Methods for Effectively Combining Group- and Individual-Level Data

Methods for Effectively Combining Group- and Individual-Level Data

A DISSERTATION PRESENTED
BY
ELIZABETH SMOOT
TO
THE DEPARTMENT OF BIOSTATISTICS

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN THE SUBJECT OF
BIOSTATISTICS

HARVARD UNIVERSITY
CAMBRIDGE, MASSACHUSETTS
FEBRUARY 2015

Methods for Effectively Combining Group- and Individual- Level Data

Abstract

In observational studies researchers often have access to multiple sources of information but ultimately choose to apply well-established statistical methods that do not take advantage of the full range of information available. In this dissertation I discuss three methods that are able to incorporate this additional data and show how using each improves the quality of the analysis.

First, in Chapters 1 and 2, I focus on methods for improving estimator efficiency in studies in which both population (group) and individual-level data is available. In such settings, the hybrid design for ecological inference efficiently combines the two sources of information; however, in practice, maximizing the likelihood is often computationally intractable. I propose and develop an alternative, computationally efficient representation of the hybrid likelihood. I then demonstrate that this approximation incurs no penalty in terms of increased bias or reduced efficiency.

Second, in Chapters 3 and 4, I highlight the problem of applying standard analyses to outcome-dependent sampling schemes in settings in which study units are cluster-correlated. I demonstrate that incorporating known outcome totals into the likelihood via inverse probability weights results in valid estimation and inference. I further discuss the applicability of outcome-dependent sampling schemes in resource-limited settings, specifically to the analysis of national ART programs in sub-Saharan Africa. I propose the cluster-stratified case-control study as a valid and logistically reasonable study design in such resource-poor settings, discuss balanced versus unbalanced sampling techniques, and address the practical trade-off between logistic considerations and statistical efficiency of cluster-stratified case-control versus case-control studies.

Finally, in Chapter 5, I demonstrate the benefit of incorporating the full-range of possible outcomes into an observational data analysis, as opposed to running the analysis on a pre-selected set of outcomes. Testing all possible outcomes for associations with the exposure inherently incorporates negative controls into the analysis and further validates a study's statistically significant results. I apply this technique to an investigation of the relationship between particulate air pollution and hospital admission causes.

Contents

1 On the Analysis of Hybrid Designs that Combine Group- and Individual-Level Data **1**

1.1 Introduction . . . 2
1.2 The Hybrid Design . . . 3
1.3 Computational Burden . . . 9
1.4 Approximating the Hybrid Likelihood . . . 10
1.5 Simulations . . . 14
1.6 Application . . . 17
1.7 Discussion . . . 21

2 Supplementary Material for: On the Analysis of Hybrid Designs that Combine Group- and Individual-Level Data **23**

2.1 Enumerating the vectors $\mathbf{N}_{\mathbf{yrs}k}$ in $\mathcal{N}_k$ . . . 23
2.2 Enumerating the vectors $\mathbf{M}_{\mathbf{rs}k}$ in $\mathcal{M}_k$ . . . 24
2.3 Approximations . . . 24
2.4 Inference . . . 24
2.5 Operating characteristics of β_2 . . . 30
2.6 Intercept Point Estimates . . . 34
2.7 Upper bound on the number of vectors $\mathbf{N}_{\mathbf{yrs}k}$ in $\mathcal{N}_k$ . . . 36
2.8 Application . . . 36

3 On the Analysis of Case-Control and Stratified Case-Control Studies in Cluster-Correlated Data Settings **38**

3.1 Introduction . . . 39
3.2 The complete data setting . . . 41
3.3 Outcome-dependent sampling . . . 42
3.4 Simulation I . . . 44
3.5 Simulation II . . . 49

3.6 Discussion 52

4 Supplementary Material for: On the Analysis of Case-Control and Stratified Case-Control Studies in Cluster-Correlated Data Settings **59**
4.1 Estimator Consistency 60
4.2 Generating group sizes 61
4.3 Operating characteristics 62

5 Short Term Exposure to Low Levels of Fine Particulate Matter and Hospital Admissions in Older Adults **66**
5.1 Background 67
5.2 Methods 68
5.3 Results 70
5.4 Discussion 72

References **80**

Author List

The following author contributed to Chapter 1: S. Haneuse.
The following author contributed to Chapter 2: S. Haneuse.
The following authors contributed to Chapter 3: J. Bobb and F. Dominici.

Listing of figures

1.2.1 Visual representations of group-level, aggregated information derived from the North Carolina birth weight data. Panels (a) and (b) together represent observed information in an aggregate data study. Panels (a), (c) and (d) collectively represent observed information in a pure ecological study. 5

1.4.1 Implementation of the compromise strategy of Section 1.4.4 which balances the use of the exact and approximate forms of the hybrid aggregate data likelihood and computational burden. Point and standard error estimates are for β_1 in model (1.1). 14

2.6.1 Point estimates of the K group-specific intercepts, β_{0k}, in the simulations described in Section 5. Shown are estimates based on the full data likelihood as well as on the exact hybrid likelihood and binomial approximate hybrid likelihood for a single draw of a hybrid aggregate data design. 35

2.8.1 Point estimates of the K=100 county-specific intercepts, β_{0k}, in the model described in Section 6. Shown are estimates based on the full data likelihood as well as on the binomial approximate hybrid likelihood for a single draw of a hybrid aggregate data design. 37

3.4.1 Distribution of clinic sizes in Malawi data. 46

4.2.1 Group sizes generated using Gamma(2, 0.5) and Gamma(0.5, 2) distributions (simulations (S1) and (S2), respectively). 61

5.2.1 Medicare enrollment for the continental-US counties included in the study overlaid with locations of PM2.5 monitoring stations within each county. 75

5.2.2 Number of days with complete $PM_{2.5}$ and temperature data by county, for the full dataset and three restricted, low-level pollution datasets. 76

5.3.1 Yearly averages of daily mean $PM_{2.5}$ measurements, by county. 76

5.3.2 Point estimates and 95% confidence intervals (CI^a) of the national average log relative risk associated with a $10\mu g/m^3$ increase in mean daily $PM_{2.5}$. Results are shown for the thirty most common diagnoses at hospitalization. Solid/open circles represent statistically significant/insignificant results, respectively. [a] The Bonferroni correction method is used to adjust CI for multiple comparisons. 77

5.3.3 Point estimates and 95% confidence intervals (CI^a) of the national average log relative risk associated with a $1\mu g/m^3$ increase in mean daily $PM_{2.5}$ from the two-stage BHM approach, for days having values of $PM_{2.5} < (1)\ 20\mu g/m^3$, $(2)\ 15\mu g/m^3$, $(3)\ 10\mu g/m^3$. Results are shown for the thirty most common diagnoses at hospitalization. Solid/open circles represent statistically significant/insignificant results, respectively. [a] The Bonferroni correction method is used to adjust CI for multiple comparisons. 78

Acknowledgments

I WOULD LIKE TO THANK my thesis advisors Prof. Sebastien Haneuse and Prof. Francesca Dominici and committee member Prof. Brent Coull for their guidance and support throughout the dissertation process, with a special thanks to Prof. Francesca Dominici for financially supporting my research. I would also like to thank Dr. Jennifer Bobb for her feedback and tutelage on my research projects and Dr. Gregory Malecha for his patient assistance with assorted C++ programming questions. I would like to thank the biostatistics department staff and in particular Jelena Follweiler, Vickie Beaulieu, and Phoebe Hackett for sharing their invaluable 'how-to' knowledge.

If you have all the details of a thousand at your finger ends,
it is odd if you can't unravel the thousand and first

Arthur Conan Doyle, *A Study in Scarlet*

1

On the Analysis of Hybrid Designs that Combine Group- and Individual-Level Data

Abstract

Ecological studies that make use of data on groups of individuals, rather than on the individuals themselves, are subject to numerous biases that cannot be resolved without some individual-level data. In the context of a rare outcome, the hybrid design for ecological inference efficiently combines group-level data with individual-level case-control data. Unfortunately, except in relatively simple settings, use of the design in practice is limited since evaluation of the hybrid likelihood is computationally prohibitive expensive. In this paper we first propose and develop an alternative representation of the hybrid likelihood. Second, based on this new representation, a series of approximations are proposed that drastically reduce computational burden. A comprehensive simulation shows that, in a broad range of scenarios, estimators based on the approximate hybrid likelihood exhibit the same operating characteristics as the exact hybrid likelihood, without any penalty in terms of increased bias or reduced efficiency. Third, in settings where the approximations may not hold, a pragmatic estimation and inference strategy is developed that uses the approximate form for some likelihood contributions and the exact form for others. The strategy gives researchers the ability to balance computational tractability with accuracy in their own settings. Finally, as a by-product of the development, we provide the first explicit characterization of the hybrid aggregate data design which

combines data from an aggregate data study (Prentice and Sheppard, 1995) with case-control samples. The methods are illustrated using data from North Carolina on births between 2007 and 2009.

1.1 Introduction

As researchers plan and conduct studies they have at their disposal a broad range of designs on which to base their data collection efforts. Typically, research studies have well-defined study units and data is collected on a sub-sample of individual units. In some settings individual-level data may not be readily-available and researchers may only have access to aggregated data on groups of individuals. When data is solely available on groups of individuals, the resulting study is commonly referred to as an ecological study [70]. With the increasing ubiquity of large administrative databases, ecological studies are often cheaper to conduct than individual-level cohort and case-control study counterparts and can also, in some cases, provide greater exposure variability and therefore greater statistical power [59]. Recent prominent examples of ecological studies in the literature include studies of the impact of air pollution on life expectancy in the U.S. [55] and China [15].

Despite the benefits, ecological studies suffer from numerous sources of bias in which the observed group-level exposure-outcome association does not accurately reflect the exposure-outcome association at the individual-level [27, 28, 56, 61, 65, 71]. Collectively, the impact of these biases is often referred to as 'ecological bias'; in the most severe case, the 'ecological fallacy' arises where conclusions drawn about the exposure-outcome association differ from those that would have been drawn had an individual-level study been conducted [52, 64, 76].

Unfortunately, any attempt to draw conclusions regarding individual-level associations solely using group-level data relies on untestable assumptions in one form or another [31]. Consequently, when scientific interest lies in individual-level associations, the only reproducible approach to avoiding ecological bias is to collect, incorporate and analyze individual-level data [30]. Over the last 20 years a number of statistical designs/methods have been proposed that combine group- and individual-level data, including hierarchical regression [26, 79], aggregate data methods [43, 44, 59], two-phase designs [6, 75, 77] and the hybrid design for ecological inference [30]. Although details differ across the designs/methods, each: (i) uses individual-level data to mitigate ecological bias, and (ii) takes advantage of the group-level data (i.e. the large sample sizes and exposure variability) to provide efficiency and power gains over designs/methods based solely on individual-level data.

In the context of a rare binary outcome, Haneuse and Bartell [29] show that the hybrid design for ecological inference provides the greatest potential for statistical efficiency. In its most general form, the hybrid design supplements group-level data with individual-level case-control data. The exact nature of the group- and individual-level data may vary, depending on the type of observed group-level information as well as on the case-control sampling scheme across the groups. The superior efficiency properties arise in part due to

the design (i.e. the case-control sampling when the outcome is rare) as well as due to estimation/inference being likelihood-based. Unfortunately, however, evaluation of the hybrid likelihood is computationally very expensive. Indeed, when the model of interest considers more than 2 or 3 risk factors the computational burden may be sufficiently prohibitive that, in practice, researchers could be tempted to simply analyze the individual-level data and forgo the efficiency gains provided by incorporating the group-level data in the analysis.

In this paper we propose a novel approach for analyzing data from the hybrid design. Towards this we first develop an alternative representation of the hybrid likelihood. We then show that much, if not all, of the computational burden can be attributed to one component of the new decomposition. A series of approximations for this component are proposed. We show that estimation/inference based on the approximate hybrid likelihood exhibits the same operating characteristics as that based on the exact hybrid likelihood while simultaneously drastically reducing computational burden. In settings where the approximations may not hold, a pragmatic strategy that balances the use of the exact and approximate hybrid likelihood representations is developed. To illustrate the ideas, concepts and methods of this paper we use data on all births in North Carolina from 2007-2009. These rich data are collected by The North Carolina State Center for Health Statistics, and are publicly available through the Odum Institute at the University of North Carolina (`http://arc.irss.unc.edu/`).

The remainder of this paper is as follows. In Section 1.2 we introduce notation and present the hybrid design and standard form of the likelihood for two general data settings: (i) a hybrid design that supplements group-level data from an aggregate data study with case-control data, and (ii) a hybrid design that supplements a pure ecological study with case-control data. To our knowledge, the hybrid aggregate data design has not been explicitly considered in the literature; the most closely-related design is the integrated aggregate data design of Martínez et al. [43, 44] which supplements an aggregate data study with a random sample of (prospectively collected) individual-level data. Section 1.3 then provides a brief demonstration of the computational burden associated with the hybrid likelihood. A novel representation of the hybrid likelihood is derived in Section 1.4, along with a series of approximations aimed at reducing computational burden. Section 3.4 presents a simulation study investigating the performance of the proposed analysis approach in a broad range of settings, and Section 1.6 provides a detailed illustration using the North Carolina birth data. Finally, the paper concludes with a discussion in Section 3.6.

1.2 The Hybrid Design

To ground the notation and exposition, consider the relationship between the risk of low birth weight (defined as a birth weight of $< 2{,}500$g) and two risk factors: the race of the baby and whether or not the mother smoked. Throughout, while numerous choices are possible, we take the births to be 'grouped' by county; in North Carolina there are K=100 counties.

1.2.1 Notation

Let R be a binary indicator of race (0/1 = white/non-white), S an indicator of whether or not the mother smoked during pregnancy (0/1 = no/yes) and Y an indicator of low birth weight status (0/1 = no/yes). Suppose interest lies in the following individual-level logistic regression model:

$$\text{logit}\, P(Y_{ki} = 1 | R_{ki}, S_{ki}) \;=\; \beta_{0k} + \beta_1 R_{ki} + \beta_2 S_{ki}, \tag{1.1}$$

where the subscript $[ki]$ indicates the i^{th} birth in the k^{th} county, for i=1, ..., N_k and k=1, ..., K. Note, model (1.1) is an individual-level model in the sense that it considers the relationship between risk factors and an outcome jointly measured on each individual birth [71]. As such, the log odds ratios β_1, and β_2 are interpreted as characterizing individual-level associations. To complete the notation, let M_{rsk} denote the number of births in the $[R, S]$=$[r, s]$ race/smoking stratum of the k^{th} county and N_{yrsk} the corresponding total number of births with Y=y.

1.2.2 A complete individual-level study

Suppose complete individual-level data is observed on all N=$\sum_{k=1}^{K} N_k$ individuals from all K groups. That is, suppose the collections $\mathbf{N_{yrs}}_k = \{N_{yrsk}; y = 0/1, r = 0/1, s = 0/1\}$ and $\mathbf{M_{rs}}_k$=$\{M_{rsk}; r = 0/1, s = 0/1\}$ are observed for each group. The top panel of Table 1.2.1 provides a summary of the notation for this data scenario. Assuming independence across groups, estimation and inference for $\beta = (\beta_{01}, \ldots, \beta_{0K}, \beta_1, \beta_2)$ could proceed straightforwardly using the following individual-level binomial likelihood:

$$\begin{aligned} L^I(\beta; \mathbf{N_{yrs}}) &= \prod_{k=1}^{K} L^I(\beta; \mathbf{N_{yrs}}_k | \mathbf{M_{rs}}_k) \\ &= \prod_{k=1}^{K} \left\{ \prod_{r=0}^{1} \prod_{s=0}^{1} \binom{M_{rsk}}{N_{1rsk}} \pi_{rsk}^{N_{1rsk}} (1 - \pi_{rsk})^{M_{rsk} - N_{1rsk}} \right\}, \end{aligned} \tag{1.2}$$

where $\mathbf{N_{yrs}}$ denotes the collection of $R/S/Y$ counts across all K groups, $\{\mathbf{N_{yrs}}_k; k = 1, \ldots, K\}$, and $\pi_{rsk} \equiv \pi_{rsk}(\beta) = P(Y = 1 | R = r, S = s, \text{Group} = k)$ is given by model (1.1). Note, one could in practice adopt additional structure on the K group-specific β_{0k} intercepts, for example assuming that they arise from some common random effects distribution which may or may not exhibit some specific spatial structure [32]. For ease of presentation, we assume that the intercept parameters are estimated without any such structure.

Figure 1.2.1: Visual representations of group-level, aggregated information derived from the North Carolina birth weight data. Panels (a) and (b) together represent observed information in an aggregate data study. Panels (a), (c) and (d) collectively represent observed information in a pure ecological study.

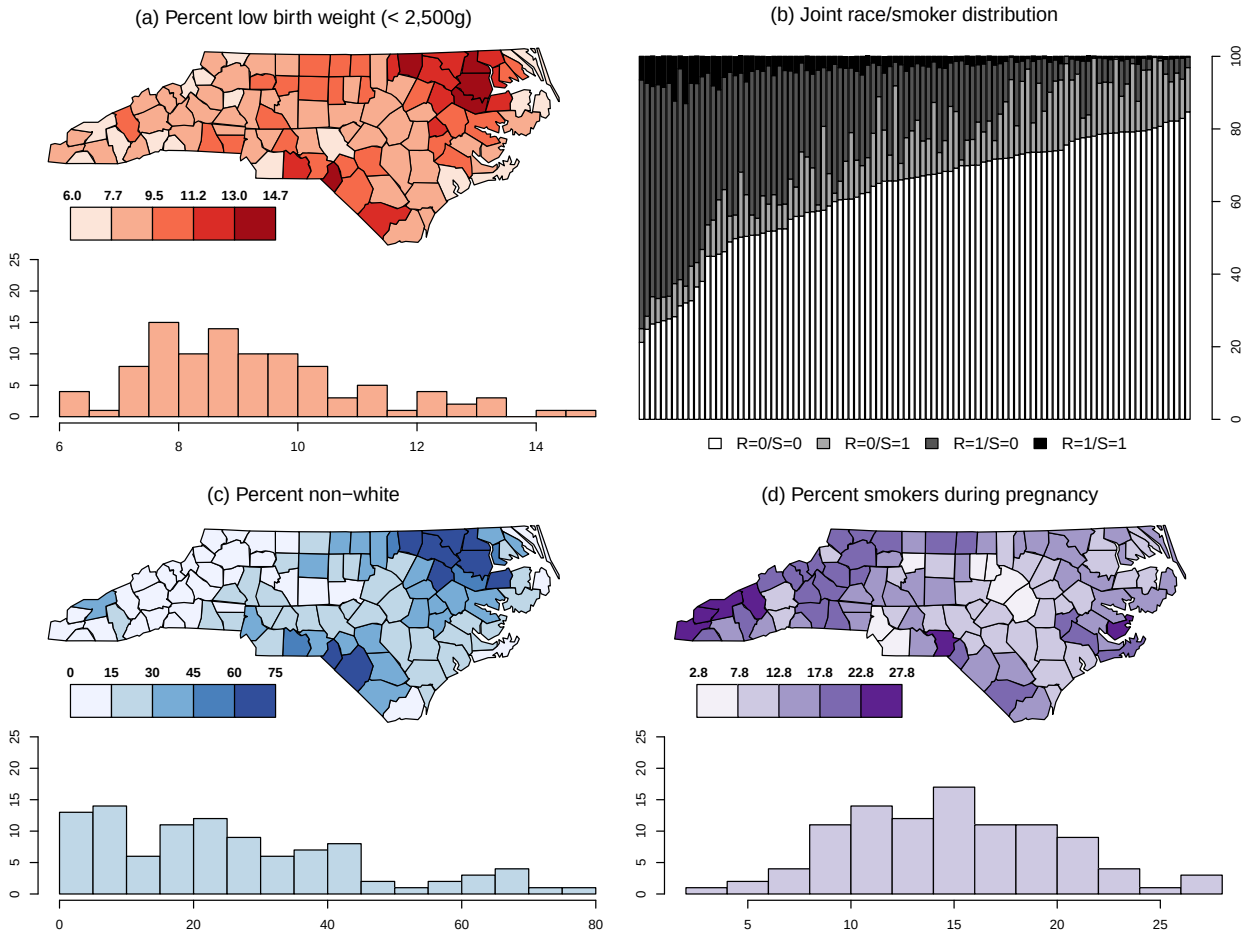


Table 1.2.1: Notation for data available under three data scenarios/designs. Shown are counts for a generic group, k. Counts within square brackets are not observed in the respective design.

I. Complete individual-level data

	Y=0	Y=1	
R=0/S=0	N_{000k}	N_{100k}	M_{00k}
R=0/S=1	N_{001k}	N_{101k}	M_{01k}
R=1/S=0	N_{010k}	N_{110k}	M_{10k}
R=1/S=1	N_{011k}	N_{111k}	M_{11k}
	N_{0k}	N_{1k}	N_k

II. Aggregate data study supplemented with case-control data

	Y=0	Y=1	
R=0/S=0	$[N_{000k}]$	$[N_{100k}]$	M_{00k}
R=0/S=1	$[N_{001k}]$	$[N_{101k}]$	M_{01k}
R=1/S=0	$[N_{010k}]$	$[N_{110k}]$	M_{10k}
R=1/S=1	$[N_{011k}]$	$[N_{111k}]$	M_{11k}
	N_{0k}	N_{1k}	N_k

	Y=0	Y=1	
R=0/S=0	n_{000k}	n_{100k}	m_{00k}
R=0/S=1	n_{001k}	n_{101k}	m_{01k}
R=1/S=0	n_{010k}	n_{110k}	m_{10k}
R=1/S=1	n_{011k}	n_{111k}	m_{11k}
	n_{0k}	n_{1k}	n_k

III. Pure ecological study supplemented with case-control data

	S=0	S=1	
R=0	$[M_{00k}]$	$[M_{01k}]$	M_{0+k}
R=1	$[M_{10k}]$	$[M_{11k}]$	M_{1+k}
	M_{+0k}	M_{+1k}	N_k

	Y=0	Y=1	
R=0/S=0	$[N_{000k}]$	$[N_{100k}]$	$[M_{00k}]$
R=0/S=1	$[N_{001k}]$	$[N_{101k}]$	$[M_{01k}]$
R=1/S=0	$[N_{010k}]$	$[N_{110k}]$	$[M_{10k}]$
R=1/S=1	$[N_{011k}]$	$[N_{111k}]$	$[M_{11k}]$
	N_{0k}	N_{1k}	N_k

	Y=0	Y=1	
R=0/S=0	n_{000k}	n_{100k}	m_{00k}
R=0/S=1	n_{001k}	n_{101k}	m_{01k}
R=1/S=0	n_{010k}	n_{110k}	m_{10k}
R=1/S=1	n_{011k}	n_{111k}	m_{11k}
	n_{0k}	n_{1k}	n_k

1.2.3 Supplementing an aggregate data design study with case-control data

In the absence of complete individual-level data, researchers may nevertheless have access to counts aggregated at the group-level. Under the aggregate data design, these data consist of the group-specific marginal outcome counts $\mathbf{N}_{\mathbf{y}k} = \{N_{0k}, N_{1k}\}$ together with the group-specific marginal covariate counts $\mathbf{M}_{\mathbf{rs}k}$. Consequently, while 'complete' information on the outcomes is observed along with 'complete' information on the marginal covariate counts, their joint distribution is not observed. In a hybrid aggregate design, these data are supplemented with a case-control sample of n_{0k} non-cases and n_{1k} cases drawn from the k^{th} group; on each of the n_k=n_{0k}+n_{1k} individuals sampled in this scheme, complete information on the joint distribution of $R/S/Y$ is retrospectively observed. The middle panel of Table 1.2.1 provides a summary of the notation for this data scenario. Note, the N_{yrsk} are within square brackets to emphasize that they are not observed.

Since complete individual-level data is not observed, estimation/inference cannot proceed using the likelihood given by (1.2). Instead, one can use the induced hybrid likelihood given by:

$$\begin{aligned} L^A(\beta; \mathbf{N}_{\mathbf{y}}, \mathbf{n}_{\mathbf{yrs}}) &= \prod_{k=1}^{K} L^A(\beta; \mathbf{N}_{\mathbf{y}k}, \mathbf{n}_{\mathbf{yrs}k} | \mathbf{M}_{\mathbf{rs}k}, \mathbf{n}_{\mathbf{y}k}) \\ &= \prod_{k=1}^{K} \sum_{\mathbf{N}_{\mathbf{yrs}k} \in \mathcal{N}_k} w(\mathbf{N}_{\mathbf{yrs}k} | \mathbf{n}_{\mathbf{yrs}k}, \mathbf{n}_{\mathbf{y}k}) L^I(\beta; \mathbf{N}_{\mathbf{yrs}k} | \mathbf{M}_{\mathbf{rs}k}) \end{aligned} \tag{1.3}$$

Intuitively, the contribution from the k^{th} group is a weighted convolution of individual-level likelihood contributions, $L^I(\beta; \mathbf{N}_{\mathbf{yrs}k} | \mathbf{M}_{\mathbf{rs}k})$, integrating over the unknown $\mathbf{N}_{\mathbf{yrs}k}$ with weights given as the product of probability distribution functions from two multivariate hypergeometric distributions:

$$w(\mathbf{N}_{\mathbf{yrs}k} | \mathbf{n}_{\mathbf{yrs}k}, \mathbf{n}_{\mathbf{y}k}) = \mathrm{HG}(\mathbf{n}_{\mathbf{0rs}k} | \mathbf{N}_{\mathbf{0rs}k}, n_{0k}) \mathrm{HG}(\mathbf{n}_{\mathbf{1rs}k} | \mathbf{N}_{\mathbf{1rs}k}, n_{1k}).$$

The set $\mathcal{N}_k$ in expression (1.3) denotes the collection of $\mathbf{N}_{\mathbf{yrs}k}$ counts that are consistent with both the aggregated group-level data, $\mathbf{N}_{\mathbf{y}k}$ and $\mathbf{M}_{\mathbf{rs}k}$, and the sampled case-control data, $\mathbf{n}_{\mathbf{yrs}k}$. The specific form of $\mathcal{N}_k$ is given in Chapter 2.1.

1.2.4 Supplementing a pure ecological study with case-control data

In some settings, researchers may not have access to the observed joint distribution of the covariates, $\mathbf{M}_{\mathbf{rs}k}$. In particular, the observed data in a pure ecological study consists of marginal totals for Y, R and S across the K groups. Using the notation developed so far, this 'pure ecological' data consists of the county-specific counts $(N_k, N_{1k}, M_{1+k}, M_{+1k})$ where N_k is the total number of births, N_{1k} is the total number of low birth weight births, M_{1+k} is the total number of non-white births, and M_{+1k} is the total number of births to mothers who smoked during pregnancy. Under the hybrid design, these data are supplemented with detailed,

individual-level data on a case-control sample of n_{0k} non-cases and n_{1k} cases drawn from the k^{th} group. The lower panel of Table 1.2.1 provides a summary of the notation for this data scenario. Note, both the N_{yrsk} and the M_{rsk} are within square brackets to emphasize that they are not observed.

As with the hybrid design of Section 1.2.3, estimation and inference cannot proceed on the basis of the individual-level likelihood given by (1.2). Again, however, the induced hybrid likelihood can be derived as the product of K group-specific weighted convolutions. In addition to integrating over the unknown N_{yrsk}, as in the supplemented aggregate data design of Section 1.2.3, one also needs to integrate over the unknown M_{rsk}. The latter requires additional parameters specific to the joint distribution of the covariates; for the setting we consider here (i.e. two binary covariates), the log odds ratio association between S and R, denoted φ_{rs}, suffices. Since this parameter will, in general, be unknown, it must be jointly estimated along with the regression parameters of interest. The resulting induced hybrid likelihood is then given by:

$$\begin{aligned} L^H(\beta, \varphi_{rs}; \mathbf{N_y}, \mathbf{n_{yrs}}) &= \prod_{k=1}^{K} L^H(\beta, \varphi_{rs}; \mathbf{N_{y}}k, \mathbf{n_{yrs}}k | \mathbf{M_{r}}+k, \mathbf{M}_{+\mathbf{s}}k, \mathbf{n_{y}}k) \\ &= \prod_{k=1}^{K} \sum_{\mathbf{M_{rs}}k \in \mathcal{M}_k} P(\mathbf{M_{rs}}k | \mathbf{M_{r}}+k, \mathbf{M}_{+\mathbf{s}}k, \varphi_{rs}) \cdot L^A(\beta; \mathbf{N_{y}}k, \mathbf{n_{yrs}}k | \mathbf{M_{rs}}k, \mathbf{n_{y}}k) \\ &= \prod_{k=1}^{K} \sum_{\mathbf{M_{rs}}k \in \mathcal{M}_k} P(\mathbf{M_{rs}}k | \mathbf{M_{r}}+k, \mathbf{M}_{+\mathbf{s}}k, \varphi_{rs}) \cdot \\ &\quad \left\{ \sum_{\mathbf{N_{yrs}}k \in \mathcal{N}_k} w(\mathbf{N_{yrs}}k | \mathbf{n_{yrs}}k, \mathbf{n_{y}}k) L^I(\beta; \mathbf{N_{yrs}}k | \mathbf{M_{rs}}k) \right\}. \end{aligned} \tag{1.4}$$

In expression (1.4), $P(\mathbf{M_{rs}}k | \mathbf{M_{r}}+k, \mathbf{M}_{+\mathbf{s}}k, \varphi_{rs})$ is the probability distribution function of an extended hypergeometric distribution [30, 38]. Furthermore, $\mathcal{M}_k$ is the set of all possible configurations of the $\mathbf{M_{rs}}k$ counts that are consistent with both the $(\mathbf{M_{r}}+k, \mathbf{M}_{+\mathbf{s}}k)$ marginal totals and the case-control counts $\mathbf{m_{rs}}k$ in the lower panel of Table 1.2.1. The specific form of $\mathcal{M}_k$ is given in Chapter 2.2.

1.2.5 Case-control sample sizes

Finally, we note that both hybrid designs in Sections 1.3 and 1.4 provide considerable flexibility regarding the sampling of case-control data across the K groups. In particular, the case-control samples sizes (n_{0k}, n_{1k}) need not be the same across all groups and in some settings one might not collect case-control data in all groups (i.e. some of the n_{0k} and/or n_{1k} might be zero). For groups with no observed case-control data, the induced group-level or ecological likelihood contribution takes on a form similar to (1.4) but without

the additional weighting by the distribution of the observed case-control data:

$$L^E(\beta, \varphi_{rs}; \mathbf{N}_{\mathbf{y}k}|\mathbf{M}_{\mathbf{r}+k}, \mathbf{M}_{+\mathbf{s}k}) = \sum_{\mathbf{M}_{\mathbf{rs}k}\in\mathcal{M}_k} P(\mathbf{M}_{\mathbf{rs}k}|\mathbf{M}_{\mathbf{r}+k}, \mathbf{M}_{+\mathbf{s}k}, \varphi_{rs}) \left\{ \sum_{\mathbf{N}_{\mathbf{yrs}k}\in\mathcal{N}_k} L^I(\beta; \mathbf{N}_{\mathbf{yrs}k}|\mathbf{M}_{\mathbf{rs}k}) \right\}. \quad (1.5)$$

1.3 Computational Burden

From expressions (1.3) and (1.4), evaluation of the hybrid likelihood requires computing a product of summations with the number of terms in the summations determined by $\mathcal{N}_k$ for the hybrid aggregate data design and $(\mathcal{M}_k, \mathcal{N}_k)$ jointly for the hybrid pure ecological design. Here we illustrate the corresponding computational burden. Specifically, from the North Carolina State Center for Health Statistics, there were a total of 387,705 births with complete vital records in North Carolina during the three-year span from 2007-2009; here, 'complete' refers to the record having no missing data on birth-county, race, infant birth weight, and mother's smoking status. Across the 100 counties, the number of births ranged from 147 to 44,076 with a median of 1,981; only seven counties had more than 10,000 births recorded.

1.3.1 Hybrid aggregate data design

From Section 1.2.3, data observed in an aggregate data study consists of group-level outcome information and group-level information on the joint covariate distribution. The top row of Figure 1.2.1 provides a visual representation of this information from the North Carolina data. Specifically, the top-left panel presents the distribution of the marginal outcome rate (percent low birth weight) across the K=100 counties. Note, using the notation of Section 1.2, these rates are $N_{1k}/N_k \times 100\%$. Overall, the low birth rate in North Carolina from 2007-2009 was 9.1% (35,406/387,705); across the 100 counties, the rates varied from 6.0% to 14.7%. The top-right panel of Figure 1.2.1 provides the joint distribution of (R, S) across the 100 counties. Specifically, each vertical bar corresponds to a single county, with the four colors indicating the four possible levels that (R, S) can take; using the notation of Section 1.2, the length of the color-specific bars is calculated as $M_{rsk}/N_k \times 100\%$ for r=0/1 and s=0/1.

To illustrate the computational burden associated with evaluating the hybrid likelihood (1.3), we drew a single stratified random sample of n_{0k}=n_{1k}=25 non-cases and cases from each county. Given the observed group- and individual-level data, a single evaluation of the hybrid likelihood requires performing the total of $\sum_{k=1}^{100} \text{size}(\mathcal{N}_k) \approx 5 \times 10^9$ calculations. We have implemented the hybrid design/likelihood in R with C as the primary computational work engine. Although details are not presented, based on simulations run on an Apple iMac with a dual-core Intel Core i5 3.6GHz processor with 8GB RAM, running Mac OS X Lion, a single evaluation of the hybrid likelihood is estimated to take approximately 21.5 days.

1.3.2 Hybrid pure ecological design

As indicated in Section 1.2.4, the joint distribution of the covariates is not observed in a pure ecological study; that is, information represented by the top-right panel of Figure 1.2.1 would not be available. Instead only marginal information on R and S, separately, is observed; the lower two panels of Figure 1.2.1 provide a visual representation of this information. Specifically, the lower-left panel provides the distribution of the marginal non-white birth rates across the 100 counties, calculated as $M_{1+k}/N_k \times 100\%$. We see that these county-specific rates varied from 0.3% to 75.1%. The lower-right panel provides analogous information on the marginal rates of smoking during pregnancy, calculated $M_{+1k}/N_k \times 100\%$; these rates varied from 2.8% to 27.8%.

Supplementing these data with a single stratified random sample of n_{0k}=n_{1k}=25 non-cases and cases from each county, we estimated that a single evaluation of the hybrid likelihood (1.4) would require more than 5×10^{12} calculations. Although details are again omitted, using the same hardware/software as the previous section, we estimated that a single evaluation of the hybrid likelihood could take up to 60 years.

1.4 Approximating the Hybrid Likelihood

From Section 1.3 it is clear that basing estimation/inference on the exact hybrid likelihood is computationally prohibitively expensive, even when interest solely lies with two binary explanatory covariates. As such, given data from a hybrid design, analysts may be tempted to solely make use of the individual-level case-control data. For example, one could simply use conditional logistic regression to estimate the log odds ratio parameters in model (1.1). Doing so, however, ignores the observed group-level data and forgoes the efficacy benefits associated with including this information in the analysis. In this section we present a novel analysis strategy for data arising from the hybrid design that makes use of an approximation to the hybrid likelihood.

1.4.1 An alternative representation of the hybrid likelihood

Consider the data set-up of hybrid aggregate data design, given by the middle row of Table 1.2.1. As indicated in Section 1.2.3, the corresponding hybrid likelihood is obtained by integrating the expression for the complete data likelihood over the distribution of the unknown $\mathbf{N}_{\mathbf{yrs}k}$. The case-control data inform this distribution by restricting the range of admissible $\mathbf{N}_{\mathbf{yrs}k}$ as well as through the weighting terms $P(\mathbf{n}_{\mathbf{yrs}k}|\mathbf{N}_{\mathbf{yrs}k}, \mathbf{n}_{\mathbf{y}k})$. Rather than viewing the n_k=n_{0k}+n_{1k} case-control samples as a subset of the broader population, an alternative is to consider them as distinct from the $N_k^* = N_k - n_k$ individuals in the k^{th} group who were not sampled. Table 1.4.1 provides a summary of the notation with a superscript '*' indicating that the counts refer to individuals not sampled by the case-control scheme. Note, the right-hand

Table 1.4.1: Notation for an alternative representation of the data available under the hybrid aggregate data design of Section 1.2.3. Shown are counts for a generic group, k. Counts within square brackets are not observed.

Individuals not sampled

	Y=0	Y=1	
R=0/S=0	$[N^*_{000k}]$	$[N^*_{100k}]$	M^*_{00k}
R=0/S=1	$[N^*_{001k}]$	$[N^*_{101k}]$	M^*_{01k}
R=1/S=0	$[N^*_{010k}]$	$[N^*_{110k}]$	M^*_{10k}
R=1/S=1	$[N^*_{011k}]$	$[N^*_{111k}]$	M^*_{11k}
	N^*_{0k}	N^*_{1k}	N^*_k

Case-control sample

	Y=0	Y=1	
R=0/S=0	n_{000k}	n_{100k}	m_{00k}
R=0/S=1	n_{001k}	n_{101k}	m_{01k}
R=1/S=0	n_{010k}	n_{110k}	m_{10k}
R=1/S=1	n_{011k}	n_{111k}	m_{11k}
	n_{0k}	n_{1k}	n_k

table is unchanged from Table 1.2.1 while the left-hand table essentially summarizes group-level aggregated information on the individuals who were not sampled.

Based on this new data representation, the hybrid aggregate data likelihood contribution by the k^{th} group can be re-written as:

$$L^A(\beta; \mathbf{N}_{yk}, \mathbf{n}_{yrsk} | \mathbf{M}_{rsk}, \mathbf{n}_{yk}) \; = L^E(\beta; \mathbf{N}^*_{yk}) \frac{\mathrm{HG}(\mathbf{m}_{rsk} | \mathbf{M}_{rsk}, n_k)}{\mathrm{HG}(\mathbf{n}_{yk} | \mathbf{N}_{yk}, n_k)} L^I(\beta; \mathbf{n}_{yrsk} | \mathbf{m}_{rsk}). \tag{1.6}$$

where $L^I(\beta; \mathbf{n}_{yrsk} | \mathbf{m}_{rsk})$ is a (naïve) prospective likelihood contribution based on the case-control data and $L^E(\beta; \mathbf{N}^*_{yk})$ is an ecological likelihood for those individuals not sampled:

$$L^E(\beta; \mathbf{N}^*_{yk}) \; = \sum_{\mathbf{N}^*_{yrsk} \in \mathcal{N}^*_k} L^I(\beta; \mathbf{N}^*_{yrsk} | \mathbf{M}^*_{rsk}), \tag{1.7}$$

where $\mathcal{N}^*_k$ denotes the collection of $\mathbf{N}^*_{yrsk}$ that are consistent with the group-level data on those not sampled $(\mathbf{N}^*_{yk}, \mathbf{M}^*_{rsk})$. The weighting in expression (1.6) by the ratio of the two (multivariate) hypergeometric distributions serves to account for the case-control sampling scheme as well as the finite population sampling from the N_k individuals in the group.

1.4.2 Approximating the aggregate data hybrid likelihood

Inspection of expression (1.6) reveals that the primary source of computational burden is $L^E(\beta; \mathbf{N}^*_{yk})$. Towards mitigating computational burden we consider approximating this component. Following Wakefield [74], who considered approximating the ecological likelihood in the setting of a single binary covariate, we propose to approximate the $L^E(\beta; \mathbf{N}^*_{yk})$ by taking the total number of events in the k^{th} group, N^*_{1k}, to be

conditionally distributed according to one of three distributions:

$$N^*_{1k}|\mathbf{M}^*_{\mathbf{rs}k} \sim \text{Binomial}\left(N^*_k,\ \sum_{r,s}\frac{M^*_{rsk}}{N^*_k}\pi_{rsk}\right) \tag{1.8}$$

$$N^*_{1k}|\mathbf{M}^*_{\mathbf{rs}k} \sim \text{Normal}\left(\sum_{r,s}M^*_{rsk}\pi_{rsk},\ \sum_{r,s}M^*_{rsk}\pi_{rsk}(1-\pi_{rsk})\right) \tag{1.9}$$

$$N^*_{1k}|\mathbf{M}^*_{\mathbf{rs}k} \sim \text{Poisson}\left(\sum_{r,s}M^*_{rsk}\pi_{rsk}\right) \tag{1.10}$$

where, as in Section 1.2.2, $\pi_{rsk} \equiv \pi_{rsk}(\beta)$ is given by model (1.1). In each of the above, the component parameters are obtained via an application of the double expectation and variance formulae. Denoting the approximate ecological likelihood contribution corresponding to any of these three distributions by $\tilde{L}^E(\beta; \mathbf{N}^*_{\mathbf{y}k})$, an approximate aggregate data hybrid likelihood contribution for the k^{th} group is:

$$\tilde{L}^A(\beta; \mathbf{N}_{\mathbf{y}k}, \mathbf{n}_{\mathbf{yrs}k}|\mathbf{M}_{\mathbf{rs}k}, \mathbf{n}_{\mathbf{y}k}) = \tilde{L}^E(\beta; \mathbf{N}^*_{\mathbf{y}k})\frac{\text{HG}(\mathbf{m}_{\mathbf{rs}k}|\mathbf{M}_{\mathbf{rs}k}, n_k)}{\text{HG}(\mathbf{n}_{\mathbf{y}k}|\mathbf{N}_{\mathbf{y}k}, n_k)}L^I(\beta; \mathbf{n}_{\mathbf{yrs}k}|\mathbf{m}_{\mathbf{rs}k}). \tag{1.11}$$

Crucially, evaluation of (1.11) no longer requires summing over the, often very large, collection of possible $\mathbf{N}^*_{\mathbf{yrs}k}$. As such, the computational burden is essentially trivial.

1.4.3 Approximating the pure ecological hybrid likelihood

The form of the pure ecological hybrid likelihood for the k^{th} group, repeated here from Section 1.2.4 for convenience, is the summation of a series of nested summations:

$$L^H(\beta, \varphi_{rs};\ \mathbf{N}_{\mathbf{y}k}, \mathbf{n}_{\mathbf{yrs}k}|\mathbf{M}_{\mathbf{r}+k}, \mathbf{M}_{+\mathbf{s}k}, \mathbf{n}_{\mathbf{y}k}) = \sum_{\mathbf{M}_{\mathbf{rs}k}\in\mathcal{M}_k} P(\mathbf{M}_{\mathbf{rs}k}|\mathbf{M}_{\mathbf{r}+k}, \mathbf{M}_{+\mathbf{s}k}, \varphi_{rs}) \cdot \left\{\sum_{\mathbf{N}_{\mathbf{yrs}k}\in\mathcal{N}_k} w(\mathbf{N}_{\mathbf{yrs}k}|\mathbf{n}_{\mathbf{yrs}k}, \mathbf{n}_{\mathbf{y}k})L^I(\beta; \mathbf{N}_{\mathbf{yrs}k}|\mathbf{M}_{\mathbf{rs}k})\right\}.$$

Unfortunately, while the nested summation corresponds to $L^A(\beta; \mathbf{N}_{\mathbf{y}k}, \mathbf{n}_{\mathbf{yrs}k}|\mathbf{M}_{\mathbf{rs}k}, \mathbf{n}_{\mathbf{y}k})$ and can therefore be approximated using the approach of Section 1.4.2, the outer summation across these approximations is not amenable to approximation. Nevertheless, even if the approximation given by

$$\tilde{L}^H(\beta, \varphi_{rs}; \mathbf{N}_{\mathbf{y}k}, \mathbf{n}_{\mathbf{yrs}k}|\mathbf{M}_{\mathbf{r}+k}, \mathbf{M}_{+\mathbf{s}k}, \mathbf{n}_{\mathbf{y}k}) = \sum_{\mathbf{M}_{\mathbf{rs}k}\in\mathcal{M}_k} P(\mathbf{M}_{\mathbf{rs}k}|\mathbf{M}_{\mathbf{r}+k}, \mathbf{M}_{+\mathbf{s}k}, \varphi_{rs})\tilde{L}^A(\beta; \mathbf{N}_{\mathbf{y}k}, \mathbf{n}_{\mathbf{yrs}k}|\mathbf{M}_{\mathbf{rs}k}, \mathbf{n}_{\mathbf{y}k}) \tag{1.12}$$

does not completely eliminate the overall computational burden, that only a single (outer) summation is required will reduce it considerably.

1.4.4 Estimation and inference

Given group- and individual-level data from K groups in a hybrid design, likelihood-based estimation and inference could proceed using expression (1.11) or (1.12) in the usual way: a likelihood could be formed by taking the product of K terms of the form (1.11) or (1.12), point estimates can be obtained by maximization and standard error estimates from the inverse of the observed information matrix. Detailed expressions for the approximate hybrid likelihood scores and Hessian terms under both the hybrid aggregate data and hybrid pure ecological designs are derived in Chapter 2.3.

In practice, use of (1.11) or (1.12) for each of the K terms will lead to the greatest reduction of computational burden although doing so may incur a trade-off in terms of statistical operating characteristics. Since the ideal is to use the exact hybrid likelihood for each group, the extent to which use of the approximate hybrid likelihood impacts estimation and inference is crucial. When the group size is large each of (2.1)-(2.1) is readily-motivated as a large sample approximation to the distribution of the number of cases, N^*_{1k}. Thus, precisely in the situations where relief of the computational burden is most needed is where the approximations are expected to be most accurate. When group sizes are small, the approximations may not be expected to hold as well. However, for these groups the computational burden may be manageable. Together, these observations suggest use of the exact form for small groups and approximate form for large groups may strike a reasonable compromise between computational tractability and accuracy. One simple strategy is to sequentially obtain MLEs and standard error estimates using an overall likelihood where:

1. All K contributions are of the approximate form, $\tilde{L}^A(\beta; \mathbf{N}_{\mathbf{y}k}, \mathbf{n}_{\mathbf{yrs}k} | \mathbf{M}_{\mathbf{rs}k}, \mathbf{n}_{\mathbf{y}k})$.

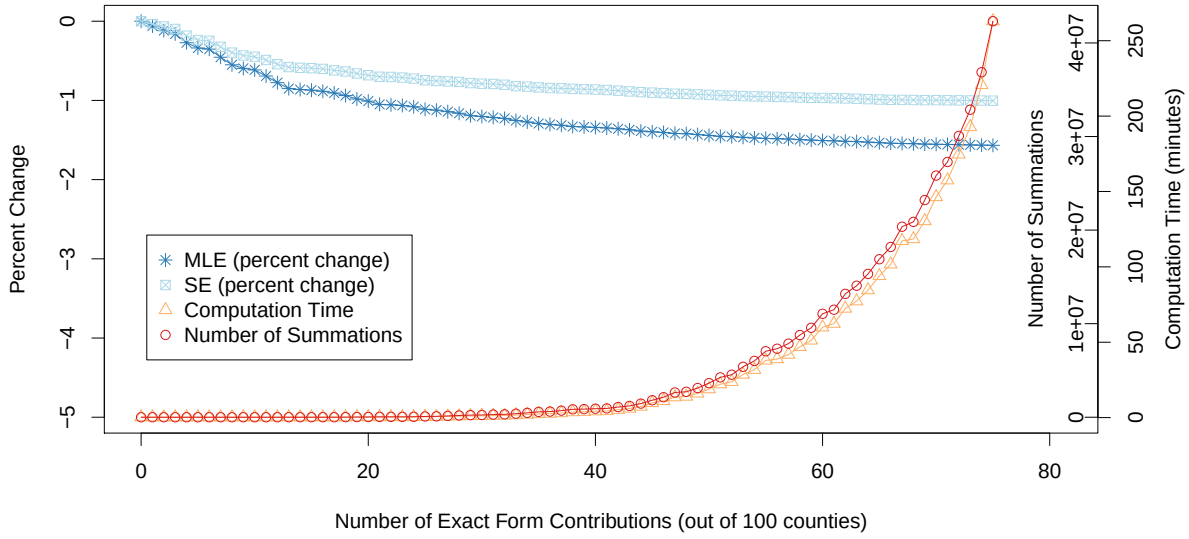
2. The group with the smallest N_{1k^*} contributes the exact form, while the remaining $K-1$ groups contribute the approximate form.

3. The two groups with the smallest N_{1k^*} contribute the exact form, while the remaining $K-2$ groups contribute the approximate form.

   ...

As one permits more and more of the contributions to be of the exact form, the computational burden will increase and the point estimates will get closer and closer to what one would have obtained by using the exact hybrid likelihood for all K contributions. Practically, one could initiate the process and stop when point estimates and standard error estimates 'converge', to some level of tolerance, in the sense that the use of additional exact likelihood contributions does not change the conclusions one draws.

Figure 1.4.1: Implementation of the compromise strategy of Section 1.4.4 which balances the use of the exact and approximate forms of the hybrid aggregate data likelihood and computational burden. Point and standard error estimates are for β_1 in model (1.1).



To illustrate this strategy, we drew a single stratified case-control sample of n_{0k}=n_{1k}=25 from each county in North Carolina and considered combining these data with data from the aggregate data design (i.e. that presented in Figures 1.2.1(a) and (b)). Figure 1.4.1 shows how the point estimates for β_1 in model (1.1) change as one modifies the balance of contributions that are of the approximate and exact form. Also shown is how the standard error estimates change, as well as the increase in computational burden in terms of the number of summations and the time taken to obtain the MLE; these were all evaluated using the same hardware/software configuration of Section 1.3. Overall, very little change is seen in the point estimates and standard error estimates; neither change by more than 2%. In contrast, the computation time quickly becomes onerous as the number of exact-form contributions is increased. In particular, the MLE computation time increased from 21 seconds when 20% of the counties were contributing the exact form of the likelihood to 1 hour and over 4 hours when the percent of counties contributing the exact form is increased to 60% and 75%, respectively. In the largest evaluation we performed, 75% of the counties contributed the exact form of the hybrid likelihood, requiring 4×10^7 summations per likelihood calculation and resulting in an MLE computation time of 4.3 hours.

1.5 Simulations

While the strategy presented in the previous section provides a pragmatic approach to using approximate and exact forms of the hybrid likelihood, the use of any approximate forms corresponds to a misspecified likelihood. As such, in contrast to estimation based on exact hybrid likelihood for all K groups, estimation

based on any of the approximations is no longer guaranteed to be consistent and/or asymptotically efficient. Furthermore, standard error estimates based on inverting the observed information matrix are not guaranteed to be valid. Consequently, use of the approximate form to mitigate computational burden may be subject to a trade-off in terms of statistical operating characteristics. To investigate this potential trade-off, we conducted a simulation study. Of specific interest are: (1) the magnitude of bias in point estimates, if any, associated with the use of the approximate likelihood and, (2) whether or not the use of the approximate likelihood impacts the efficiency gains one sees when one combines group- and individual-level data using the exact hybrid likelihood.

1.5.1 Simulation set-up

Towards addressing these questions, we initially generated 10,000 simulated datasets under the following 'baseline' scenario. For each of K=20 groups we set the group size to be N_k=2,000. Let $Q_{rk} = P(R = 1|\text{group } k)$ denote the marginal prevalence of a binary covariate R in the k^{th} group; similarly, let $Q_{sk} = P(S = 1|\text{group } k)$ denote the marginal prevalence of a binary covariate S in the k^{th} group. Values across the K groups for Q_{rk} and Q_{sk} were fixed at the quantiles of a Normal distribution with mean 0.2 and standard deviation 0.1; assignment to specific values for both Q_{rk} and Q_{sk} was randomly permuted across the 10,000 simulated datasets. Individual values for R and S were then generated as random deviates from Bernoulli(Q_{rk}) and Bernoulli(Q_{sk}), according to group membership of the individual. Given these covariate values, outcomes were generated as random draws from a Bernoulli(π_{rsk}) distribution with π_{rsk} given by model (1.1) with (β_1, β_2)=(log 1.5, log 1.25). The group-specific intercepts, β_{0k} were set such that the baseline outcome rates (i.e. π_{00k}, when R=S=0) were the quantiles from a Normal distribution with mean 0.1 and standard deviation 0.2.

We also considered six additional simulation scenarios, each modifying a single aspect of the data generating mechanism for the baseline scenario:

1. Increase the mean Q_{rk} across the K groups from 0.2 to 0.5.
2. Decrease the standard deviation of the between-group standard deviations Q_{rk} and Q_{sk} from 0.1 to 0.01.
3. Decrease the standard deviation of the between-group baseline outcome rate, π_{00k}, from 0.02 to 0.005.
4. Increase the log-odds ratio associations from (log 1.5, log 1.25) to (log 2.5, log 2.0)
5. Decrease the group sizes from N_k=2,000 to N_k=1,000 $\forall\, k$
6. Increase the number of groups from K=20 to K=40.

For each of the 10,000 simulated datasets, and under each of the 7 data scenarios, we computed aggregated totals for the outcomes and two covariates that would be observed under both an aggregate data design and a pure ecological study. We also drew a stratified random sample of n_{0k}=n_{1k}=25 non-cases and cases from each of the K groups.

1.5.2 Analyses

Combining these case-control samples with the aggregated totals simulated the hybrid aggregate data and hybrid pure ecological designs of Sections 1.2.3 and 1.2.4, respectively. For all datasets we estimated components of model (1.1) using (i) the full data estimator based on all $\sum_{k=1}^{K} N_k$ individuals, denoted $\hat{\beta}_{full}$, and (ii) the estimator obtained by performing conditional logistic regression on the (stratified) case-control data alone, denoted $\hat{\beta}_{CC}$. For simulated hybrid aggregate data designs, we considered an additional four estimators: (iii) the exact hybrid likelihood estimator, denoted $\hat{\beta}_{A}$; (iv) the approximate hybrid likelihood estimator based on the Binomial approximation, denoted $\tilde{\beta}_{A}^{Bin}$; (v) the approximate hybrid likelihood estimator based on the Normal approximation, denoted $\tilde{\beta}_{A}^{Nor}$; and, (vi) the approximate hybrid likelihood estimator based on the Poisson approximation, denoted $\tilde{\beta}_{A}^{Poi}$. Throughout, for each of the approximate hybrid likelihood estimators all K contributions were of the approximate form. For simulated hybrid pure ecological designs, we only considered the approximate hybrid likelihood estimators corresponding to (iv)-(vi) and denoted $\tilde{\beta}_{H}^{Bin}$, $\tilde{\beta}_{H}^{Nor}$ and $\tilde{\beta}_{H}^{Poi}$ respectively. The exact hybrid likelihood estimator for the pure ecological setting was not considered because of the prohibitive computational burden.

1.5.3 Results

Tables 2.5.1, 2.5.2 and 2.5.3 report operating characteristics for estimation of β_1, the log-odds ratio for race in model (1.1). Estimates of β_2, the log-odds ratio for smoking status, exhibited qualitatively similar operating characteristics; those results are provided in Web Tables D1-3.

From the upper portions of Tables 2.5.1 and Table 2.5.2, all of the estimators exhibited very low bias. Across all estimator/data scenarios, the greatest percent bias was only 2.3%. Perhaps most important for the methods of Section 1.4 is that none of the approximate hybrid likelihood estimators, under both the hybrid aggregate data and the hybrid pure ecological designs, exhibited any systematically greater bias than the exact hybrid likelihood estimator. From the middle and lower portions of Tables 2.5.1 and Table 2.5.2 we see that naïve standard error estimation for the approximate hybrid likelihood estimators was not subject to any systematic bias either. Specifically, the mean of the estimator standard error estimates based on the Hessians for the misspecified approximate likelihoods did not exhibit any systematic variation from the true standard error (calculated as the standard deviation of the 10,000 point estimates). Furthermore, in all simulation scenarios, 95% confidence intervals based on the naïve standard error estimates attained coverage probabilities very close to the nominal rate. Hence, despite the fact that the approximate hybrid

likelihood is a misspecified likelihood, estimation and inference remains valid in a broad range of data scenarios.

Finally, Table 2.5.3 reports on relative uncertainty defined as the ratio of the standard error for a given estimator to the standard error of the conditional logistic regression estimator $\hat{\beta}_{CC}$. From the first column, as expected, estimation of β_1 is substantially more efficient when the full data likelihood is used. Columns 4-6 indicate the efficiency gain associated with the combination of the group- and case-control data. Under the aggregate data scenarios considered, standard errors based on the combined data are approximately 75-80% those of the case-control data. Furthermore, there is no systematic detriment in this efficiency gain when one uses the approximate forms of the hybrid likelihood. Under the pure ecological scenarios we again see substantial efficiency gains associated with the combination of group- and case-control data. For these scenarios were were not able to evaluate the exact hybrid likelihood (due to the computational burden). As such, it is possible that analysts could enjoy even further gains through use of the exact form, although we believe the gains by having used the approximate form are important nonetheless.

1.6 Application

To further illustrate the utility of the methods in Section 1.4 we consider a more detailed analysis of risk factors for a low birth weight using the North Carolina data. Specifically, we expand on model (1.1) by considering three additional covariates: whether or not the birth was premature, defined as a birth at 37 weeks; plurality, taking on levels of a singleton birth, twins or triplets or more; whether or not the mother experienced a low weight gain during the pregnancy, defined as a weight gain of fewer than 15 pounds. Furthermore, the model is expanded to include an interaction between the mothers race and the mothers smoking status.

Restricting to the 373,438 births, across the 100 North Carolina counties, with complete data on these covariates we replicated the data that would have been observed in an aggregate data design. Across the five covariates we consider (race, smoking, premature birth, plurality, and low weight gain), an individual birth could be categorized into one of $2\times2\times2\times3\times2$=48 unique levels. Hence the observed group-level data consists of K=100 48×2 tables, each analogous to the first table in middle row of Table 1.2.1. To emulate a hybrid aggregate data study, we took a single stratified case-control sample of n_{0k}=n_{1k}=25 from each county.

Table 1.6.1 reports point and standard error estimates for the log-odds ratio parameters in the expanded model based on three analyses. The first column reports on a fit of the model using the full data (i.e. all N=373,438 individual records). The second column reports on results from a conditional logistic regression analysis of the stratified case-control sample. The final column combines the stratified case-control sample with the group-level data via an approximate hybrid aggregate data likelihood analogous to expres-

Table 1.5.1: Operating characteristics for six likelihood-based estimators of β_1 from model (1.1) using the full data and data from a hybrid aggregate data design, under the seven simulation scenarios described in Section 3.4.1. All values are based on 10,000 simulated datasets.

	Individual-level		Hybrid aggregate data likelihood			
	Full	Case-control	Exact	Binomial	Normal	Poisson
Percent bias						
Baseline	0.0	0.2	-0.2	0.4	0.3	0.5
#1	0.1	0.4	0.8	1.2	1.2	1.3
#2	-0.0	0.3	-0.1	0.4	0.4	0.6
#3	-0.1	0.3	-0.1	0.4	0.4	0.6
#4	0.0	0.2	0.6	1.0	1.2	1.2
#5	0.0	0.4	0.2	2.0	0.5	2.3
#6	0.0	-0.2	0.6	0.5	1.1	0.6
Estimated vs. true standard error $\times$ 100[a]						
Baseline	99.6	99.4	99.9	99.6	99.9	99.6
#1	98.6	99.7	99.4	99.2	99.6	99.2
#2	101.0	100.1	100.6	100.3	100.6	100.3
#3	99.1	99.0	99.5	99.2	99.5	99.2
#4	100.1	99.4	100.2	100.0	99.9	100.0
#5	99.0	98.5	98.1	97.2	99.1	97.2
#6	101.0	98.9	98.3	98.3	98.3	98.3
Coverage probability $\times$ 100[a]						
Baseline	94.9	94.8	95.0	94.9	95.0	95.0
#1	94.5	94.8	95.0	94.9	95.0	94.9
#2	95.2	94.9	95.3	95.2	95.3	95.2
#3	94.8	94.8	95.3	95.1	95.2	95.1
#4	94.9	94.9	95.2	95.1	95.0	95.0
#5	94.7	95	94.8	94.3	94.8	94.3
#6	95.5	95.0	94.7	94.7	94.7	94.8

[a] Estimated standard errors and coverage probabilities are based on the inverse of the Hessian of the corresponding (possibly misspecified) likelihood.

Table 1.5.2: Operating characteristics for six likelihood-based estimators of β_1 from model (1.1) using the full data and data from a hybrid pure ecological design, under the seven simulation scenarios described in Section 3.4.1. All values are based on 10,000 simulated datasets.

	Individual-level		Hybrid pure ecological likelihood		
	Full	Case-control	Binomial	Normal	Poisson
Percent bias					
Baseline	-0.0	0.3	0.6	0.5	0.8
#1	0.0	0.1	1.0	1.1	1.2
#2	0.0	0.1	0.5	0.3	0.7
#3	-0.1	0.1	0.4	0.3	0.6
#4	0.0	0.2	1.1	1.3	1.3
#5	-0.0	0.3	-0.4	-0.1	-0.4
#6	0.0	0.2	0.5	1.1	0.7
Estimated vs. true standard error $\times$ *100*[a]					
Baseline	98.8	100.0	98.2	98.9	98.2
#1	98.9	98.8	99.8	100.3	99.8
#2	99.2	99.4	98.3	99.1	98.3
#3	98.8	99.7	98.2	98.9	98.2
#4	99.3	99.5	98.0	98.1	98.0
#5	100.2	100.1	99.9	99.9	99.9
#6	99.5	99.5	99.3	99.7	99.3
Coverage probability[a]					
Baseline	94.7	95.0	94.5	94.7	94.5
#1	94.7	94.9	94.9	95.0	95.0
#2	94.7	95.1	94.6	94.8	94.6
#3	94.9	94.9	94.6	94.8	94.7
#4	94.6	95.0	94.3	94.4	94.4
#5	94.8	95.4	94.8	94.8	94.8
#6	95.0	95.0	95.0	95.0	95.0

[a] Estimated standard errors and coverage probabilities are based on the inverse of the Hessian of the corresponding (possibly misspecified) likelihood.

Table 1.5.3: Relative uncertainty[a] for five likelihood-based estimators of β_1 from model (1.1) under the seven simulation scenarios described in Section 3.4.1. Shown are results for both the hybrid aggregate data and hybrid pure ecological designs. All values are based on 10,000 simulated datasets.

	Individual-level		Hybrid likelihood			
	Full	Case-control	Exact[b]	Binomial	Normal	Poisson
Aggregate data						
Baseline	23.4	100	76.4	76.8	76.9	77.0
#1	24.2	100	80.3	80.6	80.6	80.7
#2	23.6	100	75.7	76.2	76.3	76.3
#3	23.8	100	76.1	76.6	76.6	76.7
#4	20.8	100	76.6	76.9	77.2	77.1
#5	34.3	100	77.6	79.0	78.1	79.2
#6	16.2	100	77.8	77.7	78.3	77.9
Pure ecological						
Baseline	23.7	100	-	78.5	78.2	78.6
#1	23.9	100	-	79.5	79.5	79.6
#2	23.8	100	-	77.2	76.9	77.3
#3	24.0	100	-	78.0	77.7	78.1
#4	20.8	100	-	78.9	79.0	79.1
#5	15.0	100	-	76.9	77.2	76.9
#6	16.6	100	-	77.4	77.7	77.5

[a] Ratio of the standard error for estimator relative to that of the case-control estimator

[b] Not considered for the pure ecological design. See Section 3.4.3.

Table 1.6.1: Point and standard error estimates for log-odds ratio parameters in extended model, described in Section 1.6, based on the North Carolina data. Estimates for the hybrid aggregate data are based on the binomial approximation to the hybrid likelihood for all K=100 counties.

	Full data		Case-control data		Hybrid aggregate data	
	Est	SE	Est	SE	Est	SE
Early birth	2.92	0.014	2.82	0.088	3.03	0.061
Number of babies						
One	1.00		1.00		1.00	
Two	2.33	0.025	2.92	0.236	2.81	0.143
Three or more	4.30	0.209	1.62	1.028	2.07	0.769
Low weight gain	0.66	0.018	0.57	0.099	0.59	0.076
Non-white	0.70	0.016	0.71	0.097	0.69	0.077
Smoker	0.90	0.024	0.97	0.107	0.94	0.085
Non-white $\times$ Smoker	-0.40	0.041	-0.36	0.215	-0.33	0.172

sion (1.11), using the binomial approximation. Note, while details are omitted, using a result presented in Chapter 2.5 we estimated that the exact hybrid likelihood for the expanded model based on the North Carolina data corresponds to a summation of approximately 10^{124} terms. Hence, for all practical purposes, the exact hybrid likelihood could not be used to perform estimation or inference.

From the point estimates based on the full data, we find increased risk of a low birth weight event associated with an early birth, an increased plurality, low weight gain by the mother, non-white race and the mother smoking during pregnancy. That the interaction is statistically significant indicates that if the mother smokes during pregnancy, the impact is somewhat less for non-white babies than white babies. From the point estimates based on the case-control data alone and the combined data analyses, the conclusions are qualitatively the same. However, the standard errors based the combined data sources are between 20-40% lower than those based on the case-control data alone. As such, use of the approximate hybrid aggregate data likelihood has resulted in substantial efficiency gains. Finally, estimates of the county-specific intercepts based on the full data likelihood and the approximate hybrid likelihood were also consistent; of the K=100 intercept estimates, the largest discrepancy between the hybrid aggregate data design estimates and the full-data point estimates is only a 6% difference. Chapter 2.6 provides a scatterplot of the two sets of estimates.

1.7 Discussion

In this paper we have proposed a pragmatic approach to efficient estimation and inference of individual-level models based on data from hybrid designs that combine group- and individual-level data. While use

of the exact hybrid likelihood will be prohibitively expensive in most settings, the proposed approximations give researchers a practical tool for making use of important group-level data that could otherwise be ignored. From a comprehensive simulation study, despite the fact that use of the approximate hybrid likelihood corresponds to use of a misspecified likelihood, estimation and inference remains valid in a broad range of data scenarios. Furthermore, over the broad range of scenarios we considered, use of the approximate form does not induce any systematic penalty in terms of the efficiency gains that one should expect when one combines the two sources of information. In short, the proposed method provides a practical approach to combining group- and individual-level data from a hybrid design without incurring any penalties in terms of bias and efficiency.

To our knowledge this paper is also the first to formally describe a hybrid aggregate data design, which supplements the aggregate data design of Prentice and Sheppard [59] with case-control data. The most closely-related design is the integrated aggregate data design of Martínez et al. [43, 44], which supplements an aggregate data design with individual-level data prospectively randomly sampled within group. Beyond the sampling of individual-level data, methods for the two designs also differ in that estimation/inference for the integrated aggregate design typically focuses on a log-linear model; in contrast, this paper considers a logistic regression model as the target of estimation/inference. Jointly, therefore, the hybrid and integrated aggregate data designs provide a comprehensive set of tools for rare and non-rare outcomes, respectively.

2

Supplementary Material for: On the Analysis of Hybrid Designs that Combine Group- and Individual-Level Data

2.1 Enumerating the vectors $\mathbf{N}_{\mathbf{yrs}k}$ in $\mathcal{N}_k$

The set $\mathcal{N}_k$ of vectors $\mathbf{N}_{\mathbf{yrs}k}$ that result in the marginal totals $\mathbf{N}_y$ and $\mathbf{M}_{rs}$ (Table 1.II) can be determined through a recursive algorithm. The basis of the recursive algorithm is the situation in which margins of a 2x2 table are known; in this instance once a single internal cell of the 2x2 table is fixed then the rest of the internal cell values are deterministic. Further, for a 2x2 table of the form in Table 2.1.1, the possible values of the internal cell N_{11} is the range between $\max(0, N_1 - M_0)$ and $\min(M_1, N_1)$, inclusive [74].

Table 2.1.1: Ecological Data

	Y=0	Y=1	
X=0	$[N_{00k}]$	$[N_{10k}]$	M_{0k}
X=1	$[N_{01k}]$	$[N_{11k}]$	M_{1k}
	N_{0k}	N_{1k}	N_k

To enumerate the vectors $\mathbf{N_{yrsk}}$ in $\mathcal{N}_k$, one collapses the problem into a series of solvable 2x2 tables:

1. The possible values of N_{111k} range between $\max(0, N_1 - (M_{00k} + M_{01k} + M_{10k}))$ and $\min(M_{11k}, N_{1k})$, inclusive.

2. For each fixed value of N_{111k}, the possible values of N_{110k} range between $\max(0, (N_1 - N_{111k}) - (M_{00k} + M_{01k}))$ and $\min(M_{10k}, N_{1k} - N_{111k})$, inclusive.

3. For each fixed value of N_{111k} and N_{110k}, the possible values of N_{101k} range between $\max(0, (N_1 - N_{111k} - N_{110k}) - M_{00k})$ and $\min(M_{01k}, N_{1k} - N_{111k} - N_{110k})$, inclusive.

4. For each fixed value of N_{111k}, N_{110k}, and N_{101k}, the value of N_{100k} is deterministic.

2.2 Enumerating the vectors $\mathbf{M_{rsk}}$ in $\mathcal{M}_k$

In Table 1.III, the values of the internal cells M_{00k}, M_{01k}, and M_{10k} are deterministic once the internal cell M_{11k} is fixed. The set $\mathcal{M}_k$ of vectors $\mathbf{M_{rsk}}$ that result in the marginal totals $\mathbf{M}_{r+k}$ and $\mathbf{M}_{+sk}$ is the set of vectors for which $M_{11k} \in [\max(0, M_{+1k} - M_{0+k}), \min(M_{1+k}, M_{+1k})]$.

2.3 Approximations

Using results from Wakefield Wakefield [74], the ecological likelihood $L^E(\beta; \mathbf{N}^*_{\mathbf{y}k})$ can be approximated by taking the total number of events in the k^{th} group, N^*_{1k}, to be conditionally distributed according to one of three distributions:

$$N^*_{1k}|\mathbf{M}^*_{\mathbf{rs}k} \sim \text{Binomial}\left(N^*_k, \sum_{r,s} \frac{M^*_{rsk}}{N^*_k}\pi_{rsk}\right)$$

$$N^*_{1k}|\mathbf{M}^*_{\mathbf{rs}k} \sim \text{Normal}\left(\sum_{r,s} M^*_{rsk}\pi_{rsk}, \sum_{r,s} M^*_{rsk}\pi_{rsk}(1-\pi_{rsk})\right)$$

$$N^*_{1k}|\mathbf{M}^*_{\mathbf{rs}k} \sim \text{Poisson}\left(\sum_{r,s} M^*_{rsk}\pi_{rsk}\right)$$

The three different approximations were nearly indistinguishable in terms of efficiency and bias under a series of simulations, including those presented in Web Appendix Section E.

2.4 Inference

Under the alternative decomposition, the hybrid likelihood is given in equation 6.

In the approximated hybrid likelihood $L^E(\beta; \mathbf{N}^*_{\mathbf{y}k})$ is replaced with $\tilde{L}^E(\beta; \mathbf{N}^*_{\mathbf{y}k})$, where $\tilde{L}^E(\beta; \mathbf{N}^*_{\mathbf{y}k})$ assumes the total number of events in the k^{th} group, N^*_{1k}, is conditionally distributed according to one of three distributions presented in Section 4.2

The log likelihood is:

$$\begin{aligned} l^A(\beta; \mathbf{N}_{\mathbf{y}k}, \mathbf{n}_{\mathbf{yrs}k} | \mathbf{M}_{\mathbf{rs}k}, \mathbf{n}_{\mathbf{y}k}) =& l^E(\beta; \mathbf{N}^*_{\mathbf{y}k}) + l^I(\beta; \mathbf{n}_{\mathbf{yrs}k} | \mathbf{m}_{\mathbf{rs}k}) + \\ & log\left(\mathrm{HG}(\mathbf{m}_{\mathbf{rs}k} | \mathbf{M}_{\mathbf{rs}k}, n_k)\right) - log\left(\mathrm{HG}(\mathbf{n}_{\mathbf{y}k} | \mathbf{N}_{\mathbf{y}k}, n_k)\right) \end{aligned}$$

with corresponding gradient and Hessian matrices:

$$\begin{aligned} \nabla_\beta l^A(\beta; \mathbf{N}_{\mathbf{y}k}, \mathbf{n}_{\mathbf{yrs}k} | \mathbf{M}_{\mathbf{rs}k}, \mathbf{n}_{\mathbf{y}k}) =& \nabla_\beta l^E(\beta; \mathbf{N}^*_{\mathbf{y}k}) + \nabla_\beta l^I(\beta; \mathbf{n}_{\mathbf{yrs}k} | \mathbf{m}_{\mathbf{rs}k}) \\ \mathrm{H}_\beta l^A(\beta; \mathbf{N}_{\mathbf{y}k}, \mathbf{n}_{\mathbf{yrs}k} | \mathbf{M}_{\mathbf{rs}k}, \mathbf{n}_{\mathbf{y}k}) =& \mathrm{H}_\beta l^E(\beta; \mathbf{N}^*_{\mathbf{y}k}) + \mathrm{H}_\beta l^I(\beta; \mathbf{n}_{\mathbf{yrs}k} | \mathbf{m}_{\mathbf{rs}k}) \end{aligned}$$

The analytic gradient and Hessian are presented below in matrix form, with 1_4 representing a vector of ones, $\mathcal{D}(\cdot)$ defined to be an operator that converts a vector to a diagonal matrix, and

$$X = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix}$$

Define the 4×1 vectors $\{\pi^\circ_k, \pi^{\circ\circ}_k, \pi^*_k, \pi^{**}_k, \pi^{***}_k, \pi^{****}_k\}$ such that:

$$\begin{aligned} \pi^\circ_{rsk} &= m^*_{rsk}\pi_{rsk} \\ \pi^{\circ\circ}_{rsk} &= m^*_{rsk}\pi_{rsk}(1 - \pi_{rsk}) \end{aligned}$$

$$\begin{aligned} \pi^*_{rsk} &= M^*_{rsk}\pi_{rsk} \\ \pi^{**}_{rsk} &= M^*_{rsk}\pi_{rsk}(1 - \pi_{rsk}) \\ \pi^{***}_{rsk} &= M^*_{rsk}\pi_{rsk}(1 - \pi_{rsk})(1 - 2\pi_{rsk}) \\ \pi^{****}_{rsk} &= M^*_{rsk}\pi_{rsk}(1 - \pi_{rsk})\left(1 - 6\pi_{rsk}(1 - \pi_{rsk})\right) \end{aligned}$$

Let

$$\lambda = 1_4^T \pi^*$$
$$\omega = 1_4^T \pi^{**}$$

Note that

$$\begin{aligned}
\nabla_\beta \lambda =& X^T \pi_k^{**} \\
\nabla_\beta \omega =& X^T \pi_k^{***} \\
\mathrm{H}_\beta \lambda =& X^T \mathcal{D}(\pi_k^{***}) X \\
\mathrm{H}_\beta \omega =& X^T \mathcal{D}(\pi_k^{****}) X
\end{aligned}$$

2.4.1 Gradient and Hessian of $l^I(\beta; \mathbf{n_{yrsk}} | \mathbf{m_{rsk}})$

$$\begin{aligned}
\nabla_\beta l^I(\beta; \mathbf{n_{yrsk}} | \mathbf{m_{rsk}}) =& X^T (n_{1k} - \pi_k^\circ) \\
\mathrm{H}_\beta l^I(\beta; \mathbf{n_{yrsk}} | \mathbf{m_{rsk}}) =& - X^T \mathcal{D}(\pi_k^{\circ\circ}) X
\end{aligned}$$

2.4.2 Gradient and Hessian of $l^E(\beta; \mathbf{N}^*_{\mathbf{yk}})$

$$\begin{aligned}
\nabla_\beta l^E(\beta; \mathbf{N}^*_{\mathbf{yk}}) =& \frac{1}{L^E(\beta; \mathbf{N}^*_{\mathbf{yk}})} \sum_{\mathbf{N_{yrsk}} \in \mathcal{N}_k^*} \left(L^I(\beta; \mathbf{N}^*_{\mathbf{yrsk}} | \mathbf{M}^*_{\mathbf{rsk}}) \cdot \nabla_\beta l^I(\beta; \mathbf{N}^*_{\mathbf{yrsk}} | \mathbf{M}^*_{\mathbf{rsk}}) \right) \\
\mathrm{H}_\beta l^E(\beta; \mathbf{N}^*_{\mathbf{yk}}) =& - \left(\nabla_\beta l^E \right) \left(\nabla_\beta l^E \right)^T + \frac{1}{L^E(\beta)} \sum_{\mathbf{N_{yrsk}} \in \mathcal{N}_k^*} \left\{ L^I(\beta) \cdot \left[\left(\nabla_\beta l^I \right) \left(\nabla_\beta l^I \right)^T + \mathrm{H}_\beta l^I(\beta) \right] \right\}
\end{aligned}$$

$$\begin{aligned}
\nabla_\beta l^I(\beta; \mathbf{N}^*_{\mathbf{yrsk}} | \mathbf{M}^*_{\mathbf{rsk}}) =& X^T (N_{1k} - \pi_k^*) \\
\mathrm{H}_\beta l^I(\beta; \mathbf{N}^*_{\mathbf{yrsk}} | \mathbf{M}^*_{\mathbf{rsk}}) =& - X^T \mathcal{D}(\pi_k^{**}) X
\end{aligned}$$

2.4.3 Gradient and Hessian of $\tilde{l}^E(\beta; \mathbf{N}^*_{\mathbf{yk}})$

Binomial Approximation

$$N^*_{1k} | \mathbf{M}^*_{\mathbf{rsk}} \sim \text{Binomial} \left(N_k^*, \frac{1}{N_k^*} \lambda \right)$$

$$\nabla_\beta \tilde{l}^E(\beta; \mathbf{N}_y^* | \mathbf{M}_x^*) = \frac{N_k^*(N_{1k}^* - \lambda)}{\lambda(N_k^* - \lambda)} \left(\nabla_\beta \lambda\right)$$
$$\mathrm{H}_\beta \tilde{l}^E(\beta; \mathbf{N}_y^* | \mathbf{M}_x^*) = \frac{N_k^*(N_{1k}^* - \lambda)}{\lambda(N_k^* - \lambda)} \left(\mathrm{H}_\beta \lambda\right) - \left(\frac{N_{1k}^*}{\lambda^2} + \frac{N_k^* - N_{1k}^*}{(N_k^* - \lambda)^2}\right) \left(\nabla_\beta \lambda\right) \left(\nabla_\beta \lambda\right)^T$$

Normal Approximation

$$N_{1k}^* | \mathbf{M}_{\mathbf{rs}k}^* \sim \text{Normal}\left(\lambda,\ \omega\right)$$

$$\nabla_\beta \tilde{l}^E(\beta; \mathbf{N}_y^* | \mathbf{M}_x^*) = \frac{(N_{1k}^* - \lambda)}{\omega} \left(\nabla_\beta \lambda\right) + \left(\frac{(N_{1k}^* - \lambda)^2}{2\omega^2} - \frac{1}{2\omega}\right) \left(\nabla_\beta \omega\right)$$

$$\begin{aligned}\mathrm{H}_\beta \tilde{l}^E(\beta; \mathbf{N}_y^* | \mathbf{M}_x^*) = &\left(\frac{(N_{1k}^* - \lambda)^2}{2\omega^2} - \frac{1}{2\omega}\right) \left(\mathrm{H}_\beta \omega\right) - \frac{(N_{1k}^* - \lambda)}{\omega^2} \left(\left(\nabla_\beta \omega\right) \left(\nabla_\beta \lambda\right)^T + \left(\nabla_\beta \lambda\right) \left(\nabla_\beta \omega\right)^T\right) \\ &- \frac{1}{\omega} \left(\nabla_\beta \lambda\right) \left(\nabla_\beta \lambda\right)^T + \frac{N_{1k}^* - \lambda}{\omega} \left(\mathrm{H}_\beta \lambda\right) - \left(\frac{(N_{1k}^* - \lambda)^2}{\omega^3} - \frac{1}{2\omega^2}\right) \left(\nabla_\beta \omega\right) \left(\nabla_\beta \omega\right)^T\end{aligned}$$

Poisson Approximation

$$N_{1k}^* | \mathbf{M}_{\mathbf{rs}k}^* \sim \text{Poisson}\left(\lambda\right)$$

$$\nabla_\beta \tilde{l}^E(\beta; \mathbf{N}_y^* | \mathbf{M}_x^*) = \left(\frac{N_{1k}^*}{\lambda} - 1\right) \left(\nabla_\beta \lambda\right)$$
$$\mathrm{H}_\beta \tilde{l}^E(\beta; \mathbf{N}_y^* | \mathbf{M}_x^*) = -\frac{N_{1k}^*}{\lambda^2} \left(\nabla_\beta \lambda\right) \left(\nabla_\beta \lambda\right)^T + \left(\frac{N_{1k}^*}{\lambda} - 1\right) \left(\mathrm{H}_\beta \lambda\right)$$

2.4.4 Gradient for the Pure Ecological Hybrid Likelihood

The pure ecological hybrid likelihood contains additional parameters: the odds ratios of covariates. In the two covariate case the hybrid likelihood is given in equation 12. In equation 12, $P(\mathbf{M}_{\mathbf{rs}k} | \mathbf{M}_{\mathbf{r}+k}, \mathbf{M}_{+\mathbf{s}k}, \varphi_{rs})$ is the probability distribution function of an extended hypergeometric distribution.

Taking $a = (\beta, \varphi_{rs})$ to represent all of the $K + 3$ parameters in the model, the gradient for the approxi-

mated log likelihood is the vector $\nabla_a \tilde{l}^H = (\nabla_\beta \tilde{l}^H, \nabla_\varphi \tilde{l}^H)$ where

$$\begin{aligned}
&\nabla_\beta \tilde{l}^H(\beta, \varphi_{rs}; \mathbf{N}_{\mathbf{y}k}, \mathbf{n}_{\mathbf{yrs}k} | \mathbf{M}_{\mathbf{r}+k}, \mathbf{M}_{+\mathbf{s}k}, \mathbf{n}_{\mathbf{y}k}) \\
&\quad = \frac{\sum_{\mathbf{M}_{\mathbf{rs}k} \in \mathcal{M}_k} P(\mathbf{M}_{\mathbf{rs}k} | \mathbf{M}_{\mathbf{r}+k}, \mathbf{M}_{+\mathbf{s}k}, \varphi_{rs}) \tilde{L}^A(\beta; \mathbf{N}_{\mathbf{y}k}, \mathbf{n}_{\mathbf{yrs}k} | \mathbf{M}_{\mathbf{rs}k}, \mathbf{n}_{\mathbf{y}k}) \nabla_\beta \tilde{l}^A(\beta; \mathbf{N}_{\mathbf{y}k}, \mathbf{n}_{\mathbf{yrs}k} | \mathbf{M}_{\mathbf{rs}k}, \mathbf{n}_{\mathbf{y}k})}{\tilde{L}^H(\beta, \varphi_{rs}; \mathbf{N}_{\mathbf{y}k}, \mathbf{n}_{\mathbf{yrs}k} | \mathbf{M}_{\mathbf{r}+k}, \mathbf{M}_{+\mathbf{s}k}, \mathbf{n}_{\mathbf{y}k})} \\
&\nabla_\varphi \tilde{l}^H(\beta, \varphi_{rs}; \mathbf{N}_{\mathbf{y}k}, \mathbf{n}_{\mathbf{yrs}k} | \mathbf{M}_{\mathbf{r}+k}, \mathbf{M}_{+\mathbf{s}k}, \mathbf{n}_{\mathbf{y}k}) \\
&\quad = \frac{\sum_{\mathbf{M}_{\mathbf{rs}k} \in \mathcal{M}_k} \nabla_{\varphi_{rs}} P(\mathbf{M}_{\mathbf{rs}k} | \mathbf{M}_{\mathbf{r}+k}, \mathbf{M}_{+\mathbf{s}k}, \varphi_{rs}) \tilde{L}^A(\beta; \mathbf{N}_{\mathbf{y}k}, \mathbf{n}_{\mathbf{yrs}k} | \mathbf{M}_{\mathbf{rs}k}, \mathbf{n}_{\mathbf{y}k})}{\tilde{L}^H(\beta, \varphi_{rs}; \mathbf{N}_{\mathbf{y}k}, \mathbf{n}_{\mathbf{yrs}k} | \mathbf{M}_{\mathbf{r}+k}, \mathbf{M}_{+\mathbf{s}k}, \mathbf{n}_{\mathbf{y}k})}
\end{aligned}$$

The extended hypergeometric distribution has pmf

$$P(\mathbf{M}_{\mathbf{rs}k} | \mathbf{M}_{\mathbf{r}+k}, \mathbf{M}_{+\mathbf{s}k}, \varphi_{rs}) = \frac{\binom{M_{0+k}}{M_{11k}} \binom{M_{1+k}}{M_{+1k} - M_{11k}} \exp(M_{11k} \varphi_{rs})}{\sum_{\mathbf{M}^*_{\mathbf{rs}k} \in \mathcal{M}_k} \binom{M_{0+k}}{M^*_{11k}} \binom{M_{1+k}}{M_{+1k} - M^*_{11k}} \exp(M^*_{11k} \varphi_{rs})}$$

The gradient $\nabla_{\varphi_{rs}} P(\mathbf{M}_{\mathbf{rs}k} | \mathbf{M}_{\mathbf{r}+k}, \mathbf{M}_{+\mathbf{s}k}, \varphi_{rs})$ is therefore

$$\begin{aligned}
\nabla_{\varphi_{rs}} P(\mathbf{M}_{\mathbf{rs}k} | \mathbf{M}_{\mathbf{r}+k}, \mathbf{M}_{+\mathbf{s}k}, \varphi_{rs}) &= P(\mathbf{M}_{\mathbf{rs}k} | \mathbf{M}_{\mathbf{r}+k}, \mathbf{M}_{+\mathbf{s}k}, \varphi_{rs}) \left(M_{11k} - \frac{\sum_{\mathbf{M}^*_{\mathbf{rs}k} \in \mathcal{M}_k} M^*_{11k} \binom{M_{0+k}}{M^*_{11k}} \binom{M_{1+k}}{M_{+1k} - M^*_{11k}} \exp(M^*_{11k} \varphi_{rs})}{\sum_{\mathbf{M}^*_{\mathbf{rs}k} \in \mathcal{M}_k} \binom{M_{0+k}}{M^*_{11k}} \binom{M_{1+k}}{M_{+1k} - M^*_{11k}} \exp(M^*_{11k} \varphi_{rs})} \right) \\
&= P(\mathbf{M}_{\mathbf{rs}k} | \mathbf{M}_{\mathbf{r}+k}, \mathbf{M}_{+\mathbf{s}k}, \varphi_{rs}) \left(M_{11k} - \frac{P(\eta | \mathbf{M}_{\mathbf{r}+k}, \mathbf{M}_{+\mathbf{s}k}, \varphi_{rs})}{f(\eta | \mathbf{M}_{\mathbf{r}+k}, \mathbf{M}_{+\mathbf{s}k}, \varphi_{rs})} \eta \right)
\end{aligned}$$

where η is the mode of the extended hypergeometric and

$$f(\eta | \mathbf{M}_{\mathbf{r}+k}, \mathbf{M}_{+\mathbf{s}k}, \varphi_{rs}) = \frac{\eta \binom{M_{0+k}}{\eta} \binom{M_{1+k}}{M_{+1k} - \eta} \exp(\eta \varphi_{rs})}{\sum_{\mathbf{M}^*_{\mathbf{rs}k} \in \mathcal{M}_k} M^*_{11k} \binom{M_{0+k}}{M^*_{11k}} \binom{M_{1+k}}{M_{+1k} - M^*_{11k}} \exp(M^*_{11k} \varphi_{rs})}$$

The mode, η, and $P(\mathbf{M}_{\mathbf{rs}k} | \mathbf{M}_{\mathbf{r}+k}, \mathbf{M}_{+\mathbf{s}k}, \varphi_{rs})$ may be computed using stable algorithms described by Liao and Rosen [41], avoiding computationally intensive and numerically unstable calculations. The value $f(\eta | \mathbf{M}_{\mathbf{r}+k}, \mathbf{M}_{+\mathbf{s}k}, \varphi_{rs})$ may be computed in a similar fashion, replacing the function

$$r(M_{11k}) = \frac{(M_{0+k} - M_{11k} + 1)(M_{+1k} - M_{11k} + 1)}{M_{11k}(M_{1+k} - M_{+1k} + M_{11k})} \exp(\varphi_{rs})$$

used to compute $P(\mathbf{M}_{\mathbf{r}\mathbf{s}k}|\mathbf{M}_{\mathbf{r}+k}, \mathbf{M}_{+\mathbf{s}k}, \varphi_{rs})$ in Liao and Rosen's algorithm with

$$r(M_{11k}) = \frac{(M_{0+k} - M_{11k} + 1)(M_{+1k} - M_{11k} + 1)}{(M_{11k} - 1)(M_{1+k} - M_{+1k} + M_{11k})} \exp(\varphi_{rs})$$

The Hessian for the approximated log likelihood is $\mathrm{H}_a \tilde{l}^H = (\nabla_\beta \tilde{l}^H, \nabla_\varphi \tilde{l}^H)$ where

$$\mathrm{H}_a \tilde{l}^H = -\left(\nabla_a \tilde{l}^H\right)\left(\nabla_a \tilde{l}^H\right)^T + \frac{1}{\tilde{L}^H} \sum_{\mathbf{M}_{\mathbf{r}\mathbf{s}k} \in \mathcal{M}_k} \mathrm{H}_{\mathbf{M}_{\mathbf{r}\mathbf{s}k}}$$

and $\mathrm{H}_{\mathbf{M}_{\mathbf{r}\mathbf{s}k}} = \left[\begin{array}{c|c} A_{\beta\beta} & A_{\beta\varphi} \\ \hline A_{\beta\varphi}^T & A_{\varphi\varphi} \end{array}\right]$ is a $(K+3) \times (K+3)$ block matrix with blocks $A_{\beta\beta}$, $A_{\beta\varphi}$, and $A_{\varphi\varphi}$ of size $(K+2) \times (K+2)$, $(K+2) \times 1$ and 1×1, respectively.

$$\begin{aligned}
A_{\beta\beta} =& P(\mathbf{M}_{\mathbf{r}\mathbf{s}k}|\mathbf{M}_{\mathbf{r}+k}, \mathbf{M}_{+\mathbf{s}k}, \varphi_{rs}) \tilde{L}^A(\beta; \mathbf{N}_{\mathbf{y}k}, \mathbf{n}_{\mathbf{y}\mathbf{r}\mathbf{s}k}|\mathbf{M}_{\mathbf{r}\mathbf{s}k}, \mathbf{n}_{\mathbf{y}k}) \cdot \\
& \left(\left(\nabla_\beta \tilde{l}^A(\beta; \mathbf{N}_{\mathbf{y}k}, \mathbf{n}_{\mathbf{y}\mathbf{r}\mathbf{s}k}|\mathbf{M}_{\mathbf{r}\mathbf{s}k}, \mathbf{n}_{\mathbf{y}k})\right)\left(\nabla_\beta \tilde{l}^A(\beta; \mathbf{N}_{\mathbf{y}k}, \mathbf{n}_{\mathbf{y}\mathbf{r}\mathbf{s}k}|\mathbf{M}_{\mathbf{r}\mathbf{s}k}, \mathbf{n}_{\mathbf{y}k})\right)^T + \mathrm{H}_\beta \tilde{l}^A(\beta; \mathbf{N}_{\mathbf{y}k}, \mathbf{n}_{\mathbf{y}\mathbf{r}\mathbf{s}k}|\mathbf{M}_{\mathbf{r}\mathbf{s}k}, \mathbf{n}_{\mathbf{y}k})\right) \\
A_{\beta\varphi} =& \nabla_{\varphi_{rs}} P(\mathbf{M}_{\mathbf{r}\mathbf{s}k}|\mathbf{M}_{\mathbf{r}+k}, \mathbf{M}_{+\mathbf{s}k}, \varphi_{rs}) \tilde{L}^A(\beta; \mathbf{N}_{\mathbf{y}k}, \mathbf{n}_{\mathbf{y}\mathbf{r}\mathbf{s}k}|\mathbf{M}_{\mathbf{r}\mathbf{s}k}, \mathbf{n}_{\mathbf{y}k}) \nabla_\beta \tilde{l}^A(\beta; \mathbf{N}_{\mathbf{y}k}, \mathbf{n}_{\mathbf{y}\mathbf{r}\mathbf{s}k}|\mathbf{M}_{\mathbf{r}\mathbf{s}k}, \mathbf{n}_{\mathbf{y}k}) \\
A_{\varphi\varphi} =& \mathrm{H}_{\varphi_{rs}} P(\mathbf{M}_{\mathbf{r}\mathbf{s}k}|\mathbf{M}_{\mathbf{r}+k}, \mathbf{M}_{+\mathbf{s}k}, \varphi_{rs}) \tilde{L}^A(\beta; \mathbf{N}_{\mathbf{y}k}, \mathbf{n}_{\mathbf{y}\mathbf{r}\mathbf{s}k}|\mathbf{M}_{\mathbf{r}\mathbf{s}k}, \mathbf{n}_{\mathbf{y}k})
\end{aligned}$$

$$\begin{aligned}
\mathrm{H}_{\varphi_{rs}} P(\mathbf{M}_{\mathbf{r}\mathbf{s}k}|\mathbf{M}_{\mathbf{r}+k}, \mathbf{M}_{+\mathbf{s}k}, \varphi_{rs}) =& P(\mathbf{M}_{\mathbf{r}\mathbf{s}k}|\mathbf{M}_{\mathbf{r}+k}, \mathbf{M}_{+\mathbf{s}k}, \varphi_{rs}) \exp(\varphi_{rs}) \left(M_{11k} + \exp(\varphi_{rs}) M_{11k}^2 \right. \\
& - \frac{P(\eta|\mathbf{M}_{\mathbf{r}+k}, \mathbf{M}_{+\mathbf{s}k}, \varphi_{rs})}{f(\eta|\mathbf{M}_{\mathbf{r}+k}, \mathbf{M}_{+\mathbf{s}k}, \varphi_{rs})} \eta - 2\exp(\varphi_{rs}) M_{11k} \frac{P(\eta|\mathbf{M}_{\mathbf{r}+k}, \mathbf{M}_{+\mathbf{s}k}, \varphi_{rs})}{f(\eta|\mathbf{M}_{\mathbf{r}+k}, \mathbf{M}_{+\mathbf{s}k}, \varphi_{rs})} \eta \\
& \left. + 2\exp(\varphi_{rs}) \left(\frac{P(\eta|\mathbf{M}_{\mathbf{r}+k}, \mathbf{M}_{+\mathbf{s}k}, \varphi_{rs})}{f(\eta|\mathbf{M}_{\mathbf{r}+k}, \mathbf{M}_{+\mathbf{s}k}, \varphi_{rs})} \eta\right)^2 - \exp(\varphi_{rs}) \frac{P(\eta|\mathbf{M}_{\mathbf{r}+k}, \mathbf{M}_{+\mathbf{s}k}, \varphi_{rs})}{f_2(\eta|\mathbf{M}_{\mathbf{r}+k}, \mathbf{M}_{+\mathbf{s}k}, \varphi_{rs})} \eta^2 \right)
\end{aligned}$$

Ratios involving the mode are again incorporated into the Hessian calculation to reduce computational intensity and avoid numerically unstable fractions. The function $f(\eta|\mathbf{M}_{\mathbf{r}+k}, \mathbf{M}_{+\mathbf{s}k}, \varphi_{rs})$ is defined as previously and

$$f_2(\eta|\mathbf{M}_{\mathbf{r}+k}, \mathbf{M}_{+\mathbf{s}k}, \varphi_{rs}) = \frac{\eta^2 \binom{M_{0+k}}{\eta} \binom{M_{1+k}}{M_{+1k}-\eta} \exp(\eta\varphi_{rs})}{\sum_{\mathbf{M}^*_{\mathbf{r}\mathbf{s}k} \in \mathcal{M}_k} M_{11k}^{*2} \binom{M_{0+k}}{M^*_{11k}} \binom{M_{1+k}}{M_{+1k}-M^*_{11k}} \exp(M^*_{11k}\varphi_{rs})}$$

2.5 Operating characteristics of β_2

Table 2.5.1: Operating characteristics for six likelihood-based estimators of β_2 from model (1) using the full data and data from a hybrid aggregate data design, under the seven simulation scenarios described in Section 5.1. All values are based on 10,000 simulated datasets.

	Individual-level		Hybrid aggregate data likelihood			
	Full	Case-control	Exact	Binomial	Normal	Poisson
Percent bias						
Baseline	0.1	0.8	-0.4	0.2	0.0	0.3
#1	-0.2	0.7	-0.6	-0.2	-0.1	0.0
#2	-0.4	0.4	-0.8	-0.2	-0.3	-0.1
#3	0.2	0.8	-0.4	0.1	0	0.3
#4	0.1	0.5	0.5	0.8	1.0	1.0
#5	-0.1	0.6	-0.4	1.3	-0.7	1.5
#6	0.1	0.4	0.2	0.2	0.7	0.3
Estimated vs. true standard error $\times$ *100*[a]						
Baseline	99.7	100.7	100.3	100.0	100.3	100.0
#1	100.4	99.4	98.2	98.0	98.2	98.0
#2	100.8	99.7	101.0	100.7	101.1	100.7
#3	99.7	99.9	99.4	99.1	99.5	99.1
#4	99.6	99.2	98.0	97.9	97.7	97.9
#5	100.3	100.1	99.3	98.5	100.4	98.5
#6	99.9	98.3	98.2	98.2	98.3	98.2
Coverage probability $\times$ *100*[a]						
Baseline	95.1	95.0	95.3	95.2	95.3	95.2
#1	95.1	95.0	94.6	94.6	94.6	94.6
#2	95.3	95.0	95.2	95.2	95.2	95.2
#3	94.9	95.0	95.0	94.9	94.9	94.9
#4	94.7	94.9	94.3	94.2	94.2	94.2
#5	95.0	95.5	95.3	95.2	95.4	95.2
#6	94.9	94.6	94.5	94.5	94.5	94.5

[a] Estimated standard errors and coverage probabilities are based on the inverse of the Hessian of the corresponding (possibly misspecified) likelihood.

Table 2.5.2: Operating characteristics for six likelihood-based estimators of β_2 from model (1) using the full data and data from a hybrid pure ecological design, under the seven simulation scenarios described in Section 5.1. All values are based on 10,000 simulated datasets.

	Individual-level		Hybrid pure ecological likelihood		
	Full	Case-control	Binomial	Normal	Poisson
Percent bias					
Baseline	0.1	-0.9	-0.6	-0.8	-0.4
#1	0.0	0.2	0.2	0.2	0.3
#2	0.1	0.6	0.2	0	0.4
#3	0.1	0	-0.3	-0.5	-0.2
#4	0.1	0.3	1	1.3	1.2
#5	0.0	0.5	-0.4	-0.1	-0.4
#6	0.1	1.3	0.8	1.3	0.9
Estimated vs. true standard error $\times$ *100*[a]					
Baseline	99.4	100.6	99.0	99.9	99.0
#1	99.8	101.0	99.5	99.9	99.5
#2	100.2	99.4	98.2	99.0	98.2
#3	100.4	100.9	99.6	100.3	99.6
#4	100.0	99.6	99.9	99.9	99.9
#5	98.9	99.3	100.4	100.4	100.4
#6	99.9	100.0	98.4	98.8	98.4
Coverage probability[a]					
Baseline	95.0	95.0	94.7	95.0	94.7
#1	94.8	95.3	94.9	94.9	94.9
#2	94.9	94.9	94.7	94.9	94.7
#3	95.0	95.4	94.9	95.1	94.9
#4	95.0	95.1	95.0	95.0	95.0
#5	94.7	94.8	95.0	95.1	95.1
#6	95.0	95.0	95.0	95.0	95.0

[a] Estimated standard errors and coverage probabilities are based on the inverse of the Hessian of the corresponding (possibly misspecified) likelihood.

Table 2.5.3: Relative uncertainty[a] for five likelihood-based estimators of β_2 from model (1) under the seven simulation scenarios described in Section 5.1. Shown are results for both the hybrid aggregate data and hybrid pure ecological designs. All values are based on 10,000 simulated datasets.

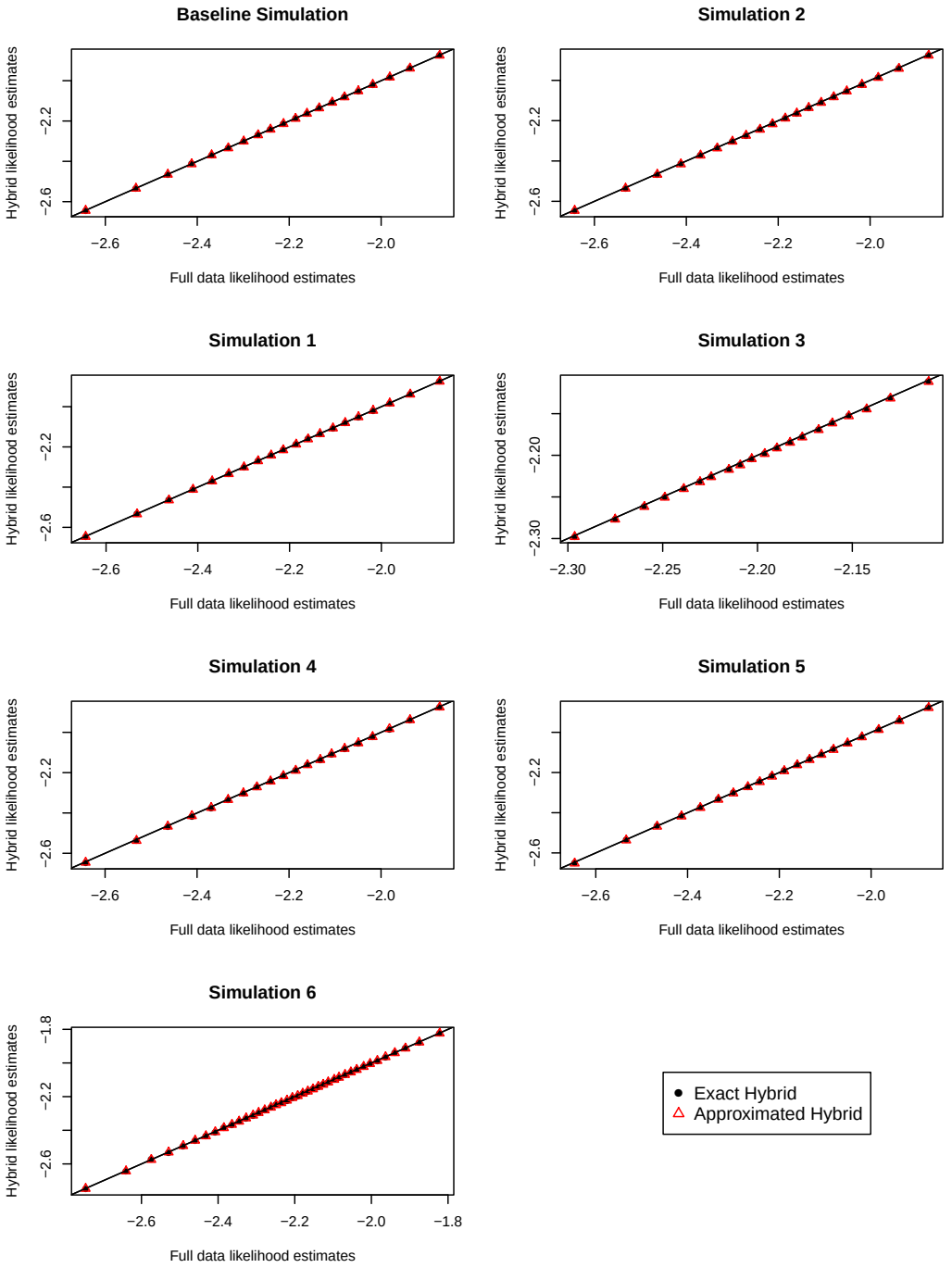
	Individual-level		Hybrid likelihood			
	Full	Case-control	Exact[b]	Binomial	Normal	Poisson
Aggregate data						
Baseline	24.5	100	77.8	78.2	78.3	78.3
#1	22.8	100	79.7	80	80.1	80.1
#2	24	100	76.2	76.7	76.8	76.8
#3	24.3	100	77.8	78.2	78.3	78.3
#4	20.9	100	79.2	79.5	79.8	79.7
#5	34.3	100	77.6	79	78.1	79.2
#6	16.6	100	78.1	78.1	78.6	78.2
Pure ecological						
Baseline	24.4	100	-	78.9	78.6	79.1
#1	23.1	100	-	80.1	80.1	80.2
#2	23.9	100	-	78.4	78.1	78.5
#3	24.3	100	-	78.8	78.5	78.9
#4	20.7	100	-	78.7	78.8	78.9
#5	15.3	100	-	76.8	77.1	76.9
#6	16.8	100	-	79.3	79.6	79.5

[a] Ratio of the standard error for estimator relative to that of the case-control estimator

[b] Not considered for the pure ecological design. See Section 5.2.

2.6 Intercept Point Estimates

Figure 2.6.1: Point estimates of the K group-specific intercepts, β_{ok}, in the simulations described in Section 5. Shown are estimates based on the full data likelihood as well as on the exact hybrid likelihood and binomial approximate hybrid likelihood for a single draw of a hybrid aggregate data design.



2.7 Upper bound on the number of vectors $\mathbf{N_{yrsk}}$ in $\mathcal{N}_k$

From combinatorics there are $\begin{pmatrix} n+r-1 \\ r-1 \end{pmatrix}$ ways to place n indistinguishable balls into r urns. Therefore for any marginal vector, $\mathbf{M_k}$, with r levels there are at most $\begin{pmatrix} N_{1k}+r-1 \\ r-1 \end{pmatrix}$ ways for the N_{1k} positive outcomes to be distributed into the r urns. In the five-covariate model example in Section 6, the five covariates are categorized into 48 possible levels. A simple upper bound for the number of summations required to calculate the hybrid likelihood is

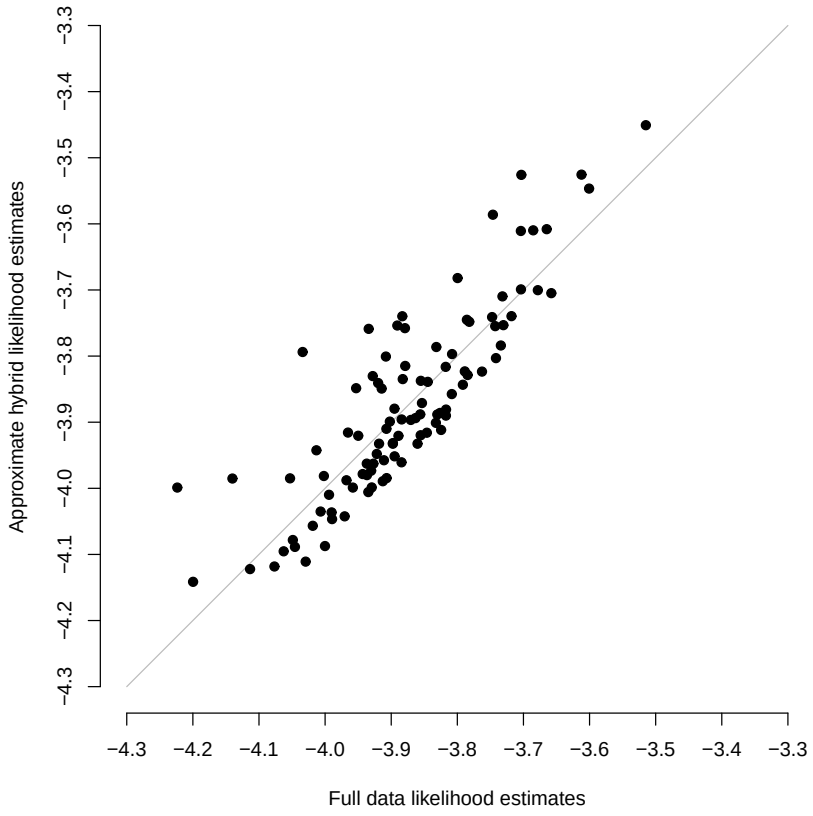
$$\sum_{k=1}^{100} \begin{pmatrix} N_{1k}+47 \\ 47 \end{pmatrix}$$

In a number of counties some of the $[P, M, R, S, W]$=$[p, m, r, s, w]$ strata had marginal counts of zero. Using the notation s_k to denote how many of the 48 strata have non-zero counts in county k, the upper bound can be reduced to

$$\sum_{k=1}^{100} \begin{pmatrix} N_{1k}+s_k-1 \\ s_k-1 \end{pmatrix} \approx 10^{124}$$

2.8 Application

Figure 2.8.1: Point estimates of the K=100 county-specific intercepts, β_{ok}, in the model described in Section 6. Shown are estimates based on the full data likelihood as well as on the binomial approximate hybrid likelihood for a single draw of a hybrid aggregate data design.



To the logician all things should be seen exactly as they are, and to underestimate one's self is as much a departure from truth as to exaggerate one's own powers.

Arthur Conan Doyle, *The Adventure of the Greek Interpreter*

3

On the Analysis of Case-Control and Stratified Case-Control Studies in Cluster-Correlated Data Settings

Abstract

In resource-limited settings, the long-term evaluation of national ART programs often relies on aggregated data, the analysis of which may be subject to ecological bias. As researchers and policy-makers consider evaluating individual-level outcomes such as treatment adherence or mortality, the well-known case-control design is appealing in that it provides efficiency gains over random sampling. In the context that motivates this paper, valid estimation and inference requires acknowledging any clustering although, to our knowledge, no statistical methods have been published for the analysis of case-control data for which the underlying population exhibits clustering. Furthermore, in the specific context of an ongoing collaboration in Malawi, rather than performing case-control sampling across all clinics, case-control sampling within clinics has been suggested as a more practical strategy. To our knowledge, while similar outcome-dependent sampling schemes have been described in the literature, such a cluster-stratified case-control design is new. In this paper we describe this design, discuss balanced versus unbalanced sampling techniques, and also provide a general approach to analyzing case-control and cluster-stratified case-control studies in

cluster-correlated settings based on inverse-probability weighted GEE. Inference is based on a robust sandwich estimator with correlation parameters estimated to ensure appropriate accounting of the outcome-dependent sampling scheme. Comprehensive simulations, based in part on real data on n=87,776 program registrants in Malawi between 2005-2007, are conducted to evaluate small-sample operating characteristics as well as potential trade-offs associated with standard case-control sampling or cluster-stratification.

3.1 Introduction

In dealing with the global HIV/AIDS epidemic, national antiretroviral treatment (ART) programs in the developing world are often designed to be simple, standardized, and decentralized to ensure as broad and efficient coverage as possible [24]. Despite limited resources, monitoring and evaluation of these programs is critical for short-term administrative goals, including resource allocation, as well as long-term success in ending the epidemic. Of particular concern, for example, is the ability to centrally monitor patient retention rates, as treatment discontinuation is both an inefficient use of scarce resources and may lead to a wider spread of drug resistant HIV strains that would threaten the overall efficacy of ART programs [33]. Typically, a single treatment program is developed nationally but responsibility for screening and monitoring patients is delegated to individual treatment centers. One such program was initiated in 2004 in Malawi, a southern African nation for which HIV/AIDS is the leading cause of death among adults aged 15-49 [80]. Overall, the goals of the Malawian ART program are to reduce population-level mortality and morbidity due to HIV/AIDS, as well as to increase the percent of HIV-positive adults physically capable of staying in the workforce [34].

While decentralized national ART programs have been shown to be incredibly effective in facilitating a rapid scale-up of treatment coverage in affected populations, a drawback is that the quality of data available for analyses geared towards program monitoring and evaluation is typically limited [3]. In the current Malawian program, for example, patient-level information is recorded on paper-based 'mastercards', locally-stored at the clinic at which the patient received treatment. Every three months, a representative from the Malawian Ministry of Health and Population (MOHP) conducts a supervision visit, at which time all patients enrolled in the prior three months are assigned to a 'quarterly-clinic cohort'. While detailed patient-level data including age, gender, WHO stage, date of registration, date of starting the first-line antiretroviral regimen and current status is recorded at the clinic, all of this information is aggregated by the MOHP representative at the time of the visit. That is, all patient-level records are reduced to a single quarterly-clinic record with cumulative admission counts, cluster-level covariate data and total outcome tallies, which is then entered in a centralized electronic system [34]. As such the data available for analyses by the MHOP consists of a series of aggregated quarterly-clinic cluster-level records which, collectively, constitute an ecological study [46].

In practice, analyses based on cluster-level data, such as that routinely collected in Malawi, are subject to

a range of potential biases, an umbrella term for which is *ecological bias* [27]. While the statistical literature is rich with methods for analyzing cluster-level data, if one is to avoid making untestable assumptions, the only reliable approach to overcoming ecological bias is to collect and analyze patient-level data [29]. In the long-term the Malawian MOHP is pursuing a strategy which hinges on an electronic system for the storage of patient-level data [18]. In the meantime, however, that detailed patient-level data is not available on a routine basis presents a significant dilemma for monitoring and evaluation of their national ART program. One possible solution is to focus data collection efforts on a judiciously-chosen sub-sample of the population of patient registrants. In practice, there are a huge number of ways in which this can be achieved. When the outcome of interest is binary and (relatively) rare, the case-control study design is well-known to be highly efficient relative to random sampling [5]. Such a design could be readily-implemented in the Malawian context although valid inference would require acknowledging the clustering of program registrants within clinic. To our knowledge, however, no statistical methods have been published specifically for the analysis of case-control data in settings for which the underlying population of interest exhibits clustering.

Building on the standard case-control design, a number of outcome-dependent sampling schemes have been proposed for correlated binary data settings including: designs for longitudinally measured binary outcomes [66–68]; family-based case-control sampling in genetic studies, where proband cases are selected at random and the remaining members of their family taken as 'controls' [47, 48, 50]; and, cluster-based sampling schemes in which a subset of clusters is chosen on the basis of the observed outcome rates and detailed information (retrospectively) collected on all study units within each of the chosen clusters [12]. Another design, which has been specifically considered in the Malawian context, is one where case-control sampling is performed within each clinic. This design has appeal in that it would provide some patient-level data from each clinic (critical for on-going monitoring and evaluation), while requiring considerably less on-going coordination between the MHOP and the clinics themselves. Implementing a standard case-control design the Malawian context, for example, would require the MHOP to identify specific cases/non-cases and then communicate to each clinic which records would need to be extracted; in practice, it would be far simpler to ask each clinic to prepare a pre-specified number of cases and non-cases. We refer to a design in which case-control sampling is performed within each cluster as a *cluster-stratified case-control* design. To our knowledge this specific design has not been described or considered in the literature. One related design is the hybrid design of [30] which supplements an ecological study with case-control data drawn from the same population. Although the authors propose an analysis approach based on hierarchical models to explicitly account for cluster-correlation, the overarching design is limited in its practical use because of the severe computational burden associated with the corresponding likelihood [73].

To summarize, motivated by an on-going collaboration in Malawi, this paper seeks to address two important gaps in the literature. The first is the analysis of case-control data in contexts where valid inference requires acknowledging clustering of study units. The second is the consideration and valid analysis of data

arising from a cluster-stratified case-control design which, to our knowledge, has not been described in the literature. The remainder of this paper is structured as follows. In Section 3.2 we briefly review the analysis of cluster-correlated binary response data using generalized estimating equations (GEE) in complete data settings. Section 3.3 outlines the proposed approach for the analysis of cluster-correlated case-control and cluster-stratified case-control data, based on inverse-probability weighting with an arbitrary user-specified working correlation structure. In Section 3.4 we present a comprehensive simulation study conducted to investigate small-sample operating characteristics of the proposed estimation/inferential procedure, with a focus on understanding potential trade-offs associated with cluster-stratification versus not. Section 3.5 provides additional insight specific to the Malawian context via a simulation study based on a real dataset of n=87,776 program registrants obtained from a one-time nationwide survey of the ART program between 2005-2007. Finally, Section 3.6 concludes the paper with a Discussion and avenues for future work.

3.2 The complete data setting

To formalize the context this paper considers we develop some notation and outline estimation and inference in the complete data setting. Towards this, suppose that study units in the population of interest can be classified into one of a set of K mutually exclusive groups or clusters. Let N_k denote the number of study units in the k^{th} cluster, $k = 1, \ldots, K$. Furthermore, let Y_{ki} denote the outcome for the i^{th} study unit in the k^{th} cluster and X_{ki} a corresponding $p \times 1$ vector of explanatory variables/risk factors. Suppose that the conditional mean response for the i^{th} study unit in the k^{th} cluster is given by $E[Y_{ki}|X_{ki}] = \mu_{ki}$, related to X_{ki} via a link function $g(\cdot)$; that is, $g(\mu_{ki}) = X_{ki}^T\beta$, with β, a $1 \times p$ vector of regression parameters, being the primary target for estimation and inference. Finally, let $\mathbf{Y}_k$ denote the $N_k \times 1$ vector of outcome responses for the k^{th} cluster, $\boldsymbol{\mu}_k$ the corresponding mean vector and $\mathbf{X}_k$ the $N_k \times p$ matrix of explanatory variables.

3.2.1 Estimation and inference via GEE

In complete data settings, where (Y_{ki}, X_{ki}) is observed for all N_k study units in the k^{th} cluster, analyses are typically performed using either generalized linear mixed models [e.g. 7] or generalized estimating equations [GEE; 40]. In the former, cluster-specific random effects are introduced into the model for μ_{ki} to account for correlation due to clustering, beyond that accounted for by covariates in X_{ki}. Typically some assumption is imposed on the distribution of the random effects across the population of clusters which then permits estimation and inference to be based on an integrated likelihood. The introduction of random effects into the mean model changes the interpretation of β, however, and concern is sometimes raised regarding the robustness of results to the specific choice of random effects distribution [36, 49, 51].

In GEE, estimation of β is achieved by solving a system of p equations given by:

$$\sum_{k=1}^{K} \mathbf{U}_k \;=\; \sum_{k=1}^{K} \mathbf{D}_k^T \mathbf{V}_k^{-1} \left(\mathbf{Y}_k - \boldsymbol{\mu}_k \right) \;=\; 0, \tag{3.1}$$

where $\mathbf{D}_k = \frac{\partial \mu_k}{\partial \beta}$ is an $N_k \times p$ matrix of partial derivatives and $\mathbf{V}_k$ is the working variance-covariance matrix for $\mathbf{Y}_k - \boldsymbol{\mu}_k$. Typically the working variance-covariance matrix is decomposed as $\mathbf{V}_k = A_k^{1/2} C_k(\alpha) A_k^{1/2}$, where A_k is an $N_k \times N_k$ diagonal matrix with $Var[Y_{ki}]$ in the $[i,i]^{th}$ entry and $C_k(\alpha)$ is the working correlation structure, specified as a function of some dispersion parameter α. In cluster-correlated data settings, an analyst might adopt a working independence correlation structure (i.e. set the $[i,j]^{th}$ element of C_k to be zero) which removes the need to consider any additional dispersion parameters. Another common choice in cluster-correlated data settings is to adopt a working exchangeable correlation structure (i.e. set the $[i,j]^{th}$ element of C_k to equal α), in which case the dispersion parameter must be estimated simultaneously with β [57].

Given estimates of (β, α), denoted $(\hat{\beta}, \hat{\alpha})$, standard errors can be obtained via empirical evaluation of the asymptotic variance-covariance matrix, which has the 'sandwich' form:

$$V[\hat{\beta}] \;=\; \mathbf{B}^{-1} \left(\sum_{k=1}^{K} \mathbf{U}_k \mathbf{U}_k^T \right) \mathbf{B}^{-1} \tag{3.2}$$

where $\mathbf{B} = \sum_{k=1}^{K} \mathbf{D}_k^T \mathbf{V}_k^{-1} \mathbf{D}_k$ is the model-based 'information' matrix corresponding to (3.1), evaluated at $\hat{\beta}$. Key to the broad appeal of GEE is that inference is 'robust' in the sense that expression (3.2) yields valid standard errors (asymptotically, at least), regardless of the choice of working correlation structure.

3.3 Outcome-dependent sampling

Estimation and inference based on (3.1) and (3.2) relies on the observed clusters being a random sample from the (underlying) population of clusters and the N_k study units in the k^{th} cluster being a complete enumeration of the cluster. Here we consider two outcome-dependent sampling schemes. The first is a traditional case-control study in which the population of $N = \sum N_k$ patients is stratified into two groups: N_1 cases and $N_0 = N - N_1$ non-cases. Under case-control sampling random sub-samples of n_1 cases and n_0 non-cases are identified and detailed covariate information retrospectively ascertained [5, 58]. In the context of this paper, each of these sub-samples would be drawn without regard to cluster membership although each study unit retains their cluster membership nonetheless. Hence, in practice, once the n_1 cases are drawn, their clinic membership must be identified and any clinic for which a non-zero number of cases were drawn must be communicated with. The same is true for the n_0 non-cases.

The second design we consider is one in which case-control sampling is performed within each cluster.

This requires stratifying each clinic into two outcomes groups: N_{1k} cases and N_{0k} = N_k - N_{0k} non-cases. For the k^{th} cluster, $n_{1k} \leq N_{1k}$ cases and $n_{0k} \leq N_{0k}$ non-cases are drawn and detailed covariate information retrospectively ascertained.

3.3.1 Estimation and inference via weighted GEE

To account for the outcome-dependent sampling in both the case-control and cluster-stratified case-control designs, we consider performing estimation and inference via weighted GEE [62, 63]. Towards this, suppose n_k study units are selected from the k^{th} cluster. Let $\widetilde{\mathbf{D}}_k$ denote the $n_k \times$ p sub-matrix of $\mathbf{D}_k$ that corresponds to these patients. Furthermore, let $\widetilde{\mathbf{Y}}_k$, $\widetilde{\mathbf{V}}_k$ and $\widetilde{\boldsymbol{\mu}}_k$ denote the corresponding sub-vectors/matrix of their "complete data" counterparts. Let $\widetilde{\mathbf{W}}_k$ denote an $n_k \times n_k$ diagonal matrix, with the $[i,i]^{th}$ entry, $W_{k,ii}$, the inverse-probability of selection for the i^{th} selected unit. Under case-control sampling, $W_{k,ii} = N_1/n_1$ if the i^{th} unit selected in the k^{th} cluster was a case and N_0/n_0 if they were a non-case. Under cluster-stratified case-control sampling these values are N_{1k}/n_{1k} and N_{0k}/n_{0k}, respectively. Consistent estimates of β are then obtained by solving the $p \times 1$ system of weighted equations given by:

$$\sum_{k=1}^{K} \widetilde{\mathbf{U}}_k = \sum_{k=1}^{K} \widetilde{\mathbf{D}}_k^T \widetilde{\mathbf{V}}_k^{-1} \widetilde{\mathbf{W}}_k \left(\widetilde{\mathbf{Y}}_k - \widetilde{\boldsymbol{\mu}}_k \right) = 0. \tag{3.3}$$

Furthermore, valid inference can be performed on the basis of an empirical evaluation of the asymptotic variance covariance matrix given by:

$$V[\hat{\beta}_w] = \widetilde{\mathbf{B}}^{-1} \left(\sum_{k=1}^{K} \widetilde{\mathbf{U}}_k \widetilde{\mathbf{U}}_k^T \right) \widetilde{\mathbf{B}}^{-1} \tag{3.4}$$

where $\widetilde{\mathbf{B}} = \sum_{k=1}^{K} \widetilde{\mathbf{D}}_k^T \widetilde{\mathbf{V}}_k^{-1} \widetilde{\mathbf{W}}_k \widetilde{\mathbf{D}}_k$. Consistency of $\hat{\beta}_w$ is shown in Chapter 4.1.

In practice, $\hat{\beta}_w$ may be computed iteratively using the Newton-Raphson algorithm, setting $\hat{\beta}_w^{t+1} = \hat{\beta}_w^t - \left[\frac{\partial U(\beta)}{\partial \beta} \big|_{\beta = \hat{\beta}_w^t} \right]^{-1} U(\hat{\beta}_w^t)$ until convergence. When the working correlation structure, $C_k(\alpha)$, is not working independence, the correlation parameters must be updated after each step t, with $\hat{\alpha}_{wGEE}^{t+1}(\hat{\beta}_w^t)$ then used in the calculation of $\hat{\beta}_w^{t+1}$, as described by [57]. The wGEE estimator for the correlation parameters α solve a second set of estimating equations that are based on the 'sample correlation' $Z_{kij}^t = r_{ki}^t r_{kj}^t$, where r_{ki}^t are the Pearson correlation coefficients corresponding to parameter estimates $\hat{\beta}_w^t$: $r_{ki}^t = \frac{\left(Y_{ki} - \mu_{ki}(\hat{\beta}_w^t)\right)}{\sqrt{\mu_{ki}(\hat{\beta}_w^t)\left(1-\mu_{ki}(\hat{\beta}_w^t)\right)}}$. Defining Z_k^t to be the $n_k(n_k-1)/2$ vector of correlations, $Z_k^t = \{Z_{kij}^t : i < j\}$ and $\delta_k(\alpha) = \{E[Z_{kij}^t] : i < j\}$, the estimating equations for the correlation parameters are

$$\sum_{k=1}^{K} D_k^{*T} V_k^{*-1} W_k^* \left(Z_k^t - \delta_k(\alpha) \right) = 0,$$

where $D_k^* = \partial\delta_k/\partial\alpha$, V_k^* is an $n_k(n_k-1)/2 \times n_k(n_k-1)/2$ working covariance matrix (typically taken to be the identity matrix for ease of computation) and W^* is a symmetric matrix of inverse probability weights where the ij^{th} element is the joint probability of sampling both individual i and individual j, $W^*_{k[ij]} = P(R_{ki} = 1, R_{kj} = 1)^{-1}$, and is dependant on the sampling design. For example, under stratified simple random sampling, $W^*_{k[ij]} = \frac{N_k}{n_k} \cdot \frac{N_k-1}{n_k-1}$, while under cluster-stratified case-control sampling,

$$W^*_{k[ij]} = \begin{cases} \frac{N_{k0}}{n_{k0}} \cdot \frac{N_{k1}}{n_{k1}} & \text{if } y_{ki} \neq y_{kj} \\ \\ \frac{N_{ky_{ki}}}{n_{ky_{ki}}} \cdot \frac{N_{ky_{ki}}-1}{n_{ky_{ki}}-1} & \text{if } y_{ki} = y_{kj} \end{cases}$$

When $C_k(\alpha)$ is the exchangeable correlation structure (*i.e.* $C_{k[ij]}(\alpha) = 1$ if $i = j$ and ρ if $i \neq j$), $E[Z^t_{kij}] = \rho$ for any $i < j$, and $D_k^* = \partial\delta_k/\partial\alpha$ is therefore a vector of ones. Taking the working covariance matrix V^* to be the identity matrix, $D_k^{*^T} V_k^{*^{-1}} W_k^* \left(Z_k^t - \delta_k(\alpha)\right) = 1^T W_k^* \left(Z_k^t - \delta_k(\alpha)\right) = \sum_{i<j} W^*_{k[ij]}(Z^t_{kij} - \rho)$, and solving the estimating equations gives the closed-form estimate of ρ given $\hat{\beta}^t_w$:

$$\hat{\rho}^{t+1} = \frac{\sum_{k=1}^K \sum_{i<j} W^*_{k[ij]} r^t_{ki} r^t_{kj}}{\sum_{k=1}^K \sum_{i<j} W^*_{k[ij]}}$$

The estimate $\hat{\rho}^{t+1}$ is then used at step $t+1$ of the Newton-Raphson algorithm to find $\hat{\beta}^{t+1}_w$.

3.4 Simulation I

In this section we report on a simulation study evaluating operating characteristics of the proposed wGEE estimators for cluster-correlated data under an outcome-dependent sampling scheme as well as an investigation of the effect of design choice on efficiency. Specifically, we investigate: (i) the small-sample properties of the wGEE estimators, (ii) the magnitude of efficiency gains, if any, associated with the outcome-dependent sampling schemes compared to random sampling schemes as well as cluster-stratified sampling schemes compared to unstratified sampling schemes, (iii) the effect of stratification by cluster on estimator operating characteristics, and (iv) the effect of cluster-size distributions on estimator operating characteristics.

3.4.1 Simulation set-up

At the outset, we initially generated 10,000 simulated datasets with one binary outcome variable, Y, and three covariates– a patient-level continuous variable X^w_1, a patient-level binary variable X^w_2, and a cluster-level binary variable X^b_3– under the following 'baseline' scenario referred to as S0. For each of $K = 100$ clusters we set the cluster size to be $N_k = 2,000$. Individual values of X^w_1 were drawn from a Normal$(\mu_{x^w_1 k}, \sigma^2_{x^w_1})$

distribution, with $\mu_{x_1^w k}$ fixed at the quantiles of a Normal$(35, 4^2)$ distribution and $\sigma_{x_1^w} = 10$. Let $Q_{x_2^w k} = P(X_2^w = 1|\text{cluster } k)$ denote the marginal prevalence of covariate X_2^w in the k^{th} cluster. Values across the K clusters for $Q_{x_2^w k}$ were fixed at the quantiles of a Normal$(0.2, 0.05^2)$ distribution, with assignment to specific values for $Q_{x_2^w k}$ randomly permuted across the 10,000 simulated datasets. Individual values for X_2^w were generated as random deviates from a Bernoulli$(Q_{x_2^w k})$ distribution, according to cluster membership of the individual. The percentage of clusters with binary covariate $X_3^b = 1$ was set at 30%; that is, for each dataset 30 clusters were randomly assigned $X_3^b = 1$. Outcomes were generated from a mixed effects model with fixed intercept parameter $\beta_0^* = logit(0.1)$ and cluster-specific random intercepts, b_k, generated from a Normal$(0, 0.5^2)$ distribution. Given these covariate values, outcomes were random draws from a Bernoulli(π_{ki}) distribution with π_{ki} given by

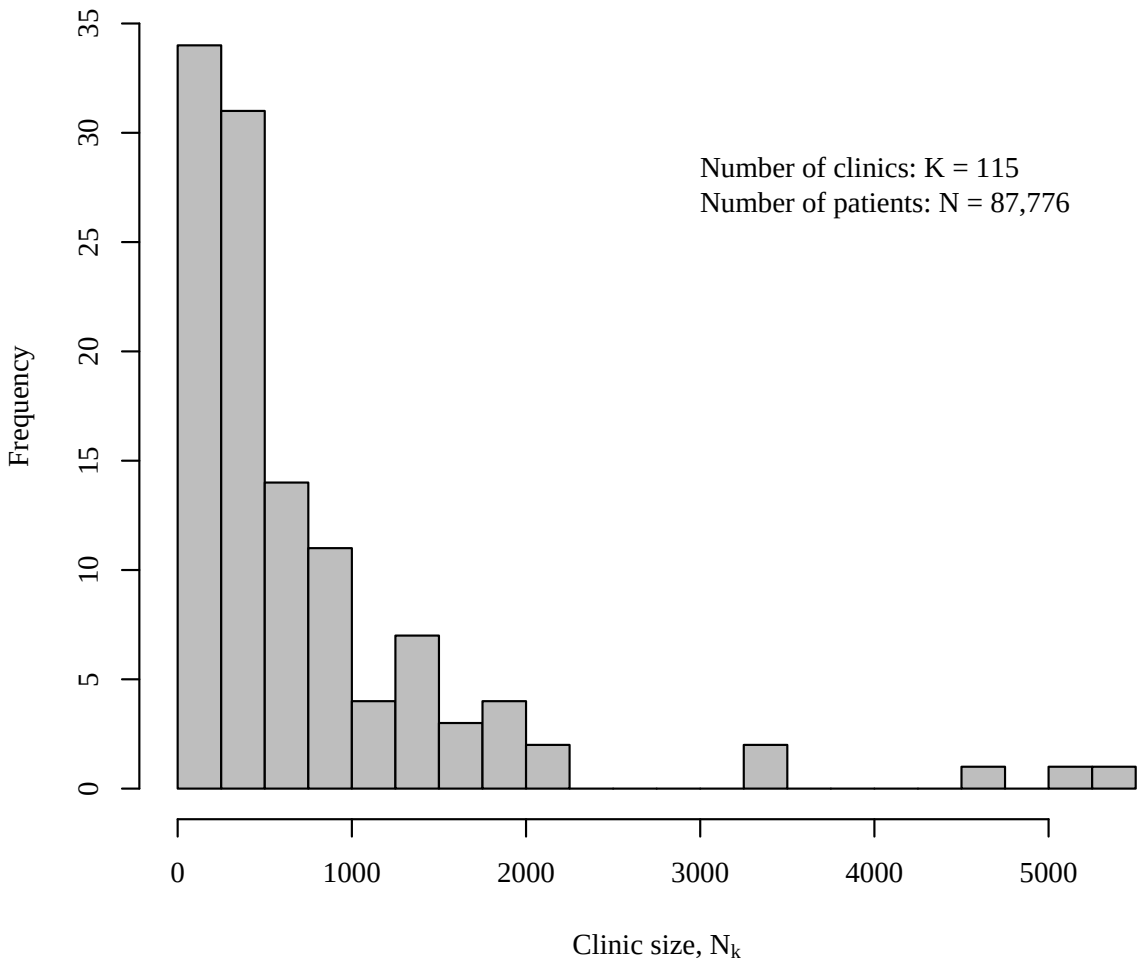
$$logit\ P(Y_{ki} = 1|X_{ki}, k) = (\beta_0^* + b_k) + \beta_1^* X_{1ki}^w + \beta_2^* X_{2ki}^w + \beta_3^* X_{3ki}^b, \tag{3.5}$$

where $(\beta_1^*, \beta_2^*, \beta_3^*)$=(log 1.03, log 1.25, log 1.5).

As seen in Figure 3.4.1, Malawi's ART clinics are not balanced in terms of clinic size, with over half treating fewer than 500 individuals and 20% of clinics treating more than 1,000 individuals, including five clinics that treat more than 3,000 individuals. Taking a balanced stratified case-control sample from the Malawi clinics would therefore result in a large variation across the inverse-probability-sampling weights. When weights vary greatly across clusters, the values of $\hat{\beta}_w$ that solve (3.3) tend to be driven by information from the clusters with large inverse-probability weights, reducing the effective sample size of the sampled data. To investigate this, we considered two additional cluster-size scenarios, both of which had an overall population total of 200,000 individuals: simulation scenario S1 in which cluster sizes N_k ranged from 70 to 2,520 individuals and simulation scenario S2 in which cluster sizes ranged from 51 to 6,075 individuals. Cluster sizes for these simulations were generated once using a Gamma$(2, 0.5)$ and a Gamma$(0.5, 2)$ distribution, respectively. A detailed description of the cluster-size algorithm is given inChapter 4.2, along with histograms of the resulting cluster sizes.

Finally, one data feature not captured in the scenarios described above is the interplay between the within-cluster variation and the between-cluster variation and estimator efficiency. To investigate this, we repeated each of the three simulation scenarios, i.e. (S0, S1, S2), increasing the between-versus-within variation in X_1^w. Specifically, we drew individual values from a Normal$(\mu_{x_1^w k}, \sigma^2_{x_1^w})$ distribution, with $\mu_{x_1^w k}$ fixed at the quantiles of a Normal$(35, 10^2)$ distribution and $\sigma_{x_1^w} = 4$. The motivating interest behind these greater within-cluster variation simulations was in determining the effect of between-cluster variability on estimators based on cluster-stratified samples.

Figure 3.4.1: Distribution of clinic sizes in Malawi data.



3.4.2 Design

For each of the 10,000 simulated datasets, generated under each of the six simulation scenarios detailed above, we sampled 4,000 individuals (2% of the total population) from the population under six different designs, drawing:

(i) a random sample (RS) of $n = 4,000$ individuals from the population

(ii) a case-control sample (CC) of $n_0 = n_1 = 2{,}000$ non-cases and cases from the population

(iii) a balanced cluster-stratified random sample (BSRS) of $n_k = 40$ individuals from each of the K clusters

(iv) a balanced cluster-stratified case-control sample (BSCC) of $n_{0k} = n_{1k} = 20$ non-cases and cases from each of the K clusters

(v) an unbalanced cluster-stratified random sample (USRS) of $n_k = 100$ individuals from the 15 largest clusters, $n_k = 20$ individuals from the 45 smallest clusters, and $n_k = 20$ individuals from all other clusters

(vi) an unbalanced cluster-stratified case-control sample (USCC) of $n_{0k} = n_{1k} = 50$ non-cases and cases from the 15 largest clusters, $n_{0k} = n_{1k} = 10$ non-cases and cases from the 45 smallest clusters, and $n_{0k} = n_{1k} = 20$ non-cases and cases from all other clusters

In the event that fewer than n_{1k} cases were observed in a clinic, additional controls were sampled from the clinic to ensure a sample size of 4,000.

3.4.3 Analyses

For the purposes of these simulations, interest lies in the marginal model

$$logit\ P(Y_{ki} = 1|X_{ki}, k) = \beta_0 + \beta_1 X^w_{1ki} + \beta_2 X^w_{2ki} + \beta_3 X^b_{3ki} \tag{3.6}$$

For each of the designs in Section 3.4.2 we estimated $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)$ by fitting (3.6) with a (naïve) GLM and by using wGEE, the approach proposed in Section 3.3.1. For the latter we considered both an independent and an exchangeable correlation structure, with the exception of the full-data model, for which only a working-independence correlation structure is assumed for computational efficiency.

Across all analyses we computed a series of operating characteristics: percent bias, relative uncertainty, coverage of Wald-based 95% confidence intervals, and the SE-to-SD ratio. Percent bias of an estimate, $\hat{\beta}$, was calculated as $(\hat{\beta} - \beta)/\beta \times 100$ where the 'truth' was taken to be the mean of the full-data GEE point estimates assuming an independent covariance structure: $\beta = (\beta_0, \beta_1, \beta_2, \beta_3) = (-2.09, 0.03, 0.21, 0.38)$. Relative uncertainty is defined as the ratio of the standard errors of two estimators and can be interpreted as the relative difference in widths of the corresponding Wald-based 95% confidence intervals. Throughout the wGEE estimator based on the CC sample and assuming a working independence correlation structure was taken as the referent. Coverage is the percentage of the 10,000 95% confidence intervals that contain the 'truth', that is $\hat{\beta}$ from the full analyses. Finally, the SE-to-SD ratio is the ratio of the mean of the 10,000 standard error estimates to the standard error of the 10,000 point estimates; values less/greater than 1.0 indicate that the standard error is under/over estimated.

3.4.4 Results

Tables 3.4.1, 3.4.2 and 3.4.3 report on operating characteristics for the GLM and wGEE estimators of β, respectively. Based on these results we draw several conclusions. First, from Table 3.4.1, naïve GLM point estimates are subject to bias under CC and BSCC/USCC sampling, as expected. In particular, the intercept

parameter estimate $\hat{\beta}_0$ was biased (-43.5 to -52.8%) under the three case-control sampling designs, as is expected with case-control sampling, while the patient-level parameter estimate $\hat{\beta}_2$ presented mild bias (2.8 to 4.4%) and the cluster-level parameter estimate $\hat{\beta}_3$ exhibited substantial bias (-127.4 to -143.9%) under the BSCC and USCC sampling schemes. Further, as expected, naïve GLM variance estimates are generally underestimated; in the complete data setting resulting naïve confidence intervals are between 9 and 72% as wide as the robust confidence intervals. Further, across the random sampling schemes (RS, BSRS, USRS), confidence intervals for the cluster-level parameter β_3 are between 42 and 60% too small, and under a case-control design underestimation in most apparent in simulation scenario S2, in which confidence intervals are between 40 and 95% too small. Note, the apparent overestimation of the BSCC/USCC variance estimates is of limited interest due to the large bias observed in the point estimates.

From Table 3.4.2, the proposed wGEE estimator corrects both the bias observed under the CC sampling scheme and the underestimation of the variance. The wGEE method of estimation results in absolute percent bias less than 5.9 across all parameters and all sampling schemes. Further, the 95% confidence-interval coverage rates range from 91.1 to 95.0% under the outcome-dependent sampling schemes and across all simulations, with an average coverage rate of 93.7%. Point estimates, the ratio of standard error estimates to the standard deviation of the observed point estimates, and confidence-interval coverage rates are presented in tabular format in Table 5.3.1 of Chapter 4.3. The ratio of the wGEE standard error estimates to the standard deviation of the observed point estimates averages 97.2% and ranges from 0.90 to 1.00 (Table 5.3.1).

Third, as is expected, case-control sampling provided large efficiency gains over random sampling in all three simulations. The relative uncertainty of case-control designs compared to the corresponding random sampling designs (CC vs. RS, BSCC vs. BSRS, and USCC vs. USRS) ranges from $89.9/117.4 \cdot 100 = 76.6$ to $100.0/101.4 \cdot 100 = 98.6$ percent, with an average relative uncertainty of 92.5% (Table 5.3.2 of Chapter 4.3).

Fourth, stratification in the sampling design influences estimator efficiency; parameter estimates for variables with little or no within-cluster variability tend to benefit from stratification, while those for variables with more within-cluster variability than between-cluster variability may be harmed by stratification. In simulation scenario S0 and under a working-exchangeable correlation structure for example, the relative uncertainty of the estimate of the cluster-level parameter β_3 from a BSCC sample compared to a CC sample is 86.5, while the estimation of the within-cluster parameters are slightly less efficient under BSCC sampling than under CC sampling, with relative uncertainties of 102.6 and 101.1 for β_1 and β_2, respectively (Table 5.3.3 of Chapter 4.3).

Fifth, the effect of stratification on estimator efficiency is highly dependent on variability in the cluster sizes N_k. In simulation scenario S0 and under a working-exchangeable correlation structure, in which all clusters are equally-sized, the relative uncertainty of point estimates from the BSCC design compared to the

CC design range from 86.5 for the cluster-level parameter β_3 to 102.6 for the individual-level continuous-variable parameter β_1. In simulation scenario S1, where cluster sizes ranged from 70 to 2,520, the relative uncertainty of point estimates from the BSCC and USCC designs compared to the CC design ranged from 96.1 for β_3 to 122.2 for β_2. Finally, in simulation scenario S2, with cluster sizes ranging from 51 to 6,097, point estimates from the BSCC were uniformly less efficient than those from the CC design, with relative uncertainties ranging from 117.8 to 172.6 (Table 5.3.3 of Chapter 4.3).

Sixth, when cluster sizes vary considerably but stratification is a logistically desirable feature of the study design, unbalanced sampling can mitigate efficiency losses. Under a working exchangeable correlation structure, the relative uncertainty of the USCC point estimates compared to the BSCC point estimates ranged from 88.3 to 91.6 in the moderate-cluster-size-variability simulation scenario S1 and from 70.6 to 83.0 in the large-cluster-size-variability simulation scenario S2 (Table 5.3.4 of Chapter 4.3). With the exception of the cluster-level parameter, β_3, the USCC point estimates remained less efficient than those for the CC study design, but the efficiency losses are mitigated; in simulation scenario S2 under a working-independence correlation structure for example, the relative uncertainty of the estimates of individual-level parameters (β_1, β_2) under a BSCC study design compared to a CC study design is (154.7, 158.0), while under the USCC study design the respective relative uncertainties drop to (114.2, 116.5). (Table 3.4.2)

Finally, the relative uncertainty of the β_1 estimate from a cluster-stratified case-control sample compared to a case-control sample is reduced when there is more between-cluster variation than within-cluster variation in X_1^w. Specifically, under a working-independence correlation structure, the relative uncertainty of the β_1 estimate from a BSCC sampling design compared to a CC sampling design is reduced from (100.2, 114.7, 154.7) (Table 3.4.2) to (86.0, 87.1, 101.2) (Table 3.4.3) in simulation scenarios (S0, S1, S2), respectively, when the individual-level covariate X_1^w is modified to exhibit greater between- than within-cluster variation. Finally, the relative uncertainty of the β_1 estimate from a cluster-stratified case-control sample compared to a case-control sample is reduced when there is more between-cluster variation than within-cluster variation in X_1^w. Specifically, under a working-independence correlation structure, the relative uncertainty of the β_1 estimate from a BSCC sampling design compared to a CC sampling design is reduced from (100.2, 114.7, 154.7) (Table 3.4.2) to (86.0, 87.1, 101.2) (Table 3.4.3) in simulation scenarios (S0, S1, S2), respectively, when the individual-level covariate X_1^w is modified to exhibit greater between- than within-cluster variation.

3.5 Simulation II

While the simulations of Section 3.4 consider operating characteristics in general settings, here we present a series of simulations geared specifically toward the Malawian context. Though a cluster-stratified study design may be logistically desirable in resource-limited settings such as Malawi, the results from Table 3.4.2 suggest that the statistical efficiency of a cluster-stratified study design may be less than that of a traditional

case-control study. Indeed, the clinic sizes in the Malawi ART program are quite varied (Figure 3.4.1), most closely matching the large cluster size variability in simulation scenario S2 in Section 3.4. This, in turn, suggests that naïvely choosing a balanced cluster-stratified study design may result in unnecessary estimator efficiency losses. In this section we conduct a simulation study to (i) evaluate the performance of the wGEE estimator in a real-world setting, and (ii) demonstrate that cluster-stratified study designs can provide important efficiency gains in resource-poor settings such as Malawi.

3.5.1 Data

Between 2005 and 2007, the Malawian MOHP performed a one-time cross-sectional survey of patients registered in the national ART program at that time. Here we restrict attention to $N = 87,776$ patients who were aged ≥ 16 years with complete demographic data and at least six months of follow-up. Patient characteristics are presented in Table 3.5.1. Our outcome of interest is the binary indicator, Y_{ki}, which represents 'status at six months post-registration' (0- patient was alive or had transferred out 180 days after registration / 1- patient had died, defaulted, or stopped treatment within 180 days of registration). The goal of our hypothetical study is to evaluate the relationship between a set of patient and clinic characteristics and the outcome.

3.5.2 Simulation Set-up and Analysis

Following the notation in Section 3.3, we consider a logistic model for the marginal probability $\pi_{ki} = E(Y_{ki} = 1|X_{ki})$:

$$\begin{aligned} \text{logit}(\pi_{ki}) &= \beta_0 + \beta_1 I(W_{ki} = 1/2) + \beta_2 I(W_{ki} = 3) + \beta_3 G_{ki} + \beta_4 A^*_{ki} + \\ &\quad \beta_5 I(T_{ki} = 2005) + \beta_6 I(T_{ki} = 2006) + \beta_7 R_k + \beta_8 P_k \end{aligned} \tag{3.7}$$

where W represents WHO Clinical Stage, a clinical classification of HIV/AIDS infection stage used in resource-limited settings in lieu of laboratory-based measurements such as CD4-counts [81], G is gender (0/1 male/female), A^* is age standardized so that 'zero' corresponds to age 35 and a one-unit change corresponds to a 10-year contrast, T is registration year, that is the year in which an individual enrolled in the national ART program, R is region (0/1 south/central or north), and P is clinic type (0/1 public/private). In the complete data setting, an analysis would be performed using GEE to account for cluster correlation. For the Malawian ART program, complete data collection is logistically infeasible as a general monitoring and evaluation strategy.

We generated 10,000 simulated data sets, each of size 87,776. Clinic sizes and patient covariates were fixed at the original data values. To induce correlation within clinics we specified the underlying data generating model of the conditional probability π^*_{ki} to be a random effects model identical to (3.7), with the

exception of β_0, which is replaced by $(\beta_0 + \gamma_k)$, where $\gamma_k \sim N(0, \tau^2)$. Fixed (β, τ) parameter values were taken from fitting the random effects model to the original data. For each simulated data set, random effects γ were generated from a random normal distribution with mean zero and standard deviation $\tau = 0.41$, and outcomes Y_{ki} were generated as a random Bernoulli draw with probability π^*_{ki}.

Following Section 3.4.2, for each of the 10,000 datasets we drew three case-control subsamples (CC, BSCC, USCC). The first two sampling schemes are as described in Section 3.4.2, though the total sample size of the unstratified case-control sample was adjusted to account for the number of clinics ($n = K \cdot 40 = 115 \cdot 40 = 4600$). For the unbalanced stratified case-control subsample (USCC), we sampled 5 cases and non-cases from each clinic with fewer than 250 patients and $n_{yk} = (10, 20, 40, 104)$ cases and non-cases from clinics with (250-499, 500-999, 1000-2999, 3000+) patients, respectively. In all stratified case-control samples, a total of $2 \cdot n_{yk}$ individuals were sampled per clinic, with additional controls being sampled in the rare case that a clinic had fewer than n_{yk} cases. The aim of this unbalanced sampling scheme was to decrease variation in inverse probability weights.

We again used the naïve GLM and wGEE (exchangeable and independent working correlation matrices) to fit the marginal model (3.7) to the full population and each of the six sampling designs. The "true" value of the marginal $\beta = (\beta_0, \ldots, \beta_8)$ parameters was taken to be the mean of the full-data wGEE estimates assuming an independent covariance structure. In addition to the set of operating characteristics explored in Section 3.4 we also consider power, defined as the percent of simulations for which the 95% Wald-based confidence intervals do not include zero.

3.5.3 Results

Table 3.5.2 provides results. Overall we draw parallel conclusions to those reported in Section 3.4.4. Additionally, we make the following observations. First, the Malawi simulations underscore the dangers of improper analyses applied to a cluster-stratified case-control sample. While the results from Section 3.4.4 suggest that only intercept and cluster-level parameters are subject to bias when a GLM is fit to the cluster-stratified data, Table 3.5.2 indicates that this is not universally true. In particular, the individual-level WHO stage 1/2 parameter β_1 and registration year parameter β_5 are -13.9% and -17.7% biased, respectively. All parameter estimates from GLM and unweighted GEE analyses are subject to bias if observations are correlated within strata of a stratified case-control study.

Second, the statistical drawbacks of the BSCC sampling design compared to the CC design are almost entirely removed by applying a USCC sampling design. As expected given the variation in cluster sizes, a BSCC design yields much less efficient within-cluster parameter estimates than a traditional CC design; under a working-independence correlation structure, the relative uncertainty of within-cluster parameters in the BSCC design compared to a CC design ranged from 145.8 to 174.3. In comparison, efficiency losses in the within-cluster parameters are greatly reduced when an unbalanced stratified sampling design is used;

the relative uncertainty of within-cluster parameters in the USCC design compared to the CC design ranged from 100.6 to 104.6. In addition to the USCC sampling design being nearly as efficient as the CC sampling design for within-cluster parameters, it is substantially more efficient than the CC design for the two cluster-level parameters, with relative uncertainties of 88.9 and 60.1 for the region and clinic type parameters β_7 and β_8, respectively.

Finally, the decision to collect a stratified or unstratified sample impacts statistical power, as does the design. In our simulations, a BSCC design resulted in a reduction in power to detect the significance of individual-level variables that ranged from $100.0 - 100.0 = 0$ to $78.7 - 43.7 = 35.0$ percentage points when compared to an unstratified case-control design and a $91.7 - 58.3 = 33.4$-percentage-point increase in power for the cluster-level hospital-type parameter β_8. In contrast, the USCC design lost no more than $78.7 - 74.4 = 4.3$ percentage points in power to detect statistical significance when compared to the CC design and saw a similar gain of $91.1 - 58.3 = 32.8$ percentage points in the power to detect significance in the hospital-type parameter β_8. The USCC sampling design is not only a logistically feasible sampling design for monitoring the Malawian national ART program; under wGEE analysis it is also nearly as statistically efficient as the case-control sampling design for within-cluster parameters and substantially more efficient than the case-control design for cluster-level covariates, especially for the parameter associated with the rare private-clinic covariate.

3.6 Discussion

In this paper we consider outcome dependent sampling schemes in settings for which study units are cluster-correlated. Applying naïve likelihood-based GLM estimators is inappropriate under outcome dependent sampling schemes when individuals are cluster-correlated. We have established valid estimation and inference techniques that can be applied to any case-control study for which cluster correlation exists. Due to its practical utility, the case-control study remains a widely used design; in the last five years over 100 case-control studies have been published in each of The Lancet and JAMA. Cluster correlation is frequently present in case-control study environments– cluster-correlation of individuals may occur within geographical location, hospital, and treating doctor for example– but in the vast majority of applications, researchers ignore the underlying correlation structure if it exists. As a result, case-control studies may suffer from a systemic issue of invalid inference, specifically underestimation of the variance. The wGEE methods presented in this paper provide researchers with the tools needed to properly analyze case-control studies in the presence of cluster correlation.

We additionally proposed the use of a cluster-stratified case-control study in resource-poor settings as a valid and logistically reasonable study design. We demonstrated that GLM point estimates estimated under any stratified case-control design with cluster-correlation are subject to bias. The marginal model wGEE point estimates eliminate bias and provide valid inference under stratified case-control sampling

with proper inverse probability weighting. The cluster-stratified case-control study design may be subject to a practical trade-off between logistic considerations and statistical efficiency. Efficiency gains and losses of the cluster-stratified case-control design compared to a case-control design depend on (i) the type of covariate (cluster- or individual-level) and (ii) the variation in cluster sizes. BSCC sampling may result in an extreme variation in inverse-probability weights across clusters when population cluster sizes are highly varied, which can reduce the effective sample size of a BSCC study design compared to a CC study design and cause stratification on cluster to be a less efficient study design. In deciding on a study design, one must consider the research goal; the cluster-stratified case-control design can be a more efficient design than the case-control design when cluster-level covariates are of particular interest. When the interest lies with individual-level covariates and cluster sizes are greatly varied, we recommend the use of a USCC design rather than a BSCC design. In this scenario a USCC design can dramatically reduce the efficiency losses a BSCC design would yield compared to a case-control design while maintaining the logistical feasibility of the within-cluster sampling scheme.

The methods presented in this paper will benefit from a few additional avenues of research. First, research into the optimal allocation of resources within a USCC design is needed to recommend an ideal study design. Second, inverse probability weighting techniques are generally known to be inefficient. It may be possible to adapt the wGEE to create more efficient estimators in the outcome-dependent sampling setting, along the lines of [63]. Third, two-phase designs provide a framework within which aggregated data is used to identify sub-samples of patients on whom detailed information is collected and both aggregated and individual-level data are incorporated into the analysis [8, 69, 75]. Under outcome-dependent sampling, the two-phase weighted likelihood method provides the same point estimates as a wGEE model assuming working independence. Two additional estimators that are more efficient than the weighted likelihood have been proposed for the analysis of a two-phase design: pseudo likelihood and maximum likelihood [9]. Two-phase likelihood theory could potentially be used to create more efficient estimators for outcome-dependent sampling designs in cluster-correlated data settings. Finally, conditional models may be a desirably method of dealing with cluster-correlated data, depending on the study objectives; marginal model parameter interpretations are not applicable to all studies.

Table 3.4.1: Operating characteristics for the GLM estimators of β from model (3.6) using the full data and six subsamples, under three cluster-size simulation scenarios, as described in Section 3.4.1. All values are based on 10,000 simulated datasets.

		Complete	**No Stratification**		**Cluster-Stratified**			
		Data	RS	CC	BSRS	USRS	BSCC	USCC
Percent Bias[a]								
S0	β_0	0	0.1	-43.6	-0.0	0.2	-52.8	-52.7
	β_1	0	0.2	0.2	-0.0	0.4	-1.1	-1.0
	β_2	0	0.4	-0.4	0.0	-0.4	3.2	2.8
	β_3	0	-0.2	0.1	-0.4	0.2	-143.8	-143.9
S1	β_0	0	0.3	-43.5	-0.0	0.1	-52.3	-52.6
	β_1	0	0.5	0.3	-0.1	0.2	-1.0	-0.9
	β_2	0	-0.0	0.1	-0.8	-0.1	4.4	3.6
	β_3	0	-0.0	-0.0	0.1	-0.2	-142.4	-143.9
S2	β_0	0	-0.0	-43.7	-0.1	0.1	-48.3	-52.4
	β_1	0	-0.1	0.0	-0.1	0.2	-1.0	-1.2
	β_2	0	-0.6	-0.1	-0.7	-0.0	3.0	3.7
	β_3	0	0.0	-0.0	0.2	0.4	-127.4	-142.2
SE/SD[b]								
S0	β_0	0.25	0.88	0.93	0.88	0.85	1.10	1.09
	β_1	0.41	0.95	0.94	0.95	0.94	1.04	1.03
	β_2	0.65	0.98	0.98	0.98	0.98	1.02	1.01
	β_3	0.10	0.60	0.57	0.60	0.53	3.39	3.15
S1	β_0	0.31	0.85	0.90	0.88	0.85	1.08	1.10
	β_1	0.49	0.95	0.93	0.96	0.95	1.02	1.05
	β_2	0.72	0.98	0.99	0.98	0.97	1.01	1.00
	β_3	0.13	0.53	0.51	0.60	0.53	3.02	3.15
S2	β_0	0.23	0.75	0.81	0.89	0.85	1.03	1.08
	β_1	0.38	0.88	0.89	0.96	0.94	1.00	1.02
	β_2	0.61	0.97	0.95	0.98	0.98	1.00	0.99
	β_3	0.09	0.42	0.40	0.60	0.53	1.57	2.94

[a] Percent bias of an estimate $\hat{\beta}$ relative to the truth β is defined to be $(\hat{\beta} - \beta)/\beta \times 100$. Here the 'truth' is taken to be the full data mean point estimate.

[b] SE/SD is the ratio of the mean of the 10,000 estimated standard errors divided by the standard deviation of the 10,000 point estimates.

Table 3.4.2: Operating characteristics for the wGEE estimator of β from model (3.6) using the full data and six subsamples, under three varied cluster-size simulation scenarios, as described in Section 3.4.1. All values are based on 10,000 simulated datasets. Both the independent (Ind) and exchangeable (Exch) correlation structures are used.

Greater within- than between-cluster variation in X_1^w

			No Stratification				**Cluster-Stratified**							
		Complete	RS		CC		BSRS		USRS		BSCC		USCC	
		Data	Ind	Exch	Ind	Exch	Ind	Exch	Ind	Exch	Ind	Exch	Ind	Exch
Percent Bias[a]														
	β_1	0.0	0.2	0.3	0.2	-0.3	0.1	0.2	0.6	0.7	-0.2	4.6	0.3	5.1
So	β_2	0.0	0.4	0.4	-0.5	-1.1	-0.1	-0.0	-0.8	-0.8	-0.5	4.4	-0.8	4.1
	β_3	0.0	-0.0	-0.2	0.5	-0.2	-0.2	-0.3	0.8	0.8	0.3	-1.8	0.1	-1.9
	β_1	0.0	0.5	0.4	0.2	-0.4	0.0	-0.0	0.0	0.0	0.2	4.9	0.4	5.1
S1	β_2	0.0	-0.0	-0.1	0.0	-0.5	-0.5	-0.4	-0.3	-0.2	1.1	5.9	0.5	5.4
	β_3	0.0	0.5	0.7	0.7	0.5	1.1	1.0	0.6	0.7	0.7	-1.2	0.6	-1.3
	β_1	0.0	-0.1	-0.2	0.0	-0.6	0.7	0.6	0.4	0.3	0.4	4.8	-0.1	4.3
S2	β_2	0.0	-0.6	-0.5	-0.2	-0.8	-1.3	-1.4	0.1	-0.0	-0.2	4.5	0.5	5.1
	β_3	0.0	-0.1	0.2	-0.1	0.3	-0.5	-0.5	0.3	0.7	-0.2	-2.0	-0.0	-1.6
Coverage[b]														
	β_1	94.1	94.2	94.2	95.0	94.6	94.3	94.4	94.6	94.8	94.6	92.7	94.1	92.5
So	β_2	95.0	94.3	94.4	94.4	94.4	94.4	94.3	94.7	94.5	94.4	94.4	94.5	94.6
	β_3	94.4	94.1	94.1	94.6	94.5	94.2	94.2	94.3	94.2	94.1	94.2	94.2	94.1
	β_1	93.9	94.2	94.6	94.4	94.6	94.0	94.0	94.4	94.5	94.2	92.6	94.9	92.7
S1	β_2	93.8	94.2	94.4	94.6	94.7	94.1	94.1	94.6	94.5	94.2	94.3	93.8	93.7
	β_3	93.3	93.2	93.9	93.5	94.0	93.7	93.6	93.4	93.7	93.4	93.4	93.3	93.9
	β_1	92.7	93.1	93.6	94.1	94.4	92.9	92.9	93.2	93.7	92.7	91.6	92.9	91.9
S2	β_2	92.9	94.1	94.0	93.6	93.5	92.6	92.7	93.6	93.5	93.1	92.8	93.1	92.9
	β_3	91.3	91.6	93.5	91.4	93.9	91.1	91.4	91.5	92.7	91.3	91.2	91.1	92.6
Relative Uncertainty[c]														
	β_1	35.9	110.0	104.8	100.0	95.5	109.9	105.2	126.0	122.6	100.2	98.0	116.4	115.9
So	β_2	22.7	106.7	104.2	100.0	97.3	108.1	105.3	125.1	123.2	99.5	98.4	117.0	116.5
	β_3	82.5	102.5	101.8	100.0	99.0	102.1	102.1	107.9	117.4	85.3	85.6	85.8	89.9
	β_1	42.3	108.6	102.6	100.0	93.9	125.0	120.0	112.7	107.5	114.7	113.2	103.0	99.9
S1	β_2	29.6	108.9	105.7	100.0	96.3	127.3	124.2	113.8	111.1	118.9	117.6	109.4	107.6
	β_3	87.9	103.1	97.0	100.0	93.0	105.9	105.9	103.6	100.5	88.9	89.4	88.8	81.9
	β_1	52.1	111.5	101.3	100.0	88.1	165.0	159.7	122.0	113.8	154.7	152.1	114.2	107.4
S2	β_2	33.7	106.5	102.0	100.0	94.7	171.7	168.1	122.4	118.1	158.0	156.7	116.5	113.4
	β_3	92.6	101.4	83.8	100.0	80.7	113.9	113.8	104.0	89.8	94.5	95.1	93.8	79.0

[a] Percent bias of an estimate $\hat{\beta}$ relative to the truth† β is defined to be $100 \cdot (\hat{\beta} - \beta)/\beta$.

[b] Coverge is the percent of simulations for which confidence intervals include the truth† β.

[c] Relative uncertainty of an estimate $\hat{\beta}$ relative to the estimate $\hat{\beta}^*$ is defined to be $100 \cdot sd(\hat{\beta})/sd(\hat{\beta}^*)$. Here relative uncertainty is with respect to a wGEE model fit to an unstratified case-control sample, assuming an independent correlation structure.

† Here the 'truth' is taken to be the full data mean point estimate.

Table 3.4.3: Relative Uncertainty[a] for the wGEE estimator of β from model (3.6) in which greater between- than within-cluster variation is present in the covariate X_1^w, using the full data and six subsamples, under three varied cluster-size simulation scenarios, as described in Section 3.4.1. All values are based on 10,000 simulated datasets. Both the independent (Ind) and exchangeable (Exch) correlation structures are used.

Greater between- than within-cluster variation in X_1^w.

			No Stratification				**Cluster-Stratified**							
		Complete	RS		CC		BSRS		USRS		BSCC		USCC	
		Data	Ind	Exch	Ind	Exch	Ind	Exch	Ind	Exch	Ind	Exch	Ind	Exch
S0	β_1	77.6	106.0	94.8	100.0	87.0	104.3	93.9	112.3	109.7	86.0	82.4	89.0	91.4
	β_2	22.5	106.4	104.1	100.0	96.6	106.6	104.1	122.9	120.8	98.1	97.1	114.7	114.3
	β_3	81.5	103.8	98.7	100.0	94.1	102.8	98.1	109.1	113.2	85.3	83.6	86.6	89.6
S1	β_1	81.7	104.2	86.9	100.0	80.4	109.3	98.5	104.4	89.7	90.8	87.1	89.3	79.6
	β_2	28.6	107.7	104.5	100.0	96.5	123.6	120.5	111.7	108.9	115.6	114.4	103.3	101.9
	β_3	86.3	103.3	91.4	100.0	87.2	108.3	103.7	104.3	95.9	90.0	87.9	89.6	80.5
S2	β_1	88.5	103.3	74.3	100.0	67.3	122.7	110.7	109.8	83.2	101.2	98.6	92.4	71.6
	β_2	33.0	105.2	100.6	100.0	94.6	166.2	163.0	120.2	115.3	148.7	146.7	117.1	113.9
	β_3	89.3	101.2	78.3	100.0	76.0	114.5	110.4	104.1	85.3	94.4	93.7	91.6	75.3

[a] Relative uncertainty of an estimate $\hat{\beta}$ relative to the estimate $\hat{\beta}^*$ is defined to be $100 \cdot sd(\hat{\beta})/sd(\hat{\beta}^*)$. Here relative uncertainty is with respect to a wGEE model fit to an unstratified case-control sample, assuming an independent correlation structure.

Table 3.5.1: Characteristics of $N = 87,776$ patients from $K = 115$ clinics from a cross-sectional survey conducted in Malawi of patients enrolled in an ART treatment program between 01/2005 and 12/2007. Mean point estimate and naïve and robust standard errors from complete-data analyses of 10,000 simulated datasets.

			Full Analysis			
			Point Estimate		95% CI	
	N	%	$\hat{\beta}$	$\exp(\hat{\beta})$	Naïve	Robust
Status at six months						
Non-negative	70,753	80.6				
Negative	17,023	19.4				
Total	87,776	100.0				
WHO stage						
4	20,714	23.6	REF	——	——	——
3	62,244	70.9	-0.60	0.55	(-0.64, -0.56)	(-0.66, -0.55)
1/2	4,818	5.5	-1.20	0.30	(-1.31, -1.1)	(-1.39, -1.02)
Gender						
Male	34,461	39.3	REF	——	——	——
Female	53,315	60.7	-0.27	0.76	(-0.31, -0.23)	(-0.31, -0.23)
Age category[†]			-0.09	0.92	(-0.11, -0.07)	(-0.11, -0.07)
16 - 25	9,516	10.8				
26 - 35	35,185	40.1				
36 - 45	27,421	31.2				
46 - 55	11,527	13.1				
56+	4,127	4.7				
Registration year						
2007	33,039	37.6	REF	——	——	——
2006	36,594	41.7	-0.18	0.84	(-0.23, -0.13)	(-0.28, -0.08)
2005	18,143	20.7	-0.08	0.92	(-0.13, -0.04)	(-0.18, 0.01)
Region						
Central/South	74,142	84.5	REF	——	——	——
North	13,634	15.5	-0.09	0.92	(-0.13, -0.04)	(-0.34, 0.17)
Clinic type						
Public	85,999	98.0	REF	——	——	——
Private	1,777	2.0	-0.57	0.57	(-0.71, -0.42)	(-0.86, -0.27)

[†] Age is included in the model as a continuous covariate, with 'zero' corresponding to age 35 and a one-unit change corresponding to a 10-year contrast.

Table 3.5.2: Operating characteristics for the GLM- and wGEE-based estimators of β from model (3.7) using the full data and three subsamples. All values are based on 10,000 simulated datasets, where covariate values and clinics were fixed at the original data values and 10,000 outcome variable vectors were generated using the model described in Section 3.5.2. Both the exchangeable (Exch) and independent (Ind) correlation structures are used in the wGEE analyses.

		Complete	No Strat.		Cluster-Stratified			
		Data	CC		BSCC		USCC	
GLM-based Estimators								
Percent Bias[a]								
Int	β_0	0.0	-203.3		-191.0		0.0	
Who Stage 1/2	β_1	0.0	0.6		-13.9		1.2	
Who Stage 3	β_2	0.0	0.1		-3.6		0.1	
Female	β_3	0.0	-0.3		-2.7		0.5	
Age	β_4	0.0	0.1		-1.0		1.5	
Reg. Year 2005	β_5	0.0	0.7		-17.7		0.0	
Reg. Year 2006	β_6	0.0	0.3		-4.1		0.0	
North	β_7	0.0	0.3		-9.0		2.5	
Private	β_8	0.0	1.5		48.3		3.3	
wGEE-based Estimators								
		Ind	Ind	Exch	Ind	Exch	Ind	Exch
Percent Bias[a]								
Int	β_0	0.0	-0.2	3.9	-0.3	-2.6	-0.2	-2.5
Who Stage 1/2	β_1	0.0	0.5	-0.5	1.7	3.9	0.5	2.4
Who Stage 3	β_2	0.0	0.1	-0.5	0.3	2.5	0.3	2.6
Female	β_3	0.0	-0.2	-1.0	0.4	2.7	0.3	2.7
Age	β_4	0.0	0.1	-0.8	-0.6	1.7	0.1	2.7
Reg. Year 2005	β_5	0.0	1.0	-0.4	-0.5	1.3	-0.4	2.3
Reg. Year 2006	β_6	0.0	0.5	-1.0	0.7	2.7	-0.3	1.9
North	β_7	0.0	0.3	-4.3	1.1	3.5	0.2	0.4
Private	β_8	0.0	1.4	-2.2	1.1	1.6	0.1	-0.8
Relative Uncertainty[b]								
Who Stage 1/2	β_1	60.0	100.0	85.7	174.3	175.7	100.6	89.4
Who Stage 3	β_2	40.6	100.0	96.0	152.6	153.2	101.0	98.6
Female	β_3	35.7	100.0	97.0	160.7	160.7	104.3	102.9
Age	β_4	32.0	100.0	98.3	158.6	158.8	104.6	103.8
Reg. Year 2005	β_5	59.4	100.0	86.5	146.2	143.7	102.1	91.2
Reg. Year 2006	β_6	58.3	100.0	88.3	145.8	145.4	101.7	93.7
North	β_7	88.2	100.0	87.1	90.0	91.2	88.9	72.6
Private	β_8	57.6	100.0	98.3	59.5	60.0	60.1	57.3
Power[c]								
Who Stage 1/2	β_1	100.0	100.0	100.0	95.3	95.7	100.0	100.0
Who Stage 3	β_2	100.0	100.0	100.0	99.8	99.9	100.0	100.0
Female	β_3	100.0	98.9	99.0	77.6	78.6	98.0	98.6
Age	β_4	100.0	78.7	78.7	43.7	45.3	74.4	76.8
Reg. Year 2005	β_5	48.5	17.3	18.1	13.1	13.1	17.4	18.5
Reg. Year 2006	β_6	90.4	49.7	55.7	31.7	32.3	49.3	56.0
North	β_7	16.6	14.7	12.3	16.4	16.2	16.4	14.7
Private	β_8	93.0	58.3	55.7	91.7	91.5	91.1	92.5

[a] Here the truth β is taken to be the full data mean point estimate.

[b] Here relative uncertainty is with respect to a wGEE model fit to an unstratified case-control sample, assuming an independent correlation structure.

[c] Power is the percent of simulations for which confidence intervals do not include zero.

4

Supplementary Material for: On the Analysis of Case-Control and Stratified Case-Control Studies in Cluster-Correlated Data Settings

4.1 Estimator Consistency

For any subsample of the data, the inverse-probability-weighted GEE parameter estimates, $\hat{\beta}_w$, solve the pseudo score equations $\sum_{k=1}^{K} U_k^{n_k}(\beta) = 0$, where $U_k^{n_k}(\beta) = D_k^{n_k T}(\beta) \left\{V_k^{n_k}\right\}^{-1} W_k^{n_k} \left(Y_k^{n_k} - \mu_k^{n_k}(\beta)\right)$ and $W_k^{n_k}$ is an $n_k \times n_k$ diagonal matrix of inverse probability weights, $V_k^{n_k}$ is a working correlation matrix, and $D_k^{n_k}$, $Y_k^{n_k}$, and $\mu_k^{n_k}$ are the the vectors in Section 3 subset to the n_k sampled individuals. To prove that the inverse-probability-weighted GEE provides consistent estimation of the parameters under case-control sampling, we need to re-express the stratum-specific pseudo-score function in terms of the complete population as $U_k(\beta) = D_k^T(\beta) V_k^{-1} W_k R_k \left(Y_k - \mu_k(\beta)\right)$, where D_k, Y_k, and μ_k are the the complete-data vectors described in Section 3. Without loss of generality, we assume that the individuals are ordered by sampling status; that is that individuals $i = 1, \ldots, n_k$ were sampled and individuals $i = n_k + 1, \ldots, N_k$ were not sampled. W_k is an $N_k \times N_k$ diagonal matrix of inverse probability weights for all individuals, R_k is an $N_k \times N_k$ diagonal matrix of indicators with $R_{k[ii]} = 0$ if individual ki was not sampled and 1 if individual ki was sampled, and V_k is a working correlation matrix with a block-diagonal structure. Specifically, V_k is a diagonal matrix in which the diagonal elements are the square matrices $V_k^{n_k}$ and V_k^*, and the off-diagonal elements are zero. We choose this format for the working correlation matrix to ensure that $V^{-1}_{k[1:n_k,1:n_k]} = \left\{V_k^{n_k}\right\}^{-1}$. Note that the equations $\sum_{k=1}^{K} U_k^{n_k}(\beta) = 0$ and $\sum_{k=1}^{K} U_k(\beta) = 0$ are identical due to the chosen structure of the working correlation matrix and the indicator matrix R_k preventing any unsampled individual from contributing to the pseudo-score. Finally, the parameter estimates that solve the inverse probability weighted estimating equations are unbiased as

$$
\begin{aligned}
&E[D_k^T(\beta) V_k^{-1} W_k R_k \left(Y_k - \mu_k(\beta)\right) | X_k] \\
&\qquad = E[E[D_k^T(\beta) V_k^{-1} W_k R_k \left(Y_k - \mu_k(\beta)\right) | Y_k, X_k] | X_k] \\
&\qquad = E[D_k^T(\beta) V_k^{-1} W_k E[R_k | Y_k, X_k] \left(Y_k - \mu_k(\beta)\right) | X_k] \\
&\qquad = E[D_k^T(\beta) V_k^{-1} W_k W_k^{-1} \left(Y_k - \mu_k(\beta)\right) | X_k] \\
&\qquad = E[D_k^T(\beta) V_k^{-1} \left(Y_k - \mu_k(\beta)\right) | X_k] \\
&\qquad = E[U_k^\circ | X_k] \\
&\qquad = 0
\end{aligned}
$$

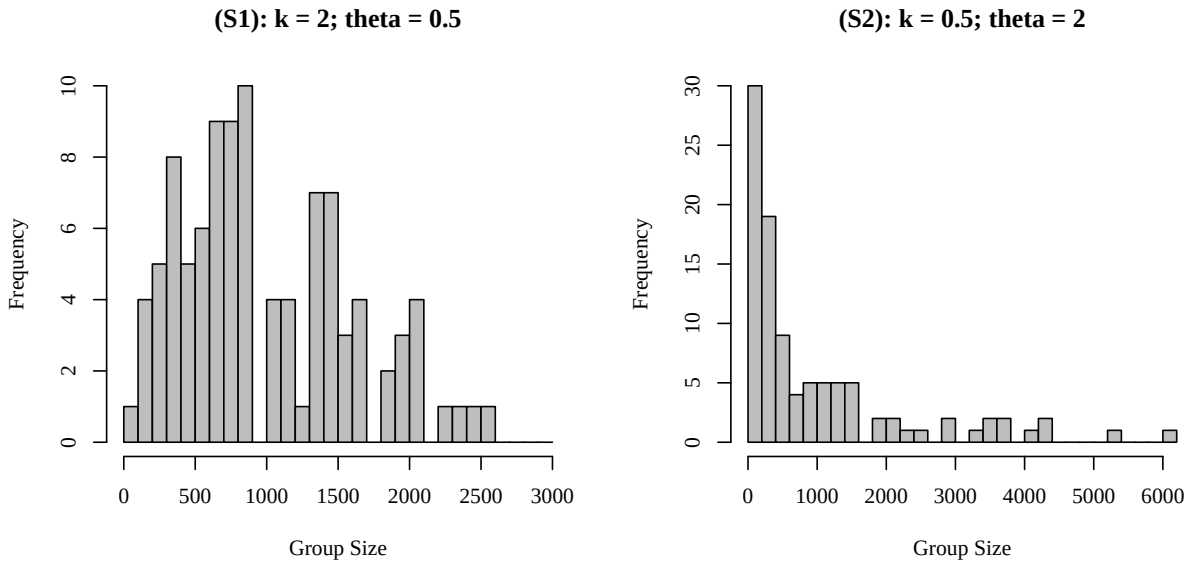
4.2 Generating group sizes using a Gamma(k, θ) distribution

We generate varied group sizes N_k by using random variables from a Gamma(k, θ) distribution to alter inverse probabiliy of sampling weights. We fixed the desired population size $N = \sum_k N_k$, the desired number of groups K, a minimum size for all of the groups N^m, and the number of individuals to be sampled per group $n_k = n$. Group sizes $N_k = N_k^* + N^m$ and for stratified random sampling, IPWs are

$$w_k = \frac{N_k^* + N^m}{n}$$

We let $w_k = cg_k$, where g_k is a vector of K random variables from a Gamma(k, θ) distribution. The population size can be rewritten as $N = \sum_k nw_k = \sum_k ncg_k \Rightarrow c = N/\left(n \sum_k g_k\right)$. The group sizes N_k are taken to be Round$(ncg_k) + N^m$. For simulations in the paper, group sizes are generated once and then are fixed at N_k for all 10,000 iterations.

Figure 4.2.1: Group sizes generated using Gamma(2, 0.5) and Gamma(0.5, 2) distributions (simulations (S1) and (S2), respectively).



4.3 Operating characteristics

Table 4.3.1: Operating characteristics for seven wGEE-based estimators of β from model (4.6) using the full data and six subsamples, under three varied group-size simulation scenarios, as described in Section 4.1. All values are based on 10,000 simulated datasets. Both the exchangeable (Exch) and independent (Ind) correlation structures are used, with the exception of the full-data model, for which only an independent correlation structure is assumed for computational efficiency.

Greater within- than between-group variation in X_1^w

			No Stratification				**Stratification**							
		Full	RS		CC		BSRS		USRS		BSCC		USCC	
			Ind	Exch	Ind	Exch	Ind	Exch	Ind	Exch	Ind	Exch	Ind	Exch
Point Estimates														
	β_0	-2.09	-2.09	-2.09	-2.09	-2.12	-2.09	-2.09	-2.10	-2.10	-2.09	-2.14	-2.09	-2.14
	β_1	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03
So	β_2	0.21	0.21	0.21	0.21	0.21	0.21	0.21	0.21	0.21	0.21	0.22	0.21	0.22
	β_3	0.38	0.38	0.38	0.39	0.38	0.38	0.38	0.39	0.39	0.39	0.38	0.38	0.38
	ρ			0.04		0.05		0.04		0.04		0.04		0.04
	β_0	-2.09	-2.10	-2.10	-2.09	-2.12	-2.09	-2.09	-2.09	-2.09	-2.10	-2.14	-2.10	-2.14
	β_1	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03
S1	β_2	0.21	0.21	0.21	0.21	0.21	0.21	0.21	0.21	0.21	0.22	0.23	0.21	0.22
	β_3	0.38	0.38	0.38	0.38	0.38	0.39	0.39	0.38	0.38	0.38	0.38	0.38	0.38
	ρ			0.04		0.05		0.04		0.04		0.04		0.04
	β_0	-2.09	-2.09	-2.09	-2.09	-2.12	-2.10	-2.10	-2.10	-2.10	-2.10	-2.14	-2.09	-2.14
	β_1	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03
S2	β_2	0.21	0.21	0.21	0.21	0.21	0.21	0.21	0.21	0.21	0.21	0.22	0.21	0.22
	β_3	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.39	0.38	0.38	0.38	0.38
	ρ			0.04		0.04		0.04		0.04		0.04		0.04
SE/SD														
	β_1	0.97	0.98	0.99	0.99	0.99	0.98	0.99	0.99	0.99	0.99	1.00	0.98	0.98
So	β_2	0.99	0.99	0.99	0.98	0.99	0.98	0.99	0.98	0.98	1.00	1.00	0.99	0.99
	β_3	0.99	0.98	0.99	0.98	0.98	0.98	0.98	0.99	0.98	0.98	0.98	0.98	0.98
	β_1	0.96	0.99	0.99	0.99	0.99	0.98	0.98	0.99	0.99	0.98	0.98	1.00	1.00
S1	β_2	0.96	0.99	0.99	0.99	1.00	0.98	0.98	0.99	0.99	0.98	0.98	0.97	0.97
	β_3	0.95	0.96	0.97	0.97	0.98	0.97	0.97	0.96	0.97	0.96	0.96	0.96	0.97
	β_1	0.90	0.95	0.96	0.98	1.00	0.95	0.95	0.97	0.97	0.93	0.94	0.96	0.96
S2	β_2	0.93	0.98	0.98	0.96	0.97	0.94	0.95	0.97	0.97	0.95	0.95	0.95	0.96
	β_3	0.91	0.92	0.97	0.92	0.97	0.91	0.91	0.91	0.95	0.91	0.91	0.90	0.94
Coverage														
	β_1	94.1	94.2	94.2	94.9	94.5	94.4	94.5	94.5	94.8	94.6	92.8	94.1	92.5
So	β_2	94.9	94.2	94.3	94.4	94.5	94.4	94.3	94.7	94.4	94.4	94.4	94.6	94.6
	β_3	94.5	94.1	94.1	94.6	94.5	94.1	94.2	94.2	94.2	94.1	94.2	94.2	94.0
	β_1	94.0	94.2	94.6	94.4	94.6	94.0	94.0	94.5	94.5	94.2	92.6	94.9	92.7
S1	β_2	93.7	94.2	94.4	94.6	94.7	94.1	94.1	94.6	94.4	94.2	94.3	93.8	93.7
	β_3	93.3	93.2	93.8	93.5	94.0	93.7	93.6	93.5	93.8	93.4	93.5	93.3	93.9
	β_1	92.7	93.1	93.6	94.1	94.4	92.9	92.9	93.2	93.7	92.7	91.6	92.9	91.9
S2	β_2	92.9	94.1	94.0	93.6	93.5	92.6	92.7	93.6	93.5	93.1	92.8	93.1	92.9
	β_3	91.3	91.6	93.5	91.4	93.9	91.1	91.4	91.5	92.7	91.3	91.2	91.1	92.6

Table 4.3.2: Relative uncertainty of wGEE-based estimators of β under case-control sampling compared to random sampling.

		No Strat.		**Stratification**			
				Balanced		Unbalanced	
		Ind	Exch	Ind	Exch	Ind	Exch
So	β_1	90.9	91.1	91.2	93.1	92.4	94.5
	β_2	93.7	93.3	92.0	93.4	93.5	94.6
	β_3	97.6	97.3	83.6	83.9	79.5	76.6
S1	β_1	92.1	91.6	91.8	94.4	91.4	93.0
	β_2	91.8	91.1	93.4	94.7	96.2	96.9
	β_3	97.0	95.9	84.0	84.4	85.7	81.5
S2	β_1	89.7	87.0	93.8	95.2	93.6	94.4
	β_2	93.9	92.9	92.0	93.2	95.1	96.0
	β_3	98.6	96.4	83.0	83.5	90.3	88.0

Table 4.3.3: Relative uncertainty of estimates $\hat{\beta}$ from stratified samples compared to unstratified samples, under model (4.6). Comparisons are made between like study designs and correlation structures. For example, a stratified random sampling design (balanced or unbalanced) and working independence correlation structure compared to an unstratified random sampling design and working independence correlation structure, and a stratified case-control design and working exchangeable correlation structure compared to an unstratified case-control design and working exchangeable correlation structure. All values are based on 10,000 simulated datasets.

		Random Sampling				**Case-Control Sampling**			
		Independent		Exchangeable		Independent		Exchangeable	
		BSRS	USRS	BSRS	USRS	BSCC	USCC	BSCC	USCC
So	β_1	100.0	114.5	100.4	117.0	100.2	116.4	102.6	121.3
	β_2	101.3	117.3	101.0	118.2	99.5	117.0	101.1	119.8
	β_3	99.6	105.3	100.3	115.4	85.3	85.8	86.5	90.8
S1	β_1	115.1	103.8	116.9	104.7	114.7	103.0	120.5	106.4
	β_2	116.9	104.5	117.6	105.1	118.9	109.4	122.2	111.8
	β_3	102.6	100.4	109.1	103.5	88.9	88.8	96.1	88.0
S2	β_1	148.0	109.5	157.7	112.3	154.7	114.2	172.6	121.9
	β_2	161.2	115.0	164.9	115.8	158.0	116.5	165.4	119.7
	β_3	112.3	102.6	135.9	107.1	94.5	93.8	117.8	97.8

Table 4.3.4: Relative uncertainty of estimates $\hat{\beta}$ from unbalanced stratified samples compared to balanced stratified samples, under model (4.6). Comparisons are made between like correlation structures. All values are based on 10,000 simulated datasets.

		Random Sampling		**Case-Control Sampling**	
		Independent	Exchangeable	Independent	Exchangeable
S0	β_1	114.6	116.5	116.1	118.3
	β_2	115.7	117.0	117.6	118.5
	β_3	105.7	115.0	100.6	105.0
S1	β_1	90.2	89.6	89.8	88.3
	β_2	89.4	89.4	92.1	91.5
	β_3	97.9	94.9	99.9	91.6
S2	β_1	74.0	71.2	73.8	70.6
	β_2	71.3	70.2	73.7	72.3
	β_3	91.3	78.9	99.3	83.0

Table 4.3.5: Relative uncertainty of wGEE-based estimators of β under the assumption of an exchangeable correlation structure compared to the assumption of an independent correlation structure.

		No Strat.		**Stratification**			
		RS	CC	BSRS	USRS	BSCC	USCC
S0	β_1	95.3	95.5	95.7	97.3	97.8	99.6
	β_2	97.7	97.3	97.4	98.5	98.9	99.6
	β_3	99.3	99.0	100.0	108.8	100.4	104.8
S1	β_1	94.5	93.9	96.0	95.4	98.7	97.0
	β_2	97.0	96.3	97.5	97.6	99.0	98.3
	β_3	94.1	93.0	100.0	97.0	100.5	92.2
S2	β_1	90.9	88.1	96.8	93.2	98.3	94.1
	β_2	95.7	94.7	98.0	96.4	99.2	97.3
	β_3	82.6	80.7	99.9	86.3	100.6	84.1

I am asking you now to put everything to the test with me, and you will judge for yourselves whether the observations I have made justify the conclusions to which I have come.

Arthur Conan Doyle, *The Valley of Fear*

5

Short Term Exposure to Low Levels of Fine Particulate Matter and Hospital Admissions in Older Adults

Abstract

Background: Exposure to air pollutants adversely affects human health, but the full scope of this impact is unknown. Studies to date have largely examined the magnitude of air pollution's effect on a set of pre-specified health conditions, rather than investigating a wide spectrum of conditions making no a priory assumptions. Also, few studies have specifically examined health effects at very low levels.

Objectives: We aim to identify all possible causes for hospitalization in the older US population associated with short-term exposure to fine-particulate matter ($PM_{2.5}$) air pollution. We further investigate this relationship for levels of $PM_{2.5}$ lower $20\mu g/m^3$, $15\mu g/m^3$, and $10\mu g/m^3$, respectively.

Methods: Using a national database of daily cause-specific -hospitalizations for 232 diseases in 220 communities for 1999-2010, we estimated cause-specific relative risk of hospitalizations associated with exposure to $PM_{2.5}$ by age and geographical location. We used a two-stage Bayesian hierarchical modeling (BHM) approach to estimate the national average association between $PM_{2.5}$ and cause-specific hospitalization rates, while accounting for possible confounding by temperature, season, and secular trends. The model includes a linear term for $PM_{2.5}$ and is fit to the entire data set and to data sets that only include days with daily levels of $PM_{2.5}$ lower than $20\mu g/m^3$, $15\mu g/m^3$, and $10\mu g/m^3$, respectively.

Results: We found evidence of an association between $PM_{2.5}$ and hospital admission rates for cerebrovascular disease, for a range of cardiovascular outcomes, syncope (fainting), and fluid and electrolyte disorders. The positive and statistically significant association between hospitalization due to cardiovascular disease and air pollution remains significant at low levels of pollution, as does that between hospitalization due to syncope and air pollution. Results tend to be consistent between the two different modeling approaches.

Conclusions: Our analysis, which considered all causes of hospital admission that could be associated with exposure to $PM_{2.5}$, reinforced existing literature on the association between $PM_{2.5}$ and cardiovascular outcomes, and suggested an association with syncope, a relationship that has not been investigated previously. It also provides further evidence that these effects persists at very low levels of $PM_{2.5}$ Knowledge of the full range of health effects associated with exposure to $PM_{2.5}$ informs public health approaches to prevention.

5.1 Background

Epidemiological studies have provided evidence of an association between ambient levels of fine particulate matter ($PM_{2.5}$) and: (1) hospital admissions for respiratory and cardiovascular causes. [13, 14, 17, 39, 53]; (2) near-immediate increases in blood pressure [11]; and (3) increased cardiovascular mortality rates [23, 54, 60]. In response to concerns over more severe health risks posed by $PM_{2.5}$ compared to particulate matter less than $10\mu m$ (PM10), the United States Environmental Protection Agency (EPA) revised the National Ambient Air Quality Standards (NAAQS) in 1997 to set separate standards on annual and 24-hour levels of particulate matter (PM10) and fine particulate matter ($PM_{2.5}$), and the 24-hour $PM_{2.5}$ NAAQS were lowered from 65 $\mu g/m^3$ to 35 $\mu g/m^3$ in 2006 after a review of scientific reports. Since the establishment of the NAAQS, national PM monitoring has shown decreasing air pollution trends in the US, with the EPA reporting a 33% decrease in the national average $PM_{2.5}$ levels between 2000 and 2012 [20]. Implementation of the EPA guidelines has been costly and controversial; an understanding of the human health benefit and downstream healthcare savings resulting from lowered $PM_{2.5}$ levels is an essential component of effective national air quality policies [45].

One of the current gaps in the research needed to steer effective air pollution policy is a better understanding of all possible disease outcomes associated with low level exposure to $PM_{2.5}$, especially in susceptible populations such as the elderly. Previous studies have, for the most part, examined health effects of air pollution on few pre-selected health outcomes, such as mortality or hospitalization for cardiovascular and respiratory diseases. Additionally, no study to date has conducted a multi-site time series study restricting the analysis to days with very low levels of $PM_{2.5}$. The objectives of this paper are twofold. First, we are interested in identifying the full spectrum of hospital admission causes that are associated with short term exposure to fine particulate matter air pollution and estimating their relative risks. We suspect that there

may be hospitalization causes associated with $PM_{2.5}$ that have not yet been identified in the literature. To address this issue, we compiled a national database of cause-specific hospital admission data and $PM_{2.5}$ - concentration data at the county- and daily-level and performed a comprehensive analysis of all possible causes of hospitalization that may be associated with $PM_{2.5}$ levels. Here we followed the approach by Bobb et. al. [4] which has been used to investigate the causes of hospital admission associated with exposure to periods of extreme heat. We carefully examine the sensitivity of the results to model selection, specifically, in the approach used to adjust for confounding in time series analyses. The second objective of this paper is to investigate the extent to which the adverse health effects persist at low levels of air pollution. To assess the relationship between low-level pollution and hospitalization we re-applied our modelling approaches to a series of restricted datasets containing only low-level pollution days ($PM_{2.5} \leq 20\mu g/m^3$, $15\mu g/m^3$, and $10\mu g/m^3$).

5.2 Methods

5.2.1 Study Population

We assembled time-series data of daily cause-specific hospitalization rates using the National Medicare cohort for the years 1999-2010. These data include individual-level longitudinal data on hospitalization for all Medicare enrollees (aged $>$ 65 years) from the Centers for Medicare and Medicaid Services. Daily hospitalization rates were derived from billing claims that contain the date of service, disease classification via ICD-9 codes, age, gender, race, and zip code of residence. Following Bobb et. al. [4], we used the Agency for Healthcare Research and Quality's Clinical Classifications Software (CCS) algorithm [19] to cluster hospitalization causes represented by ICD-9 codes into 283 mutually exclusive and clinically meaningful categories. We excluded 47 categories that had no occurrences in the older Medicare population over the twelve-year study period, most of which were pregnancy- or fertility-related, and additionally excluded four categories that are by definition comprised solely of "V-codes," supplemental classification codes that are reported in conjunction with a traditional ICD-9 code and which allow reporting of circumstances and conditions influencing the individual's health status. The final number of disease causes that we considered is 232. We restricted our analysis to 220 metropolitan counties (Figure 5.2.1) that have at least two years with 33% of days having complete information on the covariates described in Section 5.2.4. In 2000, our study population comprises of 6.7 million Medicare enrollees and 3.1 million hospitalization records for all hospitalization causes combined.

5.2.2 Air pollution and meteorology data

Weather data were obtained from the National Climatic Data Center, which comprises daily weather records from monitoring stations for 1987-2012. Temperature and dew-point temperature for each county were

taken to be the daily average measurement across any monitors either belonging to that county or within 35 kilometers of the county's geographical center. Counties without available temperature and dew point temperature data were excluded from the study.

Concentrations of fine particulate matter for 1987-2014 are publicly available and were obtained from the US EPA Air Quality System (AQS) database. For each county, daily mean $PM_{2.5}$ values were taken to be the daily average measurement, averaged across all monitors belonging to that county. In many counties particulate matter is not monitored daily; counties included in this study were required to have at least two years of fine particulate matter measurements every third day.

5.2.3 Statistical analysis

Within county and age category ($65 \leq$ age < 75, $75 \leq$ age < 85, $85 \leq$ age) we tallied daily counts of individuals at-risk for hospitalization (denominator) and hospitalization by CCS category (numerator) for each date from January 1, 1999 to December 31, 2010. To estimate the association between $PM_{2.5}$ concentration and same-day hospitalization rates for each of the 232 CCS-diagnosis outcomes we fit Bayesian hierarchical models, which we describe below. We controlled for multiple testing using a Bonferroni correction to adjust confidence intervals. Our parameters of interest will be defined as (i) the log relative risk of hospital admissions associated with a $10\mu g/m^3$ increase in $PM_{2.5}$ in our full-data analyses and (ii) the log relative risk of hospital admissions associated with a $1\mu g/m^3$ increase in $PM_{2.5}$ in our low-level$PM_{2.5}$ analyses.

5.2.4 Modeling Approach

Separately for each of the 232 disease groups, we applied a 2-stage Bayesian hierarchical model (BHM) to estimate county-specific and national-average associations between day to day changes in $PM_{2.5}$ levels and hospital admission rates. Specifically, at the first stage of the BHM, we modeled the number of cause-specific hospitalizations, Y_{itg}, on day t in age group g and county i using the quasi-Poisson generalized linear model:

$$\begin{aligned} log(E[Y_{itg}]) \;=\; & log(N_{itg}) + \gamma_{ig0} + \beta^L_{i1}(\mathrm{PM}_{2.5it}/10) + \gamma'_{i2}\mathrm{dow}_t + \mathrm{ns}(\mathrm{date}_t; 8/\mathrm{year\ DF}, \gamma_{i3}) + \\ & \mathrm{ns}(\mathrm{temp}_{it}; 6\mathrm{DF}, \gamma_{i4}) + \mathrm{ns}(\overline{\mathrm{temp}}^{(3)}_{it}; 6\mathrm{DF}, \gamma_{i5}) + \mathrm{ns}(\mathrm{dptp}_{it}; 3\mathrm{DF}, \gamma_{i6}) + \\ & \mathrm{ns}(\overline{\mathrm{dptp}}^{(3)}_{it}; 3\mathrm{DF}, \gamma_{i7}) \end{aligned} \quad (5.1)$$

where $ns(\cdot)$ denotes natural cubic splines with the specified degrees of freedom (DFs) and $\gamma_{ik}(k = 3, 7)$ representing the spline coefficients and γ'_{i2} is a vector of parameters associated with the categorical covariate dow. $PM_{2.5}$ was included in the model using a linear term. To account for trend and seasonality we included cubic spline terms in the regression model. We also adjusted for day of the week by including the

categorical covariate dow, and for same day temperature (temp_{it}), the average of the previous three days' average temperature ($\overline{\text{temp}}_{it}^{(3)}$), the current day's average dew point temperature ($\overline{\text{dptp}}_{it}^{(3)}$), and the average of the previous three days' average dew point temperature (dptp_{it}) all by using smoothing splines. At the second stage of the BHM, we estimated an overall national-average effect estimate of the log-relative-risk associated with a 10 $\mu g/m^3$ increase in $PM_{2.5}$ using two-level normal independent sampling estimation [21].

5.2.5 Low-level pollution analyses

To investigate associations between hospital admissions and low levels of air pollution, we restricted our dataset first to days with $PM_{2.5} \leq 20\ \mu g/m^3$, then to $PM_{2.5} \leq 15\ \mu g/m^3$, and finally to $PM_{2.5} \leq 10\ \mu g/m^3$ and applied the two-stage BHM approach described above to each of the three restricted datasets. The number of days with complete data in the full-range $PM_{2.5}$ dataset and each of the three low-level $PM_{2.5}$ datasets is presented for each county in Figure **??**. In the low-level analyses, our parameter of interest is defined as the log-relative-risk associated with a 1 $\mu g/m^3$ increase in $PM_{2.5}$.

5.2.6 Bootstrap approach

To adjust for residual autocorrelation in the data we estimated confidence intervals using a moving block bootstrap method [42]. For each of three hundred iterations we sampled 146 blocks of 30 days each from the original dataset and pieced them together to create a new time series. The full-data and low-level $PM_{2.5}$ models described above were fit to each of the new time series. The mean and standard deviation of the three hundred point estimates were then used to create a confidence interval.

5.3 Results

5.3.1 Characteristics of the population

The geographical locations of the 220 counties included in this study are shown in Figure 5.2.1, with the color of the counties points representing the number of Medicare enrollees in the county in the year 2010. Locations of $PM_{2.5}$ monitoring stations are shown in red. Annual availability of data is presented in Table 5.3.1, including quantiles of the number of days with complete air pollution and meteorology data, percentages of high and low $PM_{2.5}$ days, and the mean daily $PM_{2.5}$. National $PM_{2.5}$ levels decreased steadily over the study period, even in counties starting the study with comparatively low levels of $PM_{2.5}$. Table 5.3.1 shows summary statistics of the number of days with complete air pollution and meteorology data, percentage of days with l$PM_{2.5}$ measurements less than 10$\mu g/m^3$, and the mean daily $PM_{2.5}$. County-specific yearly averages of mean daily $PM_{2.5}$ measurements are plotted in Figure 5.3.1. National $PM_{2.5}$ levels de-

creased steadily over the study period, even in many counties which started the study with comparatively low levels of $PM_{2.5}$.

Table 5.3.1: Population characteristics: numbers reported are the quantiles (25^{th}, 50^{th}, and 75^{th}) of the 220 county-specific annual statistics.

	Number of days with $PM_{2.5}$ and temperature data[b]			Percent of $PM_{2.5}$ days $< 10\mu g/m^3$ [c]			Mean daily $PM_{2.5}$ [d]		
Q[a] / Year	25	50	75	25	50	75	25	50	75
1999	93	162	250	25.1	38.8	50.4	12.1	14.8	17.2
2000	115	209	341	24.7	37.2	51.5	12.4	14.5	16.6
2001	118	217	350	29.9	41	53.3	11.8	14.1	15.8
2002	119	223	354	32.1	42.6	56.2	11.6	13.5	15.1
2003	116	209	349	33.4	43.6	57.4	11.2	13.1	14.9
2004	118	210	351	34.2	46.8	59.8	10.8	12.8	14.6
2005	117	200	340	30.4	40.2	55.1	11.7	13.7	15.7
2006	119	195	343	37.6	47.5	60.3	10.4	12.4	14.2
2007	118	211	346	35.7	45.6	58.8	10.8	12.6	14.5
2008	117	209	348	40.1	51.1	63.2	9.9	11.5	12.9
2009	120	227	354	48.9	60.1	70.9	9.1	10.5	11.8
2010	120	234	357	45.7	58.7	71	9	10.8	12.3

[a] Quantiles (25th, 50th, or 75th) of the 220 county-specific annual counts (b), percentages (c), or means (d).

5.3.2 Associations between risk of hospitalization and $PM_{2.5}$

Figure 5.3.2 shows the log relative risk of hospitalization associated with a 10 $\mu g/m^3$ increase in mean daily $PM_{2.5}$ for the thirty most common diagnoses at hospitalization, which includes all hospitalization causes for which we found significant associations with $PM_{2.5}$. Of the 232 disease groups, eight had statistically significant elevated risk of hospitalization associated with $PM_{2.5}$ after adjusting for multiple comparisons. Five of the statistically significant disease groups were cardiovascular; a $10\mu g/m^3$ increase in $PM_{2.5}$ was estimated to be associated with a 3.08 (95

In our analysis of low levels of $PM_{2.5}$ (Figure 5.3.3), an association between $PM_{2.5}$ and hospitalizations due to coronary atherosclerosis remained significant for all three data sets— that which includes only days with $PM_{2.5}$ lower than 20, 15 and $10\mu g/m^3$, respectively. Here we found that a one-unit increase

in $PM_{2.5}$ was associated with a 0.68% (0.01, 1.35) increase in hospitalizations at the lowest levels of $PM_{2.5}$ ($\leq 10\mu g/m^3$). Relationships between $PM_{2.5}$ and hospitalizations for transient cerebral ischemia and syncope remained significant on days for which $PM_{2.5}$ measured less than $15\mu g/m^3$, with a one-unit increase in $PM_{2.5}$ corresponding to a 0.47% (0.11, 0.82) increase in hospitalizations for transient cerebral ischemia and a 0.51% (0.18, 0.85) increase in hospitalizations for syncope. Additionally, on days for which $PM_{2.5}$ measured less than $15\mu g/m^3$ we observed a 0.35% (0.01, 0.69) increase in hospitalizations due to nonspecific chest pain per one-unit increase in $PM_{2.5}$ and a 0.28% (0.01, 0.56) increase in hospitalizations due to pneumonia.

Finally, for $PM_{2.5} \leq 20\mu g/m^3$ a one-unit increase in $PM_{2.5}$ was associated with a 0.23% (0.03, 0.43) increase in acute myocardial infarction hospitalizations, a 0.22% (0.06, 0.38) increase in hospitalizations due to acute cerebrovascular disease, a 0.21% (0.05, 0.36) increase in congestive heart failure hospitalizations, and a 0.19% (0.02, 0.37) increase in cardiac dysrhythmia hospitalizations. We additionally observed a 0.26% (0.04, 0.49) increase in hospitalizations due to pneumonia associated with a one-unit increase in $PM_{2.5}$. Further, the associations between fluid and electrolyte disorders and septicemia and $PM_{2.5}$ observed in the full-data analyses remained significant when the data was restricted to days with $PM_{2.5} \leq 20\mu g/m^3$, with a one-unit increase in $PM_{2.5}$ corresponding to a 0.26% (0.04, 0.46) increase in hospitalizations due to fluid and electrolyte disorders and a 0.37% (0.02, 0.73) decrease in hospitalizations for septicemia. Table 5.3.2 summarizes the findings from each of the four analyses, indicating cause-specific hospitalizations for which we found a significant relationship with $PM_{2.5}$.

5.4 Discussion

To our knowledge, this is the first study to investigate the full spectrum of health conditions associated with short-term exposure to ambient air pollution. Rather than targeting a small number of health effects for investigation we performed a comprehensive analysis of 232 hospitalization causes, testing for relationships with fine-particulate air pollution levels in a population of 6.7 million Medicare enrollees from 220 US counties in the years 1999—2010.

We found evidence that day to day changes of $PM_{2.5}$ are associated with day to day changes in hospitalization risk of several cardiac events, including acute myocardial infarction, coronary atherosclerosis, congestive heart failure, and acute cerebrovascular disease. These findings validate the known cardiac effects of $PM_{2.5}$ exposure, agreeing with a number of previous studies that found associations between short-term exposure to particulate-matter air pollution and cardiac events. Dominici et. al. found increases in hospitalizations for cardiovascular health outcomes—ischemic heart disease, cerebrovascular disease, heart rhythm, heart failure, and peripheral vascular disease—in the presence of acute exposure to fine-particulate air pollution [17]. Pope et. al. demonstrated excess cardiovascular morbidity in patients with underlying

coronary artery disease related to short-term exposure to increased levels of $PM_{2.5}$ [54]. In comparing eight meta-analyses and multicity studies of short-term changes in $PM_{2.5}$ exposure, Pope and Dockery estimate that for every 10 $\mu g/m^3$ increase in $PM_{2.5}$ concentration the risk of cardiovascular mortality increases by approximately 1% citeppope2006health. Brook provides a comprehensive summary of additional studies which link ambient air pollution levels to cardiovascular events [10].

We additionally identified an association between rises in $PM_{2.5}$ levels and an increased risk of syncope, a condition not previously considered in relation to ambient air pollution. Several biological pathways can lead to syncope including sudden decreases in blood pressure, a reduction in the blood's oxygen concentration, and cardiac abnormalities that lead to reduced blood flow to the brain [72]. Each of these pathways could be affected by increased levels of $PM_{2.5}$. First, while Brook and Rajagopalan and Auchincloss et. al. found that blood pressure increased in response to increases in ambient fine-particulate air pollution, Gong et. al. found systolic blood pressure to decrease in asthmatics (but increase in healthy subjects) during exposure to $PM_{2.5}$ relative to filtered air [11, 16, 25]. Second, irritant air pollutants have been shown to decrease breathing rates [35], which would reduce blood oxygen levels. Third, we found an association between increased $PM_{2.5}$ and increased risk of acute cerebrovascular disease, a series of conditions that limit blood flow to the brain. The increased risk of syncope we observed with increased $PM_{2.5}$ levels may be due to mild cardiac events. In such a case, these findings could lead to improved preventative care policies, as patients experiencing syncope on high $PM_{2.5}$ days may be diagnosed with treatable cardiac abnormalities.

We additionally noted a decreased rate of septicemia hospital admissions at elevated levels of $PM_{2.5}$. The observed negative association between $PM_{2.5}$ and sepsis may be due to an increase in the rate of cardiac events. Cardiac dysfunction occurs in 40% of sepsis patients [37], suggesting that days with increased levels of $PM_{2.5}$ may result in serious cardiac events occurring in sepsis patients. In such a case, the cardiac event would likely be recorded as the cause-of-admission rather than sepsis.

Of specific interest is the continued impact of $PM_{2.5}$ at low levels. Coronary atherosclerosis remained significantly associated with $PM_{2.5}$ even when the data was restricted to days with $PM_{2.5} < 10\ \mu g/m^3$, and transient cerebral ischemia and nonspecific chest pain were significantly associated with $PM_{2.5}$ when the data was restricted to days with $PM_{2.5} < 15\ \mu g/m^3$. Additionally, on days with $PM_{2.5} < 20\ \mu g/m^3$, the risks of hospitalization for the cardiac events acute cerebrovascular disease, acute myocardial infarction, cardiac dysrhythmias, and congestive heart failure are significantly associated with one-unit increases in $PM_{2.5}$ levels. Though we did not find statistically significant associations in our analysis of days with $PM_{2.5} < 15\ \mu g/m^3$ and $PM_{2.5} < 10\ \mu g/m^3$, the point estimates of the log relative risk of hospitalization due to acute cerebrovascular disease, acute myocardial infarction, cardiac dysrhythmias, and congestive heart failure associated with a one unit increase in $PM_{2.5}$ were nearly identical across each of the restricted datasets. Cumulatively, these results indicate that even low levels of ambient air pollution result in cardiac events in at risk populations.

Unlike other studies, we did not see significant increases in risk of hospitalization for respiratory condi-

tions at higher levels of $PM_{2.5}$. This may in part be due to our extremely conservative analysis approach; by using bootstrap confidence intervals and adjusting for multiple measures we sought to minimize the probability of reporting false positives, which may have resulted in additional failure to reject a false null hypothesis, or Type II error. In addition, we considered only same-day effects of air pollution rather than lagged effects; lagged effects may be more strongly associated with adverse respiratory health effects [22, 39].

Our analyses incorporated several methods designed to ensure accuracy in our conclusions. First, we used the conservative Bonferroni correction to control for multiple testing, reducing the probability of falsely identifying an association between one of the health effects and ambient air pollution. Second, we controlled for known associations between temperature and health effects by including multiple temperature variables in all of our models. Third, we further controlled for temporal trends and confounding by county-level factors by including spline terms in the models. Fourth, by testing all 232 hospitalization causes we incorporated negative controls into our analysis; the lack of evidence of associations for conditions with no clinically meaningful relation to air pollution further validates our results. Fifth, we included a matched analysis with the intention of creating quite conservative estimates of the associations between $PM_{2.5}$ and reasons for hospitalization.

We note that this study is subject to a number of limitations. First, the quality of the CCS-diagnosis outcome measures is limited. We included only primary admission diagnoses in our analyses, which may result in reduced power due to missing driving hospitalization causes recorded as secondary or tertiary diagnostic codes. Second, our study is subject to exposure measurement error, as we are using county-averaged ambient concentrations measured at central monitoring sites as a surrogate for personal exposure to ambient $PM_{2.5}$. Third, we consider only short-term health effects of air pollution, while long-term effects are of equal interest and may be more severe. Long-term exposure to air pollution is associated with respiratory and cardiac effects [1, 53, 54] and has recently been shown to induce cardiovascular remodeling [78]. The full extent of the mechanisms by which long-term air pollution threatens human health is of great interest but has yet to be determined. Finally, the chemical composition of fine particulate matter air pollution, which varies geographically, relates to the severity of health effects [2]. The results presented in this paper represent nationally-averaged relationships between $PM_{2.5}$ and hospitalization causes. Environmental protection policies based on nationally-averaged relationships may be too lenient in counties in which the chemical composition of air particulates is particularly dangerous to human health.

Our results provide a complete characterization of the same-day effects of increases in $PM_{2.5}$ concentrations on hospitalization risks in the older US population. Knowledge of the type and magnitude of such risks is essential for continuing to improve national air quality policies.

Figure 5.2.1: Medicare enrollment for the continental-US counties included in the study overlaid with locations of PM2.5 monitoring stations within each county.

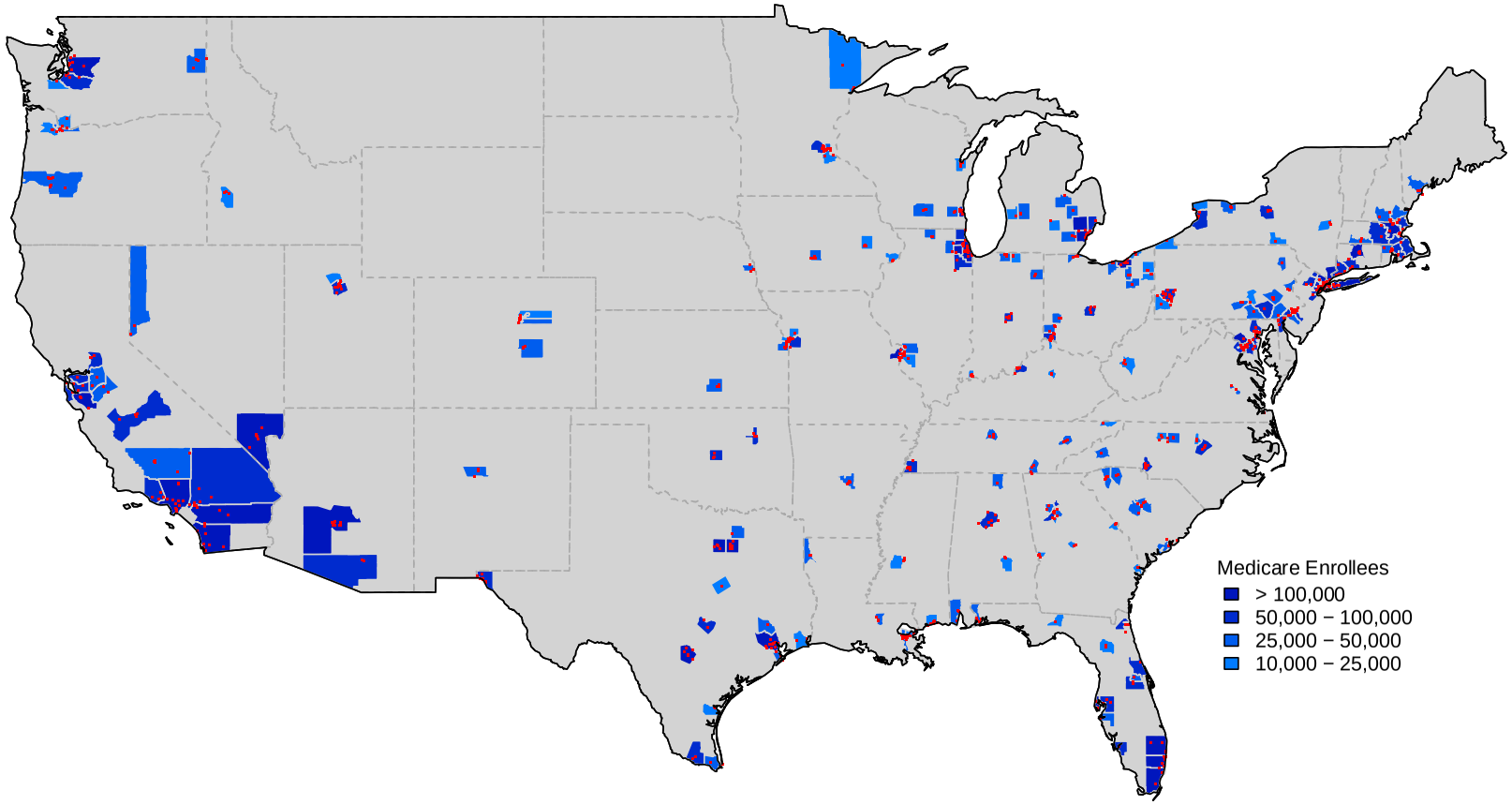


Figure 5.2.2: Number of days with complete $PM_{2.5}$ and temperature data by county, for the full dataset and three restricted, low-level pollution datasets.

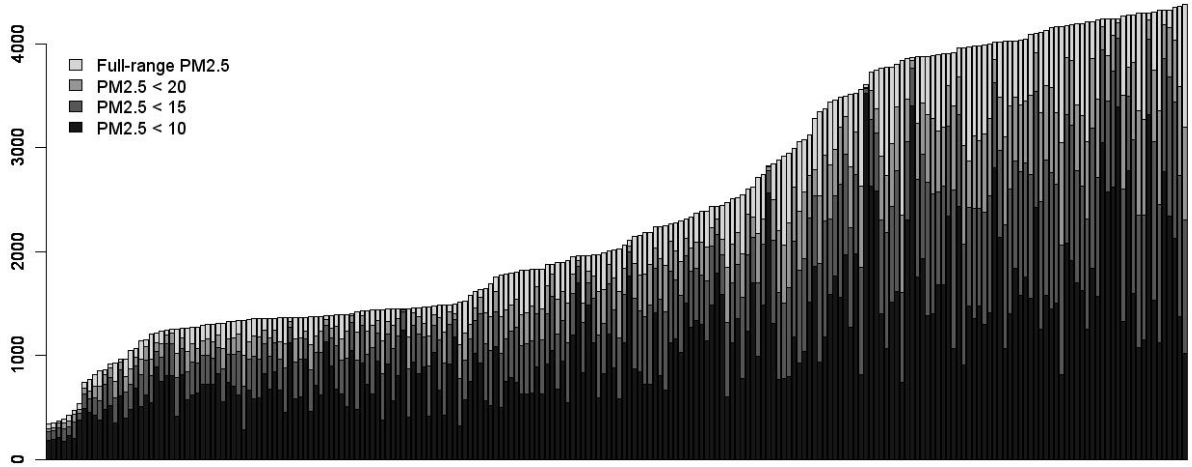


Figure 5.3.1: Yearly averages of daily mean $PM_{2.5}$ measurements, by county.

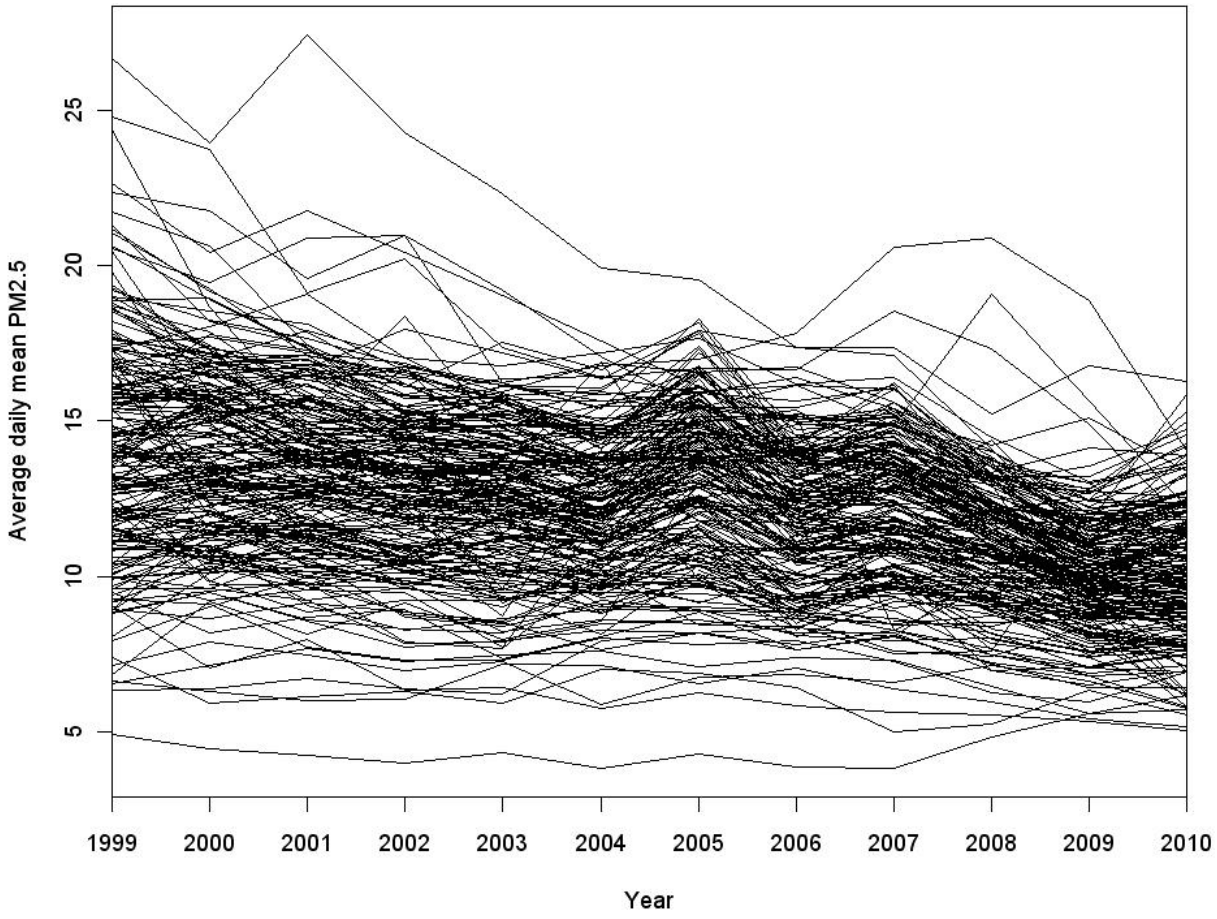


Figure 5.3.2: Point estimates and 95% confidence intervals (CI[a]) of the national average log relative risk associated with a $10\mu g/m^3$ increase in mean daily $PM_{2.5}$. Results are shown for the thirty most common diagnoses at hospitalization. Solid/open circles represent statistically significant/insignificant results, respectively. [a] The Bonferroni correction method is used to adjust CI for multiple comparisons.

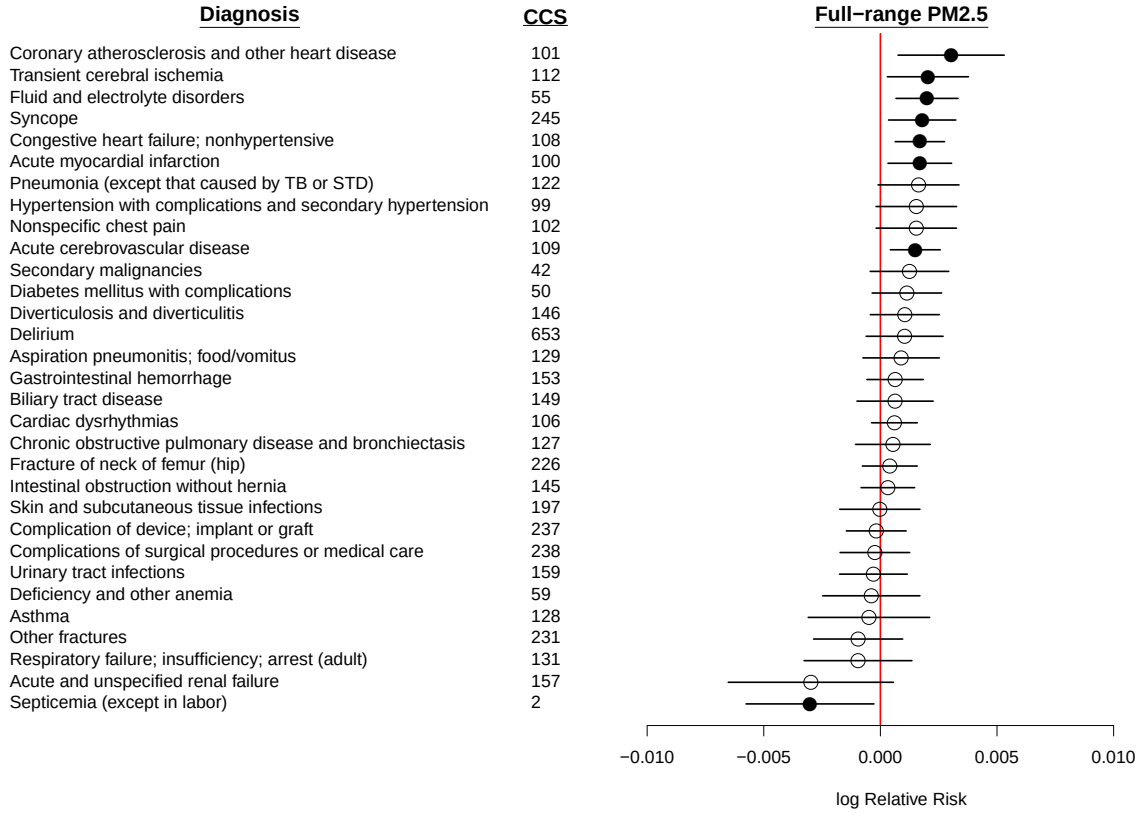


Figure 5.3.3: Point estimates and 95% confidence intervals (CI[a]) of the national average log relative risk associated with a $1\mu g/m^3$ increase in mean daily $PM_{2.5}$ from the two-stage BHM approach, for days having values of $PM_{2.5}$ < (1) $20\mu g/m^3$, (2) $15\mu g/m^3$, (3) $10\mu g/m^3$. Results are shown for the thirty most common diagnoses at hospitalization. Solid/open circles represent statistically significant/insignificant results, respectively. [a] The Bonferroni correction method is used to adjust CI for multiple comparisons.

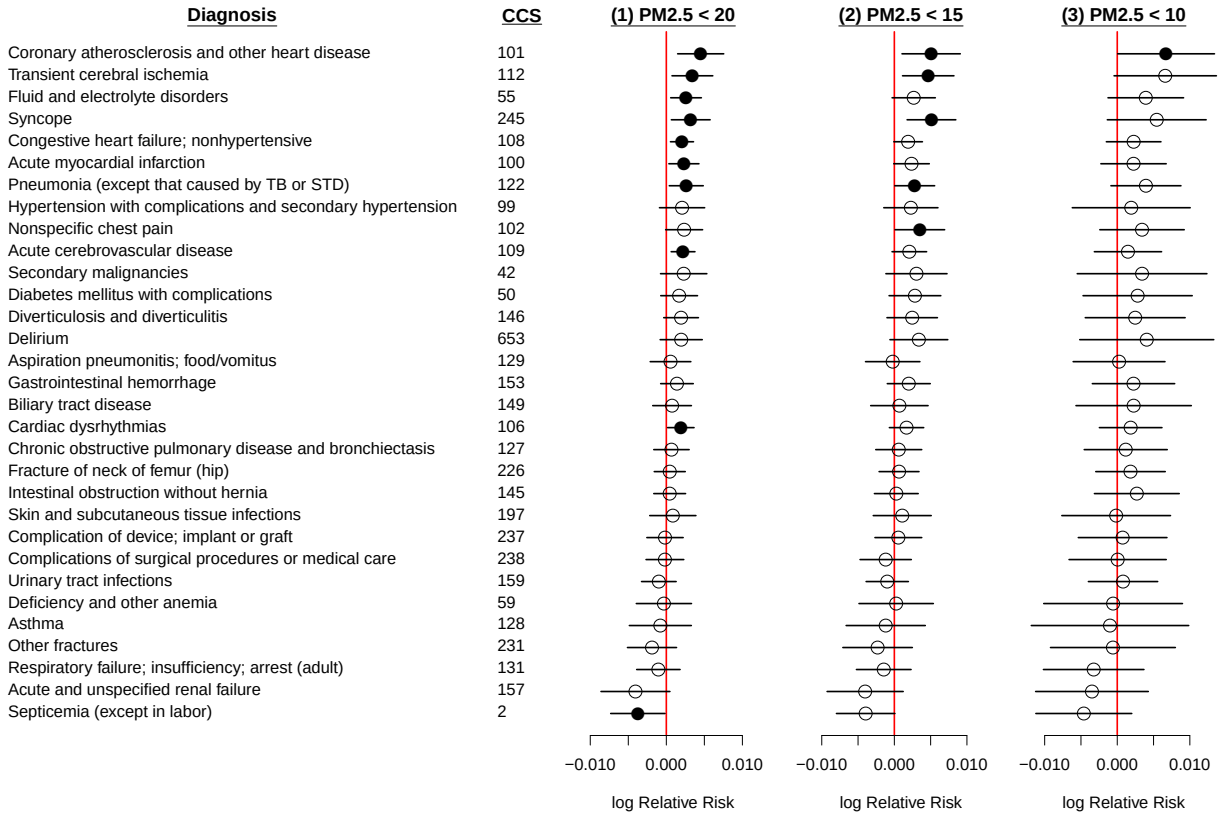


Table 5.3.2: Summary of diagnoses significantly associated with $PM_{2.5}$, by model, for thirty most common diagnoses at hospitalization. Checkmarks indicate significance a with Bonferonni correction; circles indicate significance at the $\alpha = 0.05$ level without Bonferonni correction. Diagnoses not reported in the table were not significantly associated with $PM_{2.5}$ in any analysis.

Level 1[b]	Level 2[c]	Diagnosis	Full-range $PM_{2.5}$	Low $PM_{2.5}$ ($PM_{2.5} < X\mu g/m^3$)		
				$10\mu g/m^3$	$15\mu g/m^3$	$20\mu g/m^3$
Circulatory system	Cerebrovascular disease	Acute cerebrovascular disease	✓		○	✓
		Transient cerebral ischemia	✓	○	✓	✓
	Diseases of the heart	Acute myocardial infarction	✓		○	✓
		Cardiac dysrhythmias	○		○	✓
		Congestive heart failure; nonhypertensive	✓	○	○	✓
		Coronary atherosclerosis and other heart disease	✓	✓	✓	✓
		Nonspecific chest pain	○	○	✓	○
Respiratory system	Respiratory infections	Pneumonia (except that caused by TB or STD)	○	○	✓	✓
Infectious diseases	Bacterial infection	Septicemia (except in labor)	✓	○	○	✓
Endocrine; nutritional; and metabolic diseases	Fluid and electrolyte disorders		✓	○	○	✓
Ill-defined conditions		Syncope	✓	○	✓	✓

[a] "Significance" is that of the linear $PM_{2.5}$ parameter.

[b] Level 1: Top-level diagnosis body system or condition category. CCS codes are grouped into a total of 18 broad level-1 categories.

[c] Level 2: Detailed within- body system or condition category classification of diagnoses.

References

[1] Jacques Baillargeon, Yue Wang, Yong-Fang Kuo, Holly M Holmes, and Gulshan Sharma. Temporal trends in hospitalization rates for older adults with chronic obstructive pulmonary disease. *The American journal of medicine*, 126(7):607–614, 2013.

[2] Michelle L Bell, Keita Ebisu, Roger D Peng, Jonathan M Samet, and Francesca Dominici. Hospital admissions and chemical composition of fine particle air pollution. *American journal of respiratory and critical care medicine*, 179(12):1115–1120, 2009.

[3] M Bemelmans, T Van Den Akker, N Ford, R Zachariah, A Harries, E Schouten, K Hermann, B Mwagomba, and M Massaquoi. Providing universal access to antiretroviral therapy in Thyolo, Malawi through task shifting and decentralization of HIV/AIDS care. *Tropical Medicine & International Health*, 15(12):1413–1420, 2010.

[4] Jennifer F Bobb, Ziad Obermeyer, Yun Wang, and Francesca Dominici. Cause-specific risk of hospital admission related to extreme heat in older adults. *JAMA*, 312(24):2659–2667, 2014.

[5] NE Breslow. Statistics in epidemiology: the case-control study. *Journal of the American Statistical Association*, 91(433):14–28, 1996.

[6] NE Breslow and KC Cain. Logistic regression for two-stage case-control data. *Biometrika*, 75(1): 11–20, 1988.

[7] NE Breslow and DG Clayton. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88(421):9–25, 1993.

[8] NE Breslow and R Holubkov. Maximum likelihood estimation of logistic regression parameters under two-phase, outcome-dependent sampling. *Journal of the Royal Statistical Society. Series B (Methodological)*, 59(2):447–461, 1997.

[9] Norman E Breslow and Nilanjan Chatterjee. Design and analysis of two-phase studies with binary outcome applied to wilms tumour prognosis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 48(4):457–468, 1999.

[10] R Brook. Cardiovascular effects of air pollution. *Clinical Science*, 115:175–187, 2008.

[11] Robert D Brook and Sanjay Rajagopalan. Particulate matter, air pollution, and blood pressure. *Journal of the American Society of Hypertension*, 3(5):332–350, 2009.

[12] J Cai, B Qaqish, and H Zhou. Marginal analysis for cluster-based case-control studies. *Sankhyā: The Indian Journal of Statistics, Series B*, pages 326–337, 2001.

[13] Sabit Cakmak, Robert Dales, Judith Leech, and Ling Liu. The influence of air pollution on cardiovascular and pulmonary function and exercise capacity: Canadian health measures survey (chms). *Environmental research*, 111(8):1309–1312, 2011.

[14] Chang-Chuan Chan, Kai-Jen Chuang, Lung-Chang Chien, Wen-Jone Chen, and Wei-Tien Chang. Urban air pollution and emergency admissions for cerebrovascular diseases in taipei, taiwan. *European heart journal*, 27(10):1238–1244, 2006.

[15] Yuyu Chen, Avraham Ebenstein, Michael Greenstone, and Hongbin Li. Evidence on the impact of sustained exposure to air pollution on life expectancy from China's Huai River policy. *Proceedings of the National Academy of Sciences*, 110(32):12936–12941, 2013.

[16] Ana V Diez Roux, Amy H Auchincloss, J Timothy Dvonch, Patrick L Brown, R Graham Barr, Martha L Daviglus, David C Goff Jr, Joel D Kaufman, and Marie S O'Neill. Associations between recent exposure to ambient fine particulate matter and blood pressure in the multi-ethnic study of atherosclerosis. 2008.

[17] Francesca Dominici, Roger D Peng, Michelle L Bell, Luu Pham, Aidan McDermott, Scott L Zeger, and Jonathan M Samet. Fine particulate air pollution and hospital admission for cardiovascular and respiratory diseases. *Jama*, 295(10):1127–1134, 2006.

[18] GP Douglas, OJ Gadabu, S Joukes, S Mumba, MV McKay, A Ben-Smith, A Jahn, EJ Schouten, ZL Lewis, JJ van Oosterhout, et al. Using touchscreen electronic medical record systems to support and monitor national scale-up of antiretroviral therapy in Malawi. *PLoS Medicine*, 7(8):e1000319, 2010.

[19] A Elixhauser, C Steiner, and L Palmer. Clinical classifications software (CCS), 2014. U.S. Agency for Healthcare Research and Quality., 2014. Available: http://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp (Accessed: 2014-12-20).

[20] Environmental Protection Agency. Air quality trends, 2005. Available: http://www.epa.gov/airtrends/aqtrends.html (Accessed: 2014-12-20).

[21] Philip J Everson and Carl N Morris. Inference for multivariate normal hierarchical models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(2):399–412, 2000.

[22] Annunziata Faustini, Massimo Stafoggia, Giovanna Cappai, and Francesco Forastiere. Short-term effects of air pollution in a cohort of patients with chronic obstructive pulmonary disease. *Epidemiology*, 23(6):861–879, 2012.

[23] Pedro Garrett and Elsa Casimiro. Short-term effect of fine particulate matter (pm2. 5) and ozone on daily mortality in lisbon, portugal. *Environmental Science and Pollution Research*, 18(9):1585–1592, 2011.

[24] CF Gilks, S Crowley, R Ekpini, S Gove, J Perriens, Y Souteyrand, D Sutherland, M Vitoria, T Guerma, and K De Cock. The WHO public-health approach to antiretroviral treatment against HIV in resource-limited settings. *The Lancet*, 368(9534):505–510, 2006.

[25] Henry Gong, Jr, William S Linn, Constantinos Sioutas, Sheryl L Terrell, Kenneth W Clark, Karen R Anderson, and Lester L Terrell. Controlled exposures of healthy and asthmatic volunteers to concentrated ambient fine particles in los angeles. *Inhalation toxicology*, 15(4):305–325, 2003.

[26] S Greenland. Principles of multilevel modelling. *International Journal of Epidemiology*, 29(1): 158–167, 2000.

[27] S Greenland and H Morgenstern. Ecological bias, confounding, and effect modification. *International Journal of Epidemiology*, 18(1):269–274, 1989.

[28] S Greenland and H Morgenstern. Ecological bias and confounding (letter). *International Journal of Epidemiology*, 19(3):766–767, 1990.

[29] S Haneuse and S Bartell. Designs for the combination of group-and individual-level data. *Epidemiology*, 22(3):382, 2011.

[30] S Haneuse and J Wakefield. Hierarchical models for combining ecological and case-control data. *Biometrics*, 63(1):128–136, 2007.

[31] S Haneuse and J Wakefield. The combination of ecological and case-control data. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 70(1):73–93, 2008.

[32] S Haneuse and J Wakefield. Geographic-based ecological correlation studies using supplemental case–control data. *Statistics in Medicine*, 27(6):864–887, 2008.

[33] AD Harries, DS Nyangulu, NJ Hargreaves, O Kaluwa, and FM Salaniponi. Preventing antiretroviral anarchy in sub-Saharan Africa. *The Lancet*, 358(9279):410–414, 2001.

[34] AD Harries, P Gomani, R Teck, OA de Teck, E Bakali, R Zachariah, E Libamba, A Mwansambo, F Salaniponi, and R Mpazanje. Monitoring the response to antiretroviral therapy in resource-poor settings: the Malawi model. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 98(12): 695–701, 2004.

[35] Mehdi S Hazari, William H Rowan, Darrell W Winsett, Allen D Ledbetter, Najwa Haykal-Coates, William P Watkinson, and Daniel L Costa. Potentiation of pulmonary reflex response to capsaicin 24h following whole-body acrolein exposure is mediated by trpv1. *Respiratory physiology & neurobiology*, 160(2):160–171, 2008.

[36] PJ Heagerty and BF Kurland. Misspecified maximum likelihood estimates and generalised linear mixed models. *Biometrika*, 88(4):973–985, 2001.

[37] JD Hunter and M Doddi. Sepsis and the heart. *British journal of anaesthesia*, 104(1):3–11, 2010.

[38] NL Johnson and S Kotz. *Distributions in statistics*. Wiley series in probability and mathematical statistics. John Wiley & Sons, Inc., New York, 1969.

[39] Fanny WS Ko, Wilson Tam, Tze Wai Wong, Doris PS Chan, Alvin H Tung, Christopher KW Lai, and David SC Hui. Temporal relationship between air pollutants and hospital admissions for chronic obstructive pulmonary disease in hong kong. *Thorax*, 62(9):780–785, 2007.

[40] K Liang and SL Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73 (1):13–22, 1986.

[41] JG Liao and O Rosen. Fast and stable algorithms for computing and sampling from the noncentral hypergeometric distribution. *The American Statistician*, 55(4), 2001.

[42] Regina Y Liu and Kesar Singh. Moving blocks jackknife and bootstrap capture weak dependence. *Exploring the limits of bootstrap*, 225:248, 1992.

[43] JM Martínez, J Benach, J Ginebra, FG Benavides, and Y Yasui. An integrated analysis of individual and aggregated health data using estimating equations. *The International Journal of Biostatistics*, 3(1), 2007.

[44] JM Martínez, J Benach, FG Benavides, C Muntaner, R Clèries, O Zurriaga, MA Martínez-Beneito, and Y Yasui. Improving multilevel analyses: the integrated epidemiologic design. *Epidemiology*, 20 (4):525–532, 2009.

[45] James E McCarthy and Claudia Copeland. Epa regulations: Too much, too little, or on track? Congressional Research Service, Library of Congress, 2011.

[46] H Morgenstern. Uses of ecologic analysis in epidemiologic research. *American Journal of Public Health*, 72(12):1336–1344, 1982.

[47] J Neuhaus, AJ Scott, and CJ Wild. The analysis of retrospective family studies. *Biometrika*, 89(1): 23–37, 2002.

[48] JM Neuhaus and NP Jewell. The effect of retrospective sampling on binary regression models for clustered data. *Biometrics*, pages 977–990, 1990.

[49] JM Neuhaus, WW Hauck, and JD Kalbfleisch. The effects of mixture distribution misspecification when fitting mixed-effects logistic models. *Biometrika*, 79(4):755–762, 1992.

[50] JM Neuhaus, AJ Scott, and CJ Wild. Family-specific approaches to the analysis of case–control family data. *Biometrics*, 62(2):488–494, 2006.

[51] JM Neuhaus, CE McCulloch, and R Boylan. A note on type II error under random effects misspecification in generalized linear mixed models. *Biometrics*, 67(2):pp. 654–660, 2011. ISSN 0006341X. URL http://www.jstor.org/stable/41242504.

[52] S Piantadosi, DP Byar, and S Green. The ecological fallacy. *American Journal of Epidemiology*, 127 (5):893–904, 1988.

[53] C Arden Pope, Richard T Burnett, George D Thurston, Michael J Thun, Eugenia E Calle, Daniel Krewski, and John J Godleski. Cardiovascular mortality and long-term exposure to particulate air pollution epidemiological evidence of general pathophysiological pathways of disease. *Circulation*, 109(1):71–77, 2004.

[54] C Arden Pope, Joseph B Muhlestein, Heidi T May, Dale G Renlund, Jeffrey L Anderson, and Benjamin D Horne. Ischemic heart disease events triggered by short-term exposure to fine particulate air pollution. *Circulation*, 114(23):2443–2448, 2006.

[55] C Arden Pope III, Majid Ezzati, and Douglas W Dockery. Fine-particulate air pollution and life expectancy in the United States. *New England Journal of Medicine*, 360(4):376–386, 2009.

[56] BA Portnov, J Dubnov, and M Barchana. On ecological fallacy, assessment errors stemming from misguided variable selection, and the effect of aggregation on the outcome of epidemiological study. *Journal of Exposure Science and Environmental Epidemiology*, 17(1):106–121, 2007.

[57] RL Prentice. Correlated binary regression with covariates specific to each binary observation. *Biometrics*, pages 1033–1048, 1988.

[58] RL Prentice and R Pyke. Logistic disease incidence models and case-control studies. *Biometrika*, 66 (3):403–411, 1979.

[59] RL Prentice and L Sheppard. Aggregate data studies of disease risk factors. *Biometrika*, 82(1): 113–125, 1995.

[60] David Q Rich, Haluk Ozkaynak, James Crooks, Lisa Baxter, Janet Burke, Pamela Ohman-Strickland, Kelly Thevenet-Morrison, Howard M Kipen, Junfeng Zhang, John B Kostis, et al. The triggering of myocardial infarction by fine particles is enhanced when particles are enriched in secondary species. *Environmental science & technology*, 47(16):9414–9423, 2013.

[61] S Richardson and D Hémon. Ecological bias and confounding (letter). *International Journal of Epidemiology*, 19(3):764–766, 1990.

[62] JM Robins, A Rotnitzky, and LP Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866, 1994.

[63] JM Robins, A Rotnitzky, and LP Zhao. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90(429): 106–121, 1995.

[64] WS Robinson. Ecological correlations and the behavior of individuals. *American Sociological Review*, 15(3):351–357, 1950.

[65] R Salway and J Wakefield. Sources of bias in ecological studies of non-rare events. *Environmental and Ecological Statistics*, 12(3):321–347, 2005.

[66] JS Schildcrout and PJ Heagerty. On outcome-dependent sampling designs for longitudinal binary response data with time-varying covariates. *Biostatistics*, 9(4):735–749, 2008.

[67] JS Schildcrout and PJ Rathouz. Longitudinal studies of binary response data following case–control and stratified case–control sampling: design and analysis. *Biometrics*, 66(2):365–373, 2010.

[68] JS Schildcrout, SP Garbett, and PJ Heagerty. Outcome vector dependent sampling with longitudinal continuous response data: stratified sampling based on summary statistics. *Biometrics*, 69(2): 405–416, 2013.

[69] Alastair J Scott and Chris J Wild. Fitting regression models to case-control data by maximum likelihood. *Biometrika*, 84(1):57–71, 1997.

[70] L Sheppard. Ecological study design. In *Encyclopedia of Environmetrics*, volume 2, pages 602–606. John Wiley & Sons, Ltd, New York, 2002. ISBN 9780470057339.

[71] L Sheppard. Insights on bias and information in group-level studies. *Biostatistics*, 4(2):265–278, 2003.

[72] Gunjan J Shukla and Peter J Zimetbaum. Syncope. *Circulation*, 113(16):e715–e717, 2006.

[73] E Smoot and S Haneuse. On the analysis of hybrid designs that combine group-and individual-level data. *Biometrics*, 2014.

[74] J Wakefield. Ecological inference for 2 x 2 tables. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 167(3):385–445, 2004.

[75] J Wakefield and S Haneuse. Overcoming ecologic bias using the two-phase study design. *American Journal of Epidemiology*, 167(8):908–916, 2008.

[76] J Wakefield and G Shaddick. Health-exposure modeling and the ecological fallacy. *Biostatistics*, 7(3): 438–455, 2006.

[77] JE White. A two stage design for the study of the relationship between a rare exposure and a rare disease. *American Journal of Epidemiology*, 115(1):119–128, 1982.

[78] Loren E Wold, Zhekang Ying, Kirk R Hutchinson, Markus Velten, Matthew W Gorr, Christina Velten, Dane J Youtz, Aixia Wang, Pamela A Lucchesi, Qinghua Sun, et al. Cardiovascular remodeling in response to long-term exposure to fine particulate matter air pollution. *Circulation: Heart Failure*, 5 (4):452–461, 2012.

[79] GY Wong and WM Mason. The hierarchical logistic regression model for multilevel analysis. *Journal of the American Statistical Association*, 80(391):513–524, 1985.

[80] World Health Organisation. Malawi: Summary country profile for HIV/AIDS treatment scale-up., 2005. Available: http://www.who.int/hiv/HIVCP_MWI.pdf (Accessed: 2014-11-23).

[81] World Health Organization. Interim WHO clinical staging of HIV/AIDS and HIV/AIDS case definitions for surveillance, 2005. Available: http://www.who.int/hiv/pub/guidelines/clinicalstaging.pdf (Accessed: 2014-11-23).

Colophon

This thesis was typeset using LaTeX, originally developed by Leslie Lamport and based on Donald Knuth's TeX. The body text is set in 11 point Arno Pro, designed by Robert Slimbach in the style of book types from the Aldine Press in Venice, and issued by Adobe in 2007. A template, which can be used to format a PhD thesis with this look and feel, has been released under the permissive MIT (X11) license, and can be found online at github.com/suchow/ or from the author at suchow@post.harvard.edu.