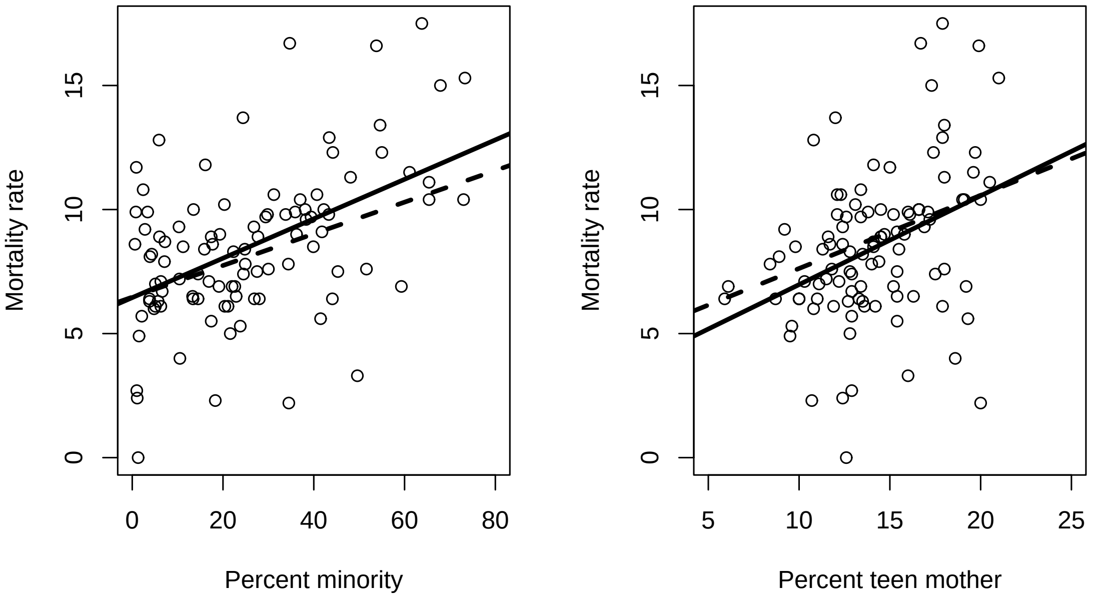

Figure 2. Ecological correlations of percent minority and percent teen mother versus crude mortality rates. Plotted line indicates least-squares fit; solid = ordinary LS and dashed = weight (by county-specific birth totals) LS.

Table 1

Notation for ecological and case-control data in a generic area

	Covariate				Ecological				Case-control		
	X=0	X=1		X/Z	Y=0	Y=1		X/Z	Y=0	Y=1	
Z=0			M_{+0}	0/0		$[N_{100}]$	$[M_{00}]$	0/0	n_{000}	n_{100}	m_{00}
Z=1		$[M_{11}]$	M_{+1}	1/0		$[N_{110}]$	$[M_{10}]$	1/0	n_{010}	n_{110}	m_{10}
	M_{0+}	M_{1+}	N	0/1		$[N_{101}]$	$[M_{01}]$	0/1	n_{001}	n_{101}	m_{01}
				1/1		$[N_{111}]$	$[M_{11}]$	1/1	n_{011}	n_{111}	m_{11}
					N_0	N_1	N		n_0	n_1	n

Table 2

Posterior summaries for the North Carolina infant mortality data. DesignA: individual-level data consist of 20 case-control samples from each of 100 counties; total n_1 = 885 and total n_0 = 1115. Design B: individual-level data consist of 100 case-control samples from each of 9 counties with at least 100 cases; total n_1 = 450 and total n_0 = 1115.

		Median (95% central credible interval)		
	Minority main effect, θ_{X}	Teen mother main effect, θ_{Z}	Interaction θ_{XZ}	Minority/Teen mother odds ratio, ϕ_{XZ}
Design A				
Individual-level	2.49 (2.33, 2.65)	1.56 (1.40, 1.74)	0.61 (0.53, 0.71)	2.03 (2.00, 2.07)
Case-control only	2.13 (1.69, 2.68)	1.34 (0.96, 1.89)	0.77 (0.45, 1.34)	-
Hybrid; $\mathbf{M}_{\mathbf{xz}}$				
Cases only	2.34 (2.04, 2.70)	1.60 (1.26, 2.00)	0.67 (0.47, 0.94)	2.03 (2.00, 2.07)
Case-control	2.34 (2.03, 2.70)	1.59 (1.26, 1.99)	0.67 (0.47, 0.94)	2.03 (2.00, 2.07)
Hybrid; $\{\mathbf{M}_{\mathbf{x}+}, \mathbf{M}_{+\mathbf{z}}\}$				
Cases only	2.34 (2.03, 2.71)	1.60 (1.25, 2.01)	0.66 (0.45, 0.98)	2.00 (1.62, 2.50)
Case-control	2.31 (1.99, 2.66)	1.54 (1.22, 1.94)	0.73 (0.51, 1.04)	1.79 (1.59, 2.01)
Design B				
Individual-level	2.49 (2.34, 2.65)	1.56 (1.40, 1.73)	0.61 (0.53, 0.72)	2.03 (2.00, 2.07)
Case-control only	2.91 (2.18, 3.91)	2.53 (1.42, 4.59)	0.47 (0.21, 1.06)	-
Hybrid; $\mathbf{M}_{\mathbf{xz}}$				
Cases only	2.59 (2.19, 3.08)	2.13 (1.42, 3.04)	0.50 (0.30, 0.83)	2.03 (2.00, 2.07)
Case-control	2.60 (2.18, 3.08)	2.12 (1.42, 3.08)	0.50 (0.30, 0.83)	2.03 (2.00, 2.07)
Hybrid; $\{\mathbf{M}_{\mathbf{x}+}, \mathbf{M}_{+\mathbf{z}}\}$				
Cases only	2.52 (2.08, 3.07)	1.98 (1.25, 3.11)	0.60 (0.28, 1.24)	1.57 (0.94, 2.70)
Case-control	2.55 (2.13, 3.07)	2.06 (1.33, 3.09)	0.54 (0.29, 1.04)	1.69 (1.13, 2.40)