Bayesian Model Expansion for Selection Bias in Epidemiology

by

Rob Trangucci

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Statistics)
in the University of Michigan
2023

Doctoral Committee:

Assistant Professor Yang Chen, Co-Chair
Associate Professor Jon Zelner, Co-Chair
Associate Professor Emily Toth Martin
Research Associate Professor Yajuan Si
Professor Stilian Stoev

Rob Trangucci
trangucc@umich.edu
ORCID iD: 0000-0002-4592-718X

DEDICATION

To my loves Shivani and Saanvi

ACKNOWLEDGMENTS

I am so thankful to both of my advisors, Yang Chen and Jon Zelner. To Yang, the first time we met I could see how much you believed in me. Your support gave me the confidence to turn germs of ideas into fully-fleshed out research projects. I have learned so much from you over the past 6 years. I'm a better statistician and researcher for having worked together. To Jon, I credit you with sparking my interest in selection bias in epidemiology. Ever since meeting at the Hungarian Pastry Shop, I've wanted to work on statistical models in epidemiology. I feel incredibly lucky to have met you, and even more lucky to have worked with you over the past 5 years. You've taught me the importance of using statistics to address real-world issues. Most importantly, thank you for making me feel like a part of your family.

I am eternally grateful to my dissertation committee members, Stilian Stoev, Yajuan Si, and Emily Martin. Your incisive questions and constructive comments on my thesis pushed me to more deeply understand my own work.

To the EpiBayes research group, thank you for listening to countless job talks, and research presentations. Your feedback always made my talks better.

To Ben Goodrich, your missing data course was the most important class I took in graduate school. The inspiration for this thesis can be directly traced back to that class.

To Andrew Gelman, thank you for being so generous with me over the past 8 years. You showed me that statistics research can be inclusive and collaborative.

To Dan and Sara, living in Ann Arbor would not have been the same without you. It was always comforting knowing you were both right around the corner.

I am deeply grateful to my parents, who always fostered my curiosity through unwavering support for my education.

I cannot overstate how thankful I am to my wife, Shivani. You are the reason I decided to get my PhD. I could not have done this without you.

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE

LIST OF TABLES

TABLE

LIST OF APPENDICES

ABSTRACT

Selection bias is a massive problem in infectious disease epidemiology that can result in needless morbidity and mortality. This bias is both subtle and ubiquitous, occurring even in randomized clinical trials. For example, medical researchers cannot randomize responses to treatment intermediate to the outcome of interest, and epidemiologists cannot force patients to report sensitive demographic information. In order to do inference in these complex scenarios, we need new classes of models that capture the scientific process of interest while accounting for how the data were observed.

In this thesis I develop theory and practice for Bayesian model expansion to mitigate and adjust for selection bias in the analysis of observational and experimental data arising in the areas of missing data, causal inference, and survey research. In the second chapter I propose a novel method to infer stratified incidence in disease surveillance data with partially-observed stratum information. Public health researchers often compare risk of disease among demographic subgroups in order to design interventions. Missingness in demographic covariates like race/ethnicity, or age group complicates this endeavor; dropping cases with missing covariates can lead to endogenous selection bias. Instead, I develop a locally-identifiable joint model for the missingness process and the disease process that allows for the missingness process to be not-missing-at-random. The model is identified by marrying spatial information in the disease data with spatial Census data. I investigate the finite-sample properties of the model via a simulation study, and apply my model to COVID-19 case data in Southeastern Michigan. I show that the burden of COVID-19 from March to July of 2020 for non-Whites relative to that of Whites is understated when cases that are missing race/ethnicity information are omitted.

In the third chapter I develop a method to point-identify vaccine efficacy (VE) against post-infection outcomes such as severe illness, and death. Policy makers need to quantify post-infection outcome VE so as to design effective vaccination strategies, but these causal estimands are typically nonidentifiable. I propose a method to identify these estimands under measurement error on infection and post-infection outcomes by taking advantage of the structure of vaccine efficacy trials; these trials are typically run across different health systems and collect pretreatment covariates related to an individual's susceptibility to infection. I show that my method not only yields identifiability of the causal estimand, but also identifies the infection measurement error parameters. I

then investigate the Type I error and power of my method via a simulation study.

In the final chapter, I propose a new Bayesian generative semiparametric model for characterizing the cumulative spatial exposure to an environmental health hazard that is not well-represented by a single point in space, like a system of wastewater canals. The model couples a dose-response model with a log-Gaussian Cox process integrated against a distance kernel with an unknown length-scale. I show that this model is well-defined, and that a simple integral approximation adequately controls the computational error. Before applying the model to survey data from Mexico, I quantify the finite-sample properties and the computational tractability of the discretization scheme in a simulation study.

CHAPTER 1

Introduction

The past several years have seen a marked increase in the volume of epidemiological data available to public health researchers due to the COVID-19 pandemic. These data, however, may not be representative of the population of interest. Even representative data may be made non-representative by seemingly innocuous research decisions to limit the analyses to a subgroup (Elwert and Winship, 2014; Lu et al., 2022). Worse yet, if researchers were to use non-representative data, their inferences may be biased, which could lead them to draw conclusions with disastrous consequences. The resulting bias in inferences is known as selection bias, reflecting the notion that how samples are selected for analysis can lead to bias.

For example, early on in the pandemic, researchers noticed that COVID-19 case data had a high rate of missing race information. This missing information frustrated efforts to compare COVID-19 case fatality rates and cumulative incidence rates by race and ethnicity; if the probability that race/ethnicity information was missing correlated with race/ethnicity, then limiting the case data to only those with complete information would bias these comparisons (Millett et al., 2020). Poor inferences on these differences could lead to health policies that are inefficient at best and harmful at worst.

This thesis investigates three scenarios in epidemiology where selection bias can arise and reduces this bias by employing model expansion and data augmentation.

1.1 Selection bias

Broadly, selection bias occurs when measurements on individuals sampled from a population systematically differ from measurements on the population as a whole. Let i index an individual in a population of size N, let the measurement of a quantity of interest be Y_i, and let the binary variable R_i be the selection indicator such that R_i equals 1 when an individual is included in the sample and 0 otherwise. Selection bias is the difference between the estimand arising from the sample, $\mathbb{E}\left[Y_i \mid R_i = 1\right]$ and the population estimand, $\mathbb{E}\left[Y_i\right]$. Straightforward algebra shows that

this difference is equal to the covariance between the selection indicator and the measurement of interest:

$$\mathbb{E}\left[Y_i \mid R_i = 1\right] - \mathbb{E}\left[Y_i\right] = \frac{\text{Cov}(Y_i, R_i)}{\mathbb{E}\left[R_i\right]}. \tag{1.1}$$

If subsequent analysis excludes some sampled units, there may be additional selection bias. Let $S_i \mid R_i = 1$ be a binary variable equal to 1 if unit i is included in subsequent analysis, 0 if the unit is excluded for analysis, and let $P(S_i = 0 \mid R_i = 0) = 1$. Then, by a repeated application of the algebra in Equation (1.1), the total selection bias is the sum of two covariance terms:

$$\mathbb{E}\left[Y_i \mid S_i = 1, R_i = 1\right] - \mathbb{E}\left[Y_i\right] = \frac{\text{Cov}(Y_i, S_i \mid R_i = 1)}{\mathbb{E}\left[S_i \mid R_i = 1\right]} + \frac{\text{Cov}(Y_i, R_i)}{\mathbb{E}\left[R_i\right]}. \tag{1.2}$$

Thus, there are two potential sources of selection bias: (1) selection bias arising from subselecting units from a sample for analysis, (i.e. $\mathbb{E}\left[Y_i \mid S_i = 1, R_i = 1\right] - \mathbb{E}\left[Y_i \mid R_i = 1\right]$) and (2) selection bias arising from selecting units from the population (i.e. $\mathbb{E}\left[Y_i \mid R_i = 1\right] - \mathbb{E}\left[Y_i\right]$) (Elwert and Winship, 2014). Equation (1.2) clarifies that even if a sample is representative of the population, i.e. $\text{Cov}(Y_i, R_i) = 0$, subsequent analysis that results in nonzero covariance between S_i and Y_i induces selection bias. For example, if S_i indicates that a unit is not missing data, analysis limited to completely observed units risks selection bias (Daniel et al., 2012).

1.2 Identifiability and selection bias

An important characteristic of parameters of statistical models is *identifiability*, or, whether data generated from a model indexed by parameter θ can be mapped uniquely to θ. Identifiability is formally defined in Rothenberg (1971) as

Definition 1.2.1 (Parameter identifiability)**.** *A parameter $\theta \in \Theta$ is identifiable if there does not exist a distinct parameter value $\theta' \in \Theta$ for which the density $f(y \mid \theta) = f(y \mid \theta')$ for all observations y.*

This asymptotic property is important to verify for statistical models because it is necessary condition for much of asymptotic theory (Keener, 2010; Lehmann and Casella, 1998). For Bayesian inference, it is important to determine which parameters are nonidentifiable in order to understand which parameters are asymptotically sensitive to priors (Gustafson, 2015).

Modifying a statistical model to account for selection bias may render an identifiable parameter nonidentifiable. This is because the process by which selection occurs introduces more unknown parameters, typically without a commensurate increase in observed information. One strategy to identify the estimand $\mathbb{E}\left[Y_i\right]$ under selection bias is by expanding the model with more information.

1.3 Model expansion

Given a covariate $X_i \in \mathcal{X}$ such that $\text{Cov}(Y_i, R_i \mid X_i = x) = 0$, the estimand $\mathbb{E}[Y_i]$ can be identified. To see why, note that when the conditional covariance is zero, $\mathbb{E}[Y_i \mid R_i = 1, X_i = x] = \mathbb{E}[Y_i \mid X_i = x]$. If one has access to the population distribution for X_i, $P(x)$, then $\mathbb{E}[Y_i] = \int_{\mathcal{X}} \mathbb{E}[Y_i \mid R_i = 1, X_i = x] \, P(dx)$. This technique is known as poststratification (Little, 1995; Gelman and Little, 1997; Lohr, 2019).

An alternative is to assume that there is a covariate Z_i, also known as an instrumental variable, that determines R_i but does not determine Y_i. The selection and observation equations can be parameterized as follows:

$$\begin{aligned} R_i &= \mathbb{1}_{\beta Z_i - V_i \geq 0} \\ Y_i &= \alpha + U_i \\ (V_i, U_i) &\sim F \end{aligned} \tag{1.3}$$

Then $P(R_i = 1 \mid Z_i = z) = P(V_i \leq \beta z \mid Z_i = z)$, and $\text{Cov}(Y_i, R_i)$ is equal to a function of the propensity score, $P(R_i = 1 \mid Z_i = z)$ and the joint distribution of U_i, V_i (Heckman et al., 2006). Explicitly:

$$\text{Cov}(Y_i, R_i) = \mathbb{E}_{Z_i}\left[\mathbb{E}[U_i \mid V_i \leq \beta z] \, P(V_i \leq \beta z_i \mid Z_i = z)\right] \tag{1.4}$$

The key assumptions that lead to identifiability of the model are: (1) that Z_i occurs in the selection equation only, (2) that Z_i has support on $\mathbb{R}$, (3) (V_i, U_i) is median zero, and (4) the errors are independent of Z_i (Heckman, 1990). If these properties hold, then $\mathbb{E}[Y_i] = \alpha$. Powell (1994) reviews semiparametric estimation of the selection equation and the joint error distribution.

1.3.1 Costs of model expansion

Model expansion is not without costs and potential pitfalls. First, without careful regularization, the decreased bias of these approaches will be offset by increased variance. Second, the additional structure of each model necessarily results in additional assumptions. To the extent these assumptions are violated, these methods may incur bias from model misspecification. Third, as Gustafson (2005) shows, when expanded identifiable models nest nonidentifiable submodels, the nonidentifiable parameter space is a measure-zero subset of the identified parameter space. As the true data generating process approaches the nonidentifiable submodel the expanded models can exhibit large mean squared error and poor coverage (akin to scenarios examined through a Frequentist lens in Andrews and Mikusheva (2015)).

1.3.2 Bayesian inference for model expansion

These pitfalls can be mitigated by the use of Bayesian inference. Bayesian inference with carefully formulated proper priors can combat increased model complexity as models expand (Gelman et al., 2013). Furthermore, Bayesian inference with proper priors can mitigate poor performance near a nonidentifiable submodel by trading a small bit of bias for a larger reduction in variance (Gustafson, 2005). Bayesian inference can also be used to weaken strict identifying assumptions by allowing for soft constraints (Gustafson, 2007).

Bayesian model expansion is the core technique linking the methods I have developed in this thesis. At base, the solutions I propose depend on jointly modeling the selection process alongside the scientific process of interest by incorporating additional data or covariates and prior information. Bayesian inference in expanded models allows for identifiability while reducing the risks of model expansion.

Next I will introduce the following chapters of the thesis.

1.4 Missing race/ethnicity information in COVID-19 cases

Chapter 2 deals with missing race and ethnicity data in COVID-19 case data. This is an ongoing problem in case data collection, and will be a problem in future pandemics. The chapter is the most straightforward demonstration of selection bias: limiting analysis to observations with no missing covariates (also known as "complete-case analysis") may result in selection bias (Elwert and Winship, 2014). If the probability of missingness of the covariate data and the outcome are influenced by the covariate values, then complete-case analyses may result in biased estimates of relative rates of disease by race/ethnicity. This type of missingness, called not-missing-at-random (NMAR) missingness, is the most pernicious missingness because it cannot be solved by modeling the observed data. Moreover, NMAR missingness for discrete covariates is an area of active research (Gómez-Rubio et al., 2019).

I developed a novel method for jointly modeling the missing covariate data and the disease data by using known population data. Importantly, this model has disease rate parameters and race/ethnicity missingness parameters that vary by race and ethnicity. The key idea is that cases of disease arise from a population with a known joint distribution of demographic covariates. Using this idea, along with the fact that cases are recorded with precise geographic information, allows us to to associate each case with fine-scale Census areas like tracts or Public Use Microdata Areas. I then employ a parametric model that allows both the missingness model parameters and the disease model parameters to be identifiable. The model can be naturally extended to include other covariate information like age and sex, and admits a hierarchical generalization to borrow

strength across differing geographic areas. We showed that the model outperforms methods that limit analysis to cases with no missing information, and multiple imputation methods that assume that missingness does not depend on race/ethnicity. Then we applied our model to COVID-19 case data in Southeastern Michigan, and show that complete-case analysis understated the relative burden of COVID-19 incidence in non-White race/ethnicities in the early pandemic.

1.5 Principal stratification for vaccine efficacy

Chapter 3 addresses an open problem in causal inference, namely how to identify a causal effect of a treatment on a secondary outcome that is conditional on a post-treatment variable. This problem arises naturally in vaccine efficacy (VE) studies (i.e. randomized placebo-controlled trials for vaccines) when inferring the post-infection outcome VE. Here too, lurks selection bias (Balke and Pearl, 1994; Frangakis and Rubin, 2002). There are many post-infection outcomes that are of interest to public health researchers: (1) binary outcomes like symptomatic disease, or severe illness, (2) ordinal outcomes like symptom severity and antibody titer, and (3) continuous outcomes like viral load or time to symptom onset. Conditional on the event that an individual is infected, vaccines may have an effect on each of these outcomes; quantifying the conditional effect is important for developing an optimal vaccination policy, choosing between several vaccine candidates, and communicating VE trial results to the public. The intuitive estimator for these effects, comparing outcomes between the vaccinated and unvaccinated *within* the subgroup that is infected, results in selection bias (Hudgens and Halloran, 2006).

In this instance, selection bias arises because the vaccinated and infected subgroup is different than the unvaccinated and infected subgroup. The validity of causal inference depends on comparing individual outcomes under vaccination and placebo (Frangakis and Rubin, 2002). For instance, let $S_i(z)$ be the binary infection outcome for an individual i that would be observed under vaccination ($z = 1$) and under placebo ($z = 0$), let $Y_i(z, S_i(z))$ be the binary indicator for severe disease given infection status $S_i(z)$, and let Z_i be the treatment assignment for that individual. The individual causal effect for vaccination on infection is $S_i(0) - S_i(1)$. This is a valid causal comparison because the individual is fixed; any unobserved variation between individuals is controlled for, and the only explanation for the difference (or lack thereof) is that the vaccine prevented (or did not prevent) infection in individual i. Randomized treatment assignment guarantees that treatment assignment is independent of the outcomes $(S_i(1), S_i(0))$, and we can estimate the expectation of this causal effect:

$$\mathbb{E}\left[S_i(0) - S_i(1)\right] = \mathbb{E}\left[S_i(0)\right] - \mathbb{E}\left[S_i(1)\right] = \mathbb{E}\left[S_i \mid Z_i = 0\right] - \mathbb{E}\left[S_i \mid Z_i = 1\right].$$

However, the conditional comparison of severe disease in infected individuals, or

$$\mathbb{E}\left[Y_i \mid S_i = 1, Z_i = 0\right] - \mathbb{E}\left[S_i \mid S_i = 1, Z_i = 1\right]$$

does not correspond to an expectation of an individual causal outcome because the left-hand expectation is taken over different individuals than the right-hand expectation. To see why, we can expand the conditional expectations. The expansion of the quantity $\mathbb{E}\left[Y_i \mid S_i = 1, Z_i = 0\right]$ in terms of potential infection outcomes:

$$\begin{aligned}\mathbb{E}\left[Y_i \mid S_i = 1, Z_i = 0\right] = &\mathbb{E}\left[Y_i(0,1) \mid S_i(1) = 1, S_i(0) = 1\right] P(S_i(1) = 1, S_i(0) = 1) \\ &+ \mathbb{E}\left[Y_i(0,1) \mid S_i(1) = 0, S_i(0) = 1\right] P(S_i(1) = 0, S_i(0) = 1)\end{aligned}$$

shows that the expectation is a mixture of outcomes between people infected under both vaccination and placebo (the "always-infected" group), and people infected under placebo and protected under vaccination (the "protected" group). The expansion of $\mathbb{E}\left[S_i \mid S_i = 1, Z_i = 1\right]$ is

$$\begin{aligned}\mathbb{E}\left[S_i \mid S_i = 1, Z_i = 1\right] = &\mathbb{E}\left[Y_i(1,1) \mid S_i(1) = 1, S_i(0) = 1\right] P(S_i(1) = 1, S_i(0) = 1) \\ &+ \mathbb{E}\left[Y_i(1,1) \mid S_i(1) = 1, S_i(0) = 0\right] P(S_i(1) = 1, S_i(0) = 0).\end{aligned}$$

This expansion shows that the expectation is a mixture of outcomes between people infected under both vaccination and placebo (the "always-infected" group), and people infected under vaccination and uninfected under placebo (the "harmed" group).

The difference of these expectations implicitly results in a comparison of two different groups of people in addition to a comparison two treatments. Thus, any value we observe for this difference could be attributed to the treatment effect, which is the estimand of interest, or the difference in the baseline risk of severe illness outcomes between the two groups. Conditioning on infection induces selection bias in that $\{i \mid S_i(1) = 1\} \neq \{i \mid S_i(0) = 1\}$ (Frangakis and Rubin, 2002).

The solution, called principal stratification, is to focus on the treatment effect in the group of always-infected people (Hudgens and Halloran, 2006). Because we can only ever observe one outcome for each person, however, we cannot identify the always-infected group.

In this chapter, I solve the problem probabilistically by using the structure of VE trials to our advantage. These trials are typically run across several geographically-disparate health centers and measure pretreatment covariates related to participants' susceptibility to infection. We develop a new model for vaccine efficacy that uses these characteristics, and show that our model identifies the post-infection outcome VE given verifiable conditions on the data generating process. This represents a novel contribution to causal inference literature as well as to vaccine efficacy literature. We then use our model to design several VE trials, and show that the sample sizes required for

high-powered trials are feasible.

1.6 Measuring cumulative spatial exposure to environmental hazards

Chapter 4 concerns the problem of measuring exposure to an environmental health hazard or built-environment source of disease that is extensive in space. Put another way, these hazards are not well-represented by a single point in space. Examples of such exposures are roadways, wastewater canals, and polluted waterways. There is a well-developed literature on inferring health risk from point-source exposures, like chemical plants, or nuclear waste sites (Diggle et al., 1997). In order to use these methods for extensive hazards, researchers may use the shortest distance to the hazard to use as a proxy for exposure (Cassell et al., 2018). These methods, however, will yield biased estimates of environmental risk if the individuals at risk are exposed at more than one point along the hazard. Moreover, this choice ignores the geometry of the hazard as well as the variation in risk intensity along the hazard.

To make things concrete, we will focus on an applied problem that exhibited key characteristics that are common to other problems in extensive environmental exposure. Health researchers were interested in how the risk of childhood diarrhea depended on household proximity to a system of wastewater canals in the Mezquital Valley in Mexico. To measure this risk, researchers ran a survey of households with children under 5 over two years (Contreras et al., 2017). This survey included questions about whether the child or children in the household had experienced any bouts of diarrhea in the past week. The initial data analysis, presented in Contreras et al. (2020), used a shapefile of the canal system in conjunction with GPS measurements of household location to measure the shortest distance between the household and the wastewater canal.

Selection bias arises in this scenario from the fact that points chosen for analysis under the shortest-distance model may not be representative of the risk at all points along the wastewater canal. In fact, if household locations are functions of risk at the shortest point, then inferences about risk of diarrhea using only the shortest distance to the canal will be understated. This is plausible if some houses were built after the canal system was built or that there are built-environment barriers between houses and the canal system that vary across the canal system. In this case, selection bias arises due to lack of information about the wastewater exposure process.

Chapter 4 presents a novel solution to measuring the cumulative spatial risk from environmental exposures by formulating a generative model of infection from an environmental hazard. This model involves a Poisson process of pathogen intensity along the canal, and a kernel function that measures exposure to a point along the canal at a given distance; the total spatial exposure for

a household to the wastewater canal is then an integral of these two functions across the extent of the canal. Thus, we prevent the analyst from having to choose a single point of exposure for each household, thereby mitigating selection bias. Because our method depends on approximating an integral of an unknown function, we show that the approximate integration scheme converges to the integral of the modeled function as our computational grid increases in resolution, and via simulation study show that our method is computationally feasible. We also show via simulation study that as the number of sampled households increases we learn the unknown intensity functions with decreasing integrated absolute error. Finally, we apply our model to the Mezquital Valley diarrheal illness dataset, and explore the differences between our model's inferences and those presented in Contreras et al. (2020).

CHAPTER 2

Modeling Racial/Ethnic Differences in COVID-19 Incidence with Covariates Subject to Non-Random Missingness

The contents of this chapter will appear as
Rob Trangucci, Yang Chen, and Jon Zelner, Modeling racial/ethnic differences in COVID-19 incidence with covariates subject to non-random missingness. *Forthcoming in Annals of Applied Statistics*.

Complete and detailed surveillance data[1] are critical sources of information for decision-making and communication in public health emergencies like the COVID-19 pandemic. Under ideal conditions, these data can provide an indication of emerging trends, e.g. growth in socioeconomic inequity in infection and disease risk, which can be used to craft policies and target resources. For example, after surveillance data pointed to wide racial/ethnic disparities in incidence and mortality in the COVID-19 pandemic in the United States (Millett et al., 2020; Zelner et al., 2021), policies intended to narrow the gap were put in place Office of Michigan Governor (2020); Governor Whitmer Executive Order (2020). However, without adequate information on the distribution of infection within and between different socioeconomic and race/ethnic groups, the impact of such policy measures is difficult to evaluate.

Missing covariates have long been a challenge associated with administrative datasets, such as public health surveillance data, and the scale and importance of this problem has only grown during the COVID-19 pandemic (Labgold et al., 2021; Millett et al., 2020). Covariate missingness in surveillance data may result from a variety of mechanisms, ranging from non-response on an intake form, refusal to participate in tracing interviews, or data-entry errors after these data are collected. Often this missingness is implicitly or explicitly assumed to occur at random, i.e. not

[1] Surveillance data are aggregated sets of disease cases, often subject to timely reporting requirements, meeting a common set of diagnostic characteristics so as to aid in the monitoring of disease outbreaks (Held et al., 2019).

as a function of the disease process or the attributes of individual cases. But if the process causing case data to be missing important categorical variables, e.g. age, race, sex, or neighborhood, is dependent on the disease process, excluding cases that are missing these covariates may result in estimates that are overconfident and biased. Furthermore, the direction of this bias is not easy to characterize, and may result in over- or under-estimates of group-level differences in risk as epidemic conditions shift.

In emergency situations, such as a surging pandemic, it is easy to see how the disease process itself may induce non-random missingness of covariates. For example, during a period of rapidly increasing caseloads, such as the Delta and Omicron surges of the COVID-19 pandemic, the overwhelming number of cases is likely to limit the ability of case investigators to collect data that are as detailed as those collected during lower-incidence periods. These differences may also be more pronounced when comparing wealthier and poorer jurisdictions with differential resources for case-finding and intervention. When these differential risks and resources are concentrated in communities with large proportions of non-White residents, the likelihood that the missingness of key demographic information, including race/ethnicity, will depend on the race of respondents is high. The intensity of this missingness is also likely to vary in space, reflecting numerous factors including differences in epidemic conditions as well as varying data quality across public health jurisdictions.

Both of these characteristics point to a nonignorable missing data problem, as presented in Rubin (1976). Both issues make quantifying the relative risk of infection among population strata during the COVID-19 pandemic potentially error-prone: high proportions of cases are missing demographic data, with missingness that is likely differential across population strata. It is in this scenario that omitting cases with missing demographic data may yield biased estimates of relative risk. Tools that assume ignorability, like multiple imputation methods typically do (Audigier et al., 2018), cannot correct for missingness that depends on the value of the covariate, and will thus incur bias as well. This problem is not exclusive to the challenge of characterizing sociodemographic disparities in infection risk: incomplete reporting of vaccination status may lead to difficulty in estimating risks of breakthrough COVID-19 infections among vaccinated people, and missing information on comorbid conditions increasing the risk of death may complicate efforts at estimating risks of death associated with infection.

In order to employ statistical methods that appropriately account for missingness, such as those presented in Little and Rubin (2002), one must make the modeling assumptions explicit in a joint probability model for the outcome variable, the covariate subject to missingness, and the missingness process for that covariate. When the missingness process is nonignorable, two broad classes of models can be used to encode the assumptions about the joint distribution: selection models and pattern mixture models (Little and Rubin, 2002). There is much literature on the theoretical

and practical applications of both classes of models: Diggle and Kenward (1994); Clark and Houle (2014); Roy and Daniels (2008). For a review of selection and pattern mixture models see Little (2008) and Little (1995).

We develop a novel model that accounts for nonignorable missingness of demographic covariates for which there is known population data, as in Zangeneh and Little (2022), but we take a selection model approach instead of a pattern mixture model approach. Our probabilistic model is similar to that of Stasny (1991), wherein Stasny develops a selection model for nonignorable missingness in binary survey data, though we incorporate ideas from Zangeneh and Little (2022) in using known census demographic data. Our approach is to develop a model that allows for simultaneous modeling of the disease and missingness processes, and that incorporates information on spatial clustering of risk in addition to sociodemographic risk factors. Given the ubiquity of missing categorical covariates in public health surveillance data and the generality of our model, there are many potential applications of this class of models.

2.1 Alternative approaches

Because missing data can lead to ineffective and potentially life-threatening decision-making in public health and medicine, analysis of epidemiological data subject to missingness is an area of active research. This work, however, is often focused on accounting for data missing-at-random or on imputing values of continuous covariates. Recent work focusing on accounting for missing covariates when modeling disease data in space and time like Holland et al. (2015); Gómez-Rubio et al. (2019); Baker et al. (2014) suffer from several limitations. Gómez-Rubio et al. (2019) presents a framework for joint modeling of the disease process and missingness process, which can incorporate NMAR missingness, but only for continuous covariates. When missing covariates are discrete, Gómez-Rubio et al. resort to multiple imputation, which compromises statistical efficiency gained from joint modeling, increases the computational burden, and assumes MAR missingness. Holland et al. (2015) does include a model for discrete missing covariates with an outcome model, but the missing data mechanism is assumed to be MAR. In other work, Baker et al. (2014) developed a cross-validation approach to missing data imputation, but assume MAR missingness. Recent work in applications for missing data continue to assume MAR (Aguayo et al., 2020; Labgold et al., 2021). As we argue above, assuming missingness is MAR can bias inferences; Perkins et al. (2018); Sidi and Harel (2018); Stavseth et al. (2019) explore how mistaken assumptions in the missing data model impact inferences. When missing data are inherently social in nature, MAR assumptions become even more tenuous than they might be in other settings. In the context of the COVID-19 pandemic, missingness of race/ethnicity data reflects a host of factors, including socioeconomic biases in the quality and thoroughness of public health data

systems, which effectively guarantee correlation between the race/ethnicity of the respondent and their likelihood of missing these data. This paucity of recent work in spatial epidemiology employing an NMAR missingness model for discrete missing covariates, and the urgency of improving the quality of inferences from public health surveillance data provided the primary motivation for the development of our model.

The most germane work is Labgold et al. (2021), which applies Bayesian Improved Surname Geocoding (BISG) to the estimation of race/ethnicity disparities in COVID-19 incidence using data from Fulton County, Georgia. BISG was originally developed in Elliott et al. (2009) for understanding disparities in health outcomes when race data are not available. The approach is an extension of a geocoding model for race, which generates a categorical distribution over race using the location of the unit of analysis (in Labgold et al. the unit of analysis is a case-patient notified of a positive SARS-CoV-2 test). BISG adds surname information to this categorical distribution, with the intention of more accurately imputing race. The weakness of this approach is that the imputation is not informed by the outcome model or vice versa. In the infectious disease context, the information that many cases for which one observes race or ethnicity should inform the categorical distribution for cases missing race information. Labgold et al. addresses this limitation by further modeling the misclassification rate for BISG by comparing BISG's imputed race to that of race for case-patients not missing race. The procedure, however, does not correspond to a probabilistic model, which makes it challenging to validate its implicit assumptions. Furthermore, BISG assumes that the missingness process is missing-at-random, which may not be a good assumption in the context of missing race data. Zhang et al. (2022) also accounts for missing race data in COVID-19 cases via multiple imputation, again assuming MAR missingness.

Despite the wide-ranging literature on analyzing case with missing covariates, a common practice among academic and applied researchers is to omit observations with missing covariates, performing what is known as "complete case analysis", when confronted with missing data (Eekhout et al., 2012). For example, complete case analysis has been used in studies of racial/ethnic disparities of COVID-19 burden when race/ethnicity data are missing (Millett et al., 2020; Zelner et al., 2021) despite the authors' acknowledgment of the risks inherent in dropping incomplete cases. This is an indication of the pervasiveness of the practice, in part due to the ease of performing complete case analysis in most statistical packages (e.g. the ubiquitous `na.rm=TRUE` argument in the R language) and in part due to the lack of methods available to researchers for nonignorable missingness. That complete case analysis is widely employed should galvanize methodologists to develop techniques that are more finely attuned to infectious disease epidemiology.

2.2 Considerations when imputing missing demographic information

Imputing missing demographic data presents multiple challenges at the intersection of ethics, sociology, and statistics (Kennedy et al., 2020). Kennedy et al. note that, beyond the formidable statistical challenges, dealing with missing data of this nature requires understanding how demographic categories in administrative data have changed over time, the relationship between official categories and categories that individuals use to identify themselves, and the fact that attributes like sex, gender, and race/ethnicity must be understood through an intersectional lens rather than as independent dimensions of identity. Furthermore, the authors note that to the extent that there is an imbalance in how groups are misrepresented in surveys, even imputation with uncertainty confers statistical bias and, ultimately, discrimination. The authors argue that despite these realities and being bound by data availability and antiquated study designs, researchers must take responsibility for the choices they make in handling missing data.

Omitting cases that have missing data may be mistaken as a safe practice when missingness of demographic information is assumed - implicitly or explicitly - to occur at random. However, in real-world public health surveillance data, it is unlikely that such information will be missing at random. Instead, there is a high likelihood that the rate of missingness will be correlated with the burden of disease in a community and the resources local authorities have to address it, which are in turn often reflected in race/ethnic disparities in disease outcomes. As this burden increases and the financial and material tools to find new cases dwindle, it becomes increasingly likely that race/ethnic minority groups will be subject to higher rates of missingness which are positively associated with disease risk. This intuition is reflected in our results, which show that when missingness does not occur completely random, dropping cases can result in biased estimates and overstatement of certainty in these estimates.

Furthermore, even when the data are missing completely at random - the most innocuous scenario - random variation in which cases are missing can amount to statistical bias in finite samples. Kennedy et al. note that this can serve as a form of discrimination if the conclusions from the analysis are used to draw inferences about the population and make decisions. This is particularly problematic when addressing missingness for groups that represent a small share of the observed data or overall population: In this case, dropping even a small number of cases missing data can result in diminished power to make valid inferences about group-specific risks. Because of this, making every effort to account for all sources of information on race/ethnicity, even those that are plausibly missing at random, is an ethical imperative.

Exact probabilistic imputation, like the approach we present below, avoids some, but not all, of the risks of wrongly imputing demographic characteristics associated with deterministic ap-

proaches. As Kennedy et al. point out, probabilistic imputation does not guarantee bias-free conclusions, especially if the procedure's misclassification rates are not equally distributed across demographic subgroups. For example, a model that mistakenly assumes that the missingness process is ignorable risks under-representing groups for which the baseline rate of missingness is higher or for whom missingness is positively correlated with the disease process. A well-designed procedure, with results interpreted with awareness of simplifying assumptions and their potential to induce bias, can facilitate increased relevance of study results for underrepresented groups, while also providing a more appropriate representation of uncertainty in these conclusions.

2.3 Roadmap

In this chapter, we present a new joint model that allows researchers to account for relationships between variation in the disease outcome measure of interest and the missingness process. This approach makes the flow of information and model assumptions clear, while at the same time affording researchers all the tools that have been developed to interrogate, summarize and present results from a coherent probabilistic model.

In the following sections, we will 1) Describe the justification for our approach and theoretical properties of the model, 2) Conduct a detailed simulation study to investigate the finite-sample performance of the model under several known data-generating processes, and finally, 3) Apply this model to detailed COVID-19 data from southeastern Michigan.

2.4 Methods

Suppose for each resident, indexed by n, in a large population with size E, the variable U_n is a binary random variable that represents a diagnostic test result (e.g. COVID-19 polymerase chain reaction (PCR) test), C_n is a categorical variable with J levels that encodes race/ethnicity information which may be missing for some residents, R_n is an indicator variable equal to 1 if C_n is observed and 0 otherwise, and S_n is a categorical variable encoding stratum information, like age or sex information for that resident. In other words, each resident is associated with the vector (U_n, C_n, R_n, S_n), of which U_n and R_n are assumed to be random variables, while C_n and S_n are fixed characteristics of each resident.

Let the variable Y_{ij} be the total cases in the population for which $S_n = i$ and $C_n = j$, or more explicitly,

$$Y_{ij} = \sum_{\{n \,|\, S_n = i, C_n = j\}} U_n,$$

and let E_{ij} be the total count of the population in stratum i and race/ethnicity j:

$$E_{ij} = \sum_{n=1}^{E} \mathbb{1}_{S_n=i} \mathbb{1}_{C_n=j}.$$

Let the set of test results U_n for the population be $\mathcal{U} \in \{0,1\}^E$. Define X_{ij} as the number of cases in stratum i for which race/ethnicity C_n is observed to be j and M_i as the number of cases in stratum i as the number of cases missing race/ethnicity information:

$$X_{ij}\,|\,\mathcal{U} = \textstyle\sum_{\{n\,|\,S_n=i,C_n=j,U_n=1\}} R_n \quad \text{and} \quad M_i\,|\,\mathcal{U} = \textstyle\sum_j \sum_{\{n\,|\,S_n=i,C_n=j,U_n=1\}} (1-R_n)$$

Let Y_{ij} be conditionally independent Poisson random variables:[2] $Y_{ij}|\mu_{ij} \sim \text{Poisson}(\mu_{ij})$. where we define *incidence* as

$$\mu_{ij}/E_{ij},$$

or the per-capita rate of disease. We further assume that race/ethnicity observation indicators R_n are conditionally independent Bernoulli distributed random variables with probability observing race/ethnicity information denoted as p_{ij}, which depends solely on stratum i and race/ethnicity category j. Then

$$X_{ij}|\mathcal{U} \overset{d}{=} X_{ij}|Y_{ij}, \implies \quad X_{ij}|Y_{ij}, p_{ij} \sim \text{Binomial}(Y_{ij}, p_{ij})$$

The distributional assumptions imply that marginalizing over total cases of race/ethnicity j in stratum i, Y_{ij}, yields conditionally independent Poisson random variables:

$$X_{ij}|p_{ij}, \mu_{ij} \sim \text{Poisson}(p_{ij}\mu_{ij})$$

for cases of race/ethnicity j observed with race/ethnicity and missing cases are mutually independent of X_{ij} and conditionally independent Poisson random variables:

$$M_i|(p_{i1}, \mu_{i1}), \ldots, (p_{iJ}, \mu_{iJ}) \sim \text{Poisson}(\textstyle\sum_j (1-p_{ij})\mu_{ij})$$

We show the connection between our model and the missing data modeling paradigm introduced by Rubin (1976) and further developed in Little and Rubin (2002) in appendix A.1, and also show that the model implies that the missingness can be not missing at random (NMAR) if p_{ij} vary by j.

Given that $p_{i1} = p_{i2} = \cdots = p_{iJ}$ for all i is a strong constraint, allowing the model to learn the

[2]We discuss the Poisson distribution and the conditional independence assumption in section 2.7.2.1

extent to which probability of observing race/ethnicity varies by race/ethnicity (i.e. allowing the model to learn how far missingness deviates from MAR) is the most judicious modeling choice.

2.4.1 Modeling incidence when missingness is dependent on race/ethnicity

We present a simple example of the model below, in which we assume that race/ethnicity is the only characteristic that predicts both disease incidence and the rate of missingness for individual-level race/ethnicity information. As above, we summarize population counts by age-sex stratum i and race/ethnicity category j:

$$E_{ij} = \sum_{n=1}^{E} \mathbb{1}_{S_n=i} \mathbb{1}_{C_n=j}.$$

Let $\mathbf{e}_i$ be the vector $(E_{i1}, \ldots, E_{iJ})^T$ with the j^{th}-element E_{ij}, and let

$$\mathbf{E} \in \mathbb{R}^{I \times J} \text{ such that } \mathbf{E}_{[i,:]} = \mathbf{e}_i^T.$$

If exposure to the disease is governed solely by race/ethnicity, and infection probability and exposure is constant across age-sex strata i, then we may assume that $\mu_{ij} = \lambda_j E_{ij}$ and that $p_{ij} = p_j, \forall i$. The observed data model, letting $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_J)$ and $\mathbf{p} = (p_1, \ldots, p_J)$, simplifies to the following:

$$\begin{aligned} X_{ij}|p_j, \lambda_j, E_{ij} &\sim \text{Poisson}(p_j \lambda_j E_{ij}), \\ M_i|\mathbf{p}, \boldsymbol{\lambda}, \mathbf{e}_i &\sim \text{Poisson}(\textstyle\sum_j \lambda_j (1 - p_j) E_{ij}); \end{aligned} \tag{2.1}$$

where we have made the conditioning on parameters and population counts explicit.

2.4.1.1 Identifiability properties of the model

We can show that model (2.1) is globally identifiable by appealing to Theorem 4 of Rothenberg (1971). Theorem 1 shows that the model parameters $(\mathbf{p}, \boldsymbol{\lambda})$ are globally identifiable given the observed data under minimal conditions on the parameters, and an easily verifiable condition on the population count matrix $\mathbf{E}$.

Theorem 1. *The observational model* (2.1) *is globally identifiable under the following conditions:*

(E.a) $\mathbf{E}$ *is rank* J*,*

(E.b) $\lambda_j \in (0, \infty) \forall j \in [1, \ldots, J]$*,*

(E.c) $p_j \in (0, 1) \forall j \in [1, \ldots, J]$*.*

Proof. Reparameterize the model from (λ_j, p_j) to (v_j, u_j) where $v_j = p_j\lambda_j$ and $u_j = (1-p_j)\lambda_j$. Given conditions (E.b) to (E.c), the mapping is one-to-one and onto. Let $\mathbf{u}$ be the J-vector with element j equal to u_j and let $\mathbf{v}$ be similarly defined for v_j. The reparameterized model becomes:

$$X_{ij}|v_j, E_{ij} \sim \text{Poisson}(v_j E_{ij}) \tag{2.2}$$

$$M_i|\mathbf{u}, \mathbf{e}_i \sim \text{Poisson}(\textstyle\sum_j u_j E_{ij}) \tag{2.3}$$

We know that $\hat{v}_j = \frac{\sum_i X_{ij}}{\sum_i E_{ij}}$ is unbiased for v_j. Let $\mathbf{m}$ be the I-vector with element i equal to M_i. Then

$$\mathbb{E}[\mathbf{m}] = \mathbf{E}\mathbf{u}$$

By condition (E.a)

$$\left(\mathbf{E}^T\mathbf{E}\right)^{-1}\mathbf{E}^T\mathbb{E}[\mathbf{m}] = \mathbf{u}. \tag{2.4}$$

Given that we can define unbiased estimators for $\mathbf{v}$ and $\mathbf{u}$, by Theorem 4 in Rothenberg (1971), the model is globally identifiable in (v_j, u_j). Given that our mapping from (λ_j, p_j) to (v_j, u_j) is one-to-one and onto, global identifiability in (v_j, u_j) implies global identifiability in (λ_j, p_j) because we can define an inverse mapping from (v_j, u_j) to (λ_j, p_j). □

It can also be seen that the variance-covariance matrix for the estimator for $\hat{\boldsymbol{\lambda}}$ is a sum of two components: the variance of $\hat{\mathbf{v}}$ and the variance-covariance matrix of the unbiased linear estimator for $\mathbf{u}$, $\left(\mathbf{E}^T\mathbf{E}\right)^{-1}\mathbf{E}^T\mathbf{m}$. This coincides with the Fisher information matrix, as derived in Appendix Section A.4.1, where we also show that the Fisher information matrix is positive definite under conditions (E.a) to (E.c).

2.4.1.2 Model intuition

We examine a simple setting in which there are two race/ethnicity groups (or equivalently, $J = 2$) subject to missingness. The unbiased estimator, $\hat{u}_1$, for $u_1 = (1-p_1)\lambda_1$ can be expressed in terms of a projection matrix:

$$\mathbf{P}_2 = \mathbf{e}_2(\mathbf{e}_2^T\mathbf{e}_2)^{-1}\mathbf{e}_2^T$$

which is the projection for a vector in $\mathbb{R}^I$ to the subspace spanned by $\mathbf{e}_2$. Then

$$\hat{u}_1 = \frac{\mathbf{m}^T(\mathbf{I}-\mathbf{P}_2)\mathbf{e}_1}{\|(\mathbf{I}-\mathbf{P}_2)\mathbf{e}_1\|_2^2}. \tag{2.5}$$

which can be understood as a relative measure of the strength of the covariance between the number of cases with missing race/ethnicity information and population counts for group 1 after accounting

for the variation in population counts attributable to group 2. The following estimator is unbiased for λ_1:

$$\frac{\sum_{i=1}^{I} X_{i1}}{\sum_{i=1}^{I} E_{i1}} + \hat{u}_1. \tag{2.6}$$

The first term in eq. (2.6) is the estimator for $\lambda_1 p_1$ for a Poisson distribution with rate $\lambda_1 p_1$, while the second term is a correction to account for missingness. If the covariance of $\mathbf{m}$ and a residualized $\mathbf{e}_1$ (by regressing $\mathbf{e}_1$ on $\mathbf{e}_2$) is large relative to the variance of the residualized $\mathbf{e}_1$ then the correction will be large. If, on the contrary, this quantity is small, the correction factor will be small. The estimator depends on the conditional expectation of the number of cases for each race/ethnicity being proportional to the number of residents in that category, which is a common assumption in modeling count data in epidemiology (Frome, 1983; Frome and Checkoway, 1985; Lash et al., 2021).

2.4.1.3 Bayesian inference and prior sensitivity

The two group setting in which one group's population is small compared to the other group's population motivates the careful choice of priors when doing Bayesian inference. We will show that in this setting the posterior mean for the rate of disease in the minority group is sensitive to priors. As above, let $j \in \{1, 2\}$ and let the rate of disease in group j be λ_j while the probability of observing race/ethnicity j is p_j. Under the following priors:

$$\begin{aligned} p_j &\overset{\text{iid}}{\sim} \text{Beta}(\alpha_j, \beta_j) \\ \lambda_j &\overset{\text{iid}}{\sim} \text{Gamma}(\alpha_j + \beta_j, r_j), \end{aligned} \tag{2.7}$$

$v_j = p_j \lambda_j \perp\!\!\!\perp u_j = (1 - p_j)\lambda_j$. As shown in theorem 1, we can write the observed-data likelihood in terms of u_j and v_j. If we assume, without loss of generality, that the majority group is 2 for all i (i.e. that $E_{i1} \ll E_{i2}$ for all i) we can make a likelihood approximation, detailed in Appendix Section A.4.2, that allows us to compute a closed-form approximate posterior mean and variance for λ_1, as well as the partial derivative of the posterior mean with respect to the prior rate parameter, r_1, for λ_1 when $\beta_1 = 1$, which is the second shape parameter for the beta prior over p_1 as well as part of the shape parameter for λ_1.

Let $s_1 = \sum_i \frac{m_i E_{i1}}{E_{i2}}$, $s_2 = \sum_i \frac{m_i E_{i1}^2}{E_{i2}^2}$, and $E_{+1} = \sum_i E_{i1}$. Further, let ϕ and Φ be the density and distribution function of the standard normal distribution, respectively, and $z = \frac{s_1 - u_2(r_1 + E_{+1})}{\sqrt{s_2}}$. If

$\beta_1 = 1$, the posterior mean for λ_1 given u_2 is then

$$\mathbb{E}\left[\lambda_1 | u_2, r_1, \beta_1 = 1\right] = \frac{\alpha_1 + \sum_i x_{i1}}{r_1 + E_{+1}} + \frac{s_1 - u_2(r_1 + E_{+1})}{s_2/u_2} + \frac{u_2}{\sqrt{s_2}}\phi(z)\Phi(z)^{-1} \tag{2.8}$$

with variance:

$$\text{Var}\left(\lambda_1 | u_2, r_1, \beta_1 = 1\right) = \frac{\alpha_1 + \sum_i x_{i1}}{(r_1 + \sum_i E_{i1})^2} + \frac{u_2^2}{s_2}\left(1 - z\phi(z)\Phi(z)^{-1} - \phi(z)^2\Phi(z)^{-2}\right) \tag{2.9}$$

Like our unbiased estimator for λ_1 in eq. (2.6), the first term in eq. (2.8) is the posterior mean for the rate of a Poisson random variable with rate $\lambda_1 p_1$, while the second term is the correction for missing data. The first term of the correction, $\frac{s_1}{s_2/u_2}$[3] can be seen as an approximate least squares estimator for u_1 scaled by the weighted average of m_i by dividing top and bottom by $\sum_i \frac{E_{i1}^2}{E_{i2}^2}$. The estimate is shrunk towards zero with magnitude dependent on u_2 and r_1. In fact, for an increase in u_2 the posterior conditional mean is shrunk towards zero, while an increase in r_1 similarly shrinks the posterior mean towards zero. This agrees with intuition that as the prior rate parameter for λ_1 increases, the prior mean decreases, and so too does the posterior mean. This can be seen from the partial derivative of the posterior mean with respect to r_1, which we show in appendix A.4.2 to be $\frac{\partial \mathbb{E}[\lambda_1 | u_2, r_1, \beta_1 = 1]}{\partial r_1} = -\text{Var}\left(\lambda_1 | u_2, r_1, \beta_1 = 1\right)$. The magnitude of the derivative is equal to that of the variance, eq. (2.9), which implies that the posterior mean is sensitive to r_1. This sensitivity does not decline as $E_{i1}, E_{i2} \to \infty$ such that group 1 remains a minority to group 2. Suppose that we take $E_{i1}, E_{i2} \to \infty$ such that $\frac{E_{i1}}{E_{i2}} = O(\frac{1}{E_{i1}})$. Then $\frac{E_{i1}^2}{E_{i2}} \to K < \infty$ for all i. Let $u_2^\star$, and $u_1^\star$ be the true data generating parameters, and let $z^\star = (u_1^\star IK - u_2^\star r_1)/\sqrt{u_2^\star IK} + \mathcal{Z}$, where $\mathcal{Z} \sim \text{N}(0, 1)$. The posterior mean and variance for λ_1 have the following convergence in distribution as E_{i1}, E_{i2} goes to infinity in the same order:

$$\mathbb{E}\left[\lambda_1 | u_2, r_1, \beta_1 = 1\right] \xrightarrow{d} v_1^\star + \frac{u_1^\star IK - r_1 u_2^\star}{IK} + \frac{\sqrt{u_2^\star}\phi(z^\star)\Phi(z^\star)^{-1}}{\sqrt{IK}} + \sqrt{\frac{u_2^\star}{IK}}\mathcal{Z},$$

$$\text{Var}\left(\lambda_1 | u_2, r_1, \beta_1 = 1\right) \xrightarrow{d} \frac{u_2^\star}{IK}\left(1 - z^\star\phi(z^\star)\Phi(z^\star)^{-1} - \phi(z^\star)^2\Phi(z^\star)^{-2}\right).$$

We can see that as $I \to \infty$, the posterior mean for λ_1 converges in probability to $v^\star + u^\star$, or the true data generating parameter, as we would expect for a globally identifiable model. However, for fixed I, the posterior mean remains both dependent on $u_2^\star$ and r_1, and, moreover, the derivative of the posterior mean with respect to r_1 can be seen to remain bounded away from zero and of the same magnitude as the posterior variance.

[3] $\frac{s_1}{I} = \sum_i \frac{m_i E_{i1}}{E_{i2}}/I$ is an approximate empirical covariance between the vector $\mathbf{m}$ and a vector with elements $\frac{E_{i1}}{E_{i2}}$ because $\frac{E_{i1}}{E_{i2}} \to 0$.

Nor does the sensitivity of the posterior mean for λ_1 to changes in β_1 diminish. We can calculate the limit for the expression $(\mathbb{E}[\lambda_1|u_2, r_1, \beta_1 = 2] - \mathbb{E}[\lambda_1|u_2, r_1, \beta_1 = 1])/\sqrt{\text{Var}(\lambda_1|u_2, r_1, \beta_1 = 1)}$, or the change in posterior mean scaled by the posterior standard deviation. The form is shown in the appendix to be

$$\frac{\sqrt{1 - z^\star\phi(z^\star)\Phi(z^\star)^{-1} - \phi(z^\star)^2\Phi(z^\star)^{-2}}}{z^\star + \phi(z^\star)\Phi(z^\star)^{-1}}.$$

which approaches 1 as $\mathbb{E}[z^\star] \to -\infty$.

This analysis implies that posterior inferences can be sensitive to both prior hyperparameters β_j and r_j for minority groups and can behave like a partially identified model asymptotically (Gustafson, 2015). In terms of classical approaches to prior sample size, like Gelman et al. (2013), a change from $p_j \sim \text{Beta}(1, 1)$ to $p_j \sim \text{Beta}(1, 2)$, or from $\lambda_j \sim \text{Gamma}(2, 100)$ to $\lambda_j \sim \text{Gamma}(3, 150)$ represents a small increase in prior information, but this can translate to large changes in the posterior mean for λ_j for small minority populations.

Despite this prior sensitivity, we show in Figure A.1 in Appendix A.4.2 that for reasonable values of cumulative incidence and race/ethnicity reporting rate for the majority group, the posterior mean dominates the maximum likelihood estimator in terms of root mean squared error for a large range of $u_1^\star$. Moreover, the asymptotic MLE for $u_1^\star$ has a non-negligible probability of being zero, whereas the posterior mean is almost surely positive. These results demonstrate the benefits of using Bayesian inference over classical maximum likelihood estimation.

2.4.2 Modeling incidence when missingness is dependent on both age-sex and race/ethnicity

Now assume that the rate of incident cases for age-sex stratum i in race/ethnicity group j, or observation (i, j), depends on fully-observed covariates, $\mathbf{z}_i \in \mathbb{R}^K$, associated with each stratum i. In the context of COVID-19, we expect that age-sex stratum will predict both exposure and probability of infection and disease given exposure, as well as the probability of race/ethnicity being recorded, so it is important to extend our model to incorporate this information. We assume that coefficients for $\mathbf{z}_i$, $\boldsymbol{\beta}$ for incidence and $\boldsymbol{\gamma}$ for race/ethnicity missingness, both in $\mathbb{R}^K$, are shared between race/ethnicity groups, which amounts to assuming there is no interaction between race and age-sex strata for predicting incidence and missingness. As above, we allow average incidence, λ_j, and log-odds of observing race/ethnicity or η_j to vary by group j. Let $\mathbf{p}_i$ be the length-J vector

with j^{th} element equal to p_{ij}. Then we can define the following observed-data model:

$$\begin{aligned} X_{ij}|\lambda_j, \mathbf{z}_i, \boldsymbol{\beta}, p_{ij}, E_{ij} &\sim \text{Poisson}(p_{ij}\lambda_j \exp(\mathbf{z}_i^T\boldsymbol{\beta})E_{ij}), \\ M_i|\boldsymbol{\lambda}, \mathbf{z}_i, \beta, \mathbf{p}_i, \mathbf{e}_i &\sim \text{Poisson}(\exp(\mathbf{z}_i^T\boldsymbol{\beta})\textstyle\sum_j(1-p_{ij})\lambda_j E_{ij}), \\ p_{ij} &= \left(1+\exp(-(\mathbf{z}_i^T\boldsymbol{\gamma}+\eta_j))\right)^{-1}; \end{aligned} \tag{2.10}$$

where the X_{ij} are independent of M_i since the missingness process is conditionally independent of the disease process. See A.3.1 for a graphical depiction of the model.

Theorem 2. *Let the model be defined as in* (2.10) *and let* $\mathbf{E}$ *be the* I *by* J *matrix where the* i*-th row is* $\mathbf{e}_i = (E_{i1}, E_{i2}, \ldots, E_{iJ})^T$*, and let* $\mathbf{Z}$ *be the* I *by* K *matrix where the* i*-th row is* $\mathbf{z}_i$*. If all of the following conditions hold:*

(S.a) $\mathbf{E}$ *is rank* J

(S.b) $\mathbf{Z}$ *is rank* K

(S.c) $I \geq J + K$

(S.d) $p_j \in (0,1)$ *for all* j

(S.e) $\lambda_j \in (0,\infty)$ *for all* j

(S.f) $e^{\mathbf{z}_i\boldsymbol{\beta}}\sum_j^J E_{ij} \in (0,\infty)$ *for all* i

(S.g) $\text{rank}\left(\left[\text{diag}(\mathbf{E}_{[:,1]})\mathbf{Z} \quad \ldots \quad \text{diag}(\mathbf{E}_{[:,J]})\mathbf{Z} \quad \mathbf{E}_{[:,1]} \quad \mathbf{E}_{[:,2]} \quad \ldots \quad \mathbf{E}_{[:,J]}\right]\right) > J + K$

the model is locally identifiable.

The proof is in Appendix A.5 and depends on showing that the model's Fisher information matrix $\mathcal{I}$ is positive definite. We use a technique employed in Mukerjee and Sutradhar (2002), which establishes a lower bound for the positive definiteness of the Fisher Information matrix via a method of moments estimator. The idea rests on the derivation of the multivariate Cramér-Rao lower bound in Rao (2002). This partially establishes that the model is regular and shows that the model is locally identifiable (Watanabe, 2009; Rothenberg, 1971).

Given section 2.4.1.3, it is important to use prior information for minority groups when possible. To that end, the following priors can be employed:

$$\begin{aligned} \lambda_j &\sim \text{LogNormal}(\mu_{\lambda_j}, s_\lambda^2) \quad \forall j \in [1,\ldots,J], \\ \eta_j &\sim \text{Normal}(\mu_{\eta_j}, s_\eta^2) \quad \forall j \in [1,\ldots,J], \\ \boldsymbol{\beta} &\sim \text{MultiNormal}(\boldsymbol{\mu}_\beta, \Sigma_\beta) \\ \boldsymbol{\gamma} &\sim \text{MultiNormal}(\boldsymbol{\mu}_\gamma, \Sigma_\gamma) \end{aligned}$$

where $\mu_{\lambda_j}, \mu_{\eta_j}, s_\lambda, s_\eta, \boldsymbol{\mu}_\beta, \Sigma_\beta, \boldsymbol{\mu}_\gamma$ and Σ_γ are known hyperparameters.

2.4.3 Modeling geographic heterogeneity in incidence and missingness

Suppose the case data are observed for more than one geographical area so we have an additional fixed categorical variable $L_n \in \{1, \ldots, G\}$ encoding the geographic area to which each resident in the population is associated. We may expect that geographical heterogeneity in incidence and race/ethnicity missingness exists between areas. For instance, with respect to the COVID-19 pandemic, we might want to allow for geographic heterogeneity in population substrata incidence and missingness because we expect that areas have different contact patterns. We can then further stratify the observations by area g as:

$$X_{igj} \,|\mathcal{U} = \textstyle\sum_{\{n \,|\, S_n=i, L_n=g, C_n=j, U_n=1\}} R_n \,,\; M_{ig} \,|\mathcal{U} = \sum_j \sum_{\{n \,|\, S_n=i, L_n=g, C_n=j, U_n=1\}} (1 - R_n),$$

and we can tabulate population counts E_{igj} as $E_{igj} = \sum_{n=1}^{E} \mathbb{1}_{S_n=i} \mathbb{1}_{C_n=j} \mathbb{1}_{L_n=g}$.

Model eq. (2.10) naturally extends to incorporate this structure. Let $\mathbf{e}_{ig}$ be the J-vector with j-th element E_{igj} for age-sex stratum i and geographic area g. Similarly define the proportions of cases in stratum i and geographic area g with observed race/ethnic information as $\mathbf{p}_{ig}$, where $\mathbf{p}_{ig}$ is the J-vector with j-th element p_{igj}. We let the covariates for stratum i vary by area g, $\mathbf{z}_{ig}$, vary by area g, and we also let the coefficients vary by g, so $\boldsymbol{\beta}_g, \boldsymbol{\gamma}_g \in \mathbb{R}^K$. Let $\boldsymbol{\lambda}_g$ be the J-vector with j-th element λ_{gj}. The observed-data model becomes

$$\begin{aligned} X_{igj} | \lambda_{gj}, \mathbf{z}_{ig}, \boldsymbol{\beta}_g, p_{igj}, E_{igj} &\sim \text{Poisson}(p_{igj} \lambda_{gj} \exp(\mathbf{z}_{ig}^T \boldsymbol{\beta}_g) E_{igj}), \\ M_{ig} | \boldsymbol{\lambda}_g, \mathbf{z}_{ig}, \boldsymbol{\beta}_g, \mathbf{p}_{ig}, \mathbf{e}_{ig} &\sim \text{Poisson}(\exp(\mathbf{z}_{ig}^T \boldsymbol{\beta}_g) \textstyle\sum_j ((1 - p_{igj}) \lambda_{gj} E_{igj}), \\ p_{igj} &= \left(1 + \exp(-(\mathbf{z}_{ig}^T \boldsymbol{\gamma}_g + \eta_{gj}))\right)^{-1}. \end{aligned} \tag{2.11}$$

See Appendix A.3.2 for a graphical depiction of the model with a table of model parameters. We can draw on the results from 2 to characterize the local identifiability of 2.11. By 2, within a geographic region g, the parameter set

$$\boldsymbol{\theta}_g = \{\boldsymbol{\lambda}_g, \boldsymbol{\eta}_g, \boldsymbol{\beta}_g, \boldsymbol{\gamma}_g\}$$

is locally identifiable provided the conditions in 2 hold. However, when data are sparse, either because there is low incidence within an area or because there is a small minority group in geographic region g, we would like to shrink our estimates for $\boldsymbol{\theta}_g$ to the global mean. Ideally we would learn the degree of shrinkage for each dimension of $\boldsymbol{\theta}_g$. This motivates a hierarchical prior for elements of $\boldsymbol{\theta}_g$.

2.4.3.1 Hierarchical priors

To that end, we may wish to incorporate area-level covariates, represented by a D-length vector $\mathbf{w}_g$, into the model for $\boldsymbol{\theta}_g$. Let $\boldsymbol{\Pi}_{\boldsymbol{\lambda}}, \boldsymbol{\Pi}_{\boldsymbol{\eta}}$ be in $\mathbb{R}^{J\times D}$ and let $\boldsymbol{\Pi}_{\boldsymbol{\beta}}, \boldsymbol{\Pi}_{\boldsymbol{\gamma}}$ be in $\mathbb{R}^{K\times D}$. A suitable model for the elements of $\boldsymbol{\theta}_g$ is:

$$\begin{aligned}
\log(\boldsymbol{\lambda}_g) &\sim \text{MultiNormal}\left(\boldsymbol{\alpha}_{\boldsymbol{\lambda}} + \boldsymbol{\Pi}_{\boldsymbol{\lambda}}\mathbf{w}_g, \boldsymbol{\Sigma}_{\boldsymbol{\lambda}}\right) \\
\boldsymbol{\eta}_g &\sim \text{MultiNormal}\left(\boldsymbol{\alpha}_{\boldsymbol{\eta}} + \boldsymbol{\Pi}_{\boldsymbol{\eta}}\mathbf{w}_g, \boldsymbol{\Sigma}_{\boldsymbol{\eta}}\right) \\
\boldsymbol{\beta}_g &\sim \text{MultiNormal}\left(\boldsymbol{\alpha}_{\boldsymbol{\beta}} + \boldsymbol{\Pi}_{\boldsymbol{\beta}}\mathbf{w}_g, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}\right) \\
\boldsymbol{\gamma}_g &\sim \text{MultiNormal}\left(\boldsymbol{\alpha}_{\boldsymbol{\gamma}} + \boldsymbol{\Pi}_{\boldsymbol{\gamma}}\mathbf{w}_g, \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}\right)
\end{aligned} \tag{2.12}$$

For a more detailed picture of how these parameters connect to eq. (2.11), see appendix A.3.2. Let the operation $\text{vec}(\mathbf{A}) : \mathbb{R}^{M\times N} \to \mathbb{R}^{MN}$ via appending the N M-length columns of $\mathbf{A}$ into an NM-length vector. Then the vector of unknown hyperparameters can be represented as

$$\begin{aligned}
\boldsymbol{\phi} = (&\text{vec}(\boldsymbol{\Pi}_{\boldsymbol{\lambda}}), \text{vec}(\boldsymbol{\Pi}_{\boldsymbol{\eta}}), \text{vec}(\boldsymbol{\Pi}_{\boldsymbol{\beta}}), \text{vec}(\boldsymbol{\Pi}_{\boldsymbol{\gamma}}), \\
&\text{vec}(\boldsymbol{\Sigma}_{\boldsymbol{\lambda}}), \text{vec}(\boldsymbol{\Sigma}_{\boldsymbol{\eta}}), \text{vec}(\boldsymbol{\Sigma}_{\boldsymbol{\beta}}), \text{vec}(\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}), \\
&\boldsymbol{\alpha}_{\boldsymbol{\lambda}}, \boldsymbol{\alpha}_{\boldsymbol{\eta}}, \boldsymbol{\alpha}_{\boldsymbol{\beta}}, \boldsymbol{\alpha}_{\boldsymbol{\gamma}})
\end{aligned}$$

We can encode our prior knowledge about the geographic heterogeneity of parameters into a joint prior over $\boldsymbol{\phi}$.

While the hierarchical prior in eq. (2.12) does not correspond to the set of priors in eq. (2.7) the results in section 2.4.1.3 suggest that posterior inferences for incidence parameters for areas with small minority groups relative to the majority groups can be sensitive to the priors over $\boldsymbol{\Sigma}_{\boldsymbol{\lambda}}, \boldsymbol{\Sigma}_{\boldsymbol{\eta}}$, and $\boldsymbol{\alpha}_{\boldsymbol{\eta}}$.

Figure A.1 in Appendix A.4.2 shows the large-population RMSE for the incidence of minority race/ethnicity cases that are missing race/ethnicity information, or u_1, under different prior scenarios. The RMSE of the posterior mean estimators are minimized when the prior mean for u_1 is close to the true parameter, when the prior for u_1 excludes prior mass near zero and the prior mean underestimates the true parameter, or when the prior mean slightly overestimates the true parameter and the prior for u_1 does not put substantial prior mass near zero. The near-zero prior behavior for u_1 can be translated to priors on $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$, and $\boldsymbol{\alpha}_{\boldsymbol{\eta}}$. By limiting the amount of prior mass in the right tail of the distribution for $\boldsymbol{\alpha}_{\boldsymbol{\eta}}$ one can limit the amount of prior mass near zero for u_1; a normal distribution with substantial mass below 5 would suffice. The prior over $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$ will also affect the tails of the marginal prior for geographic-specific parameters, and can also adversely affect shrinkage. If one uses a prior over population standard deviation with heavy tails, like a half-Cauchy, then the marginal prior for a geographic specific parameter will have substantial prior mass near zero.

If, instead, the prior over the population standard deviation hews too closely to zero, like a half-normal with a standard deviation of 0.1, then the prior will shrink geographic-specific parameters too strongly towards the overall mean. Similar considerations about shrinkage should guide priors over $\boldsymbol{\Sigma}_{\boldsymbol{\lambda}}$. For more information on techniques for prior formulation in Bayesian models see Gabry et al. (2019a); Gelman et al. (2017).

See Section 2.5.7 for more information on prior specification for population parameters.

2.4.4 Inference

We perform Bayesian inference in Stan (Team, 2021). Stan is at once a domain-specific modeling language and a suite of inference algorithms, including dynamic Hamiltonian Monte Carlo (HMC), a descendant of the No-U-Turn-Sampler (Betancourt, 2018; Hoffman et al., 2014). Stan's implementation of dynamic HMC adaptively sets the algorithm's tuning parameters (e.g. leapfrog integrator stepsize and mass matrix) during warmup iterations, which makes the sampler robust to many difficult-to-sample posteriors, such as those that arise from fitting hierarchical models like model (2.11) (Betancourt and Girolami, 2015).

We use Stan for inference because we are able to exactly marginalize over the discrete unknown cases as shown in appendix A.2. While Stan does not directly allow inference over discrete parameters, as long as the target density can be expressed as a marginalization over the discrete unknowns, Stan can sample from the posterior over the continuous parameter space and subsequently draw discrete random variables conditional on the draws of the continuous parameters.

2.5 Simulation study

In this section, we present a simulation study designed to quantify the finite-sample properties of our model under varying degrees of missingness, as well as to compare the model's performance to alternative methods of inference commonly applied to datasets with missing covariates. We chose complete-case analysis, and two different multiple imputation approaches as the comparison methods because of their prevalence among researchers. The simulation study clarifies the potential pitfalls of using such methods when analyzing data with missing covariates.

2.5.1 Population data

In our simulation study, we drew on georeferenced population data from Wayne County, Michigan, which encompasses the City of Detroit and its surrounding suburbs. The geographical areas of analysis were Public Use Microdata Areas (PUMAs), which are administrative areas defined by

the Census Bureau such that they comprise at least $100,000$ people. We aggregated Census-tract-level data from IPUMS National Historic Geographic Information System into PUMA-level counts (Manson et al., 2021). In Wayne County, there are 13 PUMAs nested within the county borders. Within each PUMA, we stratified the population by age and sex, with age in years binned in 10-year right-open intervals between 0 and 80: $[0, 10), [10, 20), \ldots, [70, 80)$ and used a single group to capture those 80 and older. We used the 2010 Decennial Census population counts as E_{ij} for each PUMA. The use of U.S. Census data constrains our race and ethnicity classification because Hispanic/Latino ethnicity is treated as mutually exclusive with race. This prevents a more nuanced modeling of a separate effects of ethnicity and race. Despite these limitations, for our simulation study we used the Census classifications to bucket the population into four groups: Black, Hispanic/Latino, Other, and White. The Black and White categories comprised people who

Table 2.1: Population summary in Wayne County, Michigan as of the 2010 Decennial Census

Race/Ethnicity	Total Pop.	Mean Age×Sex×Race/Eth. ×PUMA Pop.	Std. dev. PUMA Pop.	100× Ratio to White
Black	732801	3132	3152	81
Hispanic/Latino	95260	407	757	11
Other	90343	386	397	10
White	902180	3855	3225	100

identified as Black or White alone and not Hispanic or Latino, while the Hispanic/Latino category included anyone who identified as Hispanic and Latino. The Other category included Asians and Pacific Islanders, Native Americans and Alaska Natives, mixed race individuals, as well as people of Other races, all of whom did not identify as Hispanic or Latino. From table 2.1 we can see that in Wayne the majority of the population is White, though the Black population is of a similar order of magnitude. Hispanic/Latino people and people classified as Other are around 10% of the White population.

2.5.2 Data generating process

We simulated age-sex-stratum-specific incident cases of disease by PUMA from model 2.11, with fixed hyperparameters $\boldsymbol{\phi}$ under four scenarios that varied the proportion of cases that had fully-observed covariates: 90%, 80%, 60%, and 20%. The data were generated with two effects for sex, and nine effects for age, both with a sum to zero constraint in both the Poisson log-rate parameter and the Bernoulli log-odds parameter. More explicitly, the $\boldsymbol{\beta}_g$ parameter was decomposed into β^{sex}, and $\boldsymbol{\beta}^{\text{age}}$; $\boldsymbol{\gamma}_g$, $\boldsymbol{\alpha}_\beta$, and $\boldsymbol{\alpha}_\lambda$ were similarly decomposed. For the simulated datasets, the Poisson log-rate parameters for $\boldsymbol{\alpha}_\beta^{\text{age}}$ were fixed at values that mimicked the age pattern of relative risk of

COVID-19 cumulative incidence in the first stage of the pandemic, roughly between March 1^{st}, 2020 and July 1^{st}, 2020. The relative risk of COVID-19 for younger people was much lower compared to that of older people, especially those over 60, and we set the values of $\boldsymbol{\alpha}_{\beta}^{\text{age}}$ accordingly: $(-2.5, -2.0, 0.0, 0.0, 0.5, 0.5, 1.0, 1.0, 1.5)$ (Zelner et al., 2021). For the age pattern of the log-odds of missingness, older individuals were more likely to have race reported compared to younger ages and was thus reflected in our values for $\boldsymbol{\alpha}_{\gamma}^{\text{age}} = (-0.3, -0.3, -0.2, -0.2, -0.2, -0.1, 0.1, 0.4, 0.8)$.

In order to investigate hyperparameter inference as well as other functions of the parameters of epidemiological interest (like cumulative incidence per group at the county level or age-sex-standardized incidence) for majority and minority groups that was solely a function of missingness and not of rates of disease, we set each group's average log-rate of disease, or the elements of $\boldsymbol{\alpha}_{\lambda}$, to -4 for all simulations. We then set $\boldsymbol{\alpha}_{\eta}$, the group-wise population log-odds of observing race, to vary between scenarios according to the average proportion of cases observed with race. In order to set proportions of fully-observed cases for each race/ethnicity, we set ratios of the proportions relative to that of Whites and then varied the White proportion such that the population weighted average rate of cases with fully-observed covariates matched the population target rates of $90\%, 80\%, 60\%, 20\%$. Blacks-to-Whites was set to $\frac{0.75}{0.9}$, Hispanic/Latinos was set to 1, Other was set to $\frac{0.6}{0.9}$. The generative model for the geography-specific parameters is:

$$\begin{aligned}
\log \boldsymbol{\lambda}_g &\sim \text{MultiNormal}(\boldsymbol{\alpha}_{\lambda}, \text{diag}(\boldsymbol{\sigma}_{\lambda})) \\
\boldsymbol{\eta}_g &\sim \text{MultiNormal}(\boldsymbol{\alpha}_{\eta}, \text{diag}(\boldsymbol{\sigma}_{\eta})) \\
\boldsymbol{\beta}_g &\sim \text{MultiNormal}(\boldsymbol{\alpha}_{\beta}, \text{diag}(\boldsymbol{\sigma}_{\beta})) \\
\boldsymbol{\gamma}_g &\sim \text{MultiNormal}(\boldsymbol{\alpha}_{\gamma}, \text{diag}(\boldsymbol{\sigma}_{\gamma}))
\end{aligned} \tag{2.13}$$

with all elements of $\boldsymbol{\sigma}_{\lambda}$ and $\boldsymbol{\sigma}_{\beta}$ equal to 0.5 and all elements of $\boldsymbol{\sigma}_{\eta}$ and $\boldsymbol{\sigma}_{\gamma}$ equal to 0.3,. The elements of the hierarchical scale parameters related to cumulative disease incidence, $\boldsymbol{\sigma}_{\lambda}$ and $\boldsymbol{\sigma}_{\beta}$, were set to larger values than the parameters related to the missingness process, $\boldsymbol{\sigma}_{\eta}$ and $\boldsymbol{\sigma}_{\gamma}$, to reflect the fact that missingness of race data in Wayne County in the first wave of the pandemic was driven by local-level patient non-response and county-wide lab processing issues, while cumulative incidence was driven largely by local transmission.

Summaries of the simulated datasets are shown in Table 2.2. The differences between race in the true cumulative incidence were driven solely by the difference in age distributions between races within Wayne County. The table highlights the fact that, excluding random variation, the scenarios differ only in the observed incidence, as the disease process model as represented via hyperparameters $\boldsymbol{\alpha}_{\lambda}$ and $\boldsymbol{\alpha}_{\beta}$ remains fixed between scenarios. The variance in incidence was a function of the variance of the realizations of the geography-specific parameters $\boldsymbol{\lambda}_g$ and $\boldsymbol{\beta}_g$ driven by the population scale parameters $\boldsymbol{\sigma}_{\lambda}$ and $\boldsymbol{\sigma}_{\beta}$.

Table 2.2: The table summarizes the simulation study by missingness scenario by race/ethnicity. 200 datasets were simulated in each scenario. The column "Mean Obs." gives the average proportion of cases observed with race/ethnicity data across 200 simulated datasets. Similarly, "Mean True Inc." is the mean true incidence by group, and "Mean Obs. Inc." is the mean observed incidence by group.

Proportion cases w/ race/ethnicity	Race/Ethnicity	Mean Obs.	Std. dev.	Mean True Inc.	Std. dev.	Mean Obs. Inc.	Std. dev.
90%	Black	80.7%	(2.4%)	3.4%	(0.9%)	2.8%	(0.8%)
	Hispanic/Latino	96.7%	(0.7%)	2.4%	(0.7%)	2.3%	(0.7%)
	Other	63.9%	(3.0%)	2.6%	(0.6%)	1.7%	(0.4%)
	White	97.1%	(0.5%)	4.4%	(1.8%)	4.3%	(1.7%)
80%	Black	72.7%	(3.0%)	3.4%	(1.0%)	2.5%	(0.8%)
	Hispanic/Latino	85.0%	(2.4%)	2.4%	(0.6%)	2.1%	(0.5%)
	Other	57.4%	(3.2%)	2.6%	(0.6%)	1.5%	(0.3%)
	White	86.5%	(2.1%)	4.2%	(1.2%)	3.7%	(1.0%)
60%	Black	53.7%	(4.6%)	3.5%	(1.1%)	1.9%	(0.7%)
	Hispanic/Latino	60.3%	(4.3%)	2.4%	(0.6%)	1.5%	(0.4%)
	Other	42.3%	(3.7%)	2.6%	(0.5%)	1.1%	(0.3%)
	White	64.4%	(4.4%)	4.3%	(1.3%)	2.8%	(1.0%)
20%	Black	17.2%	(3.5%)	3.4%	(0.8%)	0.6%	(0.2%)
	Hispanic/Latino	18.4%	(3.2%)	2.4%	(0.7%)	0.4%	(0.2%)
	Other	12.9%	(2.3%)	2.7%	(0.5%)	0.4%	(0.1%)
	White	21.7%	(4.5%)	4.4%	(1.5%)	1.0%	(0.6%)

2.5.3 Inferential models

We fitted four inferential models to the simulated datasets: model (2.11), which we will refer to as the "joint" model, the "complete case" model, defined in Equation (2.15), in which cases with missing race/ethnicity are dropped, and two "multiple imputation" models in which we impute the missing/race ethnicity cases and subsequently fit the complete case model to the generated datasets. The hierarchical prior structure of the joint model matched that of the data generating model in equation 2.13, with priors over the hyperparameters:

$$\begin{aligned}
\boldsymbol{\alpha}_\lambda &\sim \text{MultiNormal}(-\mathbf{5}, \text{diag}(\mathbf{1})) &, \boldsymbol{\sigma}_\lambda &\sim \text{MultiNormal}^+(\mathbf{0}, \text{diag}(\mathbf{1})) \\
\boldsymbol{\alpha}_\eta &\sim \text{MultiNormal}(\mathbf{2}, \text{diag}(\mathbf{1})) &, \boldsymbol{\sigma}_\eta &\sim \text{MultiNormal}^+(\mathbf{0}, \text{diag}(\mathbf{1})) \\
\boldsymbol{\alpha}_\beta &\sim \text{MultiNormal}(\mathbf{0}, \text{diag}(\mathbf{1})) &, \boldsymbol{\sigma}_\beta &\sim \text{MultiNormal}^+(\mathbf{0}, \text{diag}(\mathbf{1})) \\
\boldsymbol{\alpha}_\gamma &\sim \text{MultiNormal}(\mathbf{0}, \text{diag}(\mathbf{1})) &, \boldsymbol{\sigma}_\gamma &\sim \text{MultiNormal}^+(\mathbf{0}, \text{diag}(\mathbf{0.25}))
\end{aligned} \tag{2.14}$$

A noteworthy characteristic of the priors for the hyperparameters is that the priors over $\boldsymbol{\alpha}_\lambda$ and $\boldsymbol{\alpha}_\eta$ were misspecified compared to the data-generating parameters. The true data-generating parameters fell one prior standard deviation above the prior means for $\boldsymbol{\alpha}_\lambda$, while the prior mean for $\boldsymbol{\alpha}_\eta$, which did not vary by scenario, was too large by 4 prior standard deviations in the 20% observed scenario and was too small by 1.5 standard deviations in the 90% observed scenario. This allowed us to examine the joint model's finite-sample properties for large groups and smaller groups.

2.5.3.1 Complete case model definition

The complete case model is

$$\begin{aligned} X_{igj}|\lambda_{gj}, \mathbf{z}_{ig}, \boldsymbol{\beta}_g, p_{igj}, E_{igj} &\sim \text{Poisson}(\lambda_{gj} \exp(\mathbf{z}_{ig}^T \boldsymbol{\beta}_g) E_{igj}), \\ \log \boldsymbol{\lambda}_g &\sim \text{MultiNormal}(\boldsymbol{\alpha}_\lambda, \text{diag}(\boldsymbol{\sigma}_\lambda)), \\ \boldsymbol{\beta}_g &\sim \text{MultiNormal}(\boldsymbol{\alpha}_\beta, \text{diag}(\boldsymbol{\sigma}_\beta)), \end{aligned} \tag{2.15}$$

which necessarily omits a model for the missing-race-data cases. The priors for the hyperparameters matched those in eq. (2.14) for the shared parameters between the joint model and the complete case model. We used the results of Theorem 2 to check that our PUMA-level models were locally identifiable. All 13 PUMAs satisfied the local identifiability criteria in 2.

2.5.3.2 Multiple imputation method description

1. **Ad-hoc MI**: The first multiple imputation model is an ad-hoc method which imputes missing cases using a multinomial distribution with a probability parameter equal to that of the population proportions. For example, suppose we observe m_{ig} missing cases for a certain stratum i in PUMA g, along with $\mathbf{x}_{ig}$ cases by race. In order to generate a single imputation draw, $\mathbf{y}_{ig}^{(s)}$, we draw the missing cases: $\boldsymbol{\epsilon}_{ig}^{(s)} \sim \text{Multinomial}(m_{ig}, \mathbf{e}_{ig}/\sum_j E_{igj})$ and add $\boldsymbol{\epsilon}_{ig}^{(s)}$ to $\mathbf{e}_{ig}$: $\mathbf{y}_{ig}^{(s)} = \boldsymbol{\epsilon}_{ig}^{(s)} + \mathbf{x}_{ig}$. We loop through $i \in \{1, \ldots, I\}$ to generate one complete dataset and repeat this step to generate multiple complete datasets.

2. **Gibbs MI**: The second multiple imputation model is described in Chapter 18 of Gelman et al. (2013): The method generates complete datasets using a Gibbs sampler that alternates between sampling missing cases $\boldsymbol{\epsilon}_{ig}^{(s)}|\boldsymbol{\theta}_{ig}^{(s-1)} \sim \text{Multinomial}(m_{ig}, \frac{\boldsymbol{\theta}_{ig}^{(s-1)}}{\sum_j \boldsymbol{\theta}_{ig}^{(s-1)}})$ and $\boldsymbol{\theta}^{(s)}|\mathbf{y}, \boldsymbol{\epsilon}^{(s)} \sim \text{Dirichlet}(\mathbf{1} + \mathbf{y} + \boldsymbol{\epsilon}^{(s)})$ where $\boldsymbol{\theta}^{(s)}$ is the concatenation of each $\boldsymbol{\theta}_{ig}^{(s)}$ for the Gibbs sampler iteration step s into a single vector, and $\mathbf{y}, \boldsymbol{\epsilon}^{(s)}$ are also vectors formed by concatenating $\mathbf{y}_{ig}, \boldsymbol{\epsilon}_{ig}^{(s)}$ into single vectors appropriately matching the indexing of $\boldsymbol{\theta}^{(s)}$ and $\mathbf{1}$ is an appropriately sized vector of 1s, representing the uniform prior over the simplex. We run the Gibbs sampler for 20 MCMC chains for 2,500 burn-in iterations and 2,500 samples,

which we subsequently thin by 25 steps, resulting in 5,000 total posterior samples. We then take a subset of these samples 5,000 as our completed datasets.

We generate 100 imputed datasets from each method for each simulated dataset, fit model (2.15) to each imputed dataset with Stan and combine the 100 sets of posterior draws into a single superset of posterior samples. We then compute posterior summary statistics including credible intervals for each method using the single superset of posterior samples, following advice in Zhou and Reiter (2010) which showed that proper Bayesian inference using multiple imputation must follow this procedure.

2.5.4 Estimands of interest

In order to compare the models on a common subset of parameters, we limited our comparisons to those involving the data-generating disease process parameters $\boldsymbol{\alpha}_\lambda$, and $\boldsymbol{\alpha}_\beta$. The simplest estimands against which we measured each model's inferences were $\boldsymbol{\alpha}_\lambda$, and $\boldsymbol{\sigma}_\lambda$. We were also interested in the following estimands:

$$(\exp((\boldsymbol{\alpha}_\lambda)_1 - (\boldsymbol{\alpha}_\lambda)_J), \ldots \exp((\boldsymbol{\alpha}_\lambda)_{J-1} - (\boldsymbol{\alpha}_\lambda)_J))$$

which are Wayne-county-level group-specific rates of disease relative to the rate of disease in category J; in the simulation study category J was Whites. There are several more complex estimands which have epidemiological significance, which are similar to poststratification estimators Gelman and Little (1997); Gao et al. (2021) that are functions of the PUMA-local parameters $\boldsymbol{\beta}_g$ and $\boldsymbol{\lambda}_g$, or the Poisson model coefficients for strata and rates of disease by race/ethnicity category in a geography g.

2.5.4.1 Modeled incidence

The first will be total modeled incidence for a race/ethnicity category j, or $\mathbb{I}_j$. Let $r_{igj} = \lambda_{gj} \exp(\mathbf{z}_i^T \boldsymbol{\beta}_g)$ be the rate of expected cases per person of disease in stratum i, geographical area g for category j. Then

$$\mathbb{I}_j = \frac{\sum_{i=1}^{I} \sum_{g=1}^{G} E_{igj} r_{igj}}{\sum_{i=1}^{I} \sum_{g=1}^{G} E_{igj}}$$

is the total incidence for category j. Interest often lies in relative risk ratios, or

$$\mathbb{I}_j / \mathbb{I}_J.$$

2.5.4.2 Standardized incidence

The second estimand is the standardized incidence or $\mathbb{SI}_j$. Let

$$\psi_i = \frac{\sum_{j=1}^{J}\sum_{g=1}^{G} E_{igj} r_{igj}}{\sum_{j=1}^{J}\sum_{g=1}^{G} E_{igj}}$$

be the population average incidence for a single stratum i. Then the $\mathbb{SI}_j$ for category j is:

$$\mathbb{SI}_j = \frac{\sum_{g=1}^{G}\sum_{i=1}^{I} E_{igj}\psi_i}{\sum_{g=1}^{G}\sum_{i=1}^{I} E_{igj}}.$$

The standardized incidence for race/ethnicity j quantifies the cumulative incidence based solely on race j's population distribution across strata.

2.5.4.3 Standardized incidence ratio

The third estimand is the standardized incidence ratio, denoted as the SIR in Lash et al. (2021), though not to be confused with susceptible-infected-recovered models (Keeling and Rohani, 2011), which is the ratio of the modeled incidence to standardized incidence:

$$\text{SIR}_j = \frac{\mathbb{I}_j}{\mathbb{SI}_j}.$$

The SIR_j measures how modeled cumulative incidence for a race/ethnicity category j deviates from the standardized incidence. A ratio above one indicates that race/ethnicity category j has experienced higher rates of disease than would be expected based on the population distribution across ages and sexes alone, while a ratio below one indicates the opposite. We can then derive relative estimands from $\mathbb{I}_j$, $\mathbb{SI}_j$, and SIR_j as we did using $\boldsymbol{\alpha}_\lambda$.

2.5.5 Computation

We ran Stan via the `cmdstanr` interface in R (Team, 2021; Gabry and Češnovar, 2021; R Core Team, 2021) on University of Michigan's Great Lakes Slurm High Performance Computing Cluster. For the exhaustive combination of models and datasets for the joint and complete-case models (1,600 in total), we ran four Markov chain Monte Carlo chains for 2,000 warmup iterations and 1,500 post-warmup iterations. In order to ensure that the posteriors had been sufficiently explored, for each dataset/model combination we recorded the maximum of all parameters' rank-normalized $\hat{R}$s, and the minima of bulk effective sample size and tail effective sample size divided by the total post-warmup iterations, which was 6,000 (bulk ESS efficiency, and tail ESS efficiency, respec-

tively) using the `posterior` package in `R` (Bürkner et al., 2021; R Core Team, 2021; Vehtari et al., 2020).

We generated 100 imputed datasets for each of the 800 simulated datasets for each imputation method, and subsequently ran (2.15) for 500 warmup iterations and 1,000 post-warmup iterations with four MCMC chains, resulting in 160,000 fitted four-chain Stan models between both imputation methods.

Example `R` and Stan code, including models and code to verify identifiability condition (S.g), can be found at https://github.com/rtrangucci/epi-missing-data.

2.5.6 Results

2.5.6.1 Computation

The joint and complete-case models ran with maximum rank-normalized $\hat{R}$s below 1.013. All but one model ran with bulk ESS efficiency greater than 10.0% (the 1 out of 1,600 model/data pair that violated the threshold ran with 9.7% bulk ESS efficiency) and all ran with minimum tail ESS efficiency greater than 10%. No divergent transitions were recorded, though 29 complete case models fitted to datasets generated in the 20% observed scenario needed to be rerun with a warmup-iteration target Metropolis acceptance rate of 0.995, an increase compared to the 0.95 target acceptance rate that all models were run with initially. No iterations were observed that hit maximum treedepth, which was set to 14 for all runs.

A small minority of the multiple imputation runs encountered treedepth issues, though all 160,000 model-by-imputed dataset combinations ran with bulk and tail ESS efficiencies greater than 10.0%. The CPU time required to run the multiple imputation methods was, at a minimum, ~ 42 times greater than either the joint or the complete-case models which is a clear disadvantage to multiple imputation methods. Zhou and Reiter note that for Bayesian credible intervals to achieve nominal coverage with multiple imputation many more than the typically recommended 5-20 imputed datasets are required.

2.5.6.2 Bias and root mean squared error

We made boxplots of bias for each parameter across all simulation runs S. We used the posterior mean from each model as the estimator for each estimand θ, or $\hat{\theta} = \mathbb{E}_{\theta|Y}[\theta]$, and calculated bias for a simulation run s as

$$\text{bias}(\hat{\theta}_s, \theta_s) = \hat{\theta}_s - \theta_s.$$

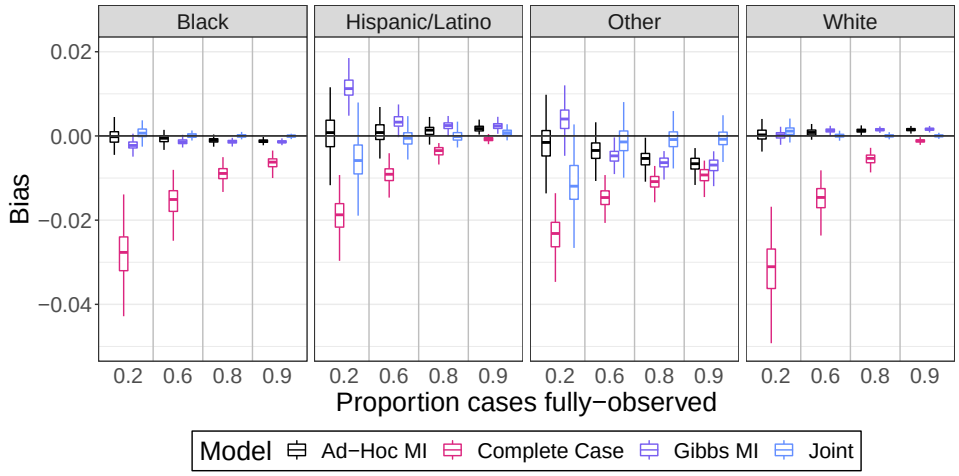


Figure 2.1: Bias across simulated datasets for the incidence, or $\mathbb{I}_j$ for Blacks, Hispanic/Latinos, Others, and Whites plotted against the proportion of cases observed with race data.

Root mean squared error was calculated as

$$\text{RMSE}(\hat{\theta}, \theta) = \sqrt{\frac{1}{S}\sum_{s=1}^{S} \text{bias}(\hat{\theta}_s, \theta_s)^2}.$$

Asymptotic 95% confidence intervals were calculated using the Delta method (Lehmann and Casella, 1998).

Bias in estimating incidence by race/ethnicity As can be seen in Figure 2.1, for Blacks and Whites, which comprise 49% and 40% of the total population in Wayne County, the bias in the posterior mean incidence estimator generated by the joint model is small across all scenarios for most simulated datasets. For Whites, the average bias in the joint model posterior mean is not significantly different than zero in the 90%, 80% and 60% scenarios, while for Blacks, there is statistically significant average bias for joint-model posterior mean incidence in all scenarios other than 80%, but it is an order of magnitude smaller than the average bias of the posterior mean estimator from the imputation methods. The complete case model, as expected, is significantly negatively biased in all scenarios. The average bias from ad-hoc multiple imputation is smallest among all methods in the 20% scenario because the data generating process, outlined in Section 2.5.2, defines the true population rate of disease for each race/ethnicity group to be the same. The distribution of missing cases by category conditional on the total missing cases is multinomial with parameter $\mathbf{e}_{ig} \odot (1 - \mathbf{p}_{ig}) / \sum_j E_{igj}(1 - p_{igj})$. When missingness is high, $(1 - p_{igj})$ is close to one, so the ad-hoc multinomial imputation procedure with parameter $\mathbf{e}_{ig} / \sum_j E_{igj}$ is approximately correct. As missingness decreases, the ad-hoc imputation parameter diverges from the data generating process

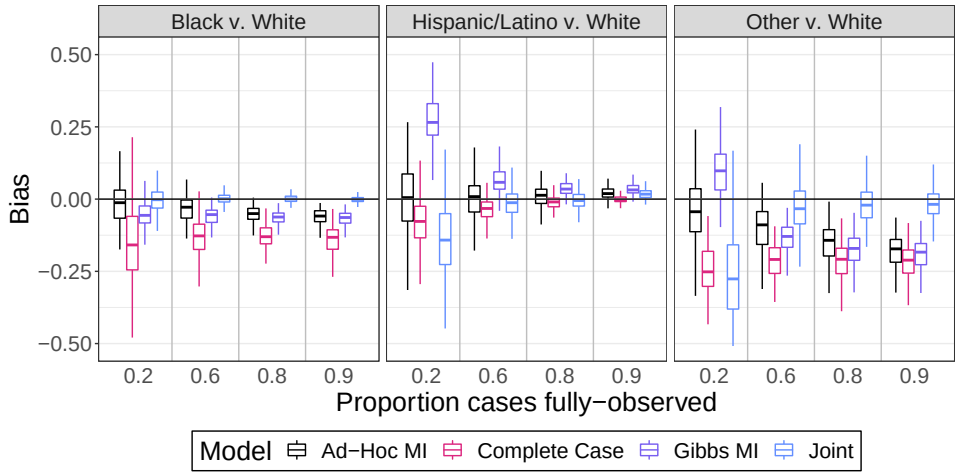


Figure 2.2: Bias across simulated datasets for the relative risk ratios, or $\mathbb{I}_j/\mathbb{I}_J$ for Blacks, Hispanic/Latinos, and Others relative to Whites plotted against the proportion of cases observed with race data.

and the bias grows. This pattern can be seen in Figure 2.2 as well. In sum, the averages of the joint model estimators' biases are sometimes more than two standard errors from zero, but the model's absolute bias is significantly smaller compared to the absolute bias of the competing estimators, with exceptions in the 20% scenario compared to the Ad-Hoc MI method.

Bias in estimating relative risk by race/ethnicity In Figure 2.2 the joint model was able to estimate the relative risk of disease with mean bias that is not significantly different from zero for Blacks vs. Whites in the 80%, 60% and 20% observed scenarios, while in the 90% scenario the mean bias is significantly nonzero, but two orders of magnitude smaller than the mean bias incurred by the complete case model's estimators. For Hispanic/Latinos and Others, there exists some mean bias in the 90%, 60%, and 20% scenarios, though in the 80% and 60% scenarios the mean bias is an order of magnitude smaller than that of the complete case analysis. Complete case analysis does yield estimators with average bias that is not significantly different from zero for the relative risk of disease for Hispanics/Latinos to Whites in the 90% observed scenario and has smaller average bias compared to the joint model's estimators. This is due to the fact that in the simulated datasets the log-odds of observing race data was equal for Whites and Hispanics/Latinos, all else being equal. The average bias from multiple imputation using Gibbs sampling is consistently nonzero across all missingness scenarios for all groups in Figure 2.2. The Gibbs multiple imputation procedure assumes the data are MAR, when the DGP is NMAR for all scenarios. This highlights the danger of using a MAR procedure when the data are NMAR. The pattern of bias is similar for $\exp((\boldsymbol{\alpha}_\lambda)_j - (\boldsymbol{\alpha}_\lambda)_J)$: the complete-case estimators are comparable in terms of mean bias to that of the joint-model estimators in the Hispanic/Latino group, while the complete-case estimators underperform

in Blacks and Others. For $\boldsymbol{\sigma}_\lambda$, the complete-case posterior mean estimators are positively biased compared to the joint-model's estimators, likely due to the fact that the complete case analysis attributes all variance in local area estimates of $\boldsymbol{\lambda}_g$ to variation in disease incidence while the joint model attributes some of the variation to variation in the observational process. The estimators from the joint model are, however, negatively biased, likely due to the fact that we have only 13 PUMAs and relatively strong Normal$^+(0, 0.5^2)$ priors that shrink towards zero on the population scale parameters $\boldsymbol{\sigma}_\lambda$.

Root mean squared error The RMSEs are shown on in Appendix A.6.1. That of the joint-model estimators for SIR$_j$ are significantly smaller (as measured shown by nonoverlapping 95% confidence intervals) than the RMSEs for the complete-case estimators in the 90%, 80%, and 60% scenarios for nearly all groups (the exception is for Hispanics/Latinos in the 60% scenario, where the RMSEs are not significantly different). In the 20% observed scenario, the RMSEs of the joint-model estimators for Blacks and Whites are smaller than those of the complete case model, but the RMSEs of the joint-model estimators for Hispanic/Latinos and Others are larger than the complete-case estimators. This is due to the fact that Hispanic/Latinos and Others are smaller populations in Wayne County, and the parameter space for the is $2\times$ as large as the complete-case model's parameter space. We also present the RMSE comparisons for the relative risk ratios and relative county-level rates, shown in figures A.3 and A.4, respectively. The relative risk ratio plots show a similar pattern to that of the SIR$_j$ estimates, with the exception of relative risk ratios for Hispanics/Latinos, for which the RMSEs of the complete-case estimators are smaller than those of the joint model's. This is due to the fact that White and Hispanics/Latinos case-patients are observed at similar relative rates across simulations because the observation ratio, or $\text{inv\_logit}((\boldsymbol{\alpha}_\eta)_j)/\text{inv\_logit}((\boldsymbol{\alpha}_\eta)_J) = 1$ for these two groups and the complete case analysis model implicitly assumes the observation ratios for all races to be exactly 1.

On the contrary, figure A.4 shows that the RMSEs for the joint-model's estimators are similar in magnitude or larger in all scenarios. While the joint-model's estimators show smaller mean biases, the variance for the estimators is much larger compared to the complete case analysis. This is again due to the fact that there are only 13 PUMAs included in the simulation study, and the fact that the dimension of the parameter space is twice as large for the joint model as that of the complete case model. The RMSEs for the ad-hoc imputation approach are small in the 20% scenario for the same reason the bias is small in the 20% scenario, but the RMSE increases as the missingness decreases. This is a clear indication that the data generating process does not agree with the imputation procedure. The RMSEs for the Gibbs imputation approach are large for the 20% scenario, likely owing to the fact that as the number of missing cases increases, the variance of the imputed datasets increased due to increased posterior uncertainty for the imputation model.

This could be an indication that more than 100 imputed datasets are necessary for the imputation procedure when missingness is high, which would accord with the observations in Zhou and Reiter (2010), though we were constrained by computational budget to use only 100 imputed datasets per simulated dataset.

2.5.6.3 Coverage and interval length

Table 2.3 summarizes the interval coverage for the complete-case model, the joint model, and the multiple imputation procedures. All intervals that follow are central $p\%$ posterior credible intervals. In the event the joint distribution of the simulated parameters and data matches the prior and the likelihood of the inferential model and we can properly draw samples from the posterior, the central $p\%$ posterior credible intervals (and any other posterior intervals, for that matter) will contain the parameter that generated the data with exactly $p\%$ probability (Cook et al., 2006). As expected, the complete-case model's 50% intervals severely under cover for all but the county-level relative rates of disease for Hispanics and Latinos compared to the rate for Whites. The ad-hoc imputation method's intervals over-covered for the population-level relative risk comparisons (as seen in A.4), while they undercovered for the standardized incidence and relative risk measures, while the Gibbs sampler imputation's intervals severely undercovered in all scenarios for all the parameters of interest. Despite the ad-hoc methods near-match to the data generating process in the 20% scenario, the intervals for incidence under-cover more than the joint model's credible intervals. The joint model's intervals are near the nominal coverage probabilities, i.e. the 50% intervals cover the true parameter value in 50% of simulations, though they do under-cover for sparsely populated groups like Others and Hispanic/Latinos, especially so with significant numbers of missing cases.

The same pattern is exhibited in the figure 2.3, which shows boxplots of the average coverage across all parameters related to the disease process for each simulated dataset for all models. The complete-case model's 50% and 80% interval coverage is about 25% and 35%, respectively, while the joint model's intervals achieve the nominal coverage probability on average. The multiple imputation methods' intervals fare a bit better though they still under-cover: the rates are near 30%-35% and 60% to 65% on average.

In appendix section A.6.2 we present table A.3, which mirrors table 2.3 but for 80% intervals. The pattern of performance is similar.

2.5.6.4 Breakdown analysis

The joint model performs well under the 90%-, 80%- and 60%-observed scenarios, but when there is a significant proportion of cases that are missing race data, like in the 20%-observed scenario, the

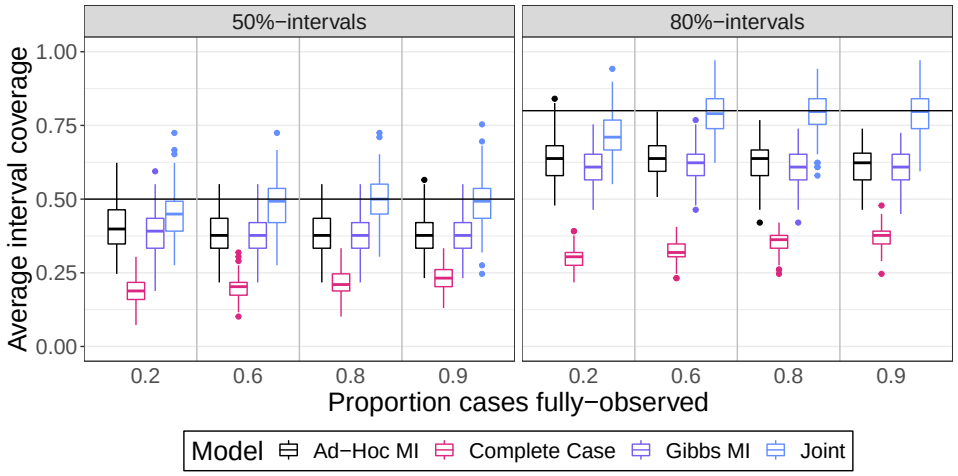


Figure 2.3: Boxplots of simulation-wise mean 50% and 80% interval coverage by observed data proportion scenario for the joint model, the complete-case model, and the multiple imputation methods. Horizontal black lines indicate the nominal coverage probability rates.

model's posterior intervals begin to undercover compared to the nominal coverage probabilities. One can see this in figure 2.3, as the interval coverage in the 60% observed scenario begin to undercover slightly as measured by the median across the 200 simulated datasets. In the 20% observed scenario, the 75^{th} quantiles of the mean parameter coverage for the full model for both the 50% and 80% intervals lie below the nominal coverage rates. This leads us to conclude that informative priors are necessary when the model is fitted to datasets that have significant numbers of cases that are missing race data. If the likelihood and prior conflict, however, these priors may have an outsized influence on the posterior estimands.

2.5.7 Prior sensitivity results

In order to test the sensitivity of model inferences to priors over population hyperparameters such as the population mean log-incidence (α_λ), or population mean log-odds of observing a specific race/ethnicity category (α_η), we used a subset of 100 simulated datasets from the 20% observed scenario. We varied the parameters of the priors over the population hyperparameters over a grid and re-estimated the quantities of interest for each prior specification. We varied one prior parameter at a time while holding the other prior parameters fixed at the values shown in eq. (2.14). The parameter values are shown in Table 2.4.

Population parameter	Prior parameter	Values
$\boldsymbol{\alpha}_{\boldsymbol{\eta}_j}$	$\mathbb{E}\left[\boldsymbol{\alpha}_{\boldsymbol{\eta}}\right]_j$	$\{0.5, 1, \mathbf{2}, 3\}\forall j$
$\boldsymbol{\alpha}_{\boldsymbol{\lambda}_j}$	$\mathbb{E}\left[\boldsymbol{\alpha}_{\boldsymbol{\lambda}}\right]_j$	$\{-3.5, -4, -4.5, \mathbf{-5}\}\forall j$
$\boldsymbol{\alpha}_{\boldsymbol{\eta}_j}$	$\mathrm{SD}(\boldsymbol{\alpha}_{\boldsymbol{\eta}})_j$	$\{0.3, 0.5, \mathbf{1}, 2, 3\}\forall j$
$\boldsymbol{\alpha}_{\boldsymbol{\lambda}_j}$	$\mathrm{SD}(\boldsymbol{\alpha}_{\boldsymbol{\lambda}})_j$	$\{0.3, 0.5, \mathbf{1}, 2, 3\}\forall j$
$\boldsymbol{\sigma}_{\boldsymbol{\eta}_j}$	$\mathbb{E}\left[\boldsymbol{\sigma}_{\boldsymbol{\eta}}\right]_j / \sqrt{2/\pi}$	$\{0.25, 0.5, \mathbf{1}, 2\}\forall j$
$\boldsymbol{\sigma}_{\boldsymbol{\lambda}_j}$	$\mathbb{E}\left[\boldsymbol{\sigma}_{\boldsymbol{\lambda}}\right]_j / \sqrt{2/\pi}$	$\{0.25, 0.5, \mathbf{1}, 2\}\forall j$

Table 2.4: Prior sensitivity simulation study prior settings.
Bold values correspond to settings used for results presented in 2.5.6. Prior parameter for $\boldsymbol{\sigma}_{\boldsymbol{\lambda}}$ and $\boldsymbol{\sigma}_{\boldsymbol{\eta}}$ is the standard deviation parameter for a half-normal distribution.

We measured 1) the sensitivity of the estimated posterior mean incidence by race/ethnic group, or $\mathbb{I}_j$, and 2) its bias. Our measure of posterior mean sensitivity to the prior mean was the change in posterior mean against a reference mean scaled by a reference standard deviation, where the reference mean and standard deviation were those obtained using the prior settings set out in Equation (2.14). Specifically, for an estimand $g(\boldsymbol{\theta})$, with a posterior over $\boldsymbol{\theta}$ $\pi_{\mathrm{b}}(\boldsymbol{\theta}|\mathrm{Data})$ under a prior with reference parameters b and a posterior $\pi_{\mathrm{a}}(\boldsymbol{\theta}|\mathrm{Data})$ under a prior with alternative parameters a:

$$\text{Posterior Z-score } = \frac{\mathbb{E}_{\pi_{\mathrm{a}}(\boldsymbol{\theta}|\mathrm{Data})}[g(\boldsymbol{\theta})] - \mathbb{E}_{\pi_{\mathrm{b}}(\boldsymbol{\theta}|\mathrm{Data})}[g(\boldsymbol{\theta})]}{\sqrt{\mathrm{Var}_{\pi_{\mathrm{b}}(\boldsymbol{\theta}|\mathrm{Data})}(g(\boldsymbol{\theta}))}}. \tag{2.16}$$

The measure of bias for a true estimand $g(\boldsymbol{\theta}^\dagger)$ is

$$\frac{\mathbb{E}_{\pi_{\mathrm{a}}(\boldsymbol{\theta}|\mathrm{Data})}[g(\boldsymbol{\theta})] - g(\boldsymbol{\theta}^\dagger)}{g(\boldsymbol{\theta}^\dagger)}. \tag{2.17}$$

Figure A.5 shows that the posterior incidence estimate is somewhat sensitive to the priors over log-population mean incidence and log-odds of observing race/ethnicity information. The right-hand column in Figure A.5 shows that as the prior mean for α_λ for the Other group differs from the true data-generating mean by 3 prior standard deviations, the posterior mean can change by roughly half a posterior standard deviation from the posterior mean under the baseline prior.

Meanwhile, the left-hand column of Figure A.5 shows the sensitivity of the posterior mean for incidence by race/ethnicity to the prior for α_η. Of interest is the posterior mean for the Other group because it is the minority group. In the 20%-observed scenario, the true α_η for the Other group is approximately 0.3, while the prior mean for α_η is 2. When the prior standard deviation is

decreased to 0.5 from 1, the prior mean is approximately 3 prior standard deviations away from the true data generating parameter, and the posterior mean decreases by about half a posterior standard deviation. Despite the fact that the posterior means can shift due to changes in the prior, however, the posterior mean never exceeds 2 posterior standard deviations, implying that the inferences do not appreciably change.

Digging deeper into the upper-left-hand plot in Figure A.6 shows that when the prior for $\boldsymbol{\alpha_\eta}$ is centered on missing-at-random missingness and the prior mean is too large compared to the true proportion of cases with observed race/ethnicity, the model over-allocates missing cases to majority groups while it under-allocates cases to minority groups. If we instead center the prior too low then we may over-allocate cases to minority groups.

The lower-left-hand plot in Figure A.5 shows a similar phenomenon when the prior reflects too-strong certainty that the data-generating process is nearly missing-at-random. When too much prior weight is allocated to near-missing-at-random $\boldsymbol{\alpha_\eta}$, the model deflates incidence for groups with higher-than-average missingness and inflates incidence for groups with lower-than-average incidence.

Figure A.6 shows that the bias is not appreciable for incidence, with the exception of the Other group when the prior for α_λ is about 3 standard deviations or more too large.

Figure A.7 shows posterior Z-score and bias plots for changes to the prior for population inter-geography standard deviation parameters for λ_{gj} and η_{gj}, or $\boldsymbol{\sigma_\lambda}$ and $\boldsymbol{\sigma_\eta}$. The posterior for incidence is not especially sensitive to the prior over these parameters.

The results of the prior sensitivity simulation study show that the model inferences for incidence are relatively robust to misspecification of priors for population hyperparameters, but that care should be taken with the prior mass apportioned to data generating processes that are centered on missing-at-random scenarios.

2.6 Application to COVID-19 case data in Wayne County, Michigan

In this section we will apply both the complete-case model and the joint model to COVID-19 case data in Wayne County from the first wave of the pandemic.

2.6.1 Data

The source of our case data is the Michigan Disease Surveillance System (MDSS) maintained by the Michigan Department of Health and Human Services (MDHHS). MDHHS's guidelines for the collection of probable COVID-19 cases is set out in Michigan Department of Health and Human

Services. (2020) as outlined in Zelner et al. (2021). We included all reported PCR-confirmed COVID-19 cases for individuals outside of state prisons with that were entered into MDSS between $2020-03-01$ through $2020-06-30$. This comprises 22,141 cases of COVID-19. We then filtered out 1,374 cases, or about 6% of the total cases, that could not be geocoded to a unique address in Wayne County. We filtered a further 74 cases for which the case patients' sex at birth was unknown, as well as 7 cases for which age was unknown. Finally, we dropped 1,398 cases which were matched to the address of a licensed nursing homes or long-term care facility (LTCF). We excluded these cases for two reasons: 1) the populations of nursing homes and LTCFs are likely not well-represented by the 2010 Census denominators and 2) the high incidence among nursing home and LTCF residents does not accord with our assumption of a Poisson process for disease cases. This results in a final dataset of 19,288 PCR-confirmed COVID-19 cases.

In total, approximately 18% of the 19,288 cases, or 3,464 cases, are missing race data. For cases that do include the race of the respondent and are not identified as Hispanic or Latino, we classify those who are identified as Asian or Hawaiian or Pacific Islander as Asian, those identified as Black/African American or Black/African American/Unknown as Black, and Caucasian and Caucasian/Unknown as White. We classify cases as Hispanic or Latino if the data field for patient ethnicity is equal to Hispanic or Latino. We classify those who identify as Native American or Alaska Native, mixed race, or other race as Other. Cases that are not missing race info but are missing patient ethnicity information are classified as the indicated race and are treated as not Hispanic or Latino.

We again have 13 PUMAs that comprise Wayne county, and 18 age by sex-at-birth strata per PUMA.

2.6.1.1 Aggregation to PUMAs

This yields 234 observations of the counts of PCR-confirmed COVID-19 cases within each race/ethnicity category, or 1,170 total observations of PUMA by age by sex-at-birth by race/ethnicity. The mean count is 13.5 while the variance is 696.9. As for observations of total counts of cases missing race and ethnicity information by PUMA by age by sex-at-birth, 6% of the 234 PUMAs have zero observed cases with missing race and ethnicity.

2.6.1.2 Population data

We added the Asian/Pacific Islander group as an additional race/ethnic category, because such individuals make up a significant fraction of the population in Wayne County, though in all other respects the PUMA-level population data is the same as in the simulation study in Subsection 2.5.1.

2.6.2 Models and priors

We fitted four of the models presented in Section 2.5.3: the joint model, the complete-case model, and the ad-hoc and Gibbs multiple imputation models. The full specification for the joint model is:

$$\begin{aligned}
X_{igj}|\lambda_{gj}, \mathbf{z}_i, \boldsymbol{\beta}_g, p_{igj}, E_{igj} &\sim \text{Poisson}(p_{igj}\lambda_{gj}\exp(\mathbf{z}_i^T\boldsymbol{\beta}_g)E_{igj}),\\
M_{ig}|\boldsymbol{\lambda}_g, \mathbf{z}_i, \boldsymbol{\beta}_g, \mathbf{p}_{ig}, \mathbf{e}_{ig} &\sim \text{Poisson}(\exp(\mathbf{z}_i^T\boldsymbol{\beta}_g)\textstyle\sum_j((1-p_{igj})\lambda_{gj}E_{igj}),\\
p_{igj} &= \left(1+\exp(-(\mathbf{z}_i^T\boldsymbol{\gamma}_g+\eta_{gj}))\right)^{-1},\\
\log\boldsymbol{\lambda}_g|\boldsymbol{\alpha}_\lambda, \boldsymbol{\sigma}_\lambda &\sim \text{MultiNormal}(\boldsymbol{\alpha}_\lambda, \text{diag}(\boldsymbol{\sigma}_\lambda^2)),\\
\boldsymbol{\eta}_g|\boldsymbol{\alpha}_\eta, \boldsymbol{\sigma}_\eta &\sim \text{MultiNormal}(\boldsymbol{\alpha}_\eta, \text{diag}(\boldsymbol{\sigma}_\eta^2)),\\
\boldsymbol{\beta}_g|\boldsymbol{\alpha}_\beta, \boldsymbol{\sigma}_\beta &\sim \text{MultiNormal}(\boldsymbol{\alpha}_\beta, \text{diag}(\boldsymbol{\sigma}_\beta^2)),\\
\boldsymbol{\gamma}_g|\boldsymbol{\alpha}_\gamma, \boldsymbol{\sigma}_\gamma &\sim \text{MultiNormal}(\boldsymbol{\alpha}_\gamma, \text{diag}(\boldsymbol{\sigma}_\gamma^2)),
\end{aligned} \tag{2.18}$$

with the same priors over the hyperparameters as in eq. (2.14) with the exception of the prior scale for $\boldsymbol{\sigma}_\gamma$ set to 1 instead of 0.5.

The full specification for the complete-case model is

$$\begin{aligned}
X_{igj}|\lambda_{gj}, \mathbf{z}_i, \boldsymbol{\beta}_g, p_{igj}, E_{igj} &\sim \text{Poisson}(\lambda_{gj}\exp(\mathbf{z}_i^T\boldsymbol{\beta}_g)E_{igj}),\\
\log\boldsymbol{\lambda}_g|\boldsymbol{\alpha}_\lambda, \boldsymbol{\sigma}_\lambda &\sim \text{MultiNormal}(\boldsymbol{\alpha}_\lambda, \text{diag}(\boldsymbol{\sigma}_\lambda^2)),\\
\boldsymbol{\beta}_g|\boldsymbol{\alpha}_\beta, \boldsymbol{\sigma}_\beta &\sim \text{MultiNormal}(\boldsymbol{\alpha}_\beta, \text{diag}(\boldsymbol{\sigma}_\beta^2)),
\end{aligned} \tag{2.19}$$

with the same priors as the joint model over the shared hyperparameters $\boldsymbol{\alpha}_\lambda, \boldsymbol{\alpha}_\beta, \boldsymbol{\sigma}_\lambda$ and $\boldsymbol{\sigma}_\beta$.

$\mathbf{z}_i$ was 9-dimensional, with the first element encoding male vs. female and the next eight elements encoding the age stratum from $[0, 10)$ to $[70, 80)$. We used a sum contrast for age and a scaled sum contrast for male vs. female. We used the results of Theorem 2 to check that our model as defined is locally identifiable for each PUMA. All 13 PUMAs meet our criteria for the model to be locally identifiable. We needed to rerun the identifiability analysis because we expanded our race/ethnicity categories by one to include Asians/Pacific Islanders as a separate group. Our construction of the $\boldsymbol{\beta}_g$ and $\boldsymbol{\gamma}_g$ is the same as in the simulation study.

2.6.2.1 Computational results

We again used `cmdstanr` as the Stan interface via `R` (Gabry and Češnovar, 2021; R Core Team, 2021). Each model was run with 8 MCMC chains with 3,000 warmup iterations, and 2,000 post-warmup iterations with a target Metropolis acceptance rate of 0.99. For the joint model, all $\hat{R}$s were less than 1.01, while the minimum bulk and tail ESS efficiencies were 0.098 and 0.200 rounded, respectively. For the complete-case model, all $\hat{R}$s were less than 1.01, while the minimum bulk and

tail ESS efficiencies were 0.156 and 0.238, respectively. All ESS efficiency numbers are rounded to three digits.

The multiple imputation methods were run for 1,000 warmup, and 2,000 post-warmup iterations for each of the 100 imputed datasets. All $\hat{R}$s were below 1.01 for each imputed datasets MCMC run, and minimum bulk and tail efficiencies exceeded 10% for the Gibbs imputation scheme while minimum bulk and tail efficiencies exceed 9% and and 10%, respectively for the ad-hoc imputation scheme. Note that the $\hat{R}$ statistics for the combined chains are typically larger than 1.01 for many parameters of interest, which can be seen in table A.6. This is due to the between-imputed-dataset variance.

2.6.3 Results and Model Comparison

2.6.3.1 Comparison of model results on completely-observed cases

Following Gelman et al. (2020) and Gabry et al. (2019a), we performed a series of graphical posterior predictive checks, or PPCs, using the `bayesplot` package (Gabry and Mahr, 2021). These involved simulating PUMA by age by sex by race case counts from the fitted models and comparing these outputs to the observed data. Along this dimension, the joint model and the complete-case model were indistinguishable in terms of errors, squared errors, and 50%, 80% and 95% interval coverage for the observed data.

These checks also revealed that the observational variance, or $\frac{1}{IJ-1}\sum_{i,j}(x_{ij} - \bar{x})^2$, $\bar{x} = \frac{1}{IJ}\sum_{i,j} x_{ij}$ and the proportion of zeros, or $\frac{1}{IJ}\sum_{i,j} \mathbb{1}_{x_{ij}=0}$, fell near the 50^{th} percentile for each model's posterior over the two statistics, which indicates that the Poisson distribution is a suitable outcome distribution for this dataset.

We also used graphical PPCs to gauge whether the model assumption that there is no interaction between race and age is reasonable. The plots are included in Appendix Section A.8.1, and show that while there were deviations from the model's posterior distribution for age by race cumulative incidence, they are small compared to the total cumulative incidence. Moreover, our interest lies in quantifying cumulative incidence by race for Wayne county instead of capturing all sources of variation in the observed data.

2.6.3.2 Posterior predictive checks on missing cases

We can compare the observed statistics for the missing cases to the joint model's posterior predictive distribution for the same statistics. The mean, variance, and proportion of age/race/sex strata with zero cases observed all fell well within the joint model's central 50% posterior intervals. A posterior predictive rootogram shown in Appendix Section A.8.2 that the tail is a bit thicker than the joint model expects, but the deviation is not extreme enough to warrant modifying the model.

2.6.3.3 Inference on epidemiological estimands

Following the results of our simulation study, the models' inferences differed for the estimands introduced in Subsection 2.5.4, like modeled incidence, standardized incidence, standardized incidence ratios, and functions of these estimands.

A comparison of the modeled incidence inferences for the joint model, the complete-case model, and the Gibbs-sampler-imputation method is shown in Figure 2.4. The most striking aspect of the figure is the elevated incidence in the Other race category across all methods. The complete case model infers uniformly lower incidence than does the joint model, which makes sense as the complete case model omits cases that are missing race/ethnicity information. The left-hand panel shows the Gibbs-imputation method imputes higher incidence for Whites, Asians/Pacific Islanders, and Hispanics/Latinos compared to the joint model. This mirrors the Gibbs performance in the simulation study as shown in Figure 2.5. The plot shows that the standardized difference in posterior means between the Gibbs imputation and the joint model is systematically greater than zero for Hispanic/Latinos and Whites, while it is systematically lower than zero for Others in the 80% observed data scenario. Visually, we can see that the understatement for incidence is more extreme for Blacks and Others than it is for Hispanics or Latinos and for Whites. Both the Gibbs and complete case intervals are shorter than the joint-model intervals.

Figure 2.8 shows the relative modeled incidence, or $\mathbb{I}_j/\mathbb{I}_J$, where J is the category for Whites. The plot shows that relative risks for all nonwhites are smaller when using complete case analysis compared to that of the joint model. The increase is most substantial for the Other race/ethnicity category, but both Blacks and Hispanic and Latinos have significant increases in relative risk.

Table A.5 in Appendix Subsection A.8.3 shows the exhaustive comparison between the two models for all of the estimands. Despite the models showing statistically significant differences for posterior means among the standardized incidence ratios, the practical differences are small for Blacks, Hispanics and Latinos and Asians and Pacific Islanders. The 80% posterior credible intervals, on the contrary, are larger on average for the joint models' inferences on the standardized incidence ratios. Whites and people of Other races are seen to have statistically significant and practically significant differences in the models' posterior mean estimators. The models' inferences differed most significantly in terms of relative incidence, as can also be seen in Figures 2.8, where for Blacks vs. Whites and Others vs. Whites the posterior 80% credible intervals do not overlap, even after taking into account Monte Carlo standard error. The joint model's posterior intervals for the epidemiological parameters of interest were wider on average, in agreement with the simulation study results.

2.6.3.4 Inference on missingness parameters

We cannot directly compare the inferences for the missingness parameters for the complete-case model to the joint model. We can, however, examine how the ratio of modeled incidences by race differs between races. In Figure 2.6, one can see that the 80% posterior credible intervals for the ratio of the complete-case model's incidence to the joint model's incidence do not include 1 for all races other than Asians and Pacific Islanders. The only groups for which the ratio of Gibbs-to-joint-model incidences exclude zero are Whites and Others, which again mirrors the pattern in Figure 2.5, though the difference is less extreme for Hispanics/Latinos and Asians/Pacific Islanders for the real-world data. The Gibbs imputation method's inference for Blacks nearly matched that of the joint model. This is not surprising when we consider the fact that between-group comparisons of the ratio of incidences reveals that non-White residents, excluding Others had missingness proportions that were near equal between groups. This can be seen from 2.6 as the posterior intervals for the Complete Case comparison overlap for Hispanics/Latinos, Blacks and Asian/Pacific Islanders. It can also be seen that the posterior intervals for Others do not overlap with any other category, and that Whites and Blacks are also do not overlap. Figure 2.7 shows the supporting evidence for NMAR missingness of race; the plot shows the Wayne-County-wide population inferences for the probability that an individual with COVID-19 of a certain race will have race reported in their case line-listing, all else being equal. This estimand is a transformation of the $\boldsymbol{\alpha}_\eta$ parameter, namely, inv_logit$(\boldsymbol{\alpha}_\eta)$. The strongest evidence for NMAR missingness exists for the Other category, whose 80% posterior credible intervals do not intersect any other category's intervals. There is also some evidence for NMAR missingness for Blacks with respect to Whites, as the 80% posterior credible intervals for the probability of completely observing race are $(0.81, 0.91)$ vs. $(0.89, 0.98)$, respectively, as shown in Table A.7 in Appendix Subsection A.8.3.

2.6.3.5 Summary

Our results largely align with those of prior analyses of racial disparities in COVID-19 incidence in the U.S. For example, Labgold et al. found a similarly large incidence among case-patients of Other race. The authors find a bias-adjusted PCR-confirmed COVID-19 rate of nearly 14% among Other race case-patients compared to rates of at most nearly 4% in Hispanic/Latino case-patients, who had the next-largest incidence among the races included in Labgold et al.'s study. The relative incidence between Others and Whites is nearly 14, which puts our 80% posterior credible interval of $(4.77, 8.11)$ in context.

Several explanations are plausible for the elevated incidence among people of Other races; Wayne county has a large Middle Eastern population and these individuals may identify themselves as not being Black, Hispanic or Latino, Asian or Pacific Islander or White. The case data

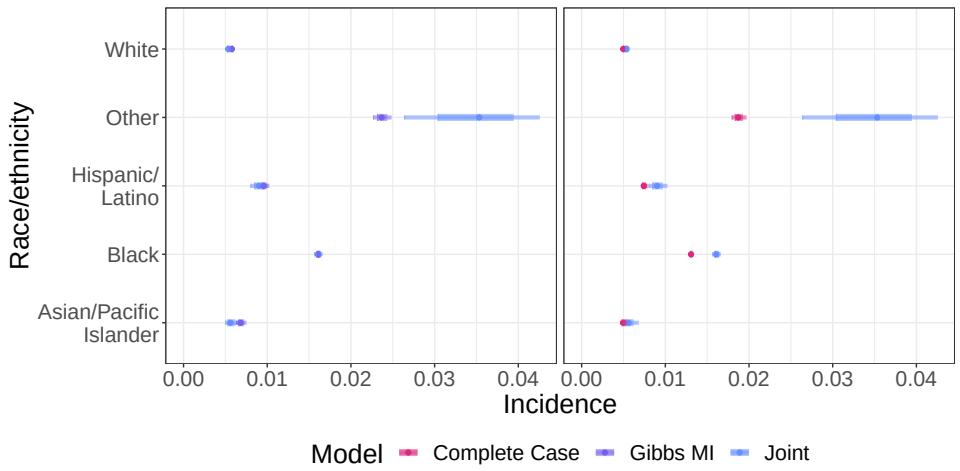


Figure 2.4: Race/ethnicity category-specific modeled incidence by model. The inner intervals are 50% and the outer intervals are 80%.

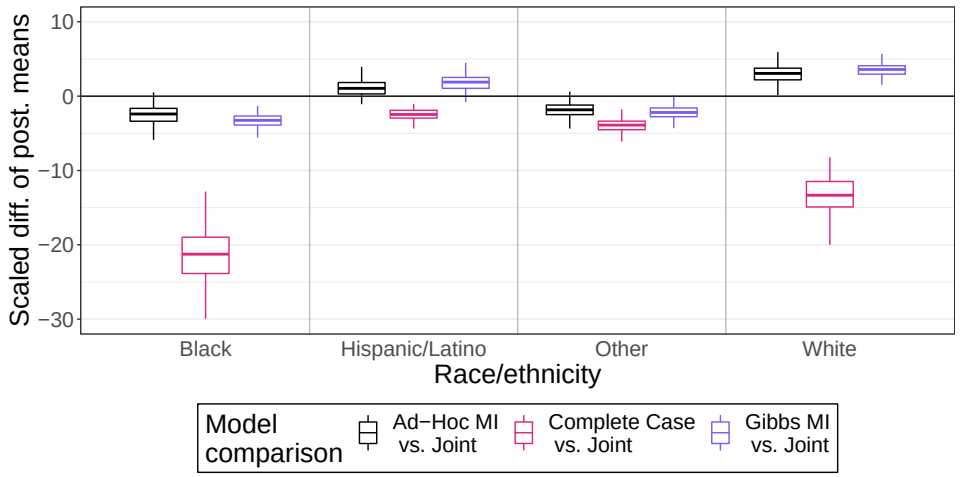


Figure 2.5: Boxplots of differences in posterior means between indicated methods and joint model scaled by pooled posterior standard deviation by race/ethnicity category-specific modeled incidence by simulated dataset for the 80% observed data scenario.

does include a field for Arab ethnicity, but the 2010 Census did not include a Middle Eastern or North African category for ethnicity. Another explanation may be our treatment of missing Hispanic/Latino ethnicity information. If many people who are identified as Other but do not have a recorded Hispanic/Latino ethnicity are truly Hispanic/Latino then our model would inflate the incidence in the Other category at the expense of the Hispanic/Latino category; given that the Other group is so small a small inflation in counts would result in a large inflation of risk.

There is strong evidence for nonignorable missingness driven by not-missing-at-random race covariates. The evidence is strongest for people of Other races. This means that omitting cases that are missing race and calculating relative risk between any race and Other would yield a biased estimate. Moreover, the size of the bias would be large because the probability of observing race for Others is low compared to the other categories; the 80% posterior credible interval is $(0.45, 077)$. There is also some evidence for NMAR missingness for Blacks with respect to Whites. Given the small number of PUMAs we modeled, there would likely be stronger evidence in favor of NMAR missingness for other race/ethnicity categories if we were to model a larger geographical area, like all of Southeastern Michigan instead of just modeling Wayne County.

While the Complete Case inferences are predictably different from the joint model's, the multiple imputation using Gibbs sampling also produced significantly different inferences. The coherence between the simulated data example and the applied data analysis suggest that multiple imputation procedures that assume MAR missingness when data are NMAR can exacerbate biases in the data by over-imputing cases for groups that are over-represented in the data because of NMAR missingness. This suggests that care must be taken when choosing an imputation procedure for missing demographic data.

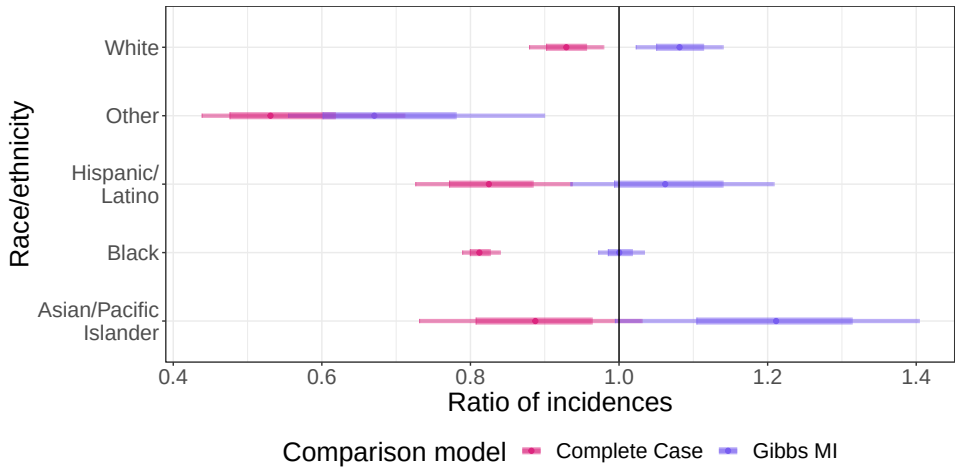


Figure 2.6: Posterior credible intervals for the ratio of modeled incidences by race/ethnicity, or $\mathbb{I}_j^{\text{CC}}/\mathbb{I}_j^{\text{J}}$ where CC stands for complete case model and J stands for the joint model. The inner and outer intervals are 50% and 80% respectively.

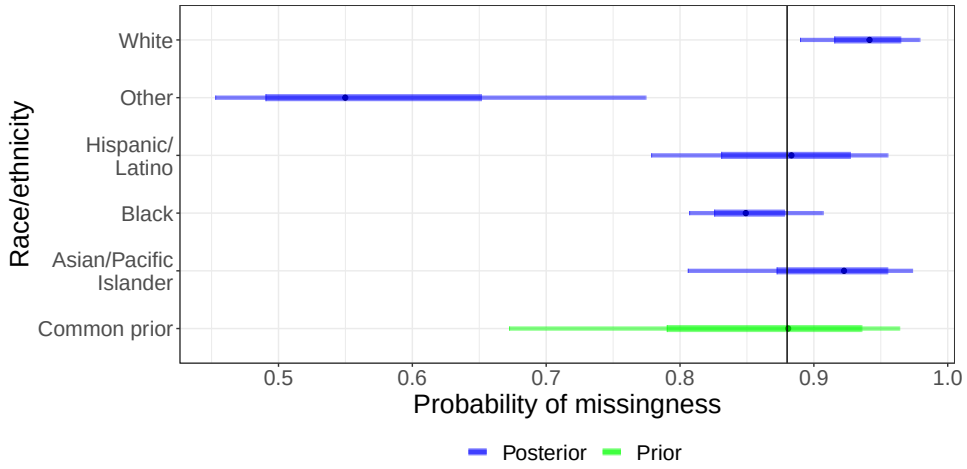


Figure 2.7: Posterior credible intervals for the population proportion of cases with fully-observed race data, all else being equal, by race/ethnicity, or inv_logit$((\boldsymbol{\alpha}_\eta)_j)$. The inner and outer intervals are 50% and 80% respectively.

Overall, our case study illustrates the large risk of bias associated with ignoring NMAR categorical data when inferring relative risks from real-world data.

2.7 Discussion

Non-random missingness of race/ethnicity covariate data is a critical challenge for the analysis of public heath data during the COVID-19 pandemic. Multiple imputation methods, which have been adopted broadly in the analysis of survey data in which the assumption of ignorability is typically reasonable (Audigier et al., 2018), may not be appropriate for the analysis of missing race/ethnicity covariates in public health surveillance data in which the possibility of not-missing-at-random (NMAR) missingness is greater.

In order to meet the needs of public health researchers to model disease data that are missing important covariates, we developed a method to jointly model the missingness process along with the disease process. Most importantly, the model can learn the extent to which the missingness process is NMAR, so our method is broadly applicable to scenarios where missingness could plausibly be NMAR, like that of missing race data.

We use a selection model formulation that combines a Poisson sampling model for the counts of disease by stratum and a conditional binomial sampling model for cases with completely observed race/ethnicity with a probability of success parameter that depends on the race/ethnicity category. Through the incorporation of known population counts from census data, the model parameters can be identified. The model can be extended to incorporate a log-linear model for incidence, a logistic model for missingness, and a hierarchy to allow for geographic heterogeneity in local parameters.

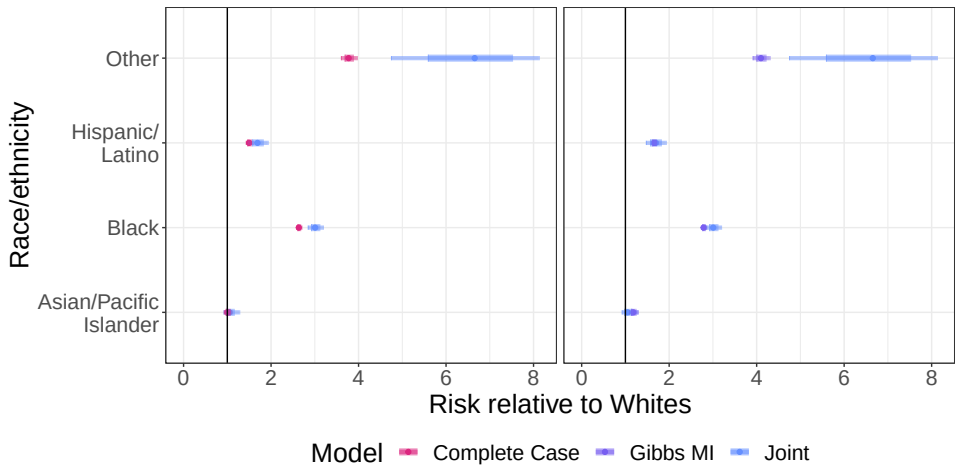


Figure 2.8: Relative risk of COVID-19. The inner intervals are 50% and the outer intervals are 80%.

Our use case is focused on missing race data in COVID-19 cases in Wayne County, Michigan from March 2020 through June 2020, which we suspect may have been NMAR. Wayne county saw the largest share of PCR-confirmed COVID-19 cases in the first wave of the pandemic, and also had a large share of cases that were missing race data, so it makes for an appropriate test bed for our method.

We ran a simulation study using Wayne county as the setting where we varied the proportion of cases with observed race from as high as 90% to as low as 20% to quantify the joint model's finite sample performance and to compare its performance against a complete-case analysis and two multiple imputation methods. The results showed that the joint model performed well in the 90%- through 40%-observed scenarios compared to the competing methods though its performance suffered in the 20% observed-data scenario. This leads us to conclude that in order to use the joint model effectively in sparse data scenarios, better priors will be needed; prior formulation for the model is an area of active research.

We then applied the models to a dataset comprising PCR-confirmed COVID-19 cases with incomplete race data from Wayne County between March 2020 through June 2020. The differences between the joint-model inferences and the multiple-imputation inferences suggest that the missingness process for race may be NMAR and that care must be taken when applying methods that assume data are MAR. Model results also suggest that cases in the Other category, which comprises those of mixed race, Native Americans, and Other races, are being undercounted in Wayne County.

2.7.1 Limitations

The biggest limitation of our analysis is the result of the joint model's dependence on census data for identifiability. This required the use of 2010 Decennial Census data, which is 10 years old, and may be systematically different than the true population distribution in Wayne County in 2020. The 2010 Decennial Census, however, is the most up-to-date source of spatially detailed population information, reflecting a broader limitation of any analysis that is dependent on decennial census data to estimate infection rates and relative risks. Because of this dependence on census population data, we were unable to model risk for race/ethnic categories that were potentially important in the Wayne County COVID-19 dataset, but for which census data were not available. MDSS collected information on Hispanic/Latino ethnicity separately from race, which resulted in missing covariate information in both categorical variables. Ideally we would have applied our method to multiple categorical covariates with missingness, but we were prevented from doing so due to the Census' coarse race and ethnicity categories. As stated in Section 2.5.1, if the Census recorded ethnicity and race separately, we would be able to model the effect of ethnicity separately from that of race and we could treat the missing ethnicity data separately from that of missing race data. Instead, we set race/ethnicity as being equal to the observed race if Hispanic/Latino ethnicity was missing, which can understate uncertainty in our posterior and could result in understating incidence for Hispanics and Latinos and overstating incidence in all other categories. We ran a separate analysis where we treated these observations as missing race; the incidence results largely agreed with the model we presented in the main text for Blacks, Latinos, and Asians, though we observed significant differences in the White and Other incidences. This analysis overstates uncertainty, because individuals for whom we observe race but not ethnicity can be only one of two categories, but our model in its current formulation treats these cases as potentially arising from any of the race/ethnicity categories. Until we have detailed 2020 Decennial Census results, we cannot model ethnicity and race separately.

We are also constrained by the mismatch of the 2010 Census question about sex and our dataset's definition of sex at birth. As Kennedy et al. (2020) argues, responses to the U.S. Census' question of sex may not correspond to sexes at birth. This mismatch can lead to bias in our parameter estimates and an understatement of uncertainty.

Another limitation of our model is that it assumes a Poisson sampling distribution for incident cases of disease. When cumulative incidence increases over time, as has occurred with COVID-19, a binomial sampling model may be more appropriate[4]. Similarly, our model assumes conditional independence between disease counts, which may not be appropriate as cumulative incidence grows[5]. Both of these reasons are why we decided to focus on the first wave of the pandemic,

[4]See appendix A.9 for an extension to a binomial likelihood

[5]See section 2.7.2.1 for more discussion

which is when the disease was relatively rare among the population of Wayne County and for which the violations of conditional independence assumption could be reasonably assumed to not lead to too much understatement of uncertainty.

2.7.2 Conclusion

Public health surveillance systems will always have to contend with missing data. Because the nature and causes of this missingness are likely to change over time and across disease systems, it is important that the methods used to address missingness are flexible and able to account for both MAR and NMAR covariates. In Michigan, missingness of categorical demographic data among COVID-19 cases has varied over the course of the pandemic. For example, some localities in our data reported as much as 40% of PCR-confirmed COVID-19 cases having missing data on race/ethnicity for the period of rapidly-increasing incidence from October 2020 to February 2021.

Our simulation study shows that complete-case analysis or naïve multiple imputation can yield uncertainty intervals that are too short to be useful and point estimators that can over- or under-state between-group relative risks. Our method represents a computationally tractable and analytically transparent alternative that performs well in many scenarios, as evidenced in our simulation studies as well as analysis of data from Wayne County, Michigan.Given the need for public health authorities to characterize risks of disease among different population groups in as close to real-time as possible, flexible, efficient methods such as ours, are urgently needed.

2.7.2.1 Extensions and future work

This work can serve as a foundation on which to build new joint-disease-missing-covariate models targeted to specific applications. Although the model presented here can give useful inferences in its own right in a variety of settings, despite its relative simplicity, domain-specific modifications may be appropriate. For example, future models could incorporate multi-level information on the public health and healthcare systems generating surveillance data to account more explicitly for contextual drivers of missingness.

The joint model can also be extended to account for infectious disease transmission dynamics and other sources of temporal and spatial autocorrelation. For example, the one-period Poisson sampling model can be extended to a time-series susceptible-infected-recovered (TSIR) model[6] or an endemic/epidemic model, both of which are discrete time analogues to classical susceptible-infected-recovered models (Held and Paul, 2012; Meyer and Held, 2014; Wakefield et al., 2019;

[6]TSIR models use a negative binomial likelihood; the code in appendix A.9 is easily extensible from binomial to negative binomial

Bauer and Wakefield, 2018; Keeling and Rohani, 2011). See Appendix A.10 for an instantiation of the model as an endemic/epidemic model.

When the disease becomes more widespread, potentially requiring a binomial likelihood, modeling the data a finer spatial resolution would make integrating over non-Poisson random variables more computationally efficient, particularly when combined with parallel computation of the likelihood. A dynamic programming implementation of the likelihood using binomial- instead of Poisson-distributed disease counts is included in appendix A.9. In order to regularize the model's inferences as the parameter space dimension increases in step with the spatial resolution, one can use a computationally-efficient log-Gaussian Cox process as a prior for the spatially-dependent parameters (Li et al., 2012; Simpson et al., 2016). Furthermore, relaxing the conditional independence between categories is possible through a latent Poisson model, as shown in Appendix A.10.3.

The dependence of the joint model on the availability of sufficiently detailed and recent census data can also be mitigated. For example, uncertainty in group-specific population denominators can be accounted using frequently updated population datasets, such as the American Community Survey, even if these data are not available at the same level of spatial granularity as decennial census data. An alternative route is to perform a "tipping point" sensitivity analysis (Liublinska and Rubin, 2014) to flag changes in census data that would lead to a substantive change in conclusions (e.g. a reversal of the sign for log relative-risk measures). Given the many degrees-of-freedom of census population data, and the critical role played by such data in population-based analyses of health and illness, this presents an interesting and important challenge that should be explored in future work.

Table 2.3: Table shows 50% posterior credible interval coverages and lengths for estimands of interest from the simulation study. Coverage proportion is calculated across 200 simulated datasets for each model/simulation scenario. Column headers for percentages (e.g. 20%) indicate the missing-data simulation scenario which corresponds to the statistic calculated in the table column; the simulation scenario corresponds to the proportion of cases observed with completely observed race covariates.

		50% interval coverage				50% mean interval length			
Parameter	Model	20%	60%	80%	90%	20%	60%	80%	90%
$\mathbb{I}_{\text{Blacks}}$	Complete Case	0.00	0.00	0.00	0.00	1e-04	2e-04	2e-04	3e-04
	Joint	0.37	0.48	0.46	0.51	2e-03	7e-04	5e-04	4e-04
	Ad-Hoc MI	0.05	0.13	0.03	0.00	3e-04	3e-04	3e-04	3e-04
	Gibbs MI	0.01	0.01	0.00	0.00	5e-04	3e-04	3e-04	3e-04
$\mathbb{I}_{\text{Hispanics/Latinos}}$	Complete Case	0.00	0.00	0.00	0.20	3e-04	5e-04	6e-04	7e-04
	Joint	0.26	0.47	0.48	0.42	6e-03	3e-03	2e-03	1e-03
	Ad-Hoc MI	0.07	0.15	0.18	0.00	9e-04	8e-04	8e-04	7e-04
	Gibbs MI	0.00	0.01	0.01	0.00	2e-03	1e-03	8e-04	7e-04
$\mathbb{I}_{\text{Others}}$	Complete Case	0.00	0.00	0.00	0.00	3e-04	5e-04	5e-04	6e-04
	Joint	0.12	0.44	0.48	0.41	6e-03	5e-03	3e-03	3e-03
	Ad-Hoc MI	0.07	0.07	0.01	0.00	1e-03	8e-04	7e-04	7e-04
	Gibbs MI	0.09	0.01	0.00	0.00	2e-03	9e-04	7e-04	7e-04
$\mathbb{I}_{\text{Whites}}$	Complete Case	0.00	0.00	0.00	0.00	1e-04	2e-04	3e-04	3e-04
	Joint	0.30	0.54	0.49	0.50	1e-03	7e-04	5e-04	4e-04
	Ad-Hoc MI	0.13	0.12	0.00	0.00	3e-04	3e-04	3e-04	3e-04
	Gibbs MI	0.14	0.01	0.00	0.00	4e-04	3e-04	3e-04	3e-04
$\mathbb{I}_{\text{Blacks}}/\mathbb{I}_{\text{Whites}}$	Complete Case	0.03	0.01	0.00	0.00	0.02	0.01	0.01	0.01
	Joint	0.48	0.54	0.52	0.48	0.06	0.03	0.02	0.01
	Ad-Hoc MI	0.04	0.10	0.01	0.00	0.01	0.01	0.01	0.01
	Gibbs MI	0.07	0.01	0.00	0.00	0.02	0.01	0.01	0.01
$\mathbb{I}_{\text{Hispanics/Latinos}}/$	Complete Case	0.09	0.08	0.33	0.53	0.03	0.02	0.02	0.02
$\mathbb{I}_{\text{Whites}}$	Joint	0.24	0.46	0.51	0.45	0.14	0.07	0.05	0.03
	Ad-Hoc MI	0.09	0.12	0.17	0.24	0.02	0.02	0.02	0.02
	Gibbs MI	0.00	0.09	0.07	0.05	0.05	0.02	0.02	0.02
$\mathbb{I}_{\text{Others}}/\mathbb{I}_{\text{Whites}}$	Complete Case	0.00	0.00	0.00	0.00	0.03	0.02	0.02	0.01
	Joint	0.12	0.43	0.47	0.41	0.16	0.12	0.09	0.07
	Ad-Hoc MI	0.09	0.06	0.00	0.00	0.02	0.02	0.02	0.02
	Gibbs MI	0.09	0.00	0.00	0.00	0.05	0.02	0.02	0.02

Table 2.5: Population summary in Wayne County, Michigan as of the 2010 Decennial Census

Race/Ethnicity	Total Pop.	Mean Age×Sex ×Race/Eth.×PUMA Pop.	Std. dev. PUMA Pop.	100× Ratio to White
Asian/Pacific Islander	45894	196	315	5
Black	732801	3132	3152	81
Hispanic/Latino	95260	407	757	11
Other	44449	190	150	5
White	902180	3855	3225	100

Table 2.6: Cumulative incidence of PCR-confirmed COVID-19 infections in Wayne County, MI from March 1, 2020 through June 30, 2020. Mean and variance for Total uses only observed-race/ethnicity cases. Mean and Variance columns rounded to zero digits.

Race/Ethnicity	Total Cases	Cumulative Incidence	Risk Relative to Whites	Mean	Variance	Prop. zero counts
Asian/Pacific Islander	229	0.005	1.0	1	3	0.55
Black	9,577	0.013	2.6	41	1904	0.02
Hispanic/Latino	708	0.007	1.5	3	34	0.37
Other	834	0.019	3.8	4	13	0.18
White	4,476	0.005	1.0	19	389	0.08
Missing	3,464	NA	NA	15	204	0.06
Total	19,288	0.011	2.1	14	697	0.24

CHAPTER 3

Principal Stratification for Vaccine Efficacy

3.1 Introduction

Phase 3 randomized, placebo-controlled clinical trials are the gold-standard by which vaccine candidates are assessed for efficacy and safety. Such trials are an important source of data about whether vaccines prevent outcomes such as infection and post-infection outcomes like secondary transmission, severe illness, or death. For example, COVID-19 vaccination trials like Polack et al. (2020) and Baden et al. (2021) measured vaccine efficacy against symptomatic disease, as well as severe illness and death. Principal stratification, developed in Frangakis and Rubin (2002), may be used to partition the intention-to-treat effect of vaccination on an outcome like hospitalization into vaccine efficacy against infection and vaccine efficacy against hospitalization given infection in the always-infected stratum; these separate effects help policy makers optimize vaccination programs, communicate with the public, allocate scarce resources, and guide future pharmaceutical therapeutic development (Lipsitch and Kahn, 2021). Methods to infer principal effects for vaccine efficacy were first developed for continuous post-infection outcomes in Gilbert et al. (2003); Jemiai et al. (2007); Shepherd et al. (2006, 2007), and further developed for binary post-infection outcomes in Hudgens and Halloran (2006).

Unfortunately, vaccine efficacy against post-infection outcomes, binary or otherwise, is not generally identifiable, even under the assumption that vaccine efficacy against infection is non-negative almost-surely (monotonicity). Moreover, the method requires that both infection and post-infection outcomes are perfectly measured. Neither monotonicity nor error-free measurements can be assumed to hold in vaccine trials. Monotonicity can be violated if a vaccine increases the per-exposure probability of infection for a participant (Gilbert et al., 2003), which is possible in influenza vaccine trials where the vaccine targets a different antigen than the circulating strain. Another way monotonicity can be violated is if vaccination increases exposure for certain participants. This can occur in a double-blinded placebo-controlled study where the vaccine is reactogenic and leads to some participants in the vaccine group becoming unblinded. Measurement error is com-

mon in vaccine trials due to the imperfect nature of diagnostic tests for infection (Kissler et al., 2021; Wang et al., 2020). Post-infection outcomes like symptoms may also be observed with error. For example, in an influenza vaccine trial, many different viruses circulate during influenza season that produce similar symptom profiles.

We develop novel methodology to point identify vaccine efficacy against binary post-infection outcomes without assuming monotonicity while allowing infection and post-infection outcomes to be misclassified. Our framework immediately generalizes to multiple treatments as we will show. We capitalize on the fact that many randomized trials for vaccines are run as multi-center trials (i.e. geographically disparate study sites) (Francis, 1982; Longini Jr et al., 2000; The FUTURE II Study Group, 2007; Halloran et al., 2010; Baden et al., 2021; Polack et al., 2020), and typically measure pretreatment covariates relevant to infection. We build on literature for identifying principal stratum effects with covariates (Rubin, 2006; Ding et al., 2011; Jiang et al., 2016), on inferring principal stratum effects in multisite randomized trials (Wang et al., 2017; Yuan et al., 2019; Luo et al., 2023), on using covariates to hone large-sample nonparametric bounds (Zhang and Rubin, 2003; Grilli and Mealli, 2008; Long and Hudgens, 2013), and on identifying causal estimands under unmeasured confounding (Miao et al., 2018; Shi et al., 2020). Our method also fits into recent literature on inferring causal estimands under measurement error (Jiang and Ding, 2020) and on identification of latent variable models (Ouyang and Xu, 2022).

We show that our method can be used to design randomized trials for comparison of multiple vaccines against a control, which will be a necessity for public health agencies in future pandemics as well as during the COVID-19 pandemic. Due to recent updates to regulatory guidance from the European Medicines Agency, the authority that authorizes pharmaceuticals in the European Union, principal effects are acceptable target estimands in randomized clinical trials and principal stratification is an acceptable analysis method for these trial data (Bornkamp et al., 2021; Lipkovich et al., 2022). This means that our methodology is directly applicable to the design and analysis of clinical trials for vaccines. As noted by several authors, vaccine efficacy against post-infection outcomes is mathematically analogous to the widely-studied survivor average treatment effects (Ding et al., 2011; Tchetgen Tchetgen, 2014; Ding and Lu, 2017), so our methodology can be readily used outside the domain of vaccine efficacy.

3.2 Vaccine efficacy in two-arm multi-center trials

Two-arm multi-center trials, or trials run in tandem across several disparate health centers where each participant is randomly assigned to receive a vaccine or a placebo, are the most common vaccine efficacy study designs. To fix ideas, we will consider the example of an influenza vaccine trial, where researchers are interested in understanding vaccine efficacy against influenza infection and

vaccine efficacy against severe illness caused by influenza infection. Crucially, it is not possible to perfectly observe influenza infection or severe illness. Instead, researchers are limited to using imperfect tests for infection, like polymerase chain reaction (PCR) tests, or serology to detect a participant's infection status. These methods measure infection with error, with varying levels of sensitivity and specificity. For example, PCRs for COVID-19 have very high specificity, but tend to have sensitivities in the range of 0.6 to 0.8 due to variation among patients in how the virus populates the nasal cavity, variation in swab quality, and viral RNA dynamics (Kissler et al., 2021; Wang et al., 2020). Depending on the severity of the post-infection outcome, these outcomes may also be mismeasured. For instance, a high proportion of participants report symptoms in vaccine efficacy studies, despite many of these participants testing negative for the target disease. In the presence of high-sensitivity tests, this necessarily means that specificity of symptoms following infection is below 1. This is because it is possible for participants to develop influenza-like severe illness from non-influenza viruses during a clinical trial. Thus our framework assumes that observed infection and severe illness are noisy proxies for true unobservable infection and severe illness states. The next section outlines the data structure for each participant.

Suppose there are n participants in the trial, and we observe the following sextuplet for each participant i: $(\tilde{S}_i, \tilde{Y}_i, Z_i, R_i, A_i, X_i)$, where Z_i is a binary variable representing treatment assignment, $\tilde{S}_i$ is binary influenza test result, and $\tilde{Y}_i$ is observed binary severe illness status. R_i is a categorical variable indicating the center with which each participant is associated, A_i is a discrete pre-treatment covariate related to infection under treatment and control, and X_i is a univariate discrete pre-treatment covariate that may combine several distinct covariates like age, sex, occupation, and pre-existing conditions. Let R_i take values from 1 to N_r, A_i take values from 1 to N_a, and X_i take values from 1 to N_x. $Z_i = 1$ for individuals assigned to receive vaccination, and $Z_i = 0$ for individuals assigned to placebo; for $\tilde{S}_i$ and $\tilde{Y}_i$, 1 indicates the presence of infection and severe illness, respectively, for individual i, while 0 indicates the absence.

Let S_i be the latent influenza infection state, and Y_i be the latent influenza-caused severe illness state for each participant.

Assumption 1 (Non-differential Misclassification Errors)**.** *Misclassification errors for $\tilde{S}_i, \tilde{Y}_i$ are conditionally independent of all else given the true values S_i, Y_i or*

$$\tilde{S}_i \perp\!\!\!\perp Z_i, S_i^{P_0}, R_i, A_i, Y_i, X_i \mid S_i, \quad \tilde{Y}_i \perp\!\!\!\perp Z_i, S_i^{P_0}, R_i, A_i, S_i, X_i \mid Y_i.$$

Under Assumption 1, we may completely characterize the distributions of the noisy outcomes $\tilde{S}_i, \tilde{Y}_i$ via the following four unknown parameters $\text{sn}_S = P(\tilde{S}_i = 1 \mid S_i = 1), \text{sp}_S = P(\tilde{S}_i = 0 \mid S_i = 0)$ and $\text{sn}_Y = P(\tilde{Y}_i = 1 \mid Y_i = 1), \text{sp}_Y = P(\tilde{Y}_i = 0 \mid Y_i = 0)$, or the respective sensitivities and specificities for infection and the post-infection outcome. This assumption can be loosened for

infection misclassification as explored in Section 3.5.

We use the Neyman-Rubin causal model to define counterfactual variables S_i, infection with influenza, and Y_i, severe illness caused by influenza, as partially-observed realizations of counterfactual outcomes (Neyman, 1923; Rubin, 1974, 1978; Holland, 1986). For an extensive review of statistical approaches to causal inference through the lens of missing data see Ding and Li (2018). Let any potential treatment plan for all n individuals in the trial be the length-n binary vector $\mathbf{z}$, where the i^{th} element is the potential treatment status of the i^{th} participant. Accordingly, each individual is associated with a binary counterfactual infection outcome, $S_i(\mathbf{z})$, and a counterfactual severe illness outcome, $Y_i(\mathbf{z}, S_i(\mathbf{z}))$, under treatment status $\mathbf{z}$. Let the observed treatment status for all n individuals in the trial be the length-n binary vector $\mathbf{Z}$, where the i^{th} element is the assigned treatment of the i^{th} participant.

Our causal model enforces the constraint that an uninfected individual cannot have severe illness caused by influenza infection. In other words, post-infection outcomes are defined such that they are *caused* by infection from a pathogen of interest (Gilbert et al., 2003; Hudgens and Halloran, 2006). Then $Y_i(\mathbf{z}, 0)$ is undefined for all $\mathbf{z}$, and is denoted as $Y_i(\mathbf{z}, 0) = \star$. $Y_i(\mathbf{z}, S_i(\mathbf{z}))$ is defined as a binary variable only when $S_i(\mathbf{z}) = 1$, or, equivalently, $Y_i(\mathbf{z}, 1)$. For the remainder of the chapter we assume that $S_i(\mathbf{z}) = S_i(\mathbf{z}')$ and $Y_i(\mathbf{z}, S_i(\mathbf{z})) = Y_i(\mathbf{z}', S_i(\mathbf{z}'))$ if $\mathbf{z}_i = \mathbf{z}'_i$. Therefore, we assume the Stable Unit Treatment Value Assumption (SUTVA) holds:

Assumption 2 (SUTVA)**.** *There is only one version of each treatment, and counterfactual outcomes are a function of only a unit's respective treatment status, z.*

SUTVA can be satisfied for vaccine efficacy trials by restrictions on participants and recruitment (Gilbert et al., 2003). Furthermore, recruited participants are a small fraction of the total population at risk of infection (Zhang et al., 2009). Thus the vector $(S_i(1), Y_i(1, S_i(1)), S_i(0), Y_i(0, S_i(0)))$ is the complete definition of counterfactual outcomes under vaccination and placebo for each individual in the trial. Given Assumption 2, the latent realized values of the counterfactual variables are as follows:

$$S_i = Z_i S_i(1) + (1 - Z_i) S_i(0), \quad Y_i = Z_i Y_i(1, S_i(1)) + (1 - Z_i) Y_i(0, S_i(0)). \tag{3.1}$$

A second assumption we will make for the rest of the chapter is that the study is a randomized experiment. This means that all trial participants have positive probabilities of being assigned to either vaccine or placebo, and that treatment assignment is unconfounded (Imbens and Rubin, 2015).

Assumption 3 (Random treatment assignment)**.** *The probability of being assigned to treatment for*

each individual lies strictly between 0 *and* 1*:*

$$0 < P(Z_i = 1 \mid S_i(1), S_i(0), Y_i(1, S(1)), Y_i(0, S_i(0))) < 1.$$

Treatment assignment is independent of all potential outcomes, or

$$S_i(1), S_i(0), Y_i(1, S_i(1)), Y_i(0, S_i(0)) \perp\!\!\!\perp Z_i.$$

With these assumptions, we can define several estimable causal estimands of interest.

Definition 3.2.1 (Vaccine efficacy against infection)**.**

$$\mathrm{VE}_S = \mathbb{E}\left[S_i(0) - S_i(1)\right] / \mathbb{E}\left[S_i(0)\right], \text{ and}$$

Definition 3.2.2 (Intention-to-treat vaccine efficacy against severe illness)**.**

$$\mathrm{VE}_{ITT} = \mathbb{E}\left[Y_i(0) - Y_i(1)\right] / \mathbb{E}\left[Y_i(0)\right].$$

With Assumptions 2 to 3, the following identities hold

$$\mathbb{E}\left[S_i(z)\right] = \mathbb{E}\left[S_i \mid Z_i = z\right], \quad \mathbb{E}\left[Y_i(z)\right] = \mathbb{E}\left[Y_i \mid Z_i = z\right].$$

These expressions, along with the identities

$$\mathbb{E}\left[S_i = 1 \mid Z_i = z\right] = \frac{\mathbb{E}\left[\tilde{S}_i = 1 \mid Z_i = z\right] + \mathrm{sp}_S - 1}{\mathrm{sp}_S + \mathrm{sn}_S - 1}, \; \mathbb{E}\left[Y_i = 1 \mid Z_i = z\right] = \frac{\mathbb{E}\left[\tilde{Y}_i = 1 \mid Y_i = z\right] + \mathrm{sp}_Y - 1}{\mathrm{sp}_Y + \mathrm{sn}_Y - 1}$$

suggest plug-in ratio estimators for each estimand if the sensitivities and specificities are known.

3.2.1 Conditional effects and principal stratification

We might be tempted to define vaccine efficacy against severe illness by comparing the rate of severe illness in vaccinated participants to that of the unvaccinated among patients who have been infected. Formally, these quantities are $\mathbb{E}\left[Y_i \mid S_i = 1, Z_i = 1\right]$ and $\mathbb{E}\left[Y_i \mid S_i = 1, Z_i = 0\right]$. However, as argued in Frangakis and Rubin (2002), the set of participants $\{i \mid S_i = 1, Z_i = 1\}$ is different from the set of participants $\{i \mid S_i = 1, Z_i = 0\}$. A causal estimand is defined such that the *only* source of variation is the change from $z = 1$ to $z = 0$, as in the numerator for VE_S: $\mathbb{E}\left[S_i(0) - S_i(1)\right]$. If we define the estimand $\mathbb{E}\left[Y_i(0) \mid S_i(0) = 1\right] - \mathbb{E}\left[Y_i(1) \mid S_i(1) = 1\right]$, this is a comparison between different individuals. This means that this quantity mixes changes between treatments and changes between the two groups being compared. If the quantity is nonzero, this

difference could potentially be explained by the difference in individual traits between the two groups rather than the difference treatment.

In terms of counterfactual outcomes, the set $\{i \mid S_i = 1, Z_i = 1\}$ includes trial participants with $(S_i(1) = 1, S_i(0) = 1)$ and those with $(S_i(1) = 1, S_i(0) = 0)$, while the set $\{i \mid S_i = 1, Z_i = 0\}$ includes those $(S_i(1) = 1, S_i(0) = 1)$ and those with $(S_i(1) = 0, S_i(0) = 1)$. It may be the case that the characteristics of patients protected by the vaccine, or $(S_i(1) = 0, S_i(0) = 1)$, are fundamentally different from the characteristics of those who are harmed by the vaccine, or $(S_i(1) = 1, S_i(0) = 0)$. Thus, a causal estimand that compares the outcomes of these two groups does not represent a valid causal estimand because the two groups may be systematically different in characteristics that predict the baseline risk of severe disease when infected, or $\mathbb{E}[Y_i(0) \mid S_i = 1]$.

The solution is to partition the participants into four strata defined by the individual's complete set of potential infection outcomes, or $(S_i(1), S_i(0))$. Then potential severe illness outcomes can be compared conditional on stratum membership so that the group in which the causal estimand is defined is homogeneous in terms of potential outcomes. Formally, let principal stratum, $S_i^{P_0}$, be defined as the ordered vector of counterfactual infection outcomes for unit i, or

$$S_i^{P_0} = (S_i(1), S_i(0)),\ S_i(z) \in \{0, 1\}$$

and let the set of all principal strata be denoted as $\mathcal{S}$. Then $\mathcal{S} \equiv \{0, 1\}^2$.

The only group in which a causal comparison can be made is the $(1, 1)$ stratum, otherwise known as the always-infected stratum. This is because it is the only stratum in which participants have a well-defined post-infection outcome under vaccination and under placebo. Thus we may define vaccine efficacy against severe illness as:

Definition 3.2.3 (Vaccine efficacy against post-infection outcome)**.**

$$\mathrm{VE}_I = 1 - \mathbb{E}\left[Y_i(1) \mid S_i^{P_0} = (1, 1)\right] / \mathbb{E}\left[Y_i(0) \mid S_i^{P_0} = (1, 1)\right].$$

The severe illness vaccine efficacy can be seen as the percent change in risk of severe illness conferred by receiving a vaccine conditional on belonging to the always-infected principal stratum. VE_I is a principal effect as defined in Frangakis and Rubin (2002) because it is conditional on a principal stratum.

The fundamental problem of causal inference (Holland, 1986), namely that we cannot observe all counterfactual outcomes for the same individual, prevents the development of a simple ratio estimator because the observed data cannot determine which individuals belong to the always-infected stratum (Hudgens and Halloran, 2006). In fact, in the next section, we will show that no

single parameter of the probabilistic model implied by Assumptions 2 to 3 can be identified by the data, and that definition 3.2.3 is also not identified by the data.

3.2.2 Identifiability of principal stratum effects

Identifiability is an asymptotic characteristic of a statistical model by which each dataset maps to a unique parameter. It is a necessary assumption for much asymptotic theory, including Bayesian posterior consistency, (Keener, 2010; Van der Vaart, 2000), and thus is useful to characterize. Formally, as defined in Rothenberg (1971):

Definition 3.2.4 (Parameter identifiability)**.** *A parameter $\theta \in \Theta$ is identifiable if there does not exist a distinct parameter value $\theta' \in \Theta$ for which the density $f(y \mid \theta) = f(y \mid \theta')$ for all observations y.*

In order to investigate this property, we must first define how the counterfactual model governs the latent variable distribution. Let $p_{syz} = P(S_i = s, Y_i = y \mid Z_i = z)$ be the observable probabilities of an infection outcome s, a post-infection outcome y given a vaccination status z, and let p_{s+z} be the marginal probability of infection given vaccination, or $P(S_i = s \mid Z_i = z)$. We will now define the parameters for the counterfactual probability model. Let $u \in \{(0,0),(1,0),(0,1),(1,1)\}$, $\theta_u = P(S_i^{P_0} = u)$, and $\beta_z^u = P(Y_i(z) = 1 \mid S_i^{P_0} = u)$. The map from the model parameters to the observable probabilities is:

$$p_{110} = \theta_{(0,1)}\beta_0^{(0,1)} + \theta_{(1,1)}\beta_0^{(1,1)}, \qquad p_{111} = \theta_{(1,0)}\beta_1^{(1,0)} + \theta_{(1,1)}\beta_1^{(1,1)}$$
$$p_{100} = \theta_{(0,1)}\left(1-\beta_0^{(0,1)}\right) + \theta_{(1,1)}\left(1-\beta_0^{(1,1)}\right), \quad p_{101} = \theta_{(1,0)}\left(1-\beta_1^{(1,0)}\right) + \theta_{(1,1)}\left(1-\beta_1^{(1,1)}\right).$$

The joint distribution of the observed data has only 4 independent quantities, but the probability model has 7 parameters. Thus, the observable probabilities do not uniquely map to counterfactual probability model parameters, and the model parameters are not identified.

In this formulation VE_I can be written in terms of the observable probabilities and three unidentified parameters $\beta_2^{(0,1)}, \beta_1^{(1,0)}, \theta_{(1,1)}$.

$$\mathrm{VE}_I = 1 - \frac{\beta_1^{(1,1)}}{\beta_0^{(1,1)}} = 1 - \frac{p_{111} - \beta_1^{(1,0)}(p_{1+1} - \theta_{(1,1)})}{p_{110} - \beta_0^{(0,1)}(p_{1+0} - \theta_{(1,1)})}. \tag{3.2}$$

In fact, the structure of the model is such that no single counterfactual model parameter is identifiable without restrictions on the principal stratum proportions. In order to identify Equation (3.2) with observed data, one needs to learn $\theta_{(1,1)}, \beta_1^{(1,0)}, \beta_0^{(0,1)}$. The form of the estimand suggests several identification strategies.

One strategy is to assume that no participant is infected under the vaccine and uninfected under the placebo. This assumption leads to $\theta_{(1,0)} = 0$, $\theta_{(1,1)} = p_{1+1}$, and $\beta_1^{(1,1)}$ as p_{111} (Gilbert et al.,

2003; Hudgens and Halloran, 2006; Zhou et al., 2016). A consequence of this assumption is that vaccine efficacy against infection is nonnegative with probability 1. VE_I then simplifies to

$$\mathrm{VE}_I = 1 - \frac{p_{111}}{p_{110} + \beta_0^{(0,1)}(p_{1+1} - p_{1+0})}, \tag{3.3}$$

If we further assume that the distribution of severe illness is mean-independent of study site, we can use the variation in infection rates by study site to identify $\beta_0^{(0,1)}$ (Jiang et al., 2016; Yuan et al., 2019).

In some trials monotonicity may be appropriate, but it is not an assumption that is made when assessing vaccine efficacy against infection (El Sahly et al., 2021). Accordingly, two separate analyses must be performed to assess the two efficacy estimands. The price for two-step procedures is two-fold: two-step estimators are often less statistically efficient, and it may be harder to communicate results to stakeholders because two sets of assumptions are needed. Assuming monotonicity also complicates the use of prior vaccine efficacy trial results in priors for future studies. If past trials have confidence intervals for vaccine efficacy against infection that include negative values, it is not clear how to incorporate this information into a prior over a parameter that excludes negative values by design.

Monotonicity may be violated if some participants are inadvertently unblinded during a vaccine trial due to post-vaccine side effects (also known as reactogenicity); participants who experience strong side effects may infer they are in the vaccine arm of the trial and may be more likely to be exposed to influenza than if they had been in the placebo group.

For example, suppose that in addition to the post-vaccination infection outcome, $S_i(z)$, there are post-vaccination binary exposure counterfactual variables $E_i(z)$ for each trial participant. Assume that exposure is a necessary condition for infection, (i.e. $S_i(z) = 0 \mid E_i(z) = 0$ with probability 1). Furthermore, assume conditional monotonicity for $S_i(z)$: $S_i(1) \leq S_i(0) \mid (E_i(1), E_i(0))$. Despite the conditional monotonicity assumption, marginalizing over $E_i^{P_0} = (E_i(1), E_i(0))$ does not preserve monotonicity:

$$P(S_i(1) = 1, S_i(0) = 0) = P(S_i(1) = 1, S_i(0) = 0 \mid E_i^{P_0} = (1, 0))P(E_i(1) = 1, E_i(0) = 0).$$

Thus $P(S_i(1) = 1, S_i(0) = 0) \neq 0$ as long as $P(E_i(1) = 1, E_i(0) = 0) \neq 0$.

This means that while it may be appropriate to assume that given an exposure to the influenza virus, the vaccine does not increase the probability of infection, if vaccination somehow increases exposure, then marginalizing over exposure will yield $\theta_{(1,0)} > 0$.

The identifiability of the estimand is further complicated by the fact that infection and post-infection outcomes are observable only through proxy variables, $\tilde{S}_i$ and $\tilde{Y}_i$ with unknown sensitiv-

ities and specificities. Let the observable probabilities, $q_{syz} = P(\tilde{S}_i = s, \tilde{Y}_i = y \mid Z_i = z)$ and be defined as

$$\begin{aligned} q_{syz} = &\, \mathrm{sn}_S^s (1-\mathrm{sn}_S)^{1-s} \mathrm{sn}_Y^y (1-\mathrm{sn}_Y)^{1-y} p_{11z} + \mathrm{sn}_S^s (1-\mathrm{sn}_S)^{1-s} \mathrm{sp}_Y^{1-y} (1-\mathrm{sp}_Y)^y p_{10z} \\ &+ \mathrm{sp}_S^{1-s} (1-\mathrm{sp}_S)^s \mathrm{sp}_Y^{1-y} (1-\mathrm{sp}_Y)^y p_{0*z}. \end{aligned}$$

Let $q_{+yz} = P(\tilde{Y}_i = y \mid Z_i = z)$, $q_{s+z} = P(\tilde{S}_i = s \mid Z_i = z)$, and $\tilde{\beta}_z^u = P(\tilde{Y}_i(z) = 1 \mid S_i^{P_0} = u)$. Under this probability model, the causal estimand of interest, or vaccine efficacy against severe illness caused by influenza, is

$$\mathrm{VE}_I = 1 - \frac{q_{+11} - (1-\mathrm{sp}_Y) - (\tilde{\beta}_1^{(1,0)} - (1-\mathrm{sp}_Y))(\frac{q_{1+1} - (1-\mathrm{sp}_S)}{\mathrm{sn}_S + \mathrm{sp}_S - 1} - \theta_{(1,1)})}{q_{+10} - (1-\mathrm{sp}_Y) - (\tilde{\beta}_0^{(0,1)} - (1-\mathrm{sp}_Y))(\frac{q_{1+0} - (1-\mathrm{sp}_S)}{\mathrm{sn}_S + \mathrm{sp}_S - 1} - \theta_{(1,1)})} \tag{3.4}$$

This expression involves nonidentifiable causal parameters $\tilde{\beta}_0^{(0,1)}, \tilde{\beta}_1^{(1,0)}, \theta_{(1,1)}$ as well as the unidentifiable measurement error parameters $\mathrm{sn}_S, \mathrm{sp}_S, \mathrm{sp}_Y$. It does not, however, involve the parameter sn_Y. This is due to the fact that for any binary random variable Q and its noisy measurement, $\tilde{Q}$, we have the following identity:

$$P(\tilde{Q} = 1) = (\mathrm{sn}_Q + \mathrm{sp}_Q - 1) P(Q = 1) + 1 - \mathrm{sp}_Q. \tag{3.5}$$

Thus, one cannot identify the estimand of interest without additional information, though it suggests that identifiability of the causal estimand can be achieved without identifying sn_Y. The next section outlines how we can use the structure of multisite randomize trials for vaccine efficacy to infer post-infection outcome vaccine efficacy.

3.2.3 Incorporating study-site and covariate information

Two factors are necessary for influenza infection (and any infection, for that matter): exposure to the pathogen and susceptibility to infection given exposure. Exposure is a post-treatment event in randomized trials, and variation in exposure can lead to variation in rates of infection, and, subsequently, principal strata. Variation in exposure can occur due to variation in disease prevalence during the duration of the trial, and multi-scale spatial variation of disease prevalence is a hallmark of infectious disease (Bauer and Wakefield, 2018). Thus, if study sites are sufficiently separated geographically, it is reasonable to expect that the study site variable is predictive of exposure during the duration of the trial. This variation in exposure should lead to variation in principal strata due to differences in exposure. Thus it is reasonable to assume that $R_i \not\perp\!\!\!\perp S_i^{P_0} \mid X_i$.

Now we turn to susceptibility to infection given exposure. In influenza trials it is common to

measure the pre-season, pre-vaccination (i.e. baseline) antibody concentrations via hemagglutination inhibition (HAI) assays or neuraminidase inhibition (NAI) assays (Monto et al., 2009). These assays are categorical measurements generated from serial two-fold dilutions of patient serum samples (Zelner et al., 2019). These measurements are related to the participants' susceptibility to infection given exposure because they measure immune markers of past infections and/or past influenza vaccinations. Given the fact that the participants will be inoculated against influenza, it is again reasonable that $A_i \not\perp\!\!\!\perp S_i^{P_0} \mid X_i$. The structure of multisite randomized trials is such that we may make two further assumptions about the joint distribution of covariate values A_i and potential post-infection outcomes $(Y_i(1), Y_i(0))$. Specifically, because patient recruitment is a tightly controlled procedure with unified inclusion criteria between study sites (Weinberger et al., 2001), it is reasonable to assume that participants' covariate values and potential post-infection outcomes are are exchangeable (Saarela et al., 2023). Formally,

Assumption 4 (Covariate homogeneity)**.** *A_i is conditionally independent of the study site and treatment receipt given the principal stratum and baseline covariates, or $A_i \perp\!\!\!\perp R_i, Z_i \mid S_i^{P_0}, X_i$,*

and

Assumption 5 (Causal Homogeneity)**.** *Conditional on principal stratum $S_i^{P_0}$ and A_i, X_i, the potential outcomes $(Y_i(1), Y_i(0))$ are independent of R_i, or $(Y_i(1), Y_i(0)) \perp\!\!\!\perp R_i \mid S_i^{P_0}, A_i, X_i$.*

Assumption 4 equates to assuming that individuals' covariate measurements A_i are exchangeable within strata defined by $(X_i = x, S_i^{P_0} = u)$, and Assumption 5 is equivalent to assuming conditional exchangeability of $Y_i(z)$ within strata defined by $(X_i = x, S_i^{P_0} = u, A_i = k)$ (Saarela et al., 2023). In the event that Assumption 4 does not hold, a parametric model for $A_i \mid R_i, S^{P_0}, X_i$ may be used.

Under Assumptions 1 to 5, we can define the joint distribution of influenza test results, reported severe illness, and pre-season antibody concentration given treatment assignment and study site membership.

Let $\theta_u^{r,x} = P(S_i^{P_0} = u \mid R_i = r, X_i = x)$ for $u \in \{(0,0),(1,0),(0,1),(1,1)\}$. Let $u_z = P(S_i(z) = 1 \mid S_i^{P_0} = u)$. Let $a_k^{u,x} = P(A_i = k \mid S_i^{P_0} = u, X_i = x)$, and $\beta_{z,k}^{u,x} = P(Y_i(z) = 1 \mid S_i^{P_0} = u, A_i = k, X_i = x)$. Further, recall $\text{sn}_S = P(\tilde{S}_i = 1 \mid S_i = 1), \text{sp}_S = P(\tilde{S}_i = 0 \mid S_i = 0)$ and $\text{sn}_Y = P(\tilde{Y}_i = 1 \mid Y_i = 1), \text{sp}_Y = P(\tilde{Y}_i = 0 \mid Y_i = 0)$. Let the observable probabilities, $q_{sk|zrx} = P(\tilde{S}_i = s, A_i = k \mid Z_i = z, R_i = r, X_i = x)$ be defined as:

$$q_{sk|zrx} = \text{sn}_S^s(1-\text{sn}_S)^{1-s}\sum_{u|u\in\mathcal{S},u_z=1} a_k^{u,x}\theta_u^{r,x} + \text{sp}_S^{1-s}(1-\text{sp}_S)^s\sum_{u|u\in\mathcal{S},u_z=0} a_k^{u,x}\theta_u^{r,x}.$$

The marginal probability of observing a positive influenza test result is the sum of the probability of truly being infected with influenza and the influenza test correctly diagnosing the infection and

the probability of being influenza-free and the influenza test incorrectly returning a positive result.

Meanwhile, the observable probabilities $q_{y|kzrx} = P(\tilde{Y}_i = y \mid Z_i = z, R_i = r, A_i = k, X_i = x)$ are defined:

$$\begin{aligned} q_{y|kzrx} &= \mathrm{sn}_Y^{y}(1-\mathrm{sn}_Y)^{1-y} \textstyle\sum_{u|u\in\mathcal{S},u_z=1} \beta_{z,k}^{u,x}\theta_u^{r,x} \\ &\quad + \mathrm{sp}_Y^{1-y}(1-\mathrm{sp}_Y)^{y}(\textstyle\sum_{u|u\in\mathcal{S},u_z=1}(1-\beta_{z,k}^{u,x})\theta_u^{r,x} + \sum_{u|u\in\mathcal{S},u_z=0}\theta_u^{r,x}). \end{aligned}$$

The marginal probability of observing severe illness caused by influenza infection is the sum over the probability of being infected, having severe illness from influenza, and reporting the illness, the probability of being either being infected, avoiding severe illness from influenza or not being infected, but in either case incorrectly reporting the illness.

In designing a clinical trial in which a primary or secondary endpoint is severe illness, limiting the definition of severe illness to encompass only the most extreme illness can at once increase the sensitivity and specificity of reporting.

3.2.3.1 Identifiability of expanded model

Given assumptions Assumptions 1 to 5, we will show that the joint variation in observed antibody concentrations and infection rates across study sites identifies the joint distribution of principal strata proportions and covariate values by study site. Then, given sufficient variation in principal strata proportions between study sites, the distribution of post-infection potential outcomes can be identified as well.

Identifiability results from arranging the observed distributions $q_{sk|zr}$ as 3-way arrays and using a modified tensor decomposition uniqueness theorem from Kruskal (1977). Kruskal's theorem defines sufficient conditions for the uniqueness of the *triple product* decomposition of L, where this product is defined in Definition 3.2.5.

Definition 3.2.5 (Array triple product)**.** *Let the array triple product with resulting array $L \in \mathbb{R}^{I\times J\times K}$ be defined between matrices $A \in \mathbb{R}^{I\times M}$, $B \in \mathbb{R}^{J\times M}$, $C \in \mathbb{R}^{K\times M}$. The operation is represented as $L = [A, B, C]$. As a result, the $(i, j, k)^{\text{th}}$ element of L, L_{ijk}, is defined the sum of three-way-products of elements a_{im}, b_{jm}, c_{km}, i.e.:*

$$L_{ijk} = \sum_{m=1}^{M} a_{im}b_{jm}c_{km}.$$

The sufficient conditions concern the *Kruskal ranks* of the matrices A, B, C, defined in Definition 3.2.6.

Definition 3.2.6 (Kruskal rank)**.** *Let the Kruskal rank of a matrix $B \in \mathbb{R}^{I \times M}$ be $k_B \in [0, 1, 2, \ldots, M]$, and let k_B be the maximum integer such that every set of k_B columns of B are linearly independent.*

Kruskal rank is stricter than matrix rank. To see why, consider a matrix with M columns of which two are repeated. At most the rank of the matrix can be $M - 1$, but Kruskal rank can be at most 1. A corollary of the definition is that if a matrix is full column rank, its Kruskal rank equals its column rank.

Let L be the 3-way array representing $q_{sk|zrx}$, where we fix $X_i = x$ for each unique value of X_i. The array's dimensions are $4 \times N_a \times N_r$ and is defined so that the $(j, k, r)^{\text{th}}$ element $P(S_i = \mathbb{1}_{j \leq 2}, A_i = k \mid Z = j - 1 \bmod 2, R_i = r, X_i = x)$. If we look at the matrix that results from fixing the third array index, also known as the *3-slab* and denoted as $L_r \in \mathbb{R}^{4 \times N_a}$, we can see a possible decomposition of this array. Let $\sum_{u_z = s} a_k^{u,x} \theta_u^{r,x}$ denote the sum over elements $u \in \{(0,0), (1,0), (0,1), (1,1)\}$ such that $P(S_i(z) = 1 \mid S_i^{P_0} = u)$. The let L_r be defined as

$$
\begin{array}{c}
\begin{matrix} (a=1) & & \cdots & & (a=N_a) \end{matrix} \\
\begin{bmatrix}
(1-\mathrm{sp}_S)\sum_{u_0=0} a_1^{u,x}\theta_u^{r,x} + \mathrm{sn}_S \sum_{u_0=1} a_1^{u,x}\theta_u^{r,x} & \cdots & (1-\mathrm{sp}_S)\sum_{u_0=0} a_{N_a}^{u,x}\theta_u^{r,x} + \mathrm{sn}_S \sum_{u_0=1} a_{N_a}^{u,x}\theta_u^{r,x} & (s=1, z=0) \\
(1-\mathrm{sp}_S)\sum_{u_1=0} a_1^{u,x}\theta_u^{r,x} + \mathrm{sn}_S \sum_{u_1=1} a_1^{u,x}\theta_u^{r,x} & \cdots & (1-\mathrm{sp}_S)\sum_{u_1=0} a_{N_a}^{u,x}\theta_u^{r,x} + \mathrm{sn}_S \sum_{u_1=1} a_{N_a}^{u,x}\theta_u^{r,x} & (s=1, z=1) \\
\mathrm{sp}_S \sum_{u_0=0} a_1^{u,x}\theta_u^{r,x} + (1-\mathrm{sn}_S)\sum_{u_0=1} a_1^{u,x}\theta_u^{r,x} & \cdots & \mathrm{sp}_S \sum_{u_0=0} a_{N_a}^{u,x}\theta_u^{r,x} + (1-\mathrm{sn}_S)\sum_{u_0=1} a_{N_a}^{u,x}\theta_u^{r,x} & (s=0, z=0) \\
\mathrm{sp}_S \sum_{u_1=0} a_1^{u,x}\theta_u^{r,x} + (1-\mathrm{sn}_S)\sum_{u_1=1} a_1^{u,x}\theta_u^{r,x} & \cdots & \mathrm{sp}_S \sum_{u_1=0} a_{N_a}^{u,x}\theta_u^{r,x} + (1-\mathrm{sn}_S)\sum_{u_1=1} a_{N_a}^{u,x}\theta_u^{r,x} & (s=0, z=1)
\end{bmatrix}
\end{array}.
$$

Elements L_{1kr} can be represented as the dot product of the vectors

$$\mathbf{v}_1 = (1 - \mathrm{sp}_S, 1 - \mathrm{sp}_S, \mathrm{sn}_S, \mathrm{sn}_S)^T,$$

and

$$\mathbf{w}_k = (a_k^{(0,0),x}\theta_{(0,0)}^{r,x}, a_k^{(1,0),x}\theta_{(1,0)}^{r,x}, a_k^{(0,1),x}\theta_{(0,1)}^{r,x}, a_k^{(1,1),x}\theta_{(1,1)}^{r,x})^T,$$

while element L_{21r} can be represented as the dot product of

$$\mathbf{v}_2 = (1 - \mathrm{sp}_S, \mathrm{sn}_S, 1 - \mathrm{sp}_S, \mathrm{sn}_S)^T,$$

and $\mathbf{w}_k$. Elements of rows 3 and 4 can be defined similarly as dot products between $\mathbf{w}_k$ and single vectors involving only sp_S and $1 - \mathrm{sn}_S$.

This structure allows us to define L as the triple product of three matrices, each of which have columns that correspond to principal strata:

$$P_2(\tilde{S} \mid Z, S^{P_0}) \in \mathbb{R}^{4 \times 4}, P_2^x(A \mid S^{P_0}) \in \mathbb{R}^{N_a \times 4}, P_2^x(S^{P_0} \mid R) \in \mathbb{R}^{4 \times N_r}.$$

The matrix $P_2(\tilde{S} \mid Z, S^{P_0})$ is defined as

$$P_2(\tilde{S} \mid Z, S^{P_0}) = \begin{array}{c} \begin{matrix} (0,0) & (1,0) & (0,1) & (1,1) \end{matrix} \\ \begin{bmatrix} 1-\text{sp}_S & 1-\text{sp}_S & \text{sn}_S & \text{sn}_S \\ 1-\text{sp}_S & \text{sn}_S & 1-\text{sp}_S & \text{sn}_S \\ \text{sp}_S & \text{sp}_S & 1-\text{sn}_S & 1-\text{sn}_S \\ \text{sp}_S & 1-\text{sn}_S & \text{sp}_S & 1-\text{sn}_S \end{bmatrix} \end{array} \begin{matrix} (s=1,z=0) \\ (s=1,z=1) \\ (s=0,z=0) \\ (s=0,z=1) \end{matrix} . \tag{3.6}$$

The matrices $P_2^x(A \mid S^{P_0}), P_2^x(S^{P_0} \mid R)$ are defined

$$P_2^x(A \mid S^{P_0}) = \begin{bmatrix} a_1^{(0,0),x} & a_1^{(1,0),x} & a_1^{(0,1),x} & a_1^{(1,1),x} \\ a_2^{(0,0),x} & a_2^{(1,0),x} & a_2^{(0,1),x} & a_2^{(1,1),x} \\ \vdots & \vdots & \vdots & \vdots \\ a_{N_a}^{(0,0),x} & a_{N_a}^{(1,0),x} & a_{N_a}^{(0,1),x} & a_{N_a}^{(1,1),x} \end{bmatrix}, \; P_2^x(S^{P_0} \mid R) = \begin{bmatrix} \theta_{(0,0)}^{r_1,x} & \theta_{(0,0)}^{r_2,x} & \cdots & \theta_{(0,0)}^{r_{N_r},x} \\ \theta_{(1,0)}^{r_1,x} & \theta_{(1,0)}^{r_2,x} & \cdots & \theta_{(1,0)}^{r_{N_r},x} \\ \theta_{(0,1)}^{r_1,x} & \theta_{(0,1)}^{r_2,x} & \cdots & \theta_{(0,1)}^{r_{N_r},x} \\ \theta_{(1,1)}^{r_1,x} & \theta_{(1,1)}^{r_2,x} & \cdots & \theta_{(1,1)}^{r_{N_r},x} \end{bmatrix}.$$

Kruskal's theorem defines sufficient conditions for the uniqueness of the triple product $L = [P_2(\tilde{S} \mid Z, S^{P_0}), P_2^x(A \mid S^{P_0}), P_2^x(S^{P_0} \mid R)^T]$:

Theorem 3 (Kruskal triple product decomposition uniqueness)**.** *Let matrices A, B, C be defined as in Definition 3.2.5, with respective ranks r_A, r_B, r_C, and let array L also be defined as in Definition 3.2.5. Suppose that $k_A \leq r_A, k_B \leq r_B$, and $k_C \leq r_C$. Then if*

$$r_A + r_B + r_C - (2M+2) \geq \begin{cases} \min(r_A - k_A, r_B - k_B) \\ \min(r_A - k_A, r_C - k_C) \end{cases},$$

$\min(k_A, k_B) + r_C \geq M + 2$*, and* $\min(k_A, k_C) + r_B \geq M + 2$ *the decomposition $L = [A, B, C]$ is unique up to column permutation matrix P and column scaling Λ, G, N such that $\Lambda G N$ is the identity matrix. In other words, L can be represented as the triple product of any three matrices $[\tilde{A}, \tilde{B}, \tilde{C}]$ such that $[\tilde{A} = AP\Lambda, \tilde{B} = BPG, \tilde{C} = CPN]$. See proof in Kruskal (1977) on page 126.*

We extend Kruskal's theorem, similarly to Allman et al. (2009), to account for the structure of matrices $P_2^x(A \mid S^{P_0})$ and $P_2^x(S^{P_0} \mid R)^T$:

Lemma 4 (Uniqueness with column and row sum conditions)**.** *Suppose B has rows that sum to 1 and C has columns that sum to 1, or $B\mathbf{1}_{R\times 1} = \mathbf{1}_{J\times 1}$, and $\mathbf{1}_{1\times K}C = \mathbf{1}_{1\times R}$. If the rank conditions in Theorem 3 on A, B, C also hold, and C is full column rank then $[A, B, C]$ is the unique triple product decomposition of array L up to a common column permutation.*

The proof of Lemma 4 is a simple extension of Theorem 3 and is shown in Appendix B.5.1.

Given that the column indices of our matrices are meaningful in that they correspond to principal strata, we need a stronger result that gives strict uniqueness of the decomposition. We show in Lemma 15 that conditions on sn_S and sp_S, namely that they both lie in the same half interval of $[0, 1]$, yields a $P_2(\tilde{S} \mid Z, S^{P_0})$ whose column domains are not invariant column permutation. This restriction, combined with the Kruskal rank conditions, yields strict identifiability of the joint distribution of $P(S_i^{P_0} = u, A_i = k \mid Z_i = z, R_i = r, X_i = x)$, as well as the distributions $P(\tilde{Y}_i(z) = 1 \mid S_i^{P_0} = u, A_i = k, X_i = x)$. Furthermore, we show that sn_S, sp_S, and sp_Y are also identified. Formally,

Theorem 5 (Identifiability of causal model parameters)**.** *Suppose that Assumptions 1 to 5 hold. If both* sn_S, sp_S *lie in* $[0, 1/2)$ *or both lie in* $(1/2, 1]$*,* $P_2^x(A \mid S^{P_0})$ *is Kruskal rank* 3 *or greater for all* x *and* $P_2^x(S^{P_0} \mid R)$ *is rank* 4 *for all* x*, then both* sn_S, sp_S *are identifiable, as are the following distributions:* $P(S_i^{P_0} = (m, n) \mid R_i = r, X_i = x), P(A_i = k \mid S_i^{P_0} = (m, n), X_i = x)$*,* $k \in \{1, \ldots, N_a\}$*,* $r \in \{1, \ldots, N_r\}$*,* $(m, n) \in \{(0, 0), (1, 0), (1, 0), (1, 1)\}$*. Furthermore, if* sn_Y *is unknown (known), distributions* $P(Y_i(z) = 1 \mid S_i^{P_0} = u, A_i = k, X_i = x)$ *are identifiable up to an unknown (known) common constant,* $r_Y = \text{sn}_Y + \text{sp}_Y - 1$*.*

Note that these are sufficient conditions for identifiability of the model parameters. In Appendix B.8, we show that a two-arm trial with at least 5 study sites, and a binary covariate yields a causal model with identifiable model parameters. The benefit of Theorem 5 is that the number of study sites required is at least 4. Reducing the number of sites likely reduces the costs of running a multisite trial more than reducing the number of covariates.

A consequence of Theorem 5 is that the marginal distribution of reported severe illness among the always-infected stratum is identified. To see why, note that the conditional counterfactual distributions for $\tilde{Y}_i(z) \mid S_i^{P_0} = (1, 1), A_i, X_i$ are identified, along with the distribution of $A_i \mid S_i^{P_0} = (1, 1), X_i$. The following corollary marginalizes over A_i to yield identifiability of the population distribution $\tilde{Y}_i(z) \mid S_i^{P_0} = (1, 1), X_i$:

Corollary 6. *By the conditions set forth in Theorem 5,* $P(A_i = k \mid S_i^{P_0} = (1, 1), X_i = x)$ *and* $P(\tilde{Y}_i(z) = 1 \mid S_i^{P_0} = (1, 1), A_i = k, X_i = x)$ *are identifiable for* $k \in \{1, \ldots, N_a\}$*. Let* $P(\tilde{Y}_i(z) = 1 \mid S_i^{P_0} = (1, 1), X_i = x) = \sum_k P(A_i = k \mid S_i^{P_0} = (1, 1), X_i = x) P(\tilde{Y}_i(z) = 1 \mid S_i^{P_0} = (1, 1), A_i = k, X_i = x)$*. Then* $P(\tilde{Y}_i(z) = 1 \mid S_i^{P_0} = (1, 1), X_i = x), P(A_i = k \mid S_i^{P_0} = (1, 1), X_i = x)$*, and* $P(\tilde{Y}(z)_i = 1 \mid S_i^{P_0} = (1, 1), A_i = k, X_i = x)$ *are identifiable.*

Given that the marginal distribution of reported severe illness for the always-infected stratum is identified, along with the identity Equation (3.5) and sp_Y we can write the estimand of interest,

the vaccine efficacy against severe illness within the always-infected stratum:

$$\begin{aligned}\mathrm{VE}_I &= 1 - \frac{\mathbb{E}\left[Y_i(1) \mid S_i^{P_0} = (1,1)\right]}{\mathbb{E}\left[Y_i(0) \mid S_i^{P_0} = (1,1)\right]} \\ &= 1 - \frac{\mathbb{E}\left[\tilde{Y}_i(1) \mid S_i^{P_0} = (1,1)\right] - (1 - \mathrm{sp}_Y)}{\mathbb{E}\left[\tilde{Y}_i(0) \mid S_i^{P_0} = (1,1)\right] - (1 - \mathrm{sp}_Y)}.\end{aligned}$$

This estimand marginalizes over the population distribution of X_i, which may be known, or may be estimated.

Moreover, the identifiability of the conditional counterfactual distributions $P(\tilde{Y}_i(z) = 1 \mid S_i^{P_0} = (1,1), A_i = k, X_i = x)$ allows for causal effect heterogeneity by covariate A_i.

Definition 3.2.7 (Conditional VE against post-infection outcome Y)**.**

$$\mathrm{VE}_I(k) = 1 - \frac{\mathbb{E}\left[\tilde{Y}_i(1) \mid S_i^{P_0} = (1,1), A_i = k\right] - (1 - \mathrm{sp}_Y)}{\mathbb{E}\left[\tilde{Y}_i(0) \mid S_i^{P_0} = (1,1), A_i = k\right] - (1 - \mathrm{sp}_Y)}$$

Again this estimand marginalizes over X_i.

The identifiability results presented in this section are in fact a special case of identifiability results related to multiarm multisite trials. Because of this, the discussion of the proof and the implications are deferred until the next section, which outlines the generalization to two or more treatments.

3.3 Vaccine efficacy in multiarm, multisite trials

Many vaccine trials involve more than two treatment arms. For example, Monto et al. (2009) compares the absolute and relative vaccine efficacy of an inactivated influenza vaccine and a live attenuated influenza vaccine against two placebo arms. The results of trials such as these can inform public health vaccine policy as well as suggest new directions for vaccine development.

Mirroring the notation presented in Section 3.2, for n total participants we observe the following sextuplet for each participant i: $(\tilde{S}_i, \tilde{Y}_i, Z_i, R_i, A_i, X_i)$. Like Section 3.2, $\tilde{S}_i, \tilde{Y}_i$ are imperfectly observed proxies for true infection status, S_i, and true severe illness status, Y_i. In contrast to Section 3.2, Z_i is a categorical variable with $N_z \geq 2$ categories representing treatment assignment. Let $Z_i \in \{z_1, \ldots, z_{N_z}\}$. The variables R_i, A_i, X_i, study site membership, pretreatment measurement of susceptibility to infection, and other pretreatment covariates, are defined as in Section 3.2. True infection status S_i and true severe illness status Y_i are assumed to be partially-observable realizations of counterfactual variables $S_i(\mathbf{z}), Y_i(\mathbf{z}, S_i(\mathbf{z}))$, where $\mathbf{z}$ is a given n-vector of treatment

assignments for each individual. The counterfactual variables are so-named because these variables are defined for *any* $\mathbf{z}$, which is possibly different from the collection of observed treatment assignments $\{Z_1, \ldots, Z_n\}$. Our causal model enforces the constraint that $Y_i(\mathbf{z}, 0) = \star$ for all $\mathbf{z}$, meaning that severe illness is caused by an influenza infection; without an influenza infection there can be no severe illness caused by influenza.

Like Section 3.2, we assume SUTVA, as is typical in vaccine trials (Gilbert et al., 2003); we also continue to assume Non-differential Misclassification Errors. This allows us to write the variables S_i, Y_i as

$$S_i = \sum_{j=1}^{N_z} S_i(z_j)\mathbb{1}_{Z_i=z_j}, \quad Y_i = \sum_{j=1}^{N_z} Y_i(z_j, S_i(z_j))\mathbb{1}_{Z_i=z_j}. \tag{3.7}$$

We continue to assume random treatment assignment in the multiarm setting.

Assumption 6 (Random treatment assignment multiple treatment)**.** *The probability of being assigned to treatment for each individual lies strictly between* 0 *and* 1*:*

$$0 < P(Z_i = z_j \mid S_i(z_1), Y_i(z_1, S(z_1)), \ldots, S_i(z_{N_z}), Y_i(z_{N_z}, S(z_{N_z}))) < 1$$

for all $z_j \in \{z_1, \ldots, z_{N_z}\}$.
Treatment assignment is independent of all potential outcomes, or

$$S_i(z_1), Y_i(z_1, S(z_1)), \ldots, S_i(z_{N_z}), Y_i(z_{N_z}, S(z_{N_z})) \perp\!\!\!\perp Z_i.$$

In keeping with the expanded set of treatments, let principal stratum, $S_i^{P_0}$, be defined as the ordered N_z-vector of counterfactual infection outcomes for unit i, or

$$S_i^{P_0} = (S_i(z_1), S_i(z_2), \ldots, S_i(z_{N_z})),\ S_i(z_j) \in \{0, 1\},\ 1 \leq j \leq N_z,$$

and let the set of all principal strata be denoted as $\mathcal{S}$. When the set of principal strata is not restricted $\mathcal{S} \equiv \{0, 1\}^{N_z}$.

It is typical to restrict the set of principal strata as the size of $\mathcal{S}$ grows because the dimension of the parameter space grows quickly. For example, monotonicity assumptions can be generalized to three treatments, as in Yuan et al. (2019) or Cheng and Small (2006). Another strain of research places strong assumptions on treatment ordering, such as Luo et al. (2023) and Wang et al. (2017). We do not make such assumptions and allow for an unrestricted space of principal strata.

However, we do impose an ordering among the elements of $\mathcal{S}$. Given that the set is a collection of binary vectors, the natural ordering of the elements of the set is the base-10 representation of the principal stratum. In order to formalize this ordering, we define a map, $\varpi_m(j)$, which generates

the m-digit binary representation of the integer j as a length-m binary vector. We also define its inverse, $\varpi_m(u)^{-1}$, where u is an element of $\mathcal{S}$.

Definition 3.3.1 (Base-10 to binary map)**.** *Let the operator ϖ_m be defined as $\varpi_m(\cdot) : j \rightarrow \{0,1\}^m, j \in \mathbb{N}, j \leq 2^m - 1$ with elements $\varpi_m(j)_i \in \{0,1\}$, so $\varpi_m(j)$ is the base-2 representation of j with m digits represented as a binary m-vector. The binary representation is indexed so the i^{th} element of the vector corresponds to the digit for 2^{i-1}. Let the inverse operator $\varpi_m^{-1}(\cdot) : \{0,1\}^m \rightarrow j$, or the binary to base-10 conversion. Let digit i of $\varpi_m(\cdot)$ represent the digit for 2^{i-1}.*

For example, $\varpi_3(4) = (0,0,1)$, $\varpi_5(4) = (0,0,1,0,0)$, and $\varpi_3((0,0,1))^{-1} = \varpi_5((0,0,1,0,0))^{-1} = 4$. In order to see how $\mathcal{S}$ is ordered, suppose that $N_z = 3$ so $\mathcal{S} \equiv \{0,1\}^3$. Then the third and fourth elements of the ordered set of principal strata are $(0,1,0)$ and $(1,1,0)$ respectively.

In the multiarm setting, we will modify our definition for the vaccine efficacy estimands. Vaccine efficacy against infection is now defined for any two treatments, z_j and z_k:

Definition 3.3.2 (Vaccine efficacy against infection z_j versus z_k)**.**

$$\text{VE}_{S,jk} = 1 - \mathbb{E}\left[S_i(z_j)\right] / \mathbb{E}\left[S_i(z_k)\right].$$

Vaccine efficacy against severe illness is well-defined for any principal stratum u and any two treatments z_j and z_k such that $P(S_i(z_j) = S_i(z_k) = 1 \mid S_i^{P_0} = u)$.

Definition 3.3.3 (Vaccine efficacy against post-infection outcome Y)**.**

$$\text{VE}_{I,jk}^u = 1 - \mathbb{E}\left[Y_i(z_j) \mid S_i^{P_0} = u\right] / \mathbb{E}\left[Y_i(z_k) \mid S_i^{P_0} = u\right].$$

$\text{VE}_{I,jk}^u$ is a principal effect as defined in Frangakis and Rubin (2002) because it is conditional on a principal stratum u. For example, when $N_z = 3$, there are 8 principal strata, three of which would admit comparisons between two treatments: $(1,1,0), (0,1,1), (1,0,1)$, and one of which would allow for comparisons between all three treatments: $(1,1,1)$. Like the two-arm setting, in which $S_i^{P_0} = (1,1)$ is the "always-infected" stratum, the stratum $S_i^{P_0} = \{1\}^{N_z}$ is the "always-infected" stratum in the multiarm trial.

To give a concrete example about how one might use the expanded definition of vaccine efficacy against severe illness, we will use Monto et al. (2009) as an example. Monto et al. (2009) treated the four-arm trial as a three-arm trial by combining the two separate placebo arms into one unified arm. Given that both placebo arms received inert treatments, albeit via different routes of administration, this is a reasonable assumption. The aim of the trial was to measure the absolute and relative

efficacies against symptomatic influenza; thus, it is of interest to infer the relative efficacy against severe illness given influenza infection for the two competing vaccines. The following causal estimand captures this effect:

$$\frac{\mathbb{E}\left[Y_i(z_2) \mid S_i^{P_0} = (1,1,1)\right] - \mathbb{E}\left[Y_i(z_3) \mid S_i^{P_0} = (1,1,1)\right]}{\mathbb{E}\left[Y_i(z_1) \mid S_i^{P_0} = (1,1,1)\right]} = \mathrm{VE}_{I,31}^{(1,1,1)} - \mathrm{VE}_{I,21}^{(1,1,1)}. \tag{3.8}$$

As in Section 3.2, we will assume Causal Homogeneity:

Assumption 7 (Causal Homogeneity for multiarm trials)**.** *Conditional on principal stratum $S_i^{P_0}, A_i$, and X_i the potential outcomes $(Y_i(z_1), \ldots, Y_i(z_{N_z}))$ are independent of R_i, or $(Y_i(z_1), \ldots, Y_i(z_{N_z})) \perp\!\!\!\perp R_i \mid S_i^{P_0}, A_i, X_i$.*

In order to expand the causal model, we shall update some of the notation introduced in Section 3.2 to the multiarm trial. Let $\theta_u^{r,x} = P(S_i^{P_0} = u \mid R_i = r, X_i = x)$ where $u \in \{0,1\}^{N_z}$, and $\beta_{j,k}^{u,x} = P(Y_i(z_j) = 1 \mid S_i^{P_0} = u, A_i = k, X_i = x)$. Let $u_j = P(S_i(z_j) = 1 \mid S_i^{P_0} = u)$. Then $\beta_{j,k}^{u,x}$ is only defined for j such that $u_j = 1$. Further, let $a_k^{u,x} = P(A_i = k \mid S_i^{P_0} = u, X_i = x)$. Let $q_{sk|jrx}$ be defined as

$$q_{sk|jrx} = \mathrm{sn}_S^s(1-\mathrm{sn}_S)^{1-s} \textstyle\sum_{u|u\in\mathcal{S},u_j=1} a_k^{u,x}\theta_u^{r,x} + \mathrm{sp}_S^{1-s}(1-\mathrm{sp}_S)^s \sum_{u|u\in\mathcal{S},u_j=0} a_k^{u,x}\theta_u^{r,x}.$$

The observable probabilities $q_{y|kjrx} = P(\tilde{Y}_i = y \mid Z_i = z_j, R_i = r, A_i = k, X_i = x)$, which marginalize over the observed infection test results, are defined:

$$\begin{aligned} q_{y|kjrx} = {} & \mathrm{sn}_Y^y(1-\mathrm{sn}_Y)^{1-y} \textstyle\sum_{u|u\in\mathcal{S},u_j=1} \beta_{j,k}^{u,x}\theta_u^{r,x} \\ & + \mathrm{sp}_Y^{1-y}(1-\mathrm{sp}_Y)^y\big(\textstyle\sum_{u|u\in\mathcal{S},u_j=1}(1-\beta_{j,k}^{u,x})\theta_u^{r,x} + \sum_{u|u\in\mathcal{S},u_j=0}\theta_u^{r,x}\big) \end{aligned}$$

3.3.1 Identifiability of multiarm, multi-site trials

Our strategy will be the same as in Section 3.2. Fixing x, we can rewrite $q_{sk|jrx}$ as a 3-way array, L and subsequently use Kruskal rank conditions to characterize the uniqueness of the array decomposition. Let L be the 3-way array representing $q_{sk|jrx}$. The array's dimensions are $2N_z \times N_a \times N_r$ and is defined so that the $(j,k,r)^{\text{th}}$ element $P(S_i = \mathbb{1}_{j\le N_z}, A_i = k \mid Z = z_{j-1 \bmod N_z+1}, R_i = r, X_i = x)$. Again, we look to the *3-slab*, denoted as $L_r \in \mathbb{R}^{2N_z\times N_a}$, to yield a possible decomposition of this array. As above, let $\sum_{u_z=s} a_k^{u,x}\theta_u^{r,x}$ denote the sum over elements $u \in \mathcal{S}$ such that

$P(S_i(z) = 1 \mid S_i^{P_0} = u)$. Let L_r be defined as

$$\begin{array}{c} (a=1) \\ \left[\begin{array}{c} (1-\mathrm{sp}_S)\sum_{u_{z_1}=0} a_1^{u,x}\theta_u^{r,x} + \mathrm{sn}_S \sum_{u_{z_1}=1} a_1^{u,x}\theta_u^{r,x} \\ (1-\mathrm{sp}_S)\sum_{u_{z_2}=0} a_1^{u,x}\theta_u^{r,x} + \mathrm{sn}_S \sum_{u_{z_2}=1} a_1^{u,x}\theta_u^{r,x} \\ \vdots \\ (1-\mathrm{sp}_S)\sum_{u_{z_{N_z}}=0} a_1^{u,x}\theta_u^{r,x} + \mathrm{sn}_S \sum_{u_{z_{N_z}}=1} a_1^{u,x}\theta_u^{r,x} \\ \mathrm{sp}_S \sum_{u_{z_1}=0} a_1^{u,x}\theta_u^{r,x} + (1-\mathrm{sn}_S)\sum_{u_{z_1}=1} a_1^{u,x}\theta_u^{r,x} \\ \mathrm{sp}_S \sum_{u_{z_2}=0} a_1^{u,x}\theta_u^{r,x} + (1-\mathrm{sn}_S)\sum_{u_{z_2}=1} a_1^{u,x}\theta_u^{r,x} \\ \vdots \\ \mathrm{sp}_S \sum_{u_{z_{N_z}}=0} a_1^{u,x}\theta_u^{r,x} + (1-\mathrm{sn}_S)\sum_{u_{z_{N_z}}=1} a_1^{u,x}\theta_u^{r,x} \end{array}\right. \end{array} \cdots \begin{array}{c} (a=N_a) \\ \left.\begin{array}{c} (1-\mathrm{sp}_S)\sum_{u_{z_1}=0} a_{N_a}^{u,x}\theta_u^{r,x} + \mathrm{sn}_S \sum_{u_{z_1}=1} a_{N_a}^{u,x}\theta_u^{r,x} \\ (1-\mathrm{sp}_S)\sum_{u_{z_2}=0} a_{N_a}^{u,x}\theta_u^{r,x} + \mathrm{sn}_S \sum_{u_{z_2}=1} a_{N_a}^{u,x}\theta_u^{r,x} \\ \vdots \\ (1-\mathrm{sp}_S)\sum_{u_{z_{N_z}}=0} a_{N_a}^{u,x}\theta_u^{r,x} + \mathrm{sn}_S \sum_{u_{z_{N_z}}=1} a_{N_a}^{u,x}\theta_u^{r,x} \\ \mathrm{sp}_S \sum_{u_{z_1}=0} a_{N_a}^{u,x}\theta_u^{r,x} + (1-\mathrm{sn}_S)\sum_{u_{z_1}=1} a_{N_a}^{u,x}\theta_u^{r,x} \\ \mathrm{sp}_S \sum_{u_{z_2}=0} a_{N_a}^{u,x}\theta_u^{r,x} + (1-\mathrm{sn}_S)\sum_{u_{z_2}=1} a_{N_a}^{u,x}\theta_u^{r,x} \\ \vdots \\ \mathrm{sp}_S \sum_{u_{z_{N_z}}=0} a_{N_a}^{u,x}\theta_u^{r,x} + (1-\mathrm{sn}_S)\sum_{u_{z_{N_z}}=1} a_{N_a}^{u,x}\theta_u^{r,x} \end{array}\right. \end{array} \begin{array}{c} \\ \left.\begin{array}{c} (s=1,z=z_1) \\ (s=1,z=z_2) \\ \vdots \\ (s=1,z=z_{N_z}) \\ (s=0,z=z_1) \\ (s=0,z=z_2) \\ \vdots \\ (s=0,z=z_{N_z}) \end{array}\right. \end{array}$$

Following the same logic as Section 3.2, we can define matrices encoding the distribution of principal strata by study site, and the distribution of pre-season titers by principal stratum. Let the matrix $P_{N_z}^x(A \mid S^{P_0})$ in $\mathbb{R}^{N_a \times 2^{N_z}}$ encode the distributions $A_i \mid S_i^{P_0}, X_i$ with $(k,j)^{\text{th}}$ element $P_{N_z}(A_i = k \mid S_i^{P_0} = \varpi_{N_z}(j-1), X_i = x)$. Let the matrix $P_{N_z}^x(S^{P_0} \mid R)$ in $\mathbb{R}^{2^{N_z} \times N_r}$ encode the distribution $S_i^{P_0} \mid R_i, X_i$ with $(k,j)^{\text{th}}$ element $P_{N_z}^x(S_i^{P_0} = \varpi_{N_z}(k-1) \mid R_i = j, X_i = x)$.

Finally, let matrix $P_{N_z}(\tilde{S} \mid Z, S^{P_0}) \in \mathbb{R}^{2N_z \times 2^{N_z}}$ with $(k,j)^{\text{th}}$ element

$$\mathrm{sn}_S^{\varpi_{N_z}(j-1)_k}(1-\mathrm{sp}_S)^{1-\varpi_{N_z}(j-1)_k}\mathbb{1}_{k \le N_z} + (1-\mathrm{sn}_S)^{\varpi_{N_z}(j-1)_{k-N_z}}\mathrm{sp}_S^{1-\varpi_{N_z}(j-1)_{k-N_z}}\mathbb{1}_{k > N_z}.$$

Let $P_{N_z}^x(S^{P_0} \mid R = r)$ be the r^{th} column of matrix $P_{N_z}^x(S^{P_0} \mid R)$. Then L_r can be represented in matrix form as

$$L_r = P_{N_z}(\tilde{S} \mid Z, S^{P_0})\mathrm{diag}(P_{N_z}^x(S^{P_0} \mid R = r))P_{N_z}^x(A \mid S^{P_0})^T$$

This structure again allows us to define L as the triple product of these three matrices, each of which have columns that correspond to principal strata:

$$L = [P_{N_z}(\tilde{S} \mid Z, S^{P_0}), P_{N_z}^x(A \mid S^{P_0}), P_{N_z}^x(S^{P_0} \mid R)^T].$$

The conditions for the identifiability of the model parameters are outlined below:

Theorem 7. *Let $N_z \ge 2$. Suppose Assumptions 1 to 2,Assumption 4, Assumption 6, and Assumption 7 hold. If both $\mathrm{sn}_S, \mathrm{sp}_S$ lie in $[0, 1/2)$ or both lie in $(1/2, 1]$, $P_{N_z}^x(A \mid S^{P_0})$ is at least Kruskal rank $2^{N_z} - 1$ and $P_{N_z}^x(S^{P_0} \mid R)$ is rank 2^{N_z} for all x then the counterfactual distributions $P(S_i^{P_0} = u \mid R_i = r, X_i = x)$, $P(A_i = k \mid S_i^{P_0} = u, X_i = x)$ are identifiable as are the quantities $\mathrm{sn}_S, \mathrm{sp}_S, \mathrm{sp}_Y, \mathrm{VE}_{I,jk}^u(k)$, and $\mathrm{VE}_{I,jk}^u$. Furthermore, if sn_Y is unknown (known), distributions $P(Y_i(z_j) = 1 \mid S_i^{P_0} = u, A_i = k, X_i = x)$ are identifiable up to an unknown (known) common constant, $r_Y = \mathrm{sn}_Y + \mathrm{sp}_Y - 1$.*

Theorem 7 allows for a more realistic model of infection measurement than Hudgens and Halloran (2006) and does not require any restrictions on the space of principal strata. The primary benefit of an unrestricted principal strata distribution is that we can jointly infer vaccine efficacy against infection and vaccine efficacy against a post-infection outcome. This will aid in designing comprehensive randomized trials for vaccine efficacy.

The proof of Theorem 7, shown in Appendix B.4 is related to the methods in Jiang et al. (2016) and Ding et al. (2011). Ding et al. (2011) addresses problems of identifiability in survivor average treatment effects, which is mathematically analogous to vaccine efficacy for post-infection outcomes, by measuring covariates that are related to the principal strata. Jiang et al. (2016) identifies principal causal effects in binary surrogate endpoint evaluations. Despite not being mathematically identical to vaccine efficacy, binary surrogacy endpoint evaluation is ultimately a problem in identification of principal causal effects. The proof of Theorem 5 is shown in the Supplementary Materials. Most importantly, the proof does not encode any restrictions on the distribution of secondary outcomes, otherwise known in our case as the post-infection outcomes. This makes the result applicable to categorical or continuous post-infection outcomes, and, more broadly, to principal stratification problems outside the scope of vaccine efficacy.

The identifiability results in Theorem 7 suggest the following so-called transparent parameterization[1]: $(\beta_{j,k}^{u,x}, \text{sp}_Y, \text{sn}_Y) \to (\tilde{p}_{j,k}^{u,x} = (\text{sn}_Y + \text{sp}_Y - 1)\beta_{j,k}^{u,x} + (1 - \text{sp}_Y), \text{sp}_Y, \text{sn}_Y)$. The quantities $\tilde{p}_{j,k}^{u,x} = P(\tilde{Y}_i = 1 \mid Z_i = z_j, S_i^{P_0} = u, A_i = k)$ and sp_Y are identified by the data, while sn_Y is not. This yields the following asymptotic identification regions for sn_Y and $\beta_{j,k}^{u,x}$:

$$\text{sn}_Y \in \left(\max_{x,u,j,k} (\tilde{p}_{j,k}^{u,x}), 1 \right), \quad \beta_{j,k}^{u,x} \in \left(\frac{\tilde{p}_{j,k}^{u,x} - (1 - \text{sp}_Y)}{\text{sp}_Y}, \frac{\tilde{p}_{j,k}^{u,x} - (1 - \text{sp}_Y)}{\max_{x,u,j,k}(\tilde{p}_{j,k}^{u,x}) + \text{sp}_Y - 1} \right) \tag{3.9}$$

This may be useful for policymakers interested in absolute risk of post-infection outcomes to forecast the burden on healthcare centers under different vaccination policies.

We will present a final corollary that will be useful in our applied examples:

Corollary 8. *Suppose in addition to Assumptions 1 to 2,Assumption 4, Assumption 6, and Assumption 7, researchers do not directly observe A_i, but instead observe a misclassified version of A_i, $\tilde{A}_i$, such that the following nondifferential error assumption holds: $\tilde{A}_i \perp\!\!\!\perp \tilde{S}_i, \tilde{Y}_i, Y_i(z_j, S(z_j)), R_i, Z_i, S_i^{P_0} \mid A_i, X_i$. If both sn_S, sp_S lie in $[0, 1/2)$ or both lie in $(1/2, 1]$, $P_{N_z}^x(\tilde{A} \mid S^{P_0})$ is at least Kruskal rank $2^{N_z} - 1$ and $P_{N_z}^x(S^{P_0} \mid R)$ is rank 2^{N_z} for all x then the counterfactual distributions $P(S_i^{P_0} = u \mid R_i = r, X_i = x)$, $P(\tilde{A}_i = k \mid S_i^{P_0} = u, X_i = x)$ are identifiable as are the quantities $\text{sn}_S, \text{sp}_S, \text{sp}_Y$, and $\text{VE}_{I,jk}^u$. Furthermore, if sn_Y is unknown (known), distributions $P(Y_i(z_j) = 1 \mid S_i^{P_0} = u, \tilde{A}_i = k, X_i = x)$ are identifiable up to an unknown*

[1] See (Gustafson, 2015) for more details on inference in partially identified Bayesian models

(known) common constant, $r_Y = \text{sn}_Y + \text{sp}_Y - 1$.

The proof, shown in Appendix B.4, follows directly from the proof of Theorem 7 and the nondifferential misclassification error assumption for A.

While misclassified $\tilde{A}$ precludes learning heterogeneous treatment effects, marginalizing over the identifiable distribution $\tilde{A}_i \mid S^{P_0}, X_i$ will yield the average post-infection vaccine efficacy.

3.3.2 Models, priors and sensitivity analyses

Under the assumptions laid out in Section 3.3 the observational model is a multinomial random variable for each study site and treatment group.

Let $\tilde{n}_{syk}(j,r,x)$ be $\sum_{i=1}^{n} \mathbb{1}_{\tilde{S}_i=s} \mathbb{1}_{\tilde{Y}_i=y} \mathbb{1}_{Z_i=z_j} \mathbb{1}_{R_i=r} \mathbb{1}_{A_i=k} \mathbb{1}_{X_i=x}$, and let the error-free partially-observed causal model probabilities $p_{syk|jrx} = P(S_i = s, Y_i = y, A_i = k \mid Z_i = z_j, R_i = r, X_i = x)$ be defined as:

$$p_{1yk|jrx} = \textstyle\sum_{u|u\in\mathcal{S},u_j=1} a_k^{u,x}\theta_u^{r,x}(\beta_{j,k}^{u,x})^y(1-\beta_{j,k}^{u,x})^{1-y}, \quad p_{0*k|jr} = \textstyle\sum_{u|u\in\mathcal{S},u_j=0} a_k^{u,x}\theta_u^{r,x}, \tag{3.10}$$

where we note that $p_{01k|jrx} = 0$ for all k, j, r. Then we define the observable joint probabilities $q_{syk|jrx} = P(\tilde{S}_i = s, \tilde{Y}_i = y, A_i = k \mid Z_i = z_j, R_i = r, X_i = x)$ as

$$\begin{aligned} q_{syk|jrx} = {} & \text{sn}_S^s(1-\text{sn}_S)^{1-s}\text{sn}_Y^y(1-\text{sn}_Y)^{1-y}p_{11kjrx} + \text{sn}_S^s(1-\text{sn}_S)^{1-s}\text{sp}_Y^{1-y}(1-\text{sp}_Y)^y p_{10kjrx} \\ & + \text{sp}_S^{1-s}(1-\text{sp}_S)^s\text{sp}_Y^{1-y}(1-\text{sp}_Y)^y p_{0*kjrx}, \end{aligned}$$

This allows us to define the observational model as:

$$\begin{aligned} &(\tilde{n}_{001}(j,r,x), \tilde{n}_{011}(j,r,x), \tilde{n}_{101}(j,r,x), \tilde{n}_{111}(j,r,x), \ldots, \tilde{n}_{00N_a}(j,r,x), \tilde{n}_{01N_a}(j,r,x), \tilde{n}_{10N_a}(j,r,x), \tilde{n}_{11N_a}(j,r,x)) \sim \\ &\text{Multinomial}(n(j,r,x) \mid q_{001|jrx}, q_{011|jrx}, q_{101|jrx}, q_{111|jrx}, \ldots, q_{00N_a|jrx}, q_{01N_a|jrx}, q_{10N_a|jrx}, q_{11N_a|jrx}), \\ &j \in \{1,\ldots,N_z\},\ r \in \{1,\ldots,N_r\},\ x \in \{1,\ldots,N_x\} \end{aligned} \tag{3.11}$$

The post-infection severe illness models can be formulated as logistic regressions:

$$\log\frac{P(Y_i(z_j)=1 \mid S_i^{P_0}=u, A_i=k, X_i=x)}{P(Y_i(z_j)=0 \mid S_i^{P_0}=u, A_i=k, X_i=x)} = \alpha_j^{u,x} + \delta_{j,k}^{u,x},\ \beta_{j,k}^{u,x} = \frac{e^{\alpha_j^{u,x}+\delta_{j,k}^{u,x}}}{1+e^{\alpha_j^{u,x}+\delta_{j,k}^{u,x}}},$$
$$\delta_{j,1}^{u,x} = 0 \,\forall\, j,\ u,\ x.$$

Deviations from Assumption 5 can be encoded as an additive term $\varepsilon_r^{u,x}$ capturing heterogeneity

between study sites:

$$\log \frac{P(Y_i(z_j) = 1 \mid S_i^{P_0} = u, A_i = k, R_i = r, X_i = x)}{P(Y_i(z_j) = 0 \mid S_i^{P_0} = u, A_i = k, R_i = r, X_i = x)} = \alpha_j^{u,x} + \delta_{j,k}^{u,x} + \varepsilon_r^{u,x},$$
$$\varepsilon_r^{u,x} \sim \text{Normal}(0, (\tau_\varepsilon^u)^2).$$

We can fix $\tau_\gamma^{u,x}$ to several values for sensitivity analysis, as developed in Jiang et al. (2016).

We may write the probability models for $S_i^{P_0} \mid R_i, X_i$ and $A_i \mid S_i^{P_0}, X_i$ as two multinomial regressions, given Assumption 4 that $A_i \perp\!\!\!\perp R_i, Z_i \mid S_i^{P_0}, X_i$.

$$\log \frac{P(S_i^{P_0} = u \mid R_i = r, X_i = x)}{P(S_i^{P_0} = u_0 \mid R_i = r, X_i = x)} = \mu_u^r + \eta_u^x + \eta_u^{r,x}$$
$$\log \frac{P(A_i = k \mid S_i^{P_0} = u, X_i = x)}{P(A_i = k_0 \mid S_i^{P_0} = u, X_i = x)} = \nu_k^u + \gamma_k^x + \gamma_k^{u,x},$$

where

$$\theta_u^{r,x} = \frac{e^{\mu_u^r + \eta_u^x + \eta_u^{r,x}}}{\sum_{w \in \mathcal{S}} e^{\mu_w^r + \eta_u^x + \eta_u^{r,x}}},\ \mu_{u_0}^r = 0 \,\forall\, r, \qquad a_k^{u,x} = \frac{e^{\nu_k^u + \gamma_k^x + \gamma_k^{u,x}}}{\sum_{m=1}^{N_a} e^{\nu_m^u + \gamma_k^x + \gamma_k^{u,x}}},\ \nu_{k_0}^u = 0 \,\forall\, u.$$

Note that $\eta_{u_0}^x, \gamma_{k_0}^x, \eta_{u_0}^{r,x}, \gamma_{k_0}^x, \gamma_{k_0}^{u,x}$ are all zero for all x. Furthermore, for given reference categories x_0, u_0, r_0, $\eta_u^{x_0}, \gamma_k^{x_0}$ are zero for all u, k, while $\eta_u^{r_0,x}$ is zero for all x, η_u^{r,x_0} is zero for all r, γ_k^{u,x_0} is zero for all u and $\gamma_k^{u_0,x}$ is zero for all x. This leads to a tidy representation of the log-odds of belonging to stratum u vs. u_0 conditional on $A_i = k, R_i = r, X_i = x$:

$$\log \frac{P(S_i^{P_0} = u \mid A_i = k, R_i = r, X_i = x)}{P(S_i^{P_0} = u_0 \mid A_i = k, R_i = r, X_i = x)} = \mu_u^r + \nu_k^u - \nu_k^{u_0} + \gamma_k^{u,x} - \gamma_k^{u_0,x} + \eta_u^x + \eta_u^{r,x}.$$

If we suspect deviations from Assumption 4, we can add an interaction between A_i and R_i:

$$\log \frac{P(A_i = k \mid R_i = r, S_i^{P_0} = u, X_i = x)}{P(A_i = k_0 \mid R_i = r, S_i^{P_0} = u, X_i = x)} = \nu_k^u + \gamma_k^x + \gamma_k^{u,x} + \epsilon_k^{u,r}, \quad \epsilon_k^{u,r} \sim \text{Normal}(0, (\tau_\epsilon^u)^2)\, \forall r. \tag{3.12}$$

In order to avoid parametric models for $A_i \mid S_i^{P_0}, X_i, S_i^{P_0} \mid R_i, X_i$, and $Y_i(z_j) \mid S_i^{P_0}, A_i, X_i$ we

can instead use categorical distributions with separate parameters for each stratum:

$$\begin{aligned} A_i \mid S_i^{P_0} = u, X_i = x &\sim \text{Categorical}(\boldsymbol{\pi}_{u,x}) \\ S_i^{P_0} \mid R_i = r, X_i = x &\sim \text{Categorical}(\boldsymbol{\rho}_{r,x}) \\ Y_i(z_j) \mid S_i^{P_0} = u, A_i = k, X_i = x &\sim \text{Bernoulli}(\beta_{j,k}^{u,x}) \\ \boldsymbol{\pi}_{u,x} &\overset{\text{iid}}{\sim} \text{Dirichlet}(\mathbf{a}),\ \forall u, x \\ \boldsymbol{\rho}_{r,x} &\overset{\text{iid}}{\sim} \text{Dirichlet}(\mathbf{b}),\ \forall r, x \\ \beta_{j,k}^{u,x} &\overset{\text{iid}}{\sim} \text{Uniform}(0,1),\ \forall u, k, x \end{aligned} \tag{3.13}$$

Given the structure of our model, Gustafson (2015) suggests that integrating out nonidentifiable parameters may improve efficiency. To do so, we reparameterize the Y_i-observation-error parameters sn_Y, sp_Y: $(\text{sn}_Y, \text{sp}_Y) \to (r_Y = \text{sn}_Y + \text{sp}_Y - 1, \text{sp}_Y)$, and put priors over $r_Y \mid \text{sp}_Y$ and sp_Y. The constraints on sn_Y, sp_Y, namely that both are greater than 0.5, yields that $r_Y \in (0, 1]$ and that $r_Y \mid \text{sp}_Y \in (\text{sp}_Y - 1/2, \text{sp}_Y]$. Finally, we let $\tilde{\beta}_{j,k}^{u,x} = r_Y \beta_{j,k}^{u,x}$ and integrate over $r_Y \mid \text{sp}_Y$. If we use uniform priors for $\beta_{j,k}^{u,x}$ and let $r_Y \mid \text{sp}_Y \sim \text{Uniform}(\text{sp}_Y - 1/2, \text{sp}_Y]$ the joint distribution for all $\tilde{\beta}_{j,k}^{u,x}$ is proportional to (with a slight abuse of notation):

$$\max\left(\text{sp}_Y - 1/2, \max_{x,u,j,k} \tilde{\beta}_{j,k}^{u,x}\right)^{1-n_\beta} - \text{sp}_Y^{1-n_\beta} \tag{3.14}$$

where n_β is the number of terms in $\beta_{j,k}^{u,x}$. For example, if $N_z = 2$, $n_\beta = 4N_xN_a$. We leave the details of the calculations to the Appendix.

3.4 Design and analysis of vaccine efficacy studies

There are several real-world applications for Theorem 7 in vaccine efficacy studies. The first is for quantifying vaccine efficacy against post-infection outcomes like severe illness, medically-attended illness or death, which is the primary motivation for the methods we have developed here. A second is to quantify the impact on vaccination on secondary transmission to household contacts. In both of these hypothetical trials, we imagine that participants are prospectively monitored for infection as well as the post-infection outcome of interest. The infection monitoring might involve regular diagnostic testing or analysis of blood specimens for signs of infection.

3.4.1 Vaccine efficacy against severe symptoms trial design

To show how our model can be used to design a vaccine efficacy study, we consider determining the sample size for two hypothetical clinical trials: one three-arm trial inspired by Monto et al. (2009), and a two-arm trial inspired by Polack et al. (2020). Monto et al. (2009) investigated vaccine efficacy against symptomatic influenza infection in a three-arm, double-blind placebo-controlled randomized trial. Polack et al. (2020) presented the results of the COVID-19 Pfizer vaccination trial, which measured vaccine efficacy against symptomatic infection using a two-arm double-blind placebo-controlled randomized trial. All trials are designed so as to jointly test the efficacy against infection and the efficacy against severe symptoms for the always-infected group.

In order to design our hypothetical trials, we simulate 100 datasets under the alternative hypothesis for each sample size and measure the proportion of datasets in which we reject the null hypothesis. For both trials, we will target a power of 0.8 against an alternative hypothesis that the vaccine efficacy against symptoms is equal to 0.6 for the always-infected stratum (i.e. $S_i^{P_0} = (1,1,1)$ and $S_i^{P_0} = (1,1)$). We reject the null when the posterior probability is 0.85 or larger that vaccine efficacy against severe illness is above 0.1 and that the vaccine efficacy against infection is greater than 0.3. We can write the rejection region for $\text{Data} = \{(\tilde{S}_i, \tilde{Y}_i, Z_i, R_i, A_i, X_i),\ 1 \leq i \leq n\}$ as $\{\text{Data} : P(\text{VE}_{I,31}^{(1,1,1)} > 0.1, \text{VE}_{S,31} > 0.3 \mid \text{Data}) \geq 0.9\}$ for the three-arm trial and $\{\text{Data} : P(\text{VE}_{I,21}^{(1,1)} > 0.1, \text{VE}_{S,21} > 0.3 \mid \text{Data}) \geq 0.85\}$ for the two-arm trial. Our decision criterion is akin to that used in Polack et al. (2020), namely that the posterior probability is greater than 0.986 that vaccine efficacy against confirmed COVID-19 is greater than 0.3. In our scenario 0.85 adequately controls the Type 1 error for a null hypothesis of no vaccine efficacy against severe illness.

We use the model defined in Equation (3.11) along with the nonparametric model in Equation (3.13); the computational details are discussed in Appendix B.7. In each trial we examine two scenarios: one in which we observe the covariate $\tilde{A}_i$, or A_i with error, and one in which we observe A_i directly. Given the results of Theorem 7, we can determine the number of study sites and the number of levels for A_i that need to be observed in order to point identify the causal estimand of interest. For the three-arm trial, we need at least 8 study sites and a covariate with at least 7 levels, while for the two-arm trial we need only 4 study sites and a covariate with at least 3 levels.

The power calculations are presented in Table 3.1, which shows power as a function of the sample size, the number of treatments, and whether A_i or $\tilde{A}_i$ was measured.

Table 3.1: Power against the alternative,$\text{VE}_S \approxeq 0.4$, $\text{VE}_{I,31}^{(1,1,1)} \approxeq 0.6$ for $N_z = 3$, and $\text{VE}_S \approxeq 0.5$, $\text{VE}_{I,21}^{(1,1)} \approxeq 0.6$ for $N_z = 2$ for sample sizes of $4{,}000$ through $120{,}000$. Scenarios in which A_i was measured with error denoted by $\tilde{A}$, A otherwise.

Trial	Measurements	4,000	20,000	40,000	80,000	120,000
3-arm	A	NA	NA	0.69	0.99	1.00
	$\tilde{A}$	NA	NA	0.34	0.88	0.98
2-arm	A	0.01	0.43	0.83	0.94	NA
	$\tilde{A}$	0.01	0.35	0.68	0.93	NA

While these results show that one needs large sample sizes to achieve 80% power for the estimands of interest in both scenarios, this is expected because the always-infected principal strata, $(1, 1, 1)$ and $(1, 1)$, are each only 3.5% of their respective populations in our simulation studies. This highlights the extent to which power calculations for our models are dependent on principal strata proportions. Furthermore, though the sample sizes are large, randomized vaccine trials of similar magnitude have been run. For example, the trial presented in Polack et al. (2020) included approximately $43{,}500$ participants. In our simulation studies, this sample size would exceed a power of 0.8 to detect joint vaccine efficacy against infection and severe symptoms. This highlights the fact that our model can be used to infer post-infection outcome vaccine efficacy from large real-world studies.

3.4.2 Household vaccination study

Consider 2-person households recruited into a vaccination study to determine the infectiousness effect, as defined in Halloran and Struchiner (1995) and further explored in VanderWeele and Tchetgen Tchetgen (2011). In other words, if one person in the pair is infected, what benefit does the other person in the household derive from the vaccination status of the infected individual? VanderWeele and Tchetgen Tchetgen (2011) considered a trial design in which exactly one member of each household is randomized between vaccination and placebo. We consider a trial in which the only source of infection for the non-randomized individual is from the individual randomized to treatment. This might be a good model for households in which one member is home-bound. Then the set of treatments for each household can be mapped to a categorical treatment: $z_1 \equiv (0, 0), z_2 \equiv (1, 0)$. Let the intermediate outcome $S_i(z_j)$ be the infection status of the randomized household member, let the set of principal strata be $\{(0, 0), (1, 0), (0, 1), (1, 1)\}$, and let the outcome $Y_i(z_j, S_i(z_j))$ be the infection status of the unvaccinated individual. The estimand of interest, vaccine efficacy against transmission, is the expected difference in outcome for the unvaccinated individual when the household member is unvaccinated vs. when the household

member is vaccinated for the set of households in the stratum $(1, 1)$:

$$\mathrm{VE}_{T,21}^{(1,1)} = \mathbb{E}\left[Y_i(z_1) \mid S_i^{P_0} = (1,1)\right] - \mathbb{E}\left[Y_i(z_2) \mid S_i^{P_0} = (1,1)\right].$$

VanderWeele and Tchetgen Tchetgen (2011) derive large-sample bounds for this effect, but we can use our method to identify this quantity. Theorem 7 shows that in order to identify this estimand under noisy infection measurements using our method, one would need at least four study sites, a relevant categorical covariate with three levels, and the sensitivity and specificity to both lie in the same half-interval of $[0, 1]$. We write the rejection region for $\mathrm{Data} = \{(\tilde{S}_i, \tilde{Y}_i, Z_i, R_i, A_i, X_i),\ 1 \leq i \leq n\}$ as $\{\mathrm{Data} : P(\mathrm{VE}_{T,21}^{(1,1)} > 0, \mathrm{VE}_{S,21} > 0.3 \mid \mathrm{Data}) \geq 0.975\}$ for the two-arm trial. The 0.975 cutoff was chosen to control the Type 1 error rate, as shown in the Supplementary Material.

Table 3.2: Power against the alternative that $\mathrm{VE}_S \approxeq 0.5$, $\mathrm{VE}_{T,21}^{(1,1)} \approxeq 0.16$. for sample sizes of $4{,}000$ to $80{,}000$. Scenarios in which A_i was measured with error denoted by $\tilde{A}$, A otherwise.

Measurements	4,000	20,000	40,000	80,000
A	0.01	0.49	0.78	0.96
$\tilde{A}$	0.01	0.52	0.87	0.97

In this example, the sample size should be understood in terms of households, rather than participants. Our method is applicable to scenarios involving partial interference, which in this case is the assumption that treatment statuses of households do not impact one another.

3.4.3 Misspecified model

A key assumption for our method is Covariate homogeneity, that the preseason antibody titers are conditionally independent of the study site indicators given covariates and principal stratum. This may not hold when A_i is a measurement of a factor causing infection; conditioning on $S_i^{P_0}$ would introduce collider bias and make A_i and R_i conditionally dependent. In order to measure the robustness of our method to deviations from this assumption, we simulated data in which A_i were not conditionally independent of R_i. To do so, we included an interaction term between A_i and R_i for the multinomial logistic regression model used to simulate A_i measurements. We generated the data under the same null and alternative scenarios as used in Section 3.4.1 We then fitted two models, the full nonparameteric model for A_i, which does not include an interaction effect between R_i, as can be seen in Equation (3.13), and a model which uses a parametric model, namely a multinomial logistic regression for A_i that *does* include an interaction term. This model is shown in Equation (3.12)

The models had similar power against the alternative hypothesis, which is detailed in Table B.4. However, with any misspecified model, it is of interest to investigate if the Type-I error is inflated. We can see in Table 3.3 that the only scenario which shows inflated Type-I error for the misspecified model is the $80{,}000$ observation scenario in which A_i is observed without noise.

Table 3.3: Size of the test,$\text{VE}_S \approxeq 0.5$, $\text{VE}^{(1,1)}_{I,21} = 0$ for $N_z = 2$ for sample sizes of $20{,}000$ through $80{,}000$. Scenarios in which A_i was measured with error denoted by $\tilde{A}$, A otherwise.

Model	Measurements	$20{,}000$	$40{,}000$	$80{,}000$
A Incorrect	A	0.03	0.04	0.28
	$\tilde{A}$	0.01	0.00	0.01
A Correct	A	0.01	0.00	0.04
	$\tilde{A}$	0.01	0.01	0.01

This likely means that the misspecified model does not protect the null asymptotically (Richardson et al., 2011). Despite this, the misspecified model does not show inflated Type I error in the $20{,}000$ and $40{,}000$ scenarios.

3.5 Discussion

Policymakers and public health experts can use vaccine efficacy for post-infection outcomes to design more precise vaccination programs. Our method makes inferring these causal estimands feasible in real-world multi-arm trials where outcomes are measured with error and vaccines cannot be assumed to have a nonnegative effect on infection for every individual. The power of our method is reflected in its flexibility to be applied to vaccine trials with multiple treatments as well as various post-infection outcomes. Although we focus on binary post-infection outcomes here, our method is readily extensible to ordinal and continuous measures, such as immune response as measured by antibody titer. Accordingly, when paired with a parametric likelihood for continuous post-infection outcomes, our method may be more statistically efficient than models identified by likelihood assumptions alone, like that of Zhang et al. (2009). Furthermore, our identifiability results are nonparametric, though we use parametric Bayesian estimators in our examples. One can use these methods to design and analyze clinical trials, as we show in Section 3.4.

3.5.1 Limitations and extensions

As shown in Section 3.4.3, there is evidence that when the conditional independence condition Assumption 4 does not hold, under the null hypothesis, the model will incorrectly asymptotically

reject. More research is necessary to develop models that are robust to misspecification. An extension to the model is to allow misclassification rates that differ by values of covariates X and by treatment assignment Z. The identifiability results readily generalize to both scenarios, and both are of interest in real-world trials. For instance, if a vaccine changes how the virus populates the nasal cavity, we might expect that PCR tests from nasopharyngeal swabs will be less sensitive in the vaccinated group. More work is needed to further generalize the procedure to categorical intermediate outcomes, which would allow for more general vaccine efficacy against transmission study designs (VanderWeele and Tchetgen Tchetgen, 2011), as well as applications beyond vaccine efficacy to noncompliance in multi-arm trials where the exclusion restriction could be violated (Cheng and Small, 2006).

CHAPTER 4

Measuring Cumulative Spatial Exposure to Environmental Hazards

4.1 Introduction

To measure the impact of environmental exposures, epidemiologists must account for both the intensity of exposure from the environmental source of interest, factors impacting individual susceptibility to infection, and other unobserved potential sources of infection. Epidemiologists often employ regression modeling to learn the relationship between exposure, susceptibility and disease risk (Bender, 2009). When a point source exposure to pathogens or other health hazard is suspected, methods have been developed to incorporate these sources of disease into a regression modeling framework (Diggle et al., 1997; Diggle and Rowlingson, 1994). Exposure to the environmental source is typically operationalized as a function of distance to the point source (Diggle, 1990), and the modeler learns how quickly the magnitude of the exposure increases as the distance to the point source decreases. For example, if there is an abnormal clustering of cancer cases near a chemical plant, we may suspect that this cluster arose in part due to exposure to hazardous chemicals.

The assumption that legitimizes the use of these methods is that the location of the exposure to be assessed is certain: the origin of the cancer risk posed by the chemical plant can reasonably be assumed to originate from a single, discrete point in space. Thus, while there may be uncertainty in the parameters governing the exposure as a function of distance to the point source, the modeler assumes that there is no uncertainty in the distance for each unit of analysis. This method has been used to quantify the risk of larynx cancer with respect to distance to an industrial incinerator Diggle (1990), the risks of various cancers in relation to petrochemical plant exposure, Calculli et al. (2010), the risk of multi-drug-resistant tuberculosis infection for individuals living near a prison in Peru, Warren et al. (2018), and to understand the risks of fast food restaurant proximity on childhood obesity Peterson and Sanchez (2018).

When the environmental source of disease is spatially extensive, such as a river, lake, canal network, or pipe system, the fundamental assumption of the point-source approach, i.e. that each unit's exposure can be summarized using a single distance from the source, no longer holds. There is no longer a plausible single point from which exposure emanates; all points that comprise the source could pose a hazard to health. In order to use the point-source method for these types of sources, we need to assume that an individual's exposure can still be summarized by their distance to a single location along the source. The typical assumption is to take the shortest distance Cassell et al. (2018); Jalava et al. (2014). This assumption will necessarily understate the uncertainty in exposure because many points along the source may contribute to exposure at a given community location. These methods also do not allow for variation in pathogen concentration along the environmental hazard, which in certain applications may not accord with reality. For instance when modeling the risk of diarrheal illness with respect to wastewater runoff, enteric pathogen concentrations decline with distance to the source of runoff Brouwer et al. (2017a). Thus units near areas of the waterway that are closer to the wastewater runoff will have higher exposure compared to units that are proximate to sections of the waterway with lower concentrations of pathogens. This can result in biased estimates of unit exposure, whereby average intensity is correctly estimated, but the risk for high-exposure units is underestimated and vice-versa for low-exposure units.

Instead, we need to take into account cumulative exposure to the environmental hazard, where every point contributes to exposure. The approach has several analogues in public health literature. The first analogue is in determining exposure to fine particulate matter ($PM_{2.5}$) and its effect on birth weight in Berrocal et al. (2011). One measurement of exposure proposed was a summation over a fixed window of time of daily $PM_{2.5}$ minus a predefined threshold, only on days that $PM_{2.5}$ exceeded the threshold. This measure accounts for the cumulative exposure to $PM_{2.5}$. When considering exposure to an environmental hazard, instead of summing over time, we can sum exposure over the spatial extent of the hazard. Another analogous technique can be seen in joint models of longitudinal outcomes and time-to-event data, summarized in Hickey et al. (2016) and further generalized in Brilleman et al., where we model the time-dependent hazard of an event dependent on parameters learned from the model for longitudinal outcomes. The form of the interdependence between the time-to-event model and the longitudinal outcome is specified by the modeler. One formulation of the joint model specifies that the log-hazard ratio of the time-to-event model at time t depends on integral over the interval $[0, t]$ of the latent parameter governing the longitudinal outcome Andrinopoulou et al. (2017), which allows the event hazard to depend on the cumulative exposure to the latent process. This corresponds to our problem setting, but substituting space for time.

The problem of assigning risk to a spatially-continuous source of disease is widespread, and the applications of a coherent model for these scenarios are myriad. The most direct application

is in modeling how infectious disease risk depends on household and occupational proximity to waterways. In multiple settings, diarrheal illness risk has been observed to cluster around rivers Thompson et al. (2015), canal systems, and other water sources. Legionella is also known to spread through water and rivers provide one route to infection Cassell et al. (2018). In order to develop an effective response to these public health risks, authorities would benefit from a detailed understanding of the intensity of disease risk at different points along waterways. This could aid in determining points at which to sample water quality, and predictive modeling could suggest groups of households to inform about the health risks. Beyond infectious disease, childhood respiratory diseases such as asthma and bronchitis have been linked to vehicle emissions Perez et al. (2012). Learning how respiratory disease risk changes as a function of cumulative exposure to freeway traffic, as well as how this exposure varies along freeway segments would be a boon to public health in this context as well. Officials could target air pollution mitigation efforts at areas where emissions concentrate and to which households have high exposure. Urban planners could use model predictions to build new housing safely away from the worst highways.

We propose a flexible, generative model for quantifying environmental exposures that naturally extends the point-source approach to account for spatially extensive sources of risk. The model addresses two key problems in the measurement of exposure from spatially extensive point sources. The first is the problem of uncertainty in the point of exposure, e.g. for a 'single hit' model in which the disease outcome is the result of a single exposure (e.g. infection), and the second in which it is impacted by the accumulation of exposure over space and/or time. To accomplish this, we allow each unit's exposure to be integrated across the entirety of the environmental hazard. We use a log-Gaussian process to parameterize the risk at distance zero to the source of risk, which accounts for differences in risk at distinct points along the source. We use Bayesian inference implemented in the software package `CmdStan` to estimate the unknown model parameters Carpenter et al. (2017), where we're able to take advantage of parallel computing of the likelihood. We demonstrate the model's ability to infer environmental exposure under different data generating processes using simulated data, and, finally, we apply our model to data collected on childhood diarrheal disease in Mezquital Valley, Mexico to show how the model yields new insights that cannot be obtained using existing methods.

4.2 Modeling environmental exposure

Suppose we observe outcome data Y_{it}, where i indexes the individual, or stratum of the population, and t designates a time interval, $t = [t_1, t_2]$. Typically Y_{it} is discrete, either representing counts of incident cases of a disease when i represents a stratum, or an indicator variable representing the event that individual i is infected with the disease of interest. These cases are observed in a spatial

domain $\mathcal{R}$ and each observed unit is associated with a location $s_i \in \mathcal{R}$. The spatial domain also contains a potential environmental source of disease, defined as a set $\mathcal{C}$, which may be modeled as a lower-dimensional manifold of $\mathcal{R}$, with associated coordinate function $\ell : \mathcal{C} \to \mathcal{R}$. Then the spatial domain of the environmental hazard is $\ell(\mathcal{C}) \subseteq \mathcal{R}$.

How can we go about learning how the location of the unit i with respect to the set of points $\mathcal{C}$ influences the risk of disease? We might fit a model to the observations, by specifying a distribution P for observations dependent on a parameter μ specific to each location i and interval t and nuisance parameters δ:

$$Y_{it} \sim P(\mu_{it}, \delta). \tag{4.1}$$

We then model how μ_{it} depends on unit location s_i with respect to the hazard $\mathcal{C}$, and also potentially on observed covariates, through a known function h and unknown parameters $\boldsymbol{\theta}$:

$$\mu_{it} = h(s_i, \mathcal{C} \mid \boldsymbol{\theta}, X_{it}).$$

The parameters are identifiable from the observed data via differential exposure to the hazard between units that vary in location. If we can learn these parameters, we'll know to what extent the environmental hazard endangers those exposed.

4.3 Existing approaches to modeling environmental exposure

When the risk factor is a point-source the domain $\mathcal{C}$ comprises a single point, c, exposure can be approximated via a function, $\mathcal{K}(d)$, of the scaled Euclidean distance of location s_i to the location of the point c, given by the function $\ell(c)$: $d_i = \|s_i - \ell(c)\|_2$. For instance, if Y_i are binary indicators of disease, we would model the outcome as a Bernoulli random variable with mean parameter μ_{it}:

$$Y_{it} \sim \text{Bernoulli}(\mu_{it}),$$

and model $h^{-1}(\mu_{it})$ as an additive decomposition of baseline risk of disease λ and the increased risk from the hazard, $f(t)\mathcal{K}(d_i/\rho)$. If we define $\mathcal{K}(d)$ to be a strictly monotone decreasing function that evaluates to 1 at $d = 0$ and tends to 0 as $d \to \infty$, then parameter ρ quantifies how quickly the risk of disease decreases as one moves away from the hazard c, and the parameter $f(t)$ gives the increased risk of disease at distance zero to the source:

$$h^{-1}(\mu_{it}) = \lambda + f(t)\mathcal{K}(d_i/\rho), \tag{4.2}$$

We impose the constraint that $\lambda, f(t) > 0 \forall t$, and h^{-1} is a link function. For a fixed distance, $\mathcal{K}$ is increasing in ρ. Some or all of the parameters that govern $h^{-1}(\mu_{it})$, ρ, λ, and $f(t)$ will typically be unknown and will need to be learned from observed data. Popular choices for $\mathcal{K}$ are the kernel of a Gaussian density and the kernel of an exponential density Warren et al. (2018); Diggle and Rowlingson (1994).

Diggle (1990) introduced the model for noninfectious disease case-control studies, namely cancer. Two independent nonhomogeneous Poisson processes with intensity functions $\lambda_{y_i=1}(s_i)$ and $\lambda_{y_i=0}(s_i)$, are assumed to govern the spatial point pattern of cases and controls respectively. If we define the intensity function for cases $\lambda_{y_i=1}(s_i) = r\lambda_{y_i=0}(s_i)(1+\alpha \exp(-\|s_i-\ell(c)\|_2^2/\rho^2))$, then α represents the incremental intensity for cases at distance 0 to the environmental hazard compared to the base rate of disease r. As $d_i \to \infty$ the risk approaches $r\lambda_{y_i=0}(s_i)$. Inference proceeds by first estimating $\lambda_{y_i=0}(s_i)$, and subsequently estimating r, α and ρ. A later paper Diggle and Rowlingson (1994) shows that inference of the nonhomogeneous Poisson process intensity function can be avoided by conditioning on the locations of the cases and controls and fitting a nonlinear binary regression. The modeled outcome is the binary event that an observation is a disease case conditional on the location of the observation. The probability of a case is μ_i:

$$\mu_i = \frac{r(1 + \alpha \exp(-\|s_i - \ell(c)\|_2^2/\rho^2))}{1 + r(1 + \alpha \exp(-\|s_i - \ell(c)\|_2^2/\rho^2))}. \tag{4.3}$$

The model can be derived by splitting a single Poisson process with intensity

$$r\lambda_{y_i=0}(s_i)(1 + \alpha \exp(-\|s_i - \ell(c)\|_2^2/\rho^2)) + \lambda_{y_i=0}(s_i)$$

into cases and controls with a location-dependent thinning probability μ_i. Then $P(Y_i = 1 \mid s_i) = \mu_i$.

The model in Diggle and Rowlingson (1994) is a special case of equation (4.2) and corresponds to modeling the odds of disease, h^{-1} as

$$h^{-1}(\mu) = \frac{\mu}{1-\mu},$$

and

$$h^{-1}(\mu_i) = r + r\alpha \exp(-\|s_i - \ell(c)\|_2^2/\rho^2) \tag{4.4}$$

$\lambda = r$, and $f(t) = r\alpha$. This equation can be rearranged to show that the environmental exposure multiplies the base odds of disease:

$$h^{-1}(\mu_i) = r(1 + \alpha \exp(-\|s_i - \ell(c)\|_2^2/\rho^2)) \tag{4.5}$$

When modeling the effects of multiple environmental hazards on a disease of interest, the odds model in equation (4.5) can be extended either multiplicatively:

$$h^{-1}(\mu_i) = r \prod_{q=1}^{Q} (1 + \alpha_q \exp(-\|s_i - \ell(c)\|_2^2/\rho_q^2)), \tag{4.6}$$

as in Diggle and Rowlingson (1994); Ramis et al. (2011), or additively:

$$h^{-1}(\mu_i) = r \, (1 + \sum_{q=1}^{Q} \alpha_q \exp(-\|s_i - \ell(c)\|_2^2/\rho_q^2)), \tag{4.7}$$

as in Biggeri and Lagazio (1999), though the authors note the inferential issues pertaining to correlated hazards.

4.3.1 Extending the model to infectious disease

The model has been successfully employed in infectious disease settings too. Warren et al. (2018) extends the point-source exposure model to infectious diseases by modeling the probability of cases of tuberculosis being multi-drug-resistant tuberculosis (MDR-TB) with respect to distance to a prison. In this case, a probit model is used, corresponding to $h^{-1}(\mu) \equiv \Phi^{-1}(\mu)$, or the inverse cumulative distribution function of the standard normal distribution.

The key difference between the model used in Warren et al. (2018) and Diggle and Rowlingson (1994) is that in Warren et al. (2018) no assumption is made about the independence of cases and noncases of MDR-TB. In Diggle and Rowlingson (1994) the cases and controls are modeled to have arisen from independent nonhomogeneous Poisson processes, which is not a valid assumption when modeling infectious disease.

4.3.2 Extensive environmental hazards

When the environmental exposure is not a point source, but is instead extensive in space, such as a river, or a lake, modelers typically resort to using the shortest distance to the source for each unit i as a proxy for exposure Cassell et al. (2018). If $\mathcal{C}$ is the set of all points that make up the source, then we can apply the same model above, defined in equation (4.2), but with d_i defined as:

$$d_i = \min_{c \in \mathcal{C}} \|s_i - \ell(c)\|_2 \tag{4.8}$$

where ℓ maps an element of $\mathcal{C}$ to its spatial coordinates. This accounting doesn't capture the true extent of the unit's dose, however, because the unit has a cumulative exposure from the entirety

of the environmental risk. This method also imposes the implicit constraint that all points on the canal impart equal exposure at distance zero, which doesn't allow the researcher to model how exposure changes as a function of c.

4.4 Cumulative exposure to extensive environmental hazards

As discussed in section 4.1, not accounting for each unit's cumulative exposure to the environmental hazard can introduce bias in estimating the risk of disease for low- and high-exposure units. In order to ameliorate this deficiency, we can extend the methodology introduced in section 4.3 by partitioning the hazard into mutually exclusive subsets C_m, $m \in \{1, \ldots, M\}$, with spatial centroids $\bar{C}_m$, and spatial areas $\Delta(C_m)$, such that $\bigcup_{m=1}^{M} C_m = \mathcal{C}$. Then we could treat each section C_m of the hazard as a separate point source:

$$h^{-1}(\mu_i) = r\,(1 + \sum_{m=1}^{M} f_m \Delta(C_m) \mathcal{K}(\|s_i - \ell(\bar{C}_m)\|_2/\rho)). \tag{4.9}$$

Further, we can jointly model the dependence between different subsets by specifying a joint distribution for the vector of exposures at distance zero, $\mathbf{f}$,

$$\mathbf{f} \sim P_f. \tag{4.10}$$

This model is nearly a direct extension of model (4.2) and it accomplishes what we could not by taking the shortest distance to the environmental hazard: the model accounts for cumulative exposure and it allows the concentration of the disease-causing agent to change along the hazard.

In infectious disease modeling, we often model how risk depends additively on exposures Crawford et al. (2019), rather than multiplicatively, like in (4.9) and (4.6). We can change the model to allow for an additive relationship between the base rate λ and the environmental exposure $\mathcal{C}$:

$$h^{-1}(\mu_i) = \lambda + \sum_{m=1}^{M} f_m \Delta(C_m) \mathcal{K}(\|s_i - \ell(\bar{C}_m)\|_2/\rho)). \tag{4.11}$$

In section 4.5 we will show how an additive decomposition of risk arises naturally from a generative model for infection and environmental exposure to a point source hazard. Then in section 4.5.3 we will show how equation (4.11) can be derived for exposure to an extensive hazard when pathogens along the hazard are distributed according to a nonhomogeneous Poisson process. Finally, we will specify a nonparametric model for P_f in line (4.10), which further extends the model to the generative scenario where pathogens along the source are distributed according to a

log-Gaussian Cox process.

4.5 A new perspective on environmental exposure

Researchers and statisticians often find value in fitting generative models to observed data because these models can tell coherent probabilistic stories about how the data arose, and allow researchers to encode scientific information about the modeled phenomenon through distributions. This then enables probabilistic model checking which can yield new models or suggest new datasets to collect that would refine or enhance the scientific insights. In the context of environmental health, generative models also allow us to examine counterfactual scenarios in order to estimate, e.g. the proportion of observed disease risk attributable to a given exposure. These and other benefits of generative modeling led us to recast the environmental hazard model as a natural consequence of a specific probabilistic story about how environmental disease data came about. This recasting allows us to extend the point-source model via an expanded generative process that yields a more detailed picture of environmental exposure.

4.5.1 Dose-response model

Our treatment begins with a model for disease called the exponential dose-response model. It is a generative model for infectious disease and, as such, can be used as a building block in a more realistic model of infection from a point-source exposure. The model is typically used to infer the dose of a pathogen that would lead to a 50% probability of infection or symptomatic disease.

First, we define some notation to be used throughout the chapter. Let a time interval, $[t_1, t_2]$, be defined for $t_1 < t_2$, and, with a slight abuse of notation, let $t = [t_1, t_2]$. Let the nonhomogeneous Poisson process rate $N_i(t)$ be the number of disease-causing pathogens to which a unit i is exposed over the time interval t. For shorthand, let $\Lambda(t) = \int_{t_1}^{t_2} \lambda(\tau) d\tau$:

$$N_i(t) \sim \text{Poisson}(\Lambda(t)).$$

Subsequently, let the number of pathogens infecting unit i, $K_i(t) \mid N_i(t)$, be conditionally binomial with probability of success parameter r_i:

$$K_i(t) \mid N_i(t) \sim \text{Binomial}(N(t), r_i).$$

Marginalizing over $N_i(t)$ yields

$$K_i(t) \sim \text{Poisson}(r_i \Lambda(t)). \tag{4.12}$$

Then the probability that unit i becomes infected over the time interval t is $1 - \exp(-r_i\Lambda(t))$, or the probability that at least one of these pathogens infects person i. The parameter r_i represents the susceptibility of unit i to the disease and may depend on covariates X_i. See Brouwer et al. (2017b) for more discussion on the merits of this model.

Given that $N_i(t)$ is a Poisson process on $\mathbb{R}^+$, $K_i(t)$ is a thinned Poisson process on $\mathbb{R}^+$ with mean measure $r_i\lambda(\tau)$. The probability of i becoming infected over the interval t is the probability of observing the first jump of a Poisson process in interval t, which is again $1 - \exp(-r_i\Lambda(t))$.

We may also model the dose of a disease causing agent as the result of a probabilistic exposure process governed by the distance to a hazardous environmental source. Recall that the environmental source of disease is denoted c and is located at $\ell(c)$. Suppose $W_i(t)$ is a second Poisson process with mean measure $f(\tau)$ that defines the pathogens emitted from this hazardous source. Again, let $F(t) = \int_{t_1}^{t_2} f(\tau)d\tau$

$$W_i(t) \mid F(t) \sim \text{Poisson}(F(t)). \tag{4.13}$$

Now we let $N_i(t) \mid W_i(t)$ be the dose that reaches unit i over t and assume that this is conditionally binomial:

$$N_i(t) \mid W_i(t) \sim \text{Binomial}(W(t), p_i).$$

If individual i is located at s_i, then we can define the probability that a pathogen reaches individual i as a function of i's distance to the source: $p_i = \mathcal{K}(\|s_i - \ell(c)\|_2/\rho)$. Marginalizing over $W_i(t)$ yields the marginal distribution for $N_i(t)$

$$N_i(t) \mid F(t) \sim \text{Poisson}(F(t)\mathcal{K}(\|s_i - \ell(c)\|_2/\rho)).$$

Using the results from equation (4.12) allows the derivation of the marginal distribution of pathogens that infect individual i.

$$K_i(t) \mid F(t) \sim \text{Poisson}(r_iF(t)\mathcal{K}(\|s_i - \ell(c)\|_2/\rho)).$$

Finally, the probability that i becomes infected from the hazardous source located at $\ell(x)$ is Bernoulli:

$$Y_{it} \mid f(t) \sim \text{Bernoulli}(1 - \exp(-r_if(t)\mathcal{K}(\|s_i - \ell(c)\|_2/\rho))). \tag{4.14}$$

Equation (4.14) is similar to equation (4.2) but with a different inverse link function:

$$g^{-1}(\theta) = -\log(1-\theta),$$

$$Y_{it} \mid \theta_{it} \sim \text{Bernoulli}(\theta_{it}) \tag{4.15}$$

$$g^{-1}(\theta_{it}) = r_i f(t) \mathcal{K}(\|s_i - \ell(c)\|_2/\rho)). \tag{4.16}$$

4.5.2 Expanding the model to include background rate of exposure

The generative model can readily accommodate a term for representing spatially-invariant background exposure if the baseline exposure is modeled as an independent homogeneous Poisson process. Let the mean measure of the background exposure process be $\lambda_b d\tau$. Then the total dose is the sum of the two exposures:

$$N_i(t) \sim \text{Poisson}(F(t)\mathcal{K}(\|s_i - \ell(c)\|_2/\rho) + \lambda_b t).$$

This then yields a model for binary infection as:

$$P(Y_{it} = 1 \mid F(t)) = 1 - \exp(-r_i(F(t)\mathcal{K}(\|s_i - \ell(c)\|_2/\rho) + \lambda_b t)). \tag{4.17}$$

The difference between the model in equation (4.3) and equation (4.17) is that the risk from the environmental hazard adds to the background risk. In equation (4.3) the odds of an observation being a disease case is modeled as:

$$\text{odds}(P(Y_{it} = 1 \mid f(t))) = r_i(1 + \alpha \exp(-\|s_i - c\|_2^2/\rho^2)). \tag{4.18}$$

so we see that risk from the environmental hazard multiplies the background rate, r_i.

4.5.2.1 Identifiability

Identifiability for a parametric model with a family of densities $f(x \mid \theta)$ over $\mathbb{R}^n$ indexed by parameter vector $\theta \in \Theta \subseteq \mathbb{R}^d$ is defined as the condition that $\theta' \neq \theta \in \Theta$ implies that $f(x \mid \theta') \neq f(x \mid \theta)$ Rothenberg (1971). Given the multiplicative structure of the parameter space, we will not be able to learn r_i without assuming a functional form. For now, we let $F(t)' = r_i F(t)$ and $\lambda_b' = r_i \lambda_b$ be the parameters of interest. Imagine we can move individual i to different locations while keeping the hazard c fixed at fixed location $\ell(c)$. Let $d_i = \sqrt{s_i - \ell(c)}$. Given the kernel function's property that $\lim_{d\to\infty} \mathcal{K}(d) \to 0$, $\lambda_b' = \lim_{d_i\to\infty} -\log(1 - P(Y_{it} = 1 \mid F(t)))/t$.

Furthermore, $F'(t) = \lim_{d\to 0} -\log(1 - P(Y_{it} = 1 \mid F(t))) - \lambda_b' t$. Finally, given that $\mathcal{K}$ is strictly monotonic, $\mathcal{K}^{-1}$ exists, and

$$\rho = \frac{d_i}{\mathcal{K}^{-1}\left((-\log(1 - P(Y_{it} = 1 \mid F(t))) - \lambda_b' t)/F'(t)\right)}.$$

4.5.3 Modeling exposure to extensive environmental hazards

If the source is extensive in space, we can extend the generative model for point source hazards. Let $\mathcal{C}$ be the set of points that make up the environmental hazard and allow the intensity of the Poisson distributed pathogens, $f(\tau)$ in equation (4.13), to depend on $c \in \mathcal{C}$: $f(c, \tau)$, and model the pathogens present at the source as a nonhomogeneous Poisson process with domain $\mathcal{C} \times \mathbb{R}^+$. Then the number of pathogens in a subset $C \subset \mathcal{C}$ is:

$$W(t) \mid F(c, t) \sim \text{Poisson}\left(\int_{C\times t} f(c, \tau) d\mathcal{C} d\tau\right). \tag{4.19}$$

If we allow $f(c, \tau)$ to be a stochastic process itself, we can model W as a doubly-stochastic Poisson process. Then the probability that unit i gets infected over time interval t from section C of the environmental hazard is:

$$Y_{it} \mid F(c, t) \sim \text{Bernoulli}\left(1 - \exp\left(-r_i \int_{C\times t} \mathcal{K}(\|s_i - \ell(c)\|_2/\rho) f(c, \tau) d\mathcal{C} d\tau\right)\right). \tag{4.20}$$

The assumption that $W(t)$ is a nonhomogeneous Poisson process ensures that $f(c, \tau)$ is integrable, and thus given the property that $0 < \mathcal{K} \leq 1$ the integral in Equation (4.20) is well-defined. A question remains, however, as to how to model $f(c, \tau)$. Our model must guarantee almost surely integrable functions $f(c, \tau)$ but be flexible so as to adapt to many different scenarios. For this reason we model $\log f(c, \tau)$ as a Gaussian process. In the next two subsections we describe the properties of Gaussian processes and doubly-stochastic Poisson processes with log-Gaussian intensities.

4.5.3.1 Gaussian processes

Gaussian processes are stochastic processes over a domain $\mathcal{V}$ where every finite-dimensional joint distribution of the stochastic process is multivariate normally distributed:

$$\mathbf{z} \sim \text{Multivariate Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

for any points v_i, $i \in [1, \ldots, N]$, with $\boldsymbol{\mu}_{[i]} = \mu(v_i)$ and $\Sigma_{[i,j]} = \sigma(v_i, v_j)$ and each element of $\mathbf{z}$ associated to a single v_i. Gaussian processes are completely characterized by the mean function

$\mu(v)$ and covariance kernel $\sigma(v, v')$. The Gaussian process represents a useful prior for unknown functions because the sample paths of a GP can be thought of random functions from $\mathcal{V} \to \mathbb{R}$. The properties of the sample paths, such as differentiability and nonlinearity, are defined by $\mu(v)$ and $\sigma(v, v')$. An example of a covariance kernel is the exponentiated quadratic kernel:

$$\sigma(v, v' \mid \alpha, \omega) = \alpha^2 \exp\left(-\frac{1}{2\omega^2}\|v - v'\|_2^2\right), \tag{4.21}$$

which generates infinitely differentiable sample paths as long as $\sigma(v, v' \mid \alpha, \omega)$ is positive definite within $\mathcal{V}$(Rasmussen and Williams, 2006) . The hyperparameter ω controls the nonlinearity of the function, in that large values of ω lead to almost-linear functions, while smaller values of ω lead to functions with many peaks and troughs over the domain. The hyperparameter α controls the marginal variance of the Gaussian random variable at a given location v. In our case, the domain of the Gaussian process is $(\mathcal{C} \times \mathbb{R}^+)$ and $v = (c, \tau)$. Next we describe how to use Gaussian processes to define the log-intensity of a Poisson process.

4.5.3.2 Log-Gaussian Cox processes

A GP prior for $\log f(c, \tau)$ is known as a log-Gaussian Cox process Møller et al. (1998). A log-Gaussian Cox process is a doubly-stochastic Poisson process where the intensity function $\lambda(c, \tau)$ is an exponentiated Gaussian process: $\lambda(c, \tau) = \exp Z(c, \tau)$, where $Z(c, \tau)$ is a Gaussian process.

A nonhomogeneous Poisson process inherits the complete randomness property from homogeneous Poisson processes when conditioning on the intensity process. Namely if $C_1, \ldots, C_k \subset (\mathcal{C} \times \mathbb{R}^+)$ and $C_i \cap C_j = \emptyset$ then $W(C_1), \ldots, W(C_k)$ are independent Poisson random variables with intensities $\int_{C_i} \lambda(c, \tau) d\mathcal{C} d\tau$. Let the collection of pairs $(C_m, t_l), m \in [1, \ldots, M], l \in [1, \ldots, L]$ be a partition of $(\mathcal{C} \times t)$. Then if $Z(c, \tau)$ is a GP with domain $(\mathcal{C} \times \mathbb{R}^+)$, with a valid covariance function σ (validity as defined in (Møller et al., 1998, p. 453)), $W(C_m, t_l) \mid Z(c, \tau) \sim$ Poisson$(\int_{C_m \times t_l} \exp Z(c, \tau) d\mathcal{C} d\tau)$, and $W(C_m, t_l) \perp\!\!\!\perp W(C_j, t_l) \mid Z(c, \tau) \forall j \neq m$. Given the intractability of $\int_{C_m \times t_l} \exp Z(c, \tau) d\mathcal{C} d\tau$, we can approximate this quantity by generating a finite-dimensional draw from a multivariate normal with $M \times L$ elements:

$$\mathbf{z} \sim \text{Multivariate Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

where $\Sigma_{i,j} = k((\bar{C}_{\lfloor i/L \rfloor + 1}, \bar{t}_{i \mod L}), (\bar{C}_{\lfloor j/L \rfloor + 1}, \bar{t}_{j \mod L}))$ where $\bar{C}_m$, $\bar{t}_l$ are the coordinates of the centroid of C_m, and t_l. Then we approximate the integral $\int_{C_m \times t_l} \exp Z(c, \tau) d\mathcal{C} d\tau$ at the centroid:

$$\int_{C_m \times t_l} \exp Z(c, \tau) d\mathcal{C} d\tau \approx \exp(\mathbf{z}_{[(m-1)L+l]}) \Delta(C_m) \Delta(t_l),$$

where $\Delta(C_m) \times \Delta(t_l)$ is the volume of element $C_m \times t_l$. Then

$$W(C_m, t_l) \mid \mathbf{z} \sim \text{Poisson}(\exp(\mathbf{z}_{[(m-1)L+l]})\Delta(C_m)\Delta(t_l)).$$

We can use the above properties to approximate the distribution of the total pathogens generated over $(\mathcal{C} \times t)$:

$$W(\mathcal{C}, t) \mid \mathbf{z} \sim \text{Poisson}\left(\sum_{m=1}^{M}\sum_{l=1}^{L}\exp(\mathbf{z}_{[(m-1)L+l]})\Delta(C_m)\Delta(t_l)\right).$$

The approximation error depends on the nature of $Z(c, \tau)$, namely how much $Z(c, \tau)$ varies within the approximation intervals. To make this precise:

$$\begin{aligned}
&\left|\int_{\mathcal{C}\times t}\exp(Z(c,\tau))d\mathcal{C}d\tau - \sum_{m=1}^{M}\sum_{l=1}^{L}\exp(Z(\bar{C}_m,\bar{t}_l))\Delta(C_m)\Delta(t_l)\right| \\
&= \left|\int_{C_m\times t_l}\exp(Z(c,\tau))d\mathcal{C}d\tau - \sum_{m=1}^{M}\sum_{l=1}^{L}\exp(Z(\bar{C}_m,\bar{t}_l))\Delta(C_m)\Delta(t_l)\right| \\
&\leq \sum_{m=1}^{M}\sum_{l=1}^{L}\left|\int_{C_m\times t_l}\exp(Z(c,\tau))d\mathcal{C}d\tau - \exp(Z(\bar{C}_m,\bar{t}_l))\Delta(C_m)\Delta(t_l)\right| \\
&\leq \sum_{m=1}^{M}\sum_{l=1}^{L}\sup_{\{(c,\tau),(c',\tau')\}\in C_m\times t_l}\left|\exp(Z(c,\tau)) - \exp(Z(c',\tau'))\right|\Delta(C_m)\Delta(t_l)
\end{aligned}$$

Thus if the sample paths do not vary much within the intervals, the integral will be well-approximated by the sum. If $Z(c, \tau)$ is s-Hölder continuous over $\mathcal{C} \times t$, then there is a bound for $|\exp(Z(c, \tau)) - \exp(Z(c', \tau'))|$:

$$|\exp(Z(c,\tau)) - \exp(Z(c',\tau'))| \leq B\|(c,\tau) - (c',\tau')\|^s,$$

for $s \in (0, 1]$; we have also used the fact that exp is 1-Hölder continuous. If we let $\Delta(C_m) = \Delta(\mathcal{C})/M$ and $\Delta(t_l) = (t_2 - t_1)/L$ for all m and l, the bound will be

$$\begin{aligned}
&\left|\int_{\mathcal{C}\times t}\exp(Z(c,\tau))d\mathcal{C}d\tau - \sum_{m=1}^{M}\sum_{l=1}^{L}\exp(Z(\bar{C}_m,\bar{t}_l))\Delta(C_m)\Delta(t_l)\right| \\
&\leq B\Delta(\mathcal{C})(t_2 - t_1)\left(\sqrt{\left(\frac{\Delta(\mathcal{C})}{M}\right)^2 + \left(\frac{t_2 - t_1}{L}\right)^2}\right)^s
\end{aligned}$$

B and s are properties of the Gaussian process, while M and L are properties of our integration

scheme. This treatment assumes that $\mathcal{C}$ is Euclidean. If the domain is non-Euclidean, there is additional error incurred from discretizing the non-Euclidean domain. We will not treat that error in this manuscript, but it is worth noting.

4.5.4 Log-Gaussian Cox process integrated exposure

With the log-Gaussian Cox process in hand, we can continue to build our model for extensive environmental hazards. If we treat section C_m of the source as if it were a point source, we may define the number of pathogens from C_m to which individual i is exposed over the time interval t_l as:

$$K(C_m, t_l) \mid \mathbf{z} \sim \text{Poisson}(\mathcal{K}\left(\tfrac{\|\ell(\bar{C}_m) - s_i\|_2}{\rho}\right) \exp(\mathbf{z}_{[m,l]})\Delta(C_m)\Delta(t_l)).$$

Given the independence of $W(C_m, t_l) \perp\!\!\!\perp W(C_j, t_l) \mid Z(c, \tau) \forall j \neq m$, we can express the total number of particles individual i is exposed to as

$$\sum_{l=1}^{L}\sum_{m=1}^{M} K(C_m) \mid \mathbf{z} \sim \text{Poisson}\left(\sum_{l=1}^{L}\sum_{m=1}^{M} \mathcal{K}\left(\tfrac{\|\ell(\bar{C}_m) - s_i\|_2}{\rho}\right) \exp(\mathbf{z}_{[m,l]})\Delta(C_m)\Delta(t_l)\right).$$

Finally, the probability that individual i becomes infected during t is

$$Y_{it} \mid \mathbf{z} \sim \text{Bernoulli}\left(1 - \exp\left(-r_i \sum_{l=1}^{L}\sum_{m=1}^{M} \mathcal{K}\left(\tfrac{\|\ell(\bar{C}_m) - s_i\|_2}{\rho}\right) \exp(\mathbf{z}_{[m,l]})\Delta(C_m)\Delta(t_l)\right)\right). \tag{4.22}$$

Using the same argument as above, we can show that as $M, L \to \infty$ our approximation converges to the integral in Equation (4.20), substituting $\exp Z(c, \tau)$ for $f(c, \tau)$, under conditions on the kernel and mean function of the Gaussian process. The proof is shown in Appendix C.

While we have presented the model in its full generality, with f nonparametrically dependent on τ, we'll make the simplifying assumption going forward that the intensity of the pathogen generation from the environmental hazard is constant in time with rate λ_e. In other words, $f(c, \tau) = f(c)\lambda_e$, so $Z(c, \tau) = Z(c) + \log(\lambda_e)$, for computational tractability. This leads to the probability model for Y_{it} being expressed as

$$Y_{it} \mid Z(c, \tau) \sim \text{Bernoulli}\left(1 - \exp\left(-r_i\, \lambda_e t \int_{\mathcal{C}} \mathcal{K}\left(\tfrac{\|\ell(c) - s_i\|_2}{\rho}\right) \exp(Z(c)) d\mathcal{C}\right)\right).$$

4.5.4.1 Computational considerations

Given the approximate exposure term:

$$\sum_{l=1}^{L}\sum_{m=1}^{M}\mathcal{K}\left(\frac{\|\ell(\bar{C}_m)-s_i\|_2}{\rho}\right)\exp(Z(\bar{C}_m,\bar{t}_l)\Delta(C_m)\Delta(t_l),$$

and the true modeled exposure

$$\int_{\mathcal{C}\times t}\mathcal{K}\left(\frac{\|\ell(c)-s_i\|_2}{\rho}\right)\exp(Z(c,\tau))d\mathcal{C}d\tau,$$

we can bound the approximation error:

$$\left|\sum_{l=1}^{L}\sum_{m=1}^{M}\mathcal{K}\left(\frac{\|\ell(\bar{C}_m)-s_i\|_2}{\rho}\right)\exp(Z(\bar{C}_m,\bar{t}_l)\Delta(C_m)\Delta(t_l)-\int_{\mathcal{C}\times t}\mathcal{K}\left(\frac{\|\ell(c)-s_i\|_2}{\rho}\right)\exp(Z(c,\tau))d\mathcal{C}d\tau\right|. \tag{4.23}$$

As shown in Appendix C, this error has an upper bound of:

$$\sum_{l=1}^{L}\sum_{m=1}^{M}\int_{C_m\times t_l}\mathcal{K}_\rho(\bar{C}_m)\left|\exp(Z(c,\tau))-\exp(Z(\bar{C}_m,\bar{t}_l)\right|d\mathcal{C}d\tau \tag{4.24}$$

$$+\sum_{l=1}^{L}\sum_{m=1}^{M}\left(\mathcal{K}\left(\frac{\inf_c\|\ell(c)-s_j\|_2}{\rho}\right)-\mathcal{K}\left(\frac{\sup_c\|\ell(c)-s_j\|_2}{\rho}\right)\right)\int_{C_m\times t_l}\exp(Z(c,\tau))d\mathcal{C}d\tau. \tag{4.25}$$

Of most consequence is the term involving differences of the distance kernel:

$$\mathcal{K}\left(\frac{\inf_c\|\ell(c)-s_j\|_2}{\rho}\right)-\mathcal{K}\left(\frac{\sup_c\|\ell(c)-s_j\|_2}{\rho}\right).$$

For units that are close to the hazard, this term is approximately $1-\mathcal{K}\left(\frac{\sup_c\|\ell(c)-s_j\|_2}{\rho}\right)$. When ρ is also close to zero this term will be near 1, leading to large integration error. Thus, when there are many units that are near the environmental hazard, the computational grid should be finer than it would need to be if few units were near the hazard in order to guarantee small approximation error.

4.5.5 Include covariates for susceptibility

Suppose we have covariates $\mathbf{x}_i\in\mathbb{R}^K$ associated with each individual i that predict the individual's susceptibility to infection. For example, these might be measurements on age, diet, or comorbidi-

ties. Then we could use a log-linear model for $\lambda_e r_i$ conditional on covariates $\mathbf{x}_i$, with $\boldsymbol{\gamma} \in \mathbb{R}^K$:

$$\lambda_e r(\mathbf{x}_i) = e^{-\boldsymbol{\gamma}^T \mathbf{x}_i}.$$

Including the background rate of exposure along with the covariates would result in the observational model:

$$Y_{it} \mid \mathbf{z} \sim \text{Bernoulli}\left(1 - \exp\left(-e^{-\boldsymbol{\gamma}^T \mathbf{x}_i} t \left(\lambda_b + \sum_{m=1}^{M} \mathcal{K}\left(\frac{\|\ell(\bar{C}_m) - s_i\|_2}{\rho}\right) \exp(\mathbf{z}_{[m]}) \Delta(C_m)\right)\right)\right). \tag{4.26}$$

4.5.6 Model Identifiability

The probability model in Equation (4.26) is not identifiable as written. To see why we expand the expression for the multivariate normal $\mathbf{z}$ as the sum of $\boldsymbol{\mu} + \boldsymbol{\eta}$:

$$Y_{it} \sim \text{Bernoulli}\left(1 - \exp\left(-e^{-\boldsymbol{\gamma}^T \mathbf{x}_i} t \left(\lambda_b + \sum_{m=1}^{M} \mathcal{K}\left(\frac{\|\ell(\bar{C}_m) - s_i\|_2}{\rho}\right) \exp(\boldsymbol{\mu}_{[m]} + \boldsymbol{\eta}_{[m]}) \Delta(C_m)\right)\right)\right)$$

$$\boldsymbol{\eta} \sim \text{Multivariate Normal}(0, \boldsymbol{\Sigma}_{\alpha,\omega}).$$

If we added a constant to $\boldsymbol{\mu}$, multiplied λ_b by the exponentiated constant, and subtracted the constant from the intercept term in γ^T we would not change the likelihood. Thus, we restrict $\boldsymbol{\mu} = 0$ and fix $\alpha = 1$, neither of which will restrict the Gaussian process from representing unknown functions (Ghosal and Roy, 2006). Furthermore, given that the total risk of infection is a product of the individual hazard of infection and the sum of the instantaneous exposure from the environmental hazard and background hazard, we cannot infer the scale of the individual hazard of infection. We fix the intercept term of $\boldsymbol{\gamma}$ to be 0. The term $e^{\boldsymbol{\gamma}^T \mathbf{x}_i}$ now models the relative risk of infection for two individuals at the same location. The identified model is

$$Y_{it} \sim \text{Bernoulli}\left(1 - \exp\left(-e^{-\boldsymbol{\gamma}^T \mathbf{x}_i} t \left(\lambda_b + \sum_{m=1}^{M} \mathcal{K}\left(\frac{\|\ell(\bar{C}_m) - s_i\|_2}{\rho}\right) \exp(\boldsymbol{\eta}_{[m]}) \Delta(C_m)\right)\right)\right) \tag{4.27}$$

$$\boldsymbol{\eta} \sim \text{Multivariate Normal}(0, \boldsymbol{\Sigma}_{1,\omega}). \tag{4.28}$$

As in 4.5.2.1, λ_b is identified as the limit as $\|\ell(\bar{C}_m) - s_i\|_2 \to \infty$, and $e^{-\boldsymbol{\gamma}^T \mathbf{x}_i}$ is identified by comparing individuals within households with different values of covariates.

4.6 Canal system simulation study

Our simulation study set up is similar to our intended application of the model: we simulate a survey of childhood diarrheal illness in households located near a system of wastewater canals within a region extending 10km horizontally and 4km vertically. This region is denoted $\mathcal{R}$. The system of canals and diarrheal illness risk is shown in Figure 4.2. The left-hand plot shows the geographic location of the canal in dashed red lines, while the flow of the wastewater is shown in solid blue arrows. There are three canal segments: the segment x_1 runs horizontally along the bottom edge of the region, segment y runs vertically through the middle of the region, and segment x_2 runs horizontally and intersects y_1 at $2\frac{2}{3}$km. In keeping with the notation developed in Section 4.5.3, the extent of the environmental hazard is the set $\mathcal{C} = \{x_1, x_2, y\}$.

Sources of wastewater are denoted as υ and are indexed by the canal segment to which they are associated; sinks are denoted δ and are similarly indexed. The diarrhea-causing pathogens along segment x_1 are generated according to a nonhomogeneous Poisson process with $\Lambda_{x_1}(c) = 0.15 + c^2/100$, while the pathogens along segment y are generated with $\Lambda_y(c) = \Lambda_{x_1}(5) + c^2/16$. Canal segment x_2 has an intensity of $\Lambda_{x_2}(c) = \Lambda_y(8/3) - 1/4 + c^2/100$. These intensities are such that $\Lambda_y(0) = \Lambda_{x_1}(5)$ and $\Lambda_y(8/3) = \Lambda_{x_2}(5)$, and are respectively indicated as $\Lambda_{x_1 \times y}$, and $\Lambda_{x_2 \times y}$.

We simulate two populations of household locations to investigate our method's sensitivity to the distribution of households. One scenario, which we term the "uniform" scenario, all houses are uniformly distributed within $\mathcal{R}$, while in the "clustered" scenario, the houses are distributed near the canal system. We simulate $200{,}000$ household locations, from which we draw simple random sample of size J, where $J \in \{500, 1000, 2000\}$. An example of the household locations for $J = 500$ under the two scenarios is shown in Figure 4.1.

For each household in the population, indexed by j and with geographic location s_j, we can define the cumulative exposure to the wastewater pathogens from a canal segment $\nu \in \mathcal{C}$ with endpoints ν_1, ν_2 as

$$\int_{\nu_1}^{\nu_2} \exp\left(-\tfrac{\|\ell_\nu(c) - s_j\|_2}{\rho}\right) \Lambda_\nu(c) dc.$$

Then the total exposure for household j from the entirety of the canal is

$$\mathcal{E}_j = \sum_{\nu \in [x_1, x_2, y]} \int_{\nu_1}^{\nu_2} \exp\left(-\tfrac{\|\ell_\nu(c) - s_j\|_2}{\rho}\right) \Lambda_\nu(c) dc. \tag{4.29}$$

Of note, we have chosen the exponential kernel, $\exp\left(-\frac{\|\ell_\nu(c) - s_j\|_2}{\rho}\right)$, for the true measure of exposure at a given distance from a differential element of the canal.

For the i-th observation within the j-th household, we observe $Y_{ij} \in \{0, 1\}$, where Y_{ij} is the binary indicator for disease. We simulate I total draws per household, which takes the values 10 and

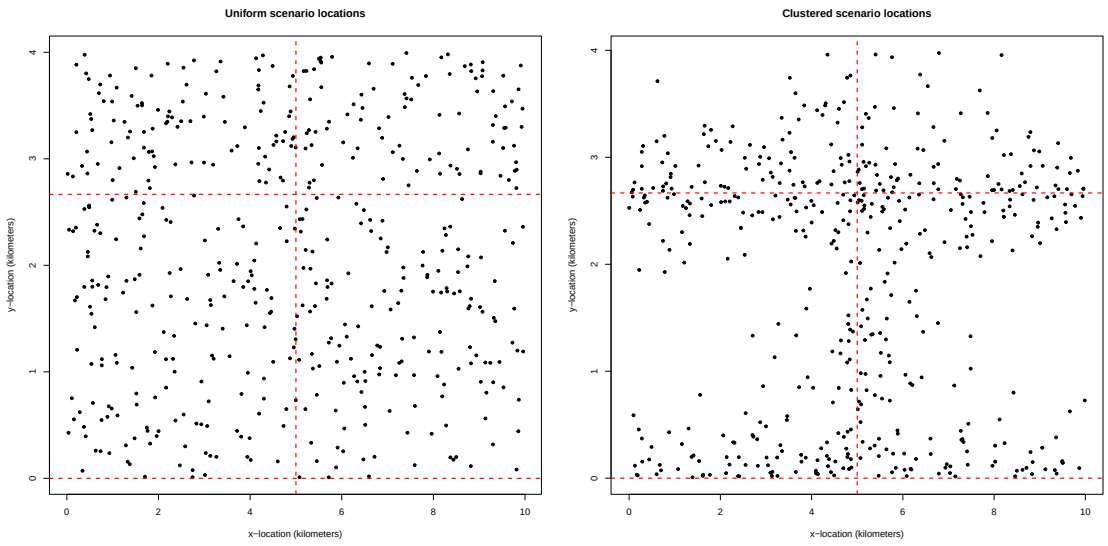


Figure 4.1: Graph shows household locations with respect to the canal segments under the uniform scenario (left) and the clustered scenario (right). Dashed lines indicate the geographic location of the canal segments. Black dots indicate the household locations.

100 for each J within each scenario. The table of simulation scenarios is shown in Appendix C.3.

These Y_{ij} are conditionally independent Bernoulli draws:

$$Y_{ij} \mid \Lambda, s_j \sim \text{Bernoulli}\left(1 - \exp\left(-\exp(X_j\,\gamma)\left(\lambda + \sum_{\nu \in [x_1, x_2, y]} \text{exposure}(\nu, s_j, \rho)\right)\right)\right),$$

with $\lambda = 0.05$, $\rho = 0.1$, $\gamma = -0.15$, and $X_j \sim \text{Normal}(0, 1)$. The integrals are numerically evaluated using Gauss-Kronod quadrature implemented in base R language's `integrate` (R Core Team, 2021).

The right-hand graph in figure 4.2 shows the function $P(Y_{ij} = 1 \mid s)$ with $X_j = 0$. The graph shows that the risk of disease concentrates close to the canal system and decays as the distance to the canal increases. Figure 4.2 also shows that the risk of disease is higher for a fixed y coordinate and an increasing x coordinate.

4.6.1 Inferential model likelihood

The inferential model is that of Equation (4.27) applied to the canal system shown in Figure 4.2. We define the finite dimensional realization of $\log \Lambda_\nu(c) = Z_\nu(c)$ as $\mathbf{z}_\nu$, with dimension M_ν, for canal segment ν. This finite-dimensional draw of the Gaussian process prior is associated with partition $\{(C_m, \bar{C}_m, \Delta(C_m)) \mid m = 1, \ldots, M_\nu\}$ such that $\bigcup_{m=1}^{M_\nu} C_m = \nu$. As in Section 4.5.3.2 the centroid of partition section m is $\bar{C}_m$, and, as such, the m^{th} element of $\mathbf{z}_\nu$ is $Z_\nu(\bar{C}_m)$. Then we can

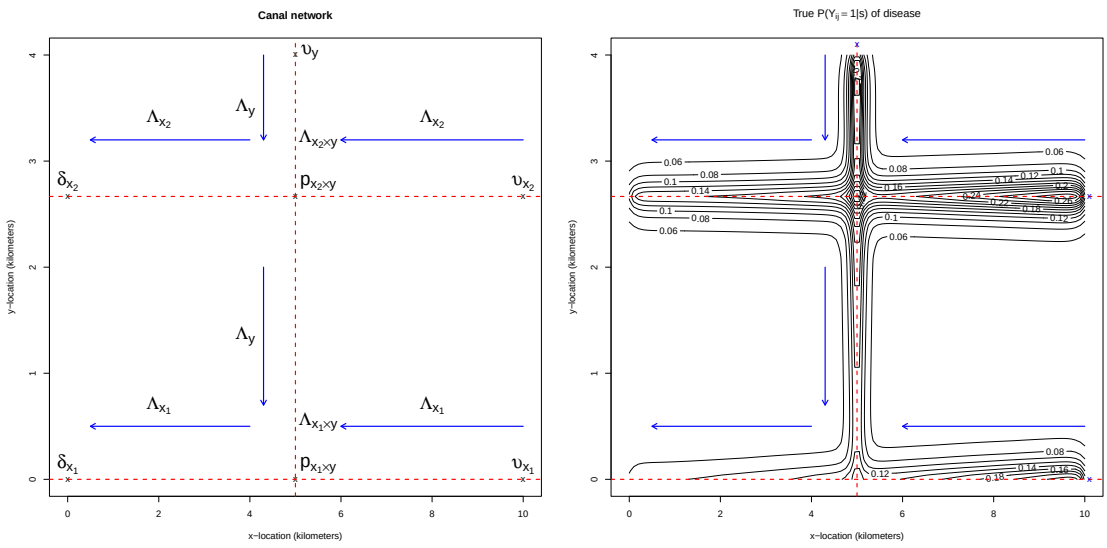


Figure 4.2: Left: Dashed lines indicate the geographic location of the canal segments x_1, x_2, y. Blue arrows indicate the flow of wastewater. Crosses indicate points of interest on the canal network: υs are sources of wastewater, δs are sinks of wastewater, and ps are canal intersections. Λs denote the intensity function of the canal segment or point. Right: True probability surface, with arrows depicting the flow of water through the canal.

define the approximate modeled exposure as:

$$F(\nu, s_j, \rho, \mathbf{z}_\nu) = \sum_{m=1}^{M_\nu} \exp\left(-\frac{\|\ell_\nu(C_m) - s_j\|_2}{\rho}\right) \exp((\mathbf{z}_\nu)_m)\Delta(C_m). \tag{4.30}$$

Let $\theta_j^{\text{environ}}$ be the total approximate exposure:

$$\theta_j^{\text{environ}} = \sum_{\nu \in \{x_1, x_2, y\}} F(\nu, s_j, \rho, \mathbf{z}_\nu). \tag{4.31}$$

We assume the functional form for the kernel, namely the exponential kernel, is known.

The full inferential model is:

$$Y_{ij} \mid x_j, \mathbf{z}_{x_1}, \mathbf{z}_{x_2}, \mathbf{z}_y \sim \text{Bernoulli}\left(1 - \exp\left(-\exp(x_j\,\gamma)\left(\lambda + \theta_j^{\text{environ}}\right)\right)\right).$$

Given the discussion in Section 4.5.4.1, we would expect larger error the clustered scenario vs. the uniform scenario. We thus use two grid sizes to investigate the impact of approximation error on our inferences. We let $M_{x_1} = M_{x_2} = M_y = M$ for $M = 40, 160$. Then $\mathbf{z}_\nu$ is in $\mathbb{R}^M$ for each ν.

The partition associated with x_1 and x_2 is

$$\{([10\,\tfrac{n-1}{M}, 10\,\tfrac{n}{M}], 10\,\tfrac{2n-1}{2M}, 10/M) \mid n = 1, \ldots, M\}. \tag{4.32}$$

Thus the n^{th} element of $\mathbf{z}_{x_1}$ is $Z_{x_1}(10\,\frac{2n-1}{M})$, while the n^{th} element of $\mathbf{z}_{x_2}$ is $Z_{x_2}(10\,\frac{2n-1}{M})$. The partition associated with $\mathbf{z}_y$ is

$$\{([\tfrac{8}{3}\,\tfrac{n-1}{M/2}, \tfrac{8}{3}\,\tfrac{n}{M/2}], \tfrac{8}{3}\tfrac{2n-1}{M}, \tfrac{16}{3M}) \mid n = 1, \ldots, M/2\}, \tag{4.33}$$

$$\{([\tfrac{8}{3} + \tfrac{4}{3}\tfrac{n-21}{M/2}, 4\,\tfrac{n-20}{M/2}], \tfrac{4}{3}\tfrac{2(n-M/2)-1}{M} + \tfrac{8}{3}, \tfrac{8}{3M}) \mid n = M/2+1, \ldots, M\}. \tag{4.34}$$

4.6.1.1 Log-intensity priors

The Gaussian process prior we use for $\mathbf{z}_{x_1}$, $\mathbf{z}_{x_2}$ and $\mathbf{z}_y$ reflects that $\Lambda_{x_1}(5) = \Lambda_y(0)$ and $\Lambda_{x_2}(5) = \Lambda_y(\frac{8}{3})$. We impose the constraint by conditioning the values of $\mathbf{z}_x$, $\mathbf{z}_{x_2}$, and $\mathbf{z}_y$ at the intersection points to be equal. This is akin to the construction of string Gaussian processes introduced in Samo and Roberts (2015), which explores the formal construction of Gaussian process priors connected at intersection points such that the Gaussian process defined at the intersection is finitely differentiable. Specifically, let $Z(p_{\nu\times\xi})$ be the value of the Gaussian field at the intersection of canal segments ν and ξ and call $p_{\nu\times\xi}$ the coordinates of the point of intersection. For instance, the intersection of x_1 and y in figure 4.2 is denoted $p_{x_1\times y}$ and is the point $(5, 0)$. The value of the field at point $p_{x_1\times y}$ would be $Z(p_{x_1\times y})$, and $Z_{x_1}((p_{x_1\times y})_1) = Z_y((p_{x_1\times y})_2) = Z(p_{x_1\times y})$. Let the mean and variance of the field at the intersection be $\mu_{x_1\times y}, \sigma^2_{x_1\times y}$.

Let $\Sigma_{x_1,\omega}$, and $\Sigma_{y,\omega}$ to be the covariance matrices associated with each Gaussian process defined on partitions Equation (4.32) and Equation (4.33) with length-scale hyperparameter set to ω. Let $\Sigma_{\nu,\omega}(p_{\nu\times\xi})$ to be the vectors of covariances associated with the centroids of the partition for ν and an intersection point $p_{\nu\times\xi}$. Then the joint prior is

$$\begin{aligned}
\mathbf{z}_{x_1} \mid Z(p_{x_1\times y}) \sim \text{Multivariate Normal}\Big(&\sigma^{-2}_{x_1\times y}\Sigma_{x_1,\omega}(p_{x_1\times y})(Z(p_{x\times y}) - \mu_{x\times y}),\\
&\Sigma_{x_1,\omega} - \sigma^{-2}_{x_1\times y}\Sigma_{x_1,\omega}(p_{x_1\times y})\Sigma_{x_1,\omega}(p_{x_1\times y})^T\Big)\\
\mathbf{z}_{y} \mid Z(p_{x_1\times y}) \sim \text{Multivariate Normal}\Big(&\sigma^{-2}_{x_1\times y}\Sigma_{y,\omega}(p_{x_1\times y})(Z(p_{x\times y}) - \mu_{x\times y}),\\
&\Sigma_{y,\omega} - \sigma^{-2}_{x_1\times y}\Sigma_{y,\omega}(p_{x_1\times y})\Sigma_{y,\omega}(p_{x_1\times y})^T\Big)\\
Z(p_{x_1\times y}) \sim \text{Normal}&(\mu_{x_1\times y}, \sigma^2_{x_1\times y})
\end{aligned}$$

The conditional distributions for the intersection of Z_{x_2} and Z_y is defined similarly. This prior

has computational benefits as well, because it allows for parallel computation of the covariance matrices Σ_y and the evaluation of the prior.

Of interest in the prior construction is that y's partition is more fine than that of x_1, x_2. Specifically, y has a discretization size of $\frac{1}{15}$km on one subsection and $\frac{2}{15}$km, while x_1 and x_2 both have grids at $\frac{1}{2}$km resolution. In order to use the same Gaussian process prior on all three canal segments, we need to scale the length-scale parameter ω on canal section y so distances are the same on y and x_1. We define $\omega_y = \omega\frac{\Delta_y}{\Delta_{x_1}}$, which allows distances in y to be scaled to be distances as measured within x_1. $\Delta_y = \frac{1}{15}$ or $\Delta 215$ depending on the subsection of y.

In order to reflect the dependence structure induced by the flow of the canal system, we put independent Normal$(0, 1)$ priors on the sources of the canal waterway, points $\upsilon_{x_1}, \upsilon_{x_2}, \upsilon_y$, in figure 4.2. Then the values at points 5 and 2 are defined as:

$$Z(p_{x_2\times y}) \mid Z(\upsilon_{x_2}), Z(\upsilon_y) \sim \text{Normal}(\mu_{x_2\times y}, \sigma^2_{x_2\times y}) \tag{4.35}$$

$$\mu_{x_2\times y} = \frac{1}{\sqrt{2}}\left(\exp\left(-\frac{d(p_{x_2\times y},\upsilon_y)^2}{2\omega^2}\right)Z(\upsilon_y) + \exp\left(-\frac{d(p_{x_2\times y},\upsilon_{x_2})^2}{2\omega^2}\right)Z(\upsilon_{x_2})\right) \tag{4.36}$$

$$\sigma_{x_2\times y} = \sqrt{1 - \tfrac{1}{2}\left(\exp\left(-\frac{d(p_{x_2\times y},\upsilon_y)^2}{\omega^2}\right) + \exp\left(-\frac{d(p_{x_2\times y},\upsilon_{x_2})^2}{\omega^2}\right)\right)}, \tag{4.37}$$

and

$$Z(p_{x_1\times y}) \mid Z(\upsilon_{x_1}), Z(\upsilon_{x_2\times y}) \sim \text{Normal}(\mu_{x_1\times y}, \sigma^2_{x_1\times y}) \tag{4.38}$$

$$\mu_{x_1\times y} = \frac{1}{\sqrt{2}}\left(\frac{1}{\sigma_{x_2\times y}}\exp\left(-\frac{d(p_{x_2\times y},p_{x_1\times y})^2}{2\omega^2}\right)(Z(p_{x_2\times y}) - \mu_{x_2\times y}) \right. \tag{4.39}$$

$$\left. + \exp\left(-\frac{d(p_{x_1\times y},\upsilon_{x_1})^2}{2\omega^2}\right)Z(\upsilon_{x_1})\right) \tag{4.40}$$

$$\sigma_{x_1\times y} = \sqrt{1 - \tfrac{1}{2}\left(\exp\left(-\frac{d(p_{x_2\times y},p_{x_1\times y})^2}{\omega^2}\right) + \exp\left(-\frac{d(p_{x_1\times y},\upsilon_{x_1})^2}{\omega^2}\right)\right)}. \tag{4.41}$$

where $d(\cdot,\cdot)$ is the distance between two points along the canal. The priors for $Z(\delta_{x_1}), Z(\delta_{x_2})$ are defined similarly.

In order to formulate priors for the parameters λ, γ, ρ, and α, we sampled from the prior predictive distribution, as advised in Gabry et al. (2019b):

$$p(y) = \int p(y \mid \boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta},$$

if the vector $\boldsymbol{\theta}$ represents a concatenation of all of the model parameters. The goal is to generate plausible observations from our model with the joint prior distribution $p(\boldsymbol{\theta})$.

For λ and γ we use independent Normal$^+(0, 0.3)$ priors, and for α we use a Gamma$(4, 1)$ prior. For ρ, we use a weakly-informative prior (Gelman et al. (2008)) of Normal$^+(0, 0.5)$.

4.6.2 Target estimands

Our model's inferential target is the cumulative exposure from the canal, defined above in Equation (4.29) for household j as $\mathcal{E}_j$. In order to measure how well our model predicts this exposure, we measured the bias, and posterior credible interval coverage for this quantity Recall the model's approximate exposure is $\theta_j^{\text{environ}}$, defined in eqs. (4.30) to (4.31). Let $\rho^\star$ be the value of ρ that generated the simulated data, in our case 0.1. Then the absolute bias in this estimand is

$$\left|\theta_j^{\text{environ}} - \mathcal{E}_j\right| = \left| \sum_{\nu \in [x_1, x_2, y]} \sum_{m=1}^{M_\nu} \int_{C_m} \left(\exp\left(-\tfrac{\|\ell_\nu(\bar{C}_m) - s_j\|_2}{\rho}\right) \exp(Z(\bar{C}_m)) \right.\right.$$
$$\left.\left. - \exp\left(-\tfrac{\|\ell_\nu(c) - s_j\|_2}{\rho^\star}\right) \Lambda_\nu(c) \right) dc \right|.$$

Let $\mathcal{K}_\rho(c) = \exp\left(-\frac{\|\ell_\nu(c) - s_j\|_2}{\rho}\right)$ The error within a partition interval C_m is

$$\begin{aligned}
&\int_{C_m} \left(\mathcal{K}_\rho(\bar{C}_m) \left(\exp(Z(\bar{C}_m)) - \Lambda_\nu(c)\right) + \Lambda_\nu(c) \left(\mathcal{K}_\rho(\bar{C}_m) - \mathcal{K}_{\rho^\star}(c)\right)\right) dc \\
&\leq \mathcal{K}_\rho(\bar{C}_m) \int_{C_m} \left|\exp(Z(\bar{C}_m)) - \Lambda_\nu(c)\right| dc + \int_{C_m} \Lambda_\nu(c) \left|\mathcal{K}_\rho(\bar{C}_m) - \mathcal{K}_{\rho^\star}(c)\right| dc \\
&\leq \mathcal{K}_\rho(\bar{C}_m) \int_{C_m} \left|\exp(Z(c)) - \Lambda_\nu(c)\right| dc + \int_{C_m} \Lambda_\nu(c) \left|\mathcal{K}_\rho(c) - \mathcal{K}_{\rho^\star}(c)\right| dc \\
&+ \mathcal{K}_\rho(\bar{C}_m) \int_{C_m} \left|\exp(Z(\bar{C}_m)) - \exp(Z(c))\right| dc + \int_{C_m} \Lambda_\nu(c) \left|\mathcal{K}_\rho(\bar{C}_m) - \mathcal{K}_\rho(c)\right| dc
\end{aligned}$$

Thus, given that $\mathcal{K}_\rho(\bar{C}_m) \leq 1$, and assuming the approximation error is small:

$$\left|\theta_j^{\text{environ}} - \mathcal{E}_j\right| \leq \sum_{\nu \in \{x_1, x_2, y\}} \int_\nu \left|\exp(Z_\nu(c)) - \Lambda_\nu(c)\right| dc + \sup_{c \in \nu} \Lambda_\nu(c) \int_\nu \left|\mathcal{K}_\rho(c) - \mathcal{K}_{\rho^\star}(c)\right| dc.$$

The upper bound on the absolute bias in the estimand is thus a function of the integrated absolute error in the intensity approximation, and the integrated error in our inference for ρ weighted by the true intensity function, and the resolution of the partition for ν. Thus it is of interest to quantify the

approximate integrated absolute and mean-squared error in $\Lambda_{x_1}, \Lambda_{x_2}, \Lambda_y$, as well as the bias for ρ.

4.6.2.1 Error estimates

The bias for a point estimator $\hat{\phi}$ with true value $\phi^\star$ is calculated as $\text{bias}(\hat{\phi}, \phi^\star) = \hat{\phi} - \phi^\star$. Our point estimator for each parameter is the posterior mean so in the results that follow, the bias for a given dataset $\mathcal{D}$ is $\text{bias}(\mathbb{E}\left[\phi \mid \mathcal{D}\right], \phi^\star)$. The expectation over datasets is approximated by the empirical mean over S simulated datasets $\{\mathcal{D}_s, s = 1, \ldots, S\}$ is $\frac{1}{S}\sum_s \text{bias}(\mathbb{E}\left[\phi \mid \mathcal{D}_s\right], \phi^\star)$. Similarly, the mean-squared error is calculated as $\frac{1}{S}\sum_s \text{bias}(\mathbb{E}\left[\phi \mid \mathcal{D}_s\right], \phi^\star)^2$. We also compute the empirical coverage of the equi-tailed 80%-credible intervals for the household-level environmental exposure for a posterior quantile function for θ_j given dataset $\mathcal{D}$ $Q_{\theta_j|\mathcal{D}}(p)$ as

$$\text{cover}\left(Q_{\theta_j|\mathcal{D}}, \theta_j^\star\right) = \mathbb{1}\left(\theta_j^\star \in \left(Q_{\theta_j|\mathcal{D}}(0.1), Q_{\theta_j|\mathcal{D}}(0.9)\right)\right).$$

Then the empirical mean coverage across simulations is given as

$$\frac{1}{S}\sum_s \frac{1}{J}\sum_j \text{cover}(Q_{\theta_j|\mathcal{D}_s}, (\theta_j)_s^\star).$$

4.6.3 Inference procedure

We run full Bayesian inference inference in CmdStanR, an implementation of the Stan modeling language and inference algorithms using dynamic Hamiltonian Monte Carlo Carpenter et al. (2017); Betancourt (2018); Gabry and Češnovar (2021). Each model was run with four Markov chain Monte Carlo chains for $2{,}000$ iterations of warmup and $2{,}000$ iterations post-warmup samples with a target Metropolis acceptance rate of 0.95 during warmup. Convergence was monitored using the Gelman-Rubin diagnostic, $\hat{R}$, Gelman and Rubin (1992); Vehtari et al. (2020). All parameters achieved $\hat{R}$ near 1 ($\max \hat{R} < 1.01$), and the minimum bulk and tail effective sample size divide by the total post-warmup samples across all parameters and simulations, was 0.07 and 0.04. While these figures are lower than the recommended 10% minimum effective sample size cutoff, we note that when $M = 160$, only 2 out of $1{,}200$ datasets had minimum tail effective sample size less than 10% of the total post-warmup sample size, while only 1 out of $1{,}200$ dataset had minimum bulk effective sample size out of total post-warmup sample size of less than 10%. For $M = 40$, the number was 4 and 1 out of $1{,}200$ for minimum tail and bulk effective sample sizes of less than 10%. There were divergent transitions during sampling for small minority of models. The total divergent transitions were small compared to the total post-warmup samples ($\approx 0.2\%$).

Given our reliance on MCMC sampling, our posterior mean and quantile estimators are Monte-Carlo estimators.

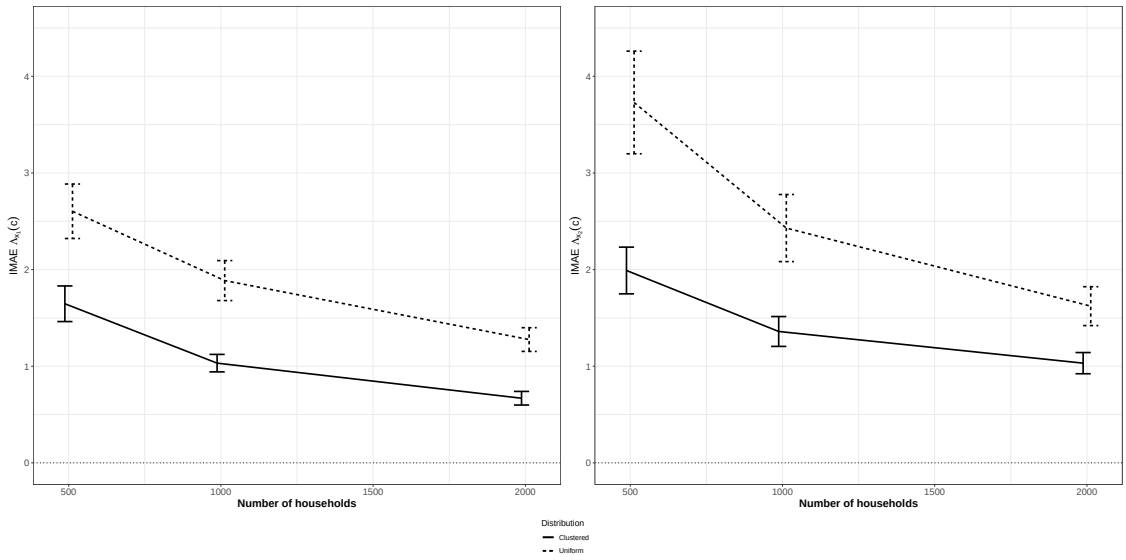


Figure 4.3: Integrated mean absolute error for Λ_{x_1} and Λ_{x_2} with ± 2 standard errors plotted as black bars, 10 observations per household, grid resolution of $M = 160$.

4.6.4 Results

The most salient result from the simulation study is that the distribution of households with respect to the canal has a large influence on the accuracy of the inferences. All results that follow use the posterior mean of the parameter as the estimator. In figure 4.4, we can see that we estimate ρ more precisely when households are clustered near the canal, but, on the contrary, we estimate λ, the spatially-invariant risk of disease, more precisely when houses are uniformly distributed. This makes sense, as the units that are most informative about λ are those that are far from the canal, and we have far fewer of those observations when houses are clustered near the canal. Naturally, we have more households that are far from the canal when the houses are arrayed uniformly on the $[0, 10] \times [0, 4]$ plot of land.

However, there is more information about the intensities $\Lambda_\nu(c)$ near the canal, so we see in Figure 4.3 that the clustered household scenario allows for smaller integrated mean absolute error compared to the uniform scenario.

We can also see in figure 4.5 that when the model is applied to either clustered or uniformly sampled households, the 80% intervals achieve the nominal coverage. The uniform scenario yields negatively biased estimates of the sample average environmental exposure, $\frac{1}{J}\sum_j \theta_j^{\text{environ}}$. This is likely due to the fact that the prior for ρ shrinks the posterior towards zero so with less information about ρ in the observed data in the uniform scenario, the prior continues to shrink the integrated risk towards zero.

Figures for 100 observations per household are shown in Appendix C, and a comparison of

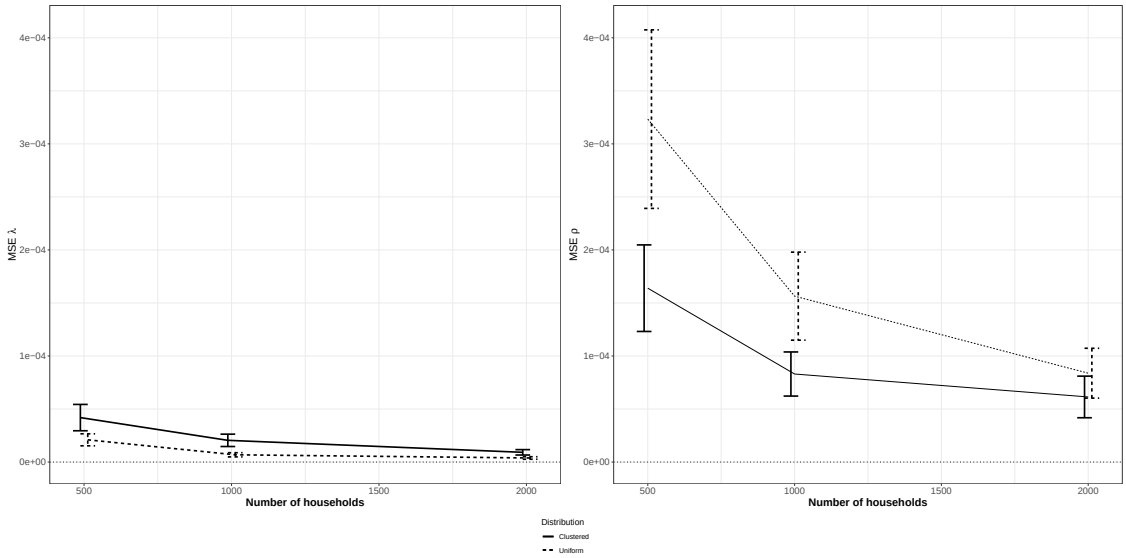


Figure 4.4: MSE for ρ and λ with ± 1.96 standard errors plotted as black bars, x-jittered for clarity on the plot for ρ, 10 observations per household, grid resolution of $M = 160$.

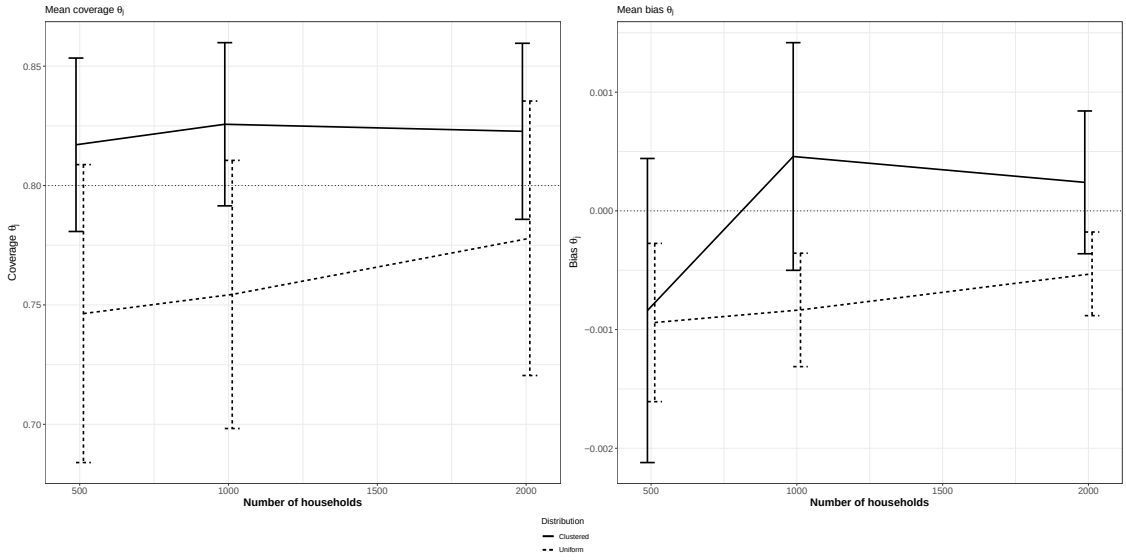


Figure 4.5: Bias and 80% interval coverage for $\theta_j^{\text{environ}}$ ± 1.96 standard errors plotted as black bars, x-jittered for clarity on both plots. The horizontal dotted line in the left plot corresponds to the nominal coverage of 50%, while the horizontal dotted line in the right plot corresponds to zero bias, 10 observations per household, grid resolution of $M = 160$.

results for grid resolution of $M = 40$ vs. $M = 160$ is presented. The comparisons show very low rates of coverage of θ_j for the 80% posterior intervals when $M = 40$. This coverage gets worse as the number of households increase. The results also show that bias in our posterior-mean estimators for θ_j persists despite increasing data sizes. These results highlight the importance of using a grid resolution that is appropriate for the problem setting. Our intuition that approximation error would be worse for the clustered scenario is borne out in Figure C.11. The picture is complicated by the fact that, on a percentage basis, the bias is worse for the uniform scenario, as seen in Figure C.12. This is due to the fact that many households in the uniform scenario have very small risk from the canal system because $\rho = 100\text{m}$ and the household locations are uniformly distributed in a 10km by 4km square.

4.7 Application

We apply our integrated risk model to survey data collected from 2017 through 2019 measuring how household proximity to a wastewater canal influences childhood diarrheal incidence in Mezquital Valley, Mexico. See Contreras et al. (2020) for more detail on data collection, and descriptive statistics. The data are longitudinal measurements of diarrheal disease in children by household. These households are located along and near the wastewater canals, and grouped into small localities. GPS coordinates were taken for each household, along with the GIS data for the canals. Privacy concerns prevent us from sharing the full map of the households.

4.7.1 Models

We model Y_{tijk}, survey responses of diarrheal illness for child i in household j at survey wave t in locality k. The model must account for changes in susceptibility due to age of the child, wealth of the household, parental education, and the intra-local correlation of exposure. We fit two models, the first of which is the model presented in subsection 4.5.5, the second of which is the model fitted in Contreras et al. (2020).

The portion of the Mezquital Valley wastewater canal system on which we are focused has 43 segments. We index these segments ν by q, of which there are Q total segments: $\nu_q, q = 1, \ldots, Q$. Let the parameters accounting for age-related differences in susceptibility be $\beta_{\texttt{age}}$, the parameter for differences in susceptibility over time be $\beta_{\texttt{wave}[t]}$, and the wealth and education-related parameters be $\beta_{\texttt{wealth}}, \beta_{\texttt{educ}}$, respectively. Let β_k be the increased exposure for locality k with respect to locality 1. The parameter ρ is defined as the spatial scale of exposure to the canal, and λ is defined as the spatially-invariant exposure to diarrheal illness.

As in Equation (4.30) we define $F(\nu, s_j, \rho, \mathbf{z}_\nu)$ to be the exposure at household location s_j to

canal segment ν for a given bandwidth ρ:

$$F(\nu, s_j, \rho, \mathbf{z}_\nu) = \sum_{m=1}^{M_\nu} \exp\left(-\frac{\|\ell_\nu(\bar{C}_m) - s_j\|_2}{\rho}\right) \exp((\mathbf{z}_\nu)_m)\Delta(C_m),$$

Let $\Sigma_{\nu,\omega}$ be the marginal covariance matrix for multivariate Gaussian random variable $\mathbf{z}_\nu$, and let $\Sigma_{\nu,\omega}(\nu_1, \nu_2)$ be the conditional covariance matrix conditional on the values of the Gaussian process at points ν_1, ν_2. Let $\mu_{\nu,\omega}(\nu_1, \nu_2)$ be the conditional mean function also dependent on the values of the random field at ν_1, ν_2. Then we may define the full inferential model as

$$\begin{aligned}
Y_{tijk} &\sim \text{Bernoulli}(1 - \exp(-\lambda_{tijk})) \\
\lambda_{tijk} &= \exp(\beta_{\texttt{age}[it]} + \beta_{\texttt{wave}[t]} + \beta_{\texttt{wealth}[j]} + \texttt{educ}_j\beta_{\texttt{educ}}) \\
&\quad \times (\exp(\beta_{\texttt{local}[k]}) + \sum_{q=1}^{Q} F(\nu_q, s_j, \rho, \mathbf{z}_\nu)) \\
\mathbf{z}_\nu \mid Z(\nu_1), Z(\nu_2) &\sim \text{GP}(\mu_{\nu,\omega}(\nu_1, \nu_2), \Sigma_{\nu,\omega}(\nu_1, \nu_2)) \\
\rho &\sim \text{Normal}^+(0, 0.5) \\
\beta_{\texttt{local}j} &\sim \text{Normal}(0, 1), \\
\omega &\sim \text{Gamma}(3.7, 0.9)
\end{aligned} \tag{4.42}$$

Distance is measured in kilometers, so we discretized the canal in 50-meter-long segments so $\Delta(C_m) = 0.05 \,\forall\, m$. This simplifies the Gaussian process construction in that we do not need to scale ω in order to define distances between canal segments with different discretization sizes. The prior for the length-scale of the Gaussian process puts 99% of its mass between 7km and 130km, which enforces the soft constraint that intensity slowly varies along the canal. The total length of the canal is about 4,400km, so a 130km length scale is still relatively local compared to the total length of the canal.

The second model we fit is a version of the minimum distance model presented in subsection 4.3.2. This model is a logistic regression with a predictor for shortest distance to the canal, and is presented in equation (4.43). We define the parameters that account for age-related differences in log-odds of disease as $\beta_{\texttt{age}}$, the parameter for differences in log-odds over time as $\beta_{\texttt{wave}[t]}$, and the wealth and education-related parameters as $\beta_{\texttt{wealth}}, \beta_{\texttt{educ}}$, respectively.

We control for the intra-household correlation of log-odds of diarrhea in house j with a parameter $\beta_{\texttt{house}[j]}$, over which we put multivariate normal prior with covariance matrix $\Sigma(\vec{s} \,|\, \alpha, \vec{\tau})$. We define $\vec{\tau}$ to be a vector of locality-specific scales for $\beta_{\texttt{house}[j]}$. If we collect all the household locations into a vector $\vec{s}$ and define the Euclidean distance between household i and j as $d_{i,j}$ then

the elements of the covariance matrix are parameterized with a Matérn 3/2 covariance kernel:

$$\Sigma(\vec{s}\,|\,\alpha,\vec{\tau})_{i,j} = \tau^2_{\texttt{local}[i]}\left(1+\frac{\sqrt{3}d_{i,j}}{\alpha}\right)\exp\left(-\frac{\sqrt{3}d_{i,j}}{\alpha}\right)\mathbb{1}\left(\texttt{local}[i]=\texttt{local}[j]\right).$$

This model allows us to take into account inter-household correlation of log-odds of diarrhea.

The full model is below:

$$\begin{aligned} Y_{tijk} &\sim \text{Bernoulli}(\mu_{tijk}),\\ \log(\text{odds}(\mu_{tijk})) &= \beta_{\texttt{age}[it]} + \beta_{\texttt{wave}[t]} + \beta_{\texttt{wealth}[j]} + \texttt{educ}_j\beta_{\texttt{educ}}\\ &\quad + \beta_{\text{canal}}\log\left(\min_{c\in\mathcal{C}}\|s_j-\ell(c)\|_2\right) + \beta_{\texttt{house}[j]},\\ \vec{\beta}_{\texttt{house}} &\sim \text{Multivariate Normal}(0,\Sigma(\vec{s}\,|\,\alpha,\vec{\tau})). \end{aligned} \tag{4.43}$$

The model will help elucidate the differences between our new method and the simpler methods currently in use. We will call this model the shortest-distance model, while we refer to our proposed model as the integrated exposure model.

4.7.2 Model inferences

The integrated exposure model infers that there is a small increased risk of diarrheal infection as distance to a point on the canal decreases, as is shown in figure 4.6. The posterior mean of ρ is 0.01 with a standard deviation of 0.006. We estimate the posterior mean of λ to be 0.016 with a standard deviation of 0.005.

From the figure we can see that exposure to wastewater canal is nearly zero when a household is located at 200 meters to the canal. On the other hand model (4.43) shows a much slower decline in risk. For instance, we show the change in odds of diarrheal illness as distance to the canal increases from ten meters to one kilometer compared to the odds of diarrheal illness at ten meters in figure 4.7. The odds of diarrhea for the integrated exposure model, given a household with location s_j, is given by

$$\exp(-(\lambda+\sum_{q=1}^{Q}F(\nu_q,s_j,\rho,\mathbf{z}_{\nu_q})))-1$$

so the change in odds for a household located at s_1 compared to s_2 is

$$\frac{\exp(-(\lambda+\sum_{q=1}^{Q}F(\nu_q,s_2,\rho,\mathbf{z}_{\nu_q})))-1}{\exp(-(\lambda+\sum_{q=1}^{Q}F(\nu_q,s_1,\rho,\mathbf{z}_{\nu_q})))-1}-1.$$

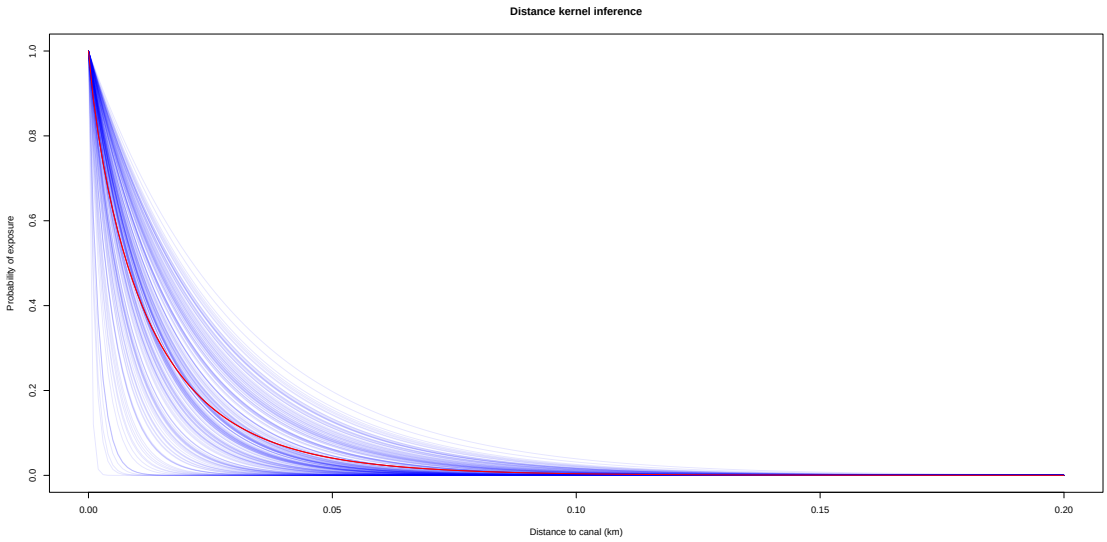


Figure 4.6: Posterior distribution for $\mathcal{K}(d/\rho)$. Red line indicates the posterior mean.

The change in odds for the shortest-distance model is given by

$$\exp\left(\beta_{\texttt{canal}} \log \frac{d_1}{d_2}\right) - 1$$

where d_1 and d_2 are the shortest distances to the canal for locations s_1 and s_2. It is clear from figure 4.7 that the reduction in odds as distance to the canal increases is more extreme for the integrated model than for the shortest-distance model.

Our integrated exposure model decocts the two processes that contribute to exposure from environmental hazards: the geometry of the hazard with respect to the at-risk population, and the variable concentration of enteric pathogens along the hazard.

4.8 Discussion

We have presented a new model for exposure to environmental sources of disease that are extensive in space. We showed how it was connected to existing methods for inferring exposure to potential point-source hazards and how our method can arise from a generative model when applying our model to infectious disease. After applying our model to five hundred simulated datasets under several different sampling scenarios we investigated the performance of our model inferences for these scenarios. We fitted a real dataset with our model and compared the results to a model using existing methodology for extensive environmental hazards, where we were able to show the benefits of our model.

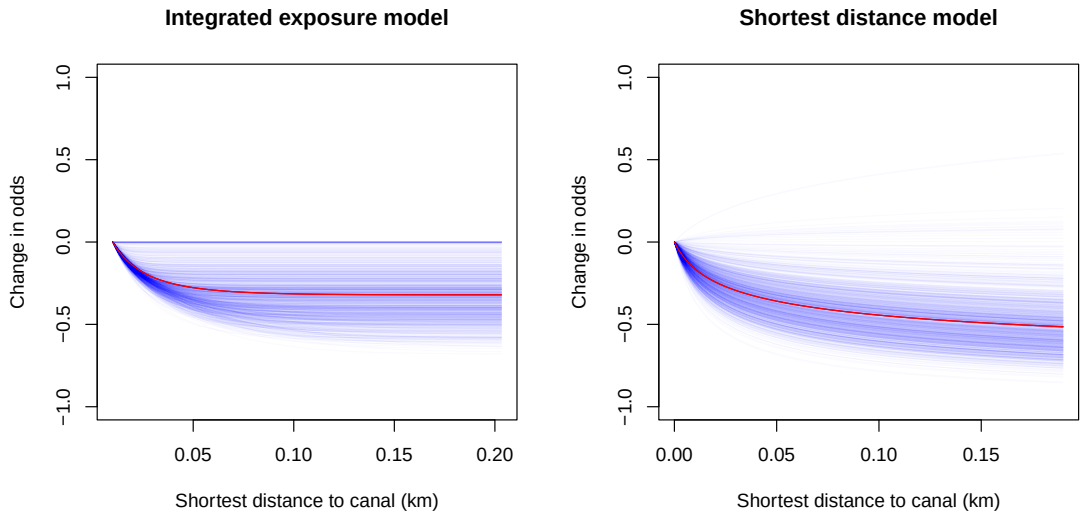


Figure 4.7: Posterior realizations of change in odds of diarrhea versus change in distance to the canal compared to 10 meters. Odds for the integrated model show the change in odds for a single household that moves laterally from the westernmost edge of the canal. Red lines indicate posterior means.

The model performed well when only 10 outcomes were observed per household. However, performance degraded when there were 100 observations per household. More investigation into why this is occurring is necessary.

Several theoretical concerns are open questions. Cotter et al. (2010) and Simpson et al. (2016) give techniques to determine the rate of convergence of posterior moments when using an approximate likelihood and Gaussian process prior. The proof in Appendix C.1 can be extended to that end. Furthermore, it is not clear whether ρ and $\exp(Z(c))$ are separately identifiable. The results of the simulation study suggest the answer may be no. More research is needed to understand if the model is identified, and, if not, how to do so.

The most exciting future work comes in extensions to the model. Several are immediately apparent. First, as presented in section 4.5.3, the model can be extended to allow the concentration of disease-causing agents to be time-varying and to be modeled using the kernel of the Gaussian process. This can be a promising new direction, and the extension can be directly applied to the Mezquital data example. There is evidence that diarrheal risk is higher in the rainy season, which might be connected to canal flooding.

In applications where environmental monitoring of health hazards is feasible, such as in air quality monitoring near roadways, we can augment our models with direct observations of concentrations of hazardous material at the source.

Our kernel specification depends only on distance to the segment or area of the environmental

hazard, but the point-source literature has investigated the use of kernels that take into account direction as well as distance. This could be useful in applications where the built environment can provide further barriers to exposure. For example, in the Mezquital dataset, some canal segments are only reachable via fields whereas other segments abut local roads lined with houses. It would be beneficial to take this information into account in a model, and modifying the kernel would be a way to do so.

APPENDIX A

Missing data appendix

A.1 Selection model derivation

Following Little et al. (2017) and Gelman et al. (2013), we wish to model the joint distribution for the data:

$$\prod_{(i,j)} f((y_{ij}, x_{ij})|\mu_{ij}, p_{ij})$$

which we have factorized according to a selection model paradigm: $f(y_{ij}|\mu_{ij})$, and $f(x_{ij}|y_{ij}, p_{ij})$, which follows from the conditional independence across i and j assumed in the generative model above. Let the vectors $\boldsymbol{\mu}_i$, $\mathbf{p}_i$ be the J-vectors with respective j^{th} elements μ_{ij} and p_{ij}. Let x_{ij}, w_{ij}, m_i be a specific realizations of X_{ij}, W_{ij}, M_i and let $\mathbf{x}_i$, $\mathbf{w}_i$ be defined as $\boldsymbol{\mu}$ was defined, and where $W_{ij} = Y_{ij} - X_{ij}$. Let $\mathbf{m}$ be the I vector with i^{th} element m_i. Then the complete data likelihood is defined as:

$$\begin{aligned} &L((\boldsymbol{\mu}_1, \mathbf{p}_1), \ldots, (\boldsymbol{\mu}_I, \mathbf{p}_I)|(\mathbf{x}_1, \mathbf{w}_1), \ldots, (\mathbf{x}_I, \mathbf{w}_I)) \\ &\qquad = \int \prod_{(i,j)} e^{-\mu_{ij}} \frac{\mu_{ij}^{y_{ij}}}{y_{ij}!} \frac{y_{ij}!}{x_{ij}! w_{ij}!} p_{ij}^{x_{ij}} (1 - p_{ij})^{w_{ij}} d\mathbf{w}_1 \ldots d\mathbf{w}_I \end{aligned}$$

which is shown in appendix A.2 to be

$$\begin{aligned} &L((\boldsymbol{\mu}_1, \mathbf{p}_1), \ldots, (\boldsymbol{\mu}_I, \mathbf{p}_I)|(\mathbf{x}_1, m_1), \ldots, (\mathbf{x}_I, m_I)) \\ &\qquad = \left(\prod_{(i,j)} e^{-p_{ij}\mu_{ij}} \frac{(p_{ij}\mu_{ij})^{x_{ij}}}{x_{ij}!} \right) \prod_i e^{-\sum_j (1-p_{ij})\mu_{ij}} \frac{(\sum_j (1 - p_{ij})\mu_{ij})^{m_i}}{m_i!} \end{aligned}$$

By the properties in Little et al. (2017) and Gelman et al. (2013) if $p_{ij} \neq p_i \forall (i, j)$ the complete data likelihood does not factorize into a term governing the observational process in Y and the

missingness process in R, viz.

$$L((\boldsymbol{\mu}_1, \mathbf{p}_1), \ldots, (\boldsymbol{\mu}_I, \mathbf{p}_I)|(\mathbf{x}_1, m_1), \ldots, (\mathbf{x}_I, m_I)) \neq$$
$$L(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_I|(\mathbf{x}_1, m_1), \ldots, (\mathbf{x}_I, m_I))L(\mathbf{p}_1, \ldots, \mathbf{p}_I|(\mathbf{x}_1, m_1), \ldots, (\mathbf{x}_I, m_I))$$

Given that equality does not hold when p_{ij} vary by j, we can say that in this case the data are not missing at random (NMAR), and thus we must model the joint distribution of observed data and missing data.

The observed data likelihood above is equivalent to the following generative model for the observed random variables X_{ij} and M_i:

$$\begin{aligned} X_{ij}|p_{ij}\mu_{ij} &\sim \text{Poisson}(p_{ij}\mu_{ij}) \\ M_i|\boldsymbol{\mu}_i, \mathbf{p}_i &\sim \text{Poisson}(\textstyle\sum_j \mu_{ij}(1 - p_{ij})) \end{aligned} \tag{A.1}$$

If we observe data for more than one geographic area, say for $g \in \{1, \ldots, G\}$, we might expect our parameters to vary across locations. For example, geographic heterogeneity in cumulative incidence has been a fundamental characteristic of the COVID-19 pandemic and many other infectious disease outbreaks and epidemics (Bilal et al., 2021; Wakefield et al., 2019). We can extend our generative model to capture geographic variation if we index our parameters with g and model them as jointly distributed under $F_{\boldsymbol{\phi}}$, with $\boldsymbol{\phi}$ as a vector of unknown hyperparameters:

$$((\boldsymbol{\mu}_{1g}, \mathbf{p}_{1g}), \ldots, (\boldsymbol{\mu}_{Ig}, \mathbf{p}_{Ig}))|\boldsymbol{\phi} \sim F_{\boldsymbol{\phi}}\ , \forall g, \tag{A.2}$$

where $\boldsymbol{\mu}_{ig}$ and $\mathbf{p}_{ig}$ are J-vectors where the j^{th} elements are equal to μ_{igj} and p_{igj}, respectively. The observed data model becomes

$$\begin{aligned} X_{igj}|p_{igj}\mu_{igj} &\sim \text{Poisson}(p_{igj}\mu_{igj}), \\ M_{ig}|\boldsymbol{\mu}_{ig}, \mathbf{p}_{ig} &\sim \text{Poisson}(\textstyle\sum_j \mu_{igj}(1 - p_{igj})), \end{aligned} \tag{A.3}$$

By extension, the joint hierarchical likelihood is:

$$\begin{aligned} \prod_g \big(&L((\boldsymbol{\mu}_{1g}, \mathbf{p}_{1g}), \ldots, (\boldsymbol{\mu}_{Ig}, \mathbf{p}_{Ig})|(\mathbf{x}_{1g}, m_{1g}), \ldots, (\mathbf{x}_{Ig}, m_{Ig})) \\ &f((\boldsymbol{\mu}_{1g}, \mathbf{p}_{1g}), \ldots, (\boldsymbol{\mu}_{Ig}, \mathbf{p}_{Ig})|\boldsymbol{\phi})\big) \end{aligned} \tag{A.4}$$

where $f((\boldsymbol{\mu}_{1g}, \mathbf{p}_{1g}), \ldots, (\boldsymbol{\mu}_{Ig}, \mathbf{p}_{Ig})|\boldsymbol{\phi})$ is the density associated with $F_{\boldsymbol{\phi}}$.

In the context of the COVID-19 case data, one might focus their analysis on a single county comprised of many smaller spatial units, with county-level parameters the target of inference, as

in (A.4) and (A.2) as $\boldsymbol{\phi}$. We will typically have prior information about the hyperparameters from data in other states or in nearby counties, so we opt to use Bayesian inference. If we represent the prior density for $\boldsymbol{\phi}$ as $h(\boldsymbol{\phi}|\boldsymbol{\tau})$ and $\boldsymbol{\tau}$ are known, the joint posterior is:

$$\begin{aligned}\pi((\boldsymbol{\mu}_{1g}, \mathbf{p}_{1g}), \ldots, (\boldsymbol{\mu}_{Ig}, \mathbf{p}_{Ig}), \boldsymbol{\phi} \big| (\mathbf{x}_{1g}, m_{1g}), \ldots, (\mathbf{x}_{Ig}, m_{Ig})) \propto \\ \Big(\prod_g \big(L((\boldsymbol{\mu}_{1g}, \mathbf{p}_{1g}), \ldots, (\boldsymbol{\mu}_{Ig}, \mathbf{p}_{Ig}) | (\mathbf{x}_{1g}, m_{1g}), \ldots, (\mathbf{x}_{Ig}, m_{Ig})) \\ f((\boldsymbol{\mu}_{1g}, \mathbf{p}_{1g}), \ldots, (\boldsymbol{\mu}_{Ig}, \mathbf{p}_{Ig}) | \boldsymbol{\phi}) \big) \Big) h(\boldsymbol{\phi}|\boldsymbol{\tau})\end{aligned} \tag{A.5}$$

Given the structure of the model, the marginal posterior for $\boldsymbol{\phi}$ is informed by the data via the terms $L((\boldsymbol{\mu}_{1g}, \mathbf{p}_{1g}), \ldots, (\boldsymbol{\mu}_{Ig}, \mathbf{p}_{Ig}) | (\mathbf{x}_{1g}, m_{1g}), \ldots, (\mathbf{x}_{Ig}, m_{Ig}))$, so it is important to understand the characteristics of the likelihood.

It can be seen that neither (A.1) nor (A.3) is identifiable as written without further assumptions.

A.2 Derivation of likelihood in Section 2.4

Here we give proof of the following property used in Section 2.4: the data generating model given by the Poisson process and Binomial selection process at the beginning of Section 2.4 results in model (A.1).

Consider two groups, $j \in [1, 2]$. As above, our fully observed likelihood gives the density for the vector of random variables, $(X_{i1}, X_{i2}, W_{i1}, W_{i2}) \forall i$, while we observe only $(X_{i1}, X_{i2}, M_i) \forall i$. Thus, we must integrate over the set of all $\{(W_{i1}, W_{i2} | W_{i1} + W_{i2} = M_i\}$.

$$\mathbb{P}\left(X_{i1} = x_{i1}, X_{i2} = x_{i2}, M_i = m_i\right) = \tag{A.6}$$

$$\mathbb{P}\left((X_{i1} = x_{i1}, W_{i1} = 0), (X_{i2} = x_{i2}, W_{i2} = 0)\right) \mathbb{1}(m_i = 0) \tag{A.7}$$

$$+ \sum_{e=0}^{m_i} \mathbb{P}\left((X_{i1} = x_{i1}, W_{i1} = e), (X_{i2} = x_{i2}, W_{i2} = (m_i - e)\right) \mathbb{1}(m_i > 0). \tag{A.8}$$

Given that $W_{ij} = Y_{ij} - X_{ij}$, this expression is equivalent to

$$\mathbb{P}\left((X_{i1} = x_{i1}, Y_{i1} = x_{i1}), (X_{i2} = x_{i2}, Y_{i2} = x_{i2})\right) \mathbb{1}(m_i = 0) \tag{A.9}$$

$$+ \sum_{e=0}^{m_i} \mathbb{P}\left((X_{i1} = x_{i1}, Y_{i1} = x_{i1} + e), (X_{i2} = x_{i2}, Y_{i2} = x_{i2} + (m_i - e)\right) \mathbb{1}(m_i > 0). \tag{A.10}$$

Each term

$$\mathbb{P}\left((X_{i1} = x_{i1}, Y_{i1} = y_{i1}), (X_{i2} = x_{i2}, Y_{i2} = y_{i2})\right)$$

decomposes to

$$\mathbb{P}\left(X_{i1}=x_{i1}|Y_{i1}=y_{i1}\right)\mathbb{P}\left(Y_{i1}=y_{i1}\right)\mathbb{P}\left(X_{i2}=x_{i2}|Y_{i2}=y_{i2}\right)\mathbb{P}\left(Y_{i2}=y_{i2}\right)$$

given the independence between Y_{i1} and Y_{i2} and the conditional independence of $X_{i1}|Y_{i1}$ and $X_{i2}|Y_{i2}$.

$$\prod_{i=1}^{I}\left(\frac{\lambda_{i1}^{x_{i1}}e^{-\lambda_{i1}}}{x_{i1}!}p_{i1}^{x_{i1}}\frac{\lambda_{i2}^{x_{i2}}e^{-\lambda_{i2}}}{x_{i2}!}p_{i2}^{x_{i2}}\right)^{\mathbb{1}(m_i=0)}$$
$$\left(\sum_{e=0}^{m_i}\frac{\lambda_{i1}^{x_{i1}+(m_i-e)}e^{-\lambda_{i1}}}{(x_{i1}+(m_i-e))!}\binom{x_{i1}+m_i-e}{x_{i1}}p_{i1}^{x_{i1}}(1-p_{i1})^{m_i-e}\frac{\lambda_{i2}^{x_{i2}+e}e^{-\lambda_{i2}}}{(x_{i2}+e)!}\binom{x_{i2}+e}{x_{i2}}p_{i2}^{x_{i2}}(1-p_{i2})^{e}\right)^{1-\mathbb{1}(m_i=0)}$$

This simplifies to

$$\prod_{i=1}^{I}\left(\frac{\mu_{i1}^{x_{i1}}e^{-\mu_{i1}}}{x_{i1}!}p_{i1}^{x_{i1}}\frac{\mu_{i2}^{x_{i2}}e^{-\mu_{i2}}}{x_{i2}!}p_{i2}^{x_{i2}}\right)$$
$$\left(\sum_{e=0}^{m_i}\frac{\left((1-p_{i1})\mu_{i1}\right)^{(m_i-e)}}{(m_i-e)!}\frac{\left((1-p_{i2})\mu_{i2}\right)^{e}}{e!}\right)$$

which, multiplying by $\frac{m_i!}{m_i!}$ and using the binomial theorem, further simplifies to

$$\prod_{i=1}^{I}\left(\frac{\mu_{i1}^{x_{i1}}e^{-\mu_{i1}}}{x_{i1}!}p_{i1}^{x_{i1}}\frac{\mu_{i2}^{x_{i2}}e^{-\mu_{i2}}}{x_{i2}!}p_{i2}^{x_{i2}}\right)$$
$$\frac{\left((1-p_{i1})\mu_{i1}+(1-p_{i2})\mu_{i2}\right)^{m_i}}{m_i!}$$

Finally we multiply by $e^{-((1-p_{i1})\mu_{i1}+(1-p_{i2})\mu_{i2})}e^{(1-p_{i1})\mu_{i1}+(1-p_{i2})\mu_{i2}}$ to yield

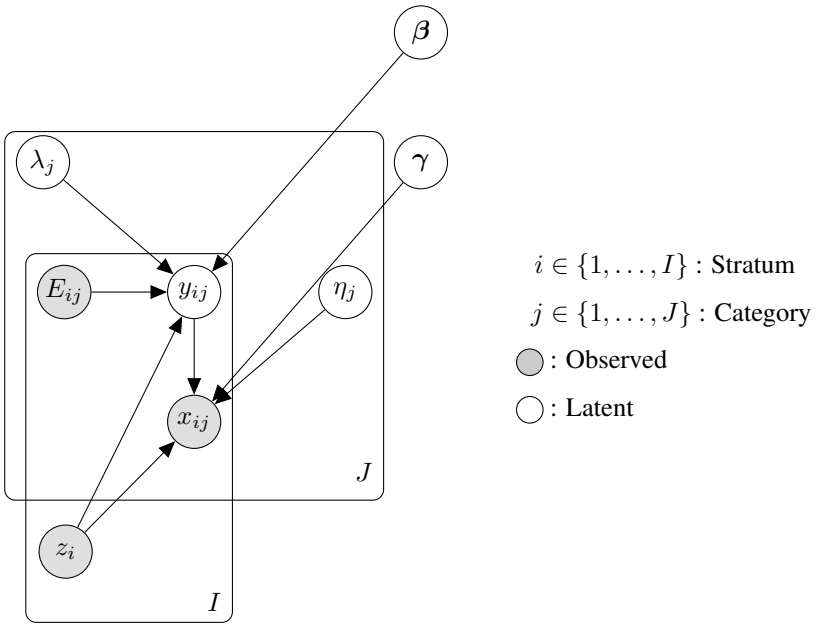
$$\prod_{i=1}^{I}\left(\frac{(\mu_{i1}p_{i1})^{x_{i1}}e^{-p_{i1}\mu_{i1}}}{x_{i1}!}\frac{(p_{i2}\mu_{i2})^{x_{i2}}e^{-p_{i2}\mu_{i2}}}{x_{i2}!}\right)\frac{e^{-((1-p_{i1})\mu_{i1}+(1-p_{i2})\mu_{i2})}\left((1-p_{i1})\mu_{i1}+(1-p_{i2})\mu_{i2}\right)^{m_i}}{m_i!}$$

which we recognize as the product of filtered Poisson random variables, and the marginally Poisson distributed cases missing stratum information.

The proof of the generalization to J groups, which can be show with induction, has been omitted.

A.3 Graphical model depictions

A.3.1 Graphical model of model with covariates

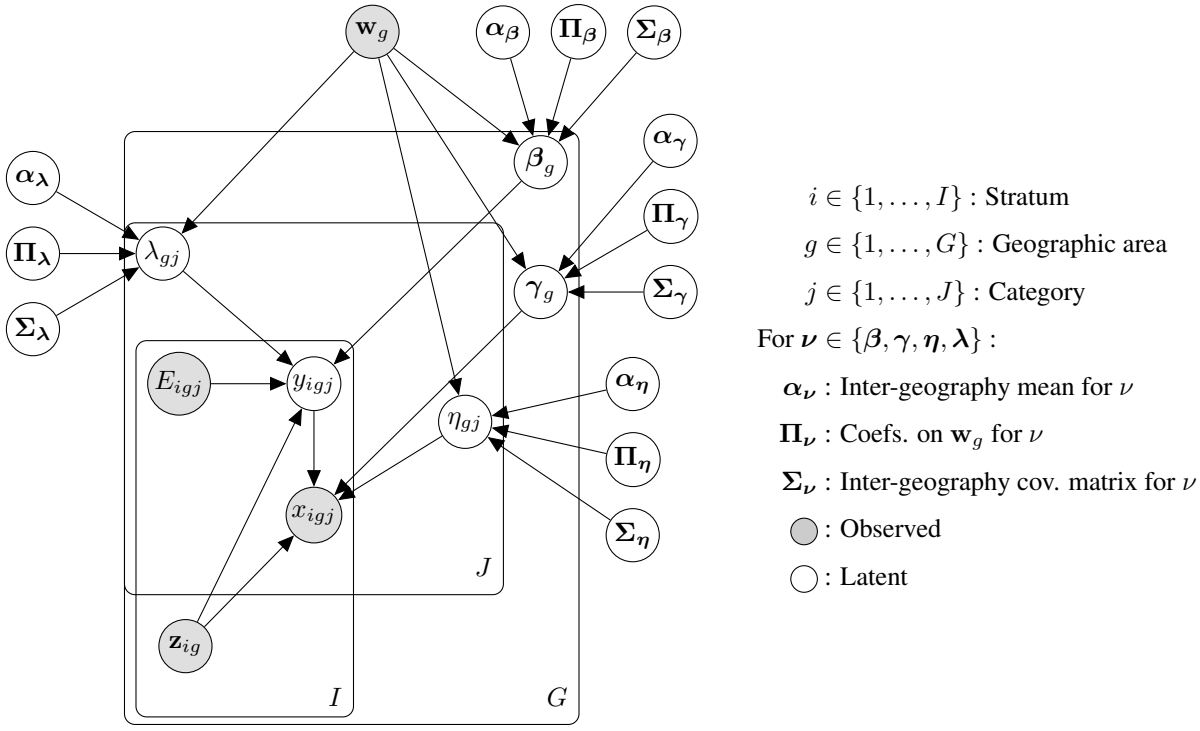


Variable	Domain	Description
y_{ij}	$\mathbb{N}_0$	Total cases
x_{ij}	$\mathbb{N}_0$	Observed cases
E_{ij}	$\mathbb{N}_0$	Observed population
$\mathbf{z}_i$	$\mathbb{R}^K$	Observed covariates
λ_j	$\mathbb{R}^+$	Per-capita rate of disease
η_j	$\mathbb{R}$	Log-odds of observing category info.
$\boldsymbol{\beta}$	$\mathbb{R}^K$	Log-relative rates of disease
$\boldsymbol{\gamma}$	$\mathbb{R}^K$	Log-odds of observing category info.

Table A.1: Table of generative model variables for Model 2.10

The parameters of interest in Table A.1 are λ_j, which give the category-specific, per-capita rates of disease, and transformations of the parameters like those enumerated in Section 2.5.4.

A.3.2 Graphical model of hierarchical model with covariates



Variable	Domain	Description
y_{igj}	$\mathbb{N}_0$	Total cases
x_{igj}	$\mathbb{N}_0$	Observed cases
E_{igj}	$\mathbb{N}_0$	Observed population
$\mathbf{z}_{ig}$	$\mathbb{R}^K$	Observed covariates
$\mathbf{w}_g$	$\mathbb{R}^D$	Observed geographic-specific covariates
λ_{gj}	$\mathbb{R}^+$	Per-capita rate of disease
η_{gj}	$\mathbb{R}$	Log-odds of observing category
$\boldsymbol{\beta}_g$	$\mathbb{R}^K$	Log-relative rates of disease
$\boldsymbol{\gamma}_g$	$\mathbb{R}^K$	Log-odds of observing category

Table A.2: Table of generative model variables for Model 2.11

The parameters of interest in Table A.2 are λ_{gj}, which give the category-specific geographical-area-specific, per-capita rates of disease, and transformations of the parameters like those enumerated in Section 2.5.4. Interest may also lie in the across-geography mean category-specific log per-capita rates of disease for category j, $\boldsymbol{\alpha}_{\boldsymbol{\lambda}}$ the coefficients on $\mathbf{w}_g$ or $\boldsymbol{\Pi}_{\boldsymbol{\lambda}}$, and the inter-geography

covariance between log-per-capita, category-specific rates of disease, Σ_λ

A.4 Lemmas and Proofs for Model Identifiability Properties

A.4.1 Fisher information positive definiteness of simple model

The following is a derivation of the Fisher information matrix $\mathcal{I}$ for model 2.1. The likelihood for the model is

$$\begin{aligned}\ell(\lambda_1, p_1, \ldots, \lambda_J, p_J) = \sum_{i=1}^{I}\sum_{j=1}^{J} -E_{ij}\lambda_j + x_{ij}\log\left(E_{ij}\lambda_j p_j\right) - \log x_{ij}! \\ + \sum_{i=1}^{I} m_i \log\left(\sum_{j=1}^{J} E_{ij}\lambda_j(1-p_j)\right) - \log m_i!\end{aligned}$$

We reparameterize as we do in 1:

$$(\lambda_j, p_j) \to (v_j, u_j) \forall j \in [1, \ldots, J] \tag{A.11}$$

where $v_j = \lambda_j p_j$ $u_j = \lambda_j(1-p_j)$. Then the likelihood is

$$\begin{aligned}\ell(v_1, u_1, \ldots, v_j, u_J) = \sum_{i=1}^{I}\sum_{j=1}^{J} -E_{ij}(u_j + v_j) + x_{ij}\log\left(E_{ij}v_j\right) - \log x_{ij}! \\ + \sum_{i=1}^{I} m_i \log\left(\sum_{j=1}^{J} E_{ij}u_j\right) - \log m_i!\end{aligned}$$

Let $\delta_{i,j}$ be the Kronecker delta function. The Fisher information matrix for the reparameterized log-likelihood is:

$$-\mathbb{E}\left[\frac{\partial \ell}{\partial v_j v_k}\right] = \frac{\sum_{i=1}^{I} E_{ij}}{v_j}\delta_{j,k} \tag{A.12}$$

$$-\mathbb{E}\left[\frac{\partial \ell}{\partial u_j u_k}\right] = \sum_{i=1}^{I} \frac{E_{ij}E_{ik}}{\sum_{m=1}^{J} E_{im}u_m} \tag{A.13}$$

$$-\mathbb{E}\left[\frac{\partial \ell}{\partial u_j v_k}\right] = 0 \tag{A.14}$$

We can arrange the Fisher information in block matrix form for all I observations:

$$\mathcal{I} = \begin{bmatrix}\mathbf{V} & 0 \\ 0 & \mathbf{U}\end{bmatrix} \tag{A.15}$$

with

$$\mathbf{U}_{jk} = \sum_{i=1}^{I} \frac{E_{ij}E_{ik}}{\sum_{m=1}^{J} E_{im}u_m} \tag{A.16}$$

$$\mathbf{V}_{jk} = \frac{\sum_{i=1}^{I} E_{ij}}{v_j}\delta_{j,k} \tag{A.17}$$

$\mathcal{I}$ is positive definite if $\mathbf{U} \succ 0$ and $\mathbf{V} \succ 0$. $\mathbf{V}$ is a diagonal matrix and is positive definite as long as all elements along the diagonal are strictly positive. The parameter constraints on u_j and v_j yield $u_j, v_j > 0$ for each j so as long as $\sum_{i=1}^{I} E_{ij} > 0$ for all j $\mathbf{V} \succ 0$.

The matrix $\mathbf{U}$ can be represented as the matrix product of three matrices. Let $\mathbf{E}$ be the $I \times J$ matrix with its $(i,j)^{\text{th}}$ entry as E_{ij}, and let $\mathbf{S}$ be a diagonal matrix defined as

$$\mathbf{S}_{jk} = \frac{1}{\sqrt{\sum_{m=1}^{J} E_{im}u_m}}\delta_{j,k} \tag{A.18}$$

Then $\mathbf{U} = \mathbf{E}^T\mathbf{S}^2\mathbf{E}$. In order for $\mathbf{U} \succ 0$ $\mathbf{SE}$ must be rank J. This will be so if $\mathbf{S}$ is invertible and if $\mathbf{E}$ is rank J. If $\det \mathbf{S} \neq 0$ $\mathbf{S}$ is invertible:

$$\det(\mathbf{S}) = (\prod_{i=1}^{I}\sum_{m=1}^{J} E_{im}u_m)^{-1/2} \tag{A.19}$$

$\mathbf{S}$ has a nonzero determinant if at least one $E_{im}u_{>}0$ for all $i \in [1, \ldots, I]$. Given the constraints on the parameter space, $p_j \in (0,1)$ and $\lambda_j > 0$ for all j ensures that $u_j > 0$ for all j. The minimal conditions for the positive definiteness of $\mathcal{I}$ are as follows:

- $\mathbf{E}$ is rank J
- $p_j \in (0,1)$
- $\lambda_j \in (0,\infty)$
- $\sum_{m=1}^{J} E_{im} > 0 \forall i$

Given estimators $\hat{v}_j = \sum_{i=1}^{I} X_{ij} / \sum_{i=1}^{I} E_{ij}$ for v_j and $\hat{\mathbf{u}} = (\mathbf{E}^T\mathbf{E})^{-1}\mathbf{E}^T\mathbf{m}$, the observed Fisher information matrix $\hat{\mathcal{I}}$ is:

$$\hat{\mathcal{I}} = \begin{bmatrix} \hat{\mathbf{V}} & 0 \\ 0 & \hat{\mathbf{U}} \end{bmatrix} \tag{A.20}$$

where

$$\hat{\mathbf{V}}_{mn} = \delta_{m,n} \frac{(\sum_{i=1}^{I} E_{ij})^2}{\sum_{i=1}^{I} X_{ij}}$$

and

$$\hat{\mathbf{U}} = \mathbf{E}^T \text{diag}(\mathbf{E}(\mathbf{E}^T\mathbf{E})^{-1}\mathbf{E}^T\mathbf{m})^{-1}\mathbf{E}$$

The inverse of $\hat{\mathcal{I}}$ is just the block diagonal matrix with $\hat{\mathbf{V}}^{-1}$ and $\hat{\mathbf{U}}^{-1}$ along the diagonal:

A.4.2 Derivation of posterior mean of λ_1 for minority group

Let $J = 2$ and let the following priors be implemented for p_j, λ_j

$$p_j \overset{\text{iid}}{\sim} \text{Beta}(\alpha_j, \beta_j)$$
$$\lambda_j \overset{\text{iid}}{\sim} \text{Gamma}(\alpha_j + \beta_j, r_j)$$

then $v_j = p_j\lambda_j \perp\!\!\!\perp u_j = (1-p_j)\lambda_j$ with $v_j \sim \text{Gamma}(\alpha_j, r_j)$ and $u_j \sim \text{Gamma}(\beta_j, r_j)$ We can write the observed-data likelihood in terms of u_j and v_j as follows:

$$\exp\left(-\sum_j u_j \sum_i E_{ij}\right) \prod_{i=1}^{I} \frac{(E_{i1}u_1 + E_{i2}u_2)^{m_i}}{m_i!} \prod_{j=1}^{J} \frac{\exp\left(-v_j \sum_i E_{ij}\right) \left(\prod_{i=1}^{I} (E_{ij})^{x_{ij}}\right) v_j^{\sum_i x_{ij}}}{\prod_{i=1}^{I} x_{ij}!}$$

If we assume that, WLOG, group 2 is the majority group for all i, or, in other words, $E_{i1} \ll E_{i2}$ for all i, then using the approximation

$$\begin{aligned}(E_{i1}u_1 + E_{i2}u_2)^{m_i} &= (E_{i2}u_2)^{m_i} \exp m_i \log\left(1 + \frac{E_{i1}}{E_{i2}u_2}u_1\right) \\ &\approxeq (E_{i2}u_2)^{m_i} \exp m_i \left(u_1 \frac{E_{i1}}{E_{i2}u_2} - u_1^2 \frac{E_{i1}^2}{2E_{i2}^2 u_2^2}\right)\end{aligned}$$

Leading to the approximate likelihood:

$$\begin{aligned}&\exp\left(-\sum_j u_j \sum_i E_{ij}\right) \left(\prod_{i=1}^{I} E_{i2}\right) u_2^{\sum_i m_i} \frac{\prod_{i=1}^{I} \exp\left(u_1 \frac{E_{i1}}{E_{i2}u_2} - u_1^2 \frac{E_{i1}^2}{2E_{i2}^2 u_2^2}\right)^{m_i}}{m_i!} \\ &\times \prod_{j=1}^{2} \frac{\exp\left(-v_j \sum_i E_{ij}\right) \left(\prod_{i=1}^{I} (E_{ij})^{x_{ij}}\right) v_j^{\sum_i x_{ij}}}{\prod_{i=1}^{I} x_{ij}!}\end{aligned}$$

When we multiply by the priors and collect terms, we get an expression separable in u_j and v_j:

$$\left(\prod_{i=1}^{I} E_{i2}\right) \exp\left(-u_2 \textstyle\sum_i E_{i2}\right) u_2^{\beta_2+\sum_i m_i-1}/m_i!$$
$$u_1^{\beta_1-1} \exp\left(-u_1^2 u_2^{-2} \textstyle\sum_i \frac{m_i E_{i1}^2}{2E_{i2}^2} + u_1\left(u_2^{-1}\sum_i \frac{m_i E_{i1}}{E_{i2}} - r_1 - \sum_i E_{i1}\right)\right)$$
$$\times \prod_{j=1}^{2} \frac{\exp\left(-v_j\left(r_j + \sum_i E_{ij}\right)\right)\left(\prod_{i=1}^{I}(E_{ij})^{x_{ij}}\right) v_j^{\alpha_j+\sum_i x_{ij}} - 1}{\prod_{i=1}^{I} x_{ij}!}$$

The approximate posterior distribution, $\pi(u_1|u_2, m)$, is a modified half-normal distribution, as introduced in Sun et al. (2021). While the functional form of the posterior mean is complicated for general values of β_1, we compute the posterior mean, variance, and local sensitivity of the posterior mean to r_1 when $\beta_1 = 1$ and $\beta_1 = 2$, which correspond to implied priors for u_1 of an exponential distribution with rate r_1, and a gamma distribution with shape 2 and rate r, respectively. Let $s_1 = \sum_i \frac{m_i E_{i1}}{E_{i2}}$, $s_2 = \sum_i \frac{m_i E_{i1}^2}{E_{i2}^2}$, and $E_{+1} = \sum_i E_{i1}$. Further, let ϕ and Φ be the density and distribution function of the standard normal distribution, respectively, and $z(u_2, r_1) = \frac{s_1 - u_2(r_1+E_{+1})}{\sqrt{s_2}}$. If $\beta_1 = 1$, the posterior mean of $u_1|u_2, m$ is

$$\mathbb{E}\left[u_1|u_2, r_1, \beta_1 = 1\right] = \frac{u_2}{\sqrt{s_2}}\left(z(u_2, r_1) + \phi(z(u_2, r_1))\Phi(z(u_2, r_1))^{-1}\right)$$

while the posterior variance is

$$\text{Var}\left(u_1|u_2, r_1, \beta_1 = 1\right) = \frac{u_2^2}{s_2}\left(1 - z(u_2, r_1)\phi(z(u_2, r_1))\Phi(z(u_2, r_1))^{-1} - \phi(z(u_2, r_1))^2\Phi(z(u_2, r_1))^{-2}\right)$$

The posterior mean for λ_1 is then

$$\frac{\alpha_1 + \sum_i x_{i1}}{r_1 + \sum_i E_{ij}} + \frac{u_2}{\sqrt{s_2}}(z(u_2, r_1) + \phi(z(u_2, r_1))\Phi(z(u_2, r_1))^{-1})$$

with variance:

$$\frac{\alpha_1 + \sum_i x_{i1}}{(r_1 + \sum_i E_{i1})^2} + \frac{u_2^2}{s_2}\left(1 - z(u_2, r_1)\phi(z(u_2, r_1))\Phi(z(u_2, r_1))^{-1} - \phi(z(u_2, r_1))^2\Phi(z(u_2, r_1))^{-2}\right)$$

The partial derivative of $\mathbb{E}\left[u_1|u_2, r_1\right]$ with respect to r_1 is

$$\frac{-u_2^2}{s_2}\Bigg(1 - \frac{1}{\sqrt{s_2}}\phi(z(u_2, r_1))\Phi(z(u_2, r_1))^{-2}$$
$$\times\left(\sqrt{s_2}\phi(z(u_2, r_1)) + (s_1 - u_2(E_{+j} + r_1))\Phi(z(u_2, r_1))\right)\Bigg)$$

which simplifies to

$$\begin{aligned}\frac{\partial \mathbb{E}\left[u_1|u_2,r_1\right]}{\partial r_1} &= \frac{-u_2^2}{s_2}\left(1 - z(u_2,r_1)\phi(z(u_2,r_1))\Phi(z(u_2,r_1))^{-1} - \phi(z(u_2,r_1))^2\Phi(z(u_2,r_1))^{-2}\right) \\ &= -\text{Var}\left(u_1|u_2,r_1\right)\end{aligned}$$

If $\beta_1 = 2$, the posterior mean of $u_1|u_2, m$ is

$$\frac{u_2}{\sqrt{s_2}}\left(z(u_2,r_1) + \frac{\Phi(z(u_2,r_1))}{\phi(z(u_2,r_1)) + z(u_2,r_1)\Phi(z(u_2,r_1))}\right)$$

while the posterior variance is

$$\frac{u_2^2}{s_2}\left(2 - \left(z(u_2,r_1) + \frac{\Phi(z(u_2,r_1))}{\phi(z(u_2,r_1)) + z(u_2,r_1)\Phi(z(u_2,r_1))}\right)\right)$$

The posterior mean for λ_1 is then

$$\frac{\alpha_1 + \sum_i x_{i1}}{r_1 + \sum_i E_{ij}} + \frac{u_2}{\sqrt{s_2}}\left(z(u_2,r_1) + \frac{\Phi(z(u_2,r_1))}{\phi(z(u_2,r_1)) + z(u_2,r_1)\Phi(z(u_2,r_1))}\right)$$

with variance:

$$\frac{\alpha_1 + \sum_i x_{i1}}{(r_1 + \sum_i E_{i1})^2} + \frac{u_2^2}{s_2}\left(2 - \left(z(u_2,r_1) + \frac{\Phi(z(u_2,r_1))}{\phi(z(u_2,r_1)) + z(u_2,r_1)\Phi(z(u_2,r_1))}\right)\right)$$

Taking the difference between $\mathbb{E}\left[u_1|u_2,r_1,\beta_1=2\right]$ and $\mathbb{E}\left[u_1|u_2,r_1,\beta_1=1\right]$ yields

$$\begin{aligned}\frac{u_2}{\sqrt{s_2}}\Bigg(&z(u_2,r_1) + \frac{\Phi(z(u_2,r_1))}{\phi(z(u_2,r_1)) + z(u_2,r_1)\Phi(z(u_2,r_1))} \\ &- \left(z(u_2,r_1) + \phi(z(u_2,r_1))\Phi(z(u_2,r_1))^{-1}\right)\Bigg)\end{aligned}$$

Algebra reveals that $\mathbb{E}\left[u_1|u_2,r_1,\beta_1=2\right] - \mathbb{E}\left[u_1|u_2,r_1,\beta_1=1\right]$ is

$$\frac{u_2}{\sqrt{s_2}}\frac{1 - z(u_2,r_1)\phi(z(u_2,r_1))\Phi(z(u_2,r_1))^{-1} - \phi(z(u_2,r_1))^2\Phi(z(u_2,r_1))^{-2})}{z(u_2,r_1) + \phi(z(u_2,r_1))\Phi(z(u_2,r_1))^{-1}}$$

or

$$\frac{\text{Var}\left(u_1|u_2,r_1,\beta_1=1\right)}{\mathbb{E}\left[u_1|u_2,r_1,\beta_1=1\right]}$$

Dividing this by the standard deviation:

$$\frac{\sqrt{\text{Var}\left(u_1|u_2, r_1, \beta_1 = 1\right)}}{\mathbb{E}\left[u_1|u_2, r_1, \beta_1 = 1\right]}$$
$$= \frac{\sqrt{1 - z(u_2, r_1)\phi(z(u_2, r_1))\Phi(z(u_2, r_1))^{-1} - \phi(z(u_2, r_1))^2\Phi(z(u_2, r_1))^{-2})}}{z(u_2, r_1) + \phi(z(u_2, r_1))\Phi(z(u_2, r_1))^{-1}}$$

If we further assume that $E_{i2}, E_{i1} \to \infty$ such that $E_{i1}/E_{i2} \to 0$ and $E_{i1}^2/E_{i2} \to K < \infty$ for all i, the posterior for u_2 converges to a point mass at $u_2^\star$, the true data generating parameter.

This can be seen from the fact that the MLE for u_2 converges to $u_2^\star$. The gradient of the log-likelihood ℓ with respect to u_2 and u_1 is:

$$\frac{\partial \ell}{\partial u_1} = -E_{+1} - \frac{s_2 u_1}{u_2^2} + \frac{s_1}{u_2}$$
$$\frac{\partial \ell}{\partial u_2} = -E_{+2} + \frac{s_2 u_1^2}{u_2^3} - \frac{s_1 u_1}{u_2^2} + \frac{m_+}{u_2}$$

Setting these equal to zero yields the following two solutions:

$$\hat{u}_1(u_2) = u_2 \frac{s_1 - u_2 E_{+1}}{s_2}$$
$$\hat{u}_2 = \frac{E_{+1}s_1 + E_{+2}s_2 \pm \sqrt{-4E_{+1}^2 m_+ s_2 + (-E_{+1}s_1 - E_{+2}s_2)^2}}{2E_{+1}^2}$$

Recall that $M_i \sim \text{Poisson}(u_1^\star E_{i1} + u_2^\star E_{i2})$. For $E_{i1}, E_{i2} \to \infty$ with $u_1^\star$ and $u_2^\star$ bounded away from zero and $< \infty$, $\frac{M_i - (u_1^\star E_{i1} + u_2^\star E_{i2})}{\sqrt{u_1^\star E_{i1} + u_2^\star E_{i2}}} \xrightarrow{d} \text{N}(0, 1)$ by the CLT. Let $\mathcal{Z} \sim \text{N}(0, 1)$, then $s_1 - u_2^\star E_{+1} \xrightarrow{d} u_1^\star IK + \sqrt{u_2^\star IK}\mathcal{Z}$, $s_2 \xrightarrow{p} u_2^\star IK$.

$$\frac{\sqrt{-4E_{+1}^2 m_+ s_2 + (-E_{+1}s_1 - E_{+2}s_2)^2}}{2E_{+1}^2} = \sqrt{\frac{(E_{+1}IKu_1 - E_{+1}^2 u_2 + E_{+2}IKu_2)^2}{4E_{+1}^4} + O_p(\frac{1}{\sqrt{E_{+2}}})}$$

so

$$\frac{E_{+1}s_1 + E_{+2}s_2 \pm \sqrt{-4E_{+1}^2 m_+ s_2 + (-E_{+1}s_1 - E_{+2}s_2)^2}}{2E_{+1}^2} \xrightarrow{p} u_2^\star$$

Finally, by Slutsky's theorem

$$\hat{u}_1 \xrightarrow{d} u_1^\star + \left(\sqrt{\frac{u_2^\star}{IK}}\right)\mathcal{Z}$$

We can calculate the asymptotic MSE of $\hat{u}_1$, assuming that when $\hat{u}_1 \leq 0$ we set $\hat{u}_1$ to be 0.

Let $\alpha = -u_1^\star \sqrt{IK/u_2^\star}$. Then bias of $\hat{u}_1$ is

$$\sqrt{u_2^\star/IK}\phi(\alpha) - u_1^\star \Phi(\alpha)$$

and the variance is

$$\sqrt{u_2^\star/IK}(1 + \alpha\phi(\alpha) - \phi(\alpha)^2 - \Phi(\alpha)) + 2\phi(\alpha)\Phi(\alpha)\sqrt{u_2/IK}u_1 + \Phi(-\alpha)\Phi(\alpha)u_1^2$$

The asymptotics above also imply: $z(u_2, r_1) \xrightarrow{d} \frac{u_1^\star IK - u_2^\star r_1}{\sqrt{u_2^\star IK}} + \mathcal{Z} = z^\star$, and the posterior mean for u_1 given $\beta = 1$ is

$$\sqrt{\frac{u_2}{IK}}\left(z^\star + \phi(z^\star)\Phi(z^\star)^{-1}\right) = u_1^\star - \frac{r_1 u_2^\star}{IK} + \frac{\sqrt{u_2^\star}\phi(z^\star)\Phi(z^\star)^{-1}}{\sqrt{IK}} + \sqrt{\frac{u_2^\star}{IK}}\,\mathcal{Z}$$

The expression $z + \phi(z)/\Phi(z) \geq 0 \forall z$, so the Bayesian posterior mean for u_1 will be positive a.s. whereas the MLE may be 0 with positive probability depending on I, K and $u_1^\star, u_2^\star$. The asymptotic posterior mean for $\beta = 2$ is

$$\sqrt{\frac{u_2}{IK}}\left(z^\star + \frac{1}{z^\star + \phi(z^\star)\Phi(z^\star)^{-1}}\right)$$

We can compare the asymptotic root mean-squared error for the MLE and the posterior mean under an exponential prior for u_1, or $\mathrm{Exp}(r_1)$ and under a $\mathrm{Gamma}(2, r_1)$ prior for u_1 for a range of values of $u_1^\star$ for $I = 15$ and $K = 1$. Note that $u_1 \sim \mathrm{Exp}(r_1)$ corresponds to $p_1 \sim \mathrm{Beta}(\alpha_1, 1)$ and $u_1 \sim \mathrm{Gamma}(2, r_1)$ corresponds to $p_1 \sim \mathrm{Beta}(\alpha_1, 2)$. We use the square root of the exact asymptotic MSE for the MLE, while we use a Monte Carlo approximation to the RMSE for the two Bayesian estimators. We assume that $u_i^\star = (1 - p_i^\star)\lambda_i^\star$ and fix $p_1^\star = 0.6$ and $p_2^\star = 0.9$, which represents a high race/ethnicity reporting rate for the majority group, and low race/ethnicity reporting rate for the minority group. We assume $\lambda_2^\star \in \{0.001, 0.009, 0.02\}$ while we examine $\lambda_1^\star$ from 0.001 to 0.05. We fix the posterior mean for u_1 at 0.01, which implies $r_1 = 100$ for the exponential prior and $r_1 = 200$ for the gamma prior.

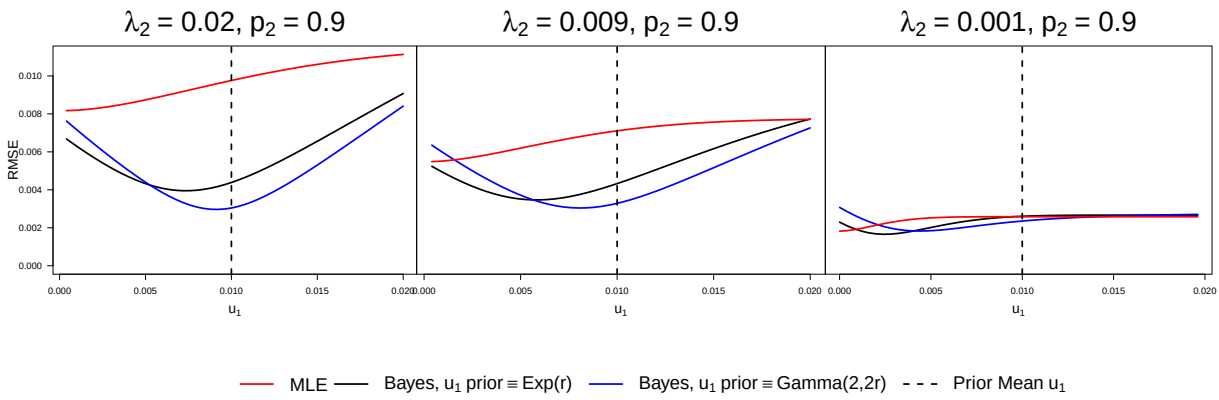


Figure A.1: Asymptotic root mean-squared error (RMSE) of posterior mean for two Bayes estimators vs. MLE. Monte Carlo approximation to RMSE for posterior means, with standard error on the order of 10^{-6} for all u_1. Note an exponential prior puts prior mass near zero while the $\text{gamma}(2, r_1)$ prior puts vanishing prior mass as $u_1 \to 0$. The y-axis represents the RMSE of the a given point estimator for certain data-generating values of u_1 and u_2. The panels of the graphs represent different true values of u_2, corresponding to $u_2 = \lambda_2 p_2$, while the x-axes represent a continuum of true values for u_1. The dashed vertical line represents the prior mean for u_1. Thus each panel of the graph shows how RMSE of each point estimator varies as u_1 increases from 4×10^{-6} to 1.9×10^{-2} given a certain value of u_2. The RMSE of the MLE, shown as the solid red line, slowly increases as u_1 increases as the variance of the MLE increases faster than the squared bias decreases. The Bayes estimators show decreasing RMSE as the prior mean for u_1 approaches the true u_1. Two conclusions can be drawn from the graphs: Both Bayes solutions dominate the MLE for reasonable values of u_1 and u_2. The exception is for small u_2 and when the prior for u_1 is several orders of magnitude too large. The second conclusion is that the Bayes estimator with gamma prior dominates the exponential-prior estimator when the prior mean for u_1 is moderately larger than the true u_1 and when the prior mean underestimates the true u_1.

Figure A.1 shows that Bayes estimators yield gains over the MLE for minority groups. Within the class of Bayes estimators, estimators derived from models with priors that put too much support near zero (e.g. an exponential prior) can shrink too the posterior mean too strongly towards zero even when the prior mean over-estimates the true parameter. Given that the near-zero behavior is driven by the prior over p_1, limiting prior mass near 1 for p_1 can yield point estimators with lower MSEs for a broad range of values for u_1.

A.4.2.1 Results when shape of gamma prior not equal to sum of beta shape parameters

Set-up Let $p \sim \text{Beta}(\alpha, \beta)$ and let $\lambda \sim \text{Gamma}(a, c)$ with $p \perp\!\!\!\perp \lambda$ So

$$f_{p,\lambda}(x, y) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{c^a}{\Gamma(a)} x^{\alpha-1}(1-x)^{\beta-1} \mathbb{1}_{x\in[0,1]} y^{a-1} e^{-cy} \mathbb{1}_{y>0}$$

Let $v = p\lambda$ and $u = (1-p)\lambda$ so $\lambda = v + u$ and $p = v/(v+u)$. Then $|J| = \left|-\frac{1}{v+u}\right|$ Then

$$f_{v,u}(v,u) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{c^a}{\Gamma(a)}(\frac{v}{v+u})^{\alpha-1}(\frac{u}{v+u})^{\beta-1}(v+u)^{a-1}e^{-c(v+u)}(v+u)^{-1}\mathbb{1}_{u>0}\mathbb{1}_{v>0}$$

which simplifies to

$$f_{v,u}(v,u) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{c^a}{\Gamma(a)}v^{\alpha-1}u^{\beta-1}(v+u)^{a-(\alpha+\beta)}e^{-c(v+u)}\mathbb{1}_{u>0}\mathbb{1}_{v>0}$$

It follows that if $a = \alpha + \beta$ then the density factorizes in v and u into two Gamma distributions with shape parameters α and β, respectively; thus $v \perp\!\!\!\perp u$.

If we want to calculate the marginal distribution of u, we can integrate over v:

$$= \int_0^\infty v^{\alpha-1}(v+u)^{a-(\alpha+\beta)}e^{-cv}dv \tag{A.21}$$

$$= \int_0^\infty v^{\alpha-1}v^{a-\alpha-\beta}(1+\frac{u}{v})^{a-(\alpha+\beta)}e^{-cv}dv \tag{A.22}$$

$$= \int_0^\infty v^{a-(\beta+1)}(1+\frac{u}{v})^{a-(\alpha+\beta)}e^{-cv}dv \tag{A.23}$$

$$= \int_\infty^0 (\frac{u}{x})^{a-(\beta+1)}(1+x)^{a-(\alpha+\beta)}e^{-c\frac{u}{x}}dx\frac{-u}{x^2} \tag{A.24}$$

$$= \int_0^\infty x^{\beta-(a+1)}(1+x)^{a-(\alpha+\beta)}e^{-c\frac{u}{x}}dxu^{a-\beta} \tag{A.25}$$

$$= \int_\infty^0 y^{(a+1)-\beta}(1+y^{-1})^{a-(\alpha+\beta)}e^{-cuy} - dyy^{-2}u^{a-\beta} \tag{A.26}$$

$$= \int_0^\infty y^{(a+1)-\beta}y^{\alpha+\beta-a}(y+1)^{a-(\alpha+\beta)}e^{-cuy} - dyy^{-2}u^{a-\beta} \tag{A.27}$$

$$= u^{a-\beta}\int_0^\infty y^{\alpha-1}(y+1)^{a-(\alpha+\beta)}e^{-cuy}dy \tag{A.28}$$

$$= u^{a-\beta}\int_0^\infty y^{\alpha-1}(y+1)^{a+1-\beta-\alpha-1}e^{-cuy}dy \tag{A.29}$$

$$= u^{a-\beta}\Gamma(\alpha)U(\alpha, a+1-\beta, cu) \tag{A.30}$$

where $U(\alpha, a+1-\beta, cu)$ is Tricomi's confluent hypergeometric function. Then the marginal distribution of u is

$$f(u) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\beta)}\frac{c^a}{\Gamma(a)}U(\alpha, a+1-\beta, cu)u^{a-1}e^{-cu}\mathbb{1}_{u>0} \tag{A.31}$$

From Oldham et al. (2008), we find that for $\alpha > 0$, which is required by the Beta distribution, $U(\alpha, a+1-\beta, cu) \to 0$ as $u \to \infty$ For behavior as $u \to 0$, if $\beta \geq a$ then $f(u) \to 0$ as

$U(\alpha, a+1-\beta, cu)$ is continuous and bounded at zero, namely:

$$U(\alpha, a+1-\beta, 0) = \frac{\Gamma(\beta - a)}{\Gamma(\alpha + \beta - a)}$$

(Oldham et al., 2008). If, instead, $\beta < a$, then $U(\alpha, a+1-\beta, 0) \to \infty$ as $u \to 0$. However, whether $\lim_{u\to 0} f(u) \to \infty$ or $\lim_{u\to 0} f(u) \to 0$ depends on the relation between β and a (see $\mathcal{S}$48:9 in Oldham et al. (2008)). If $a - \beta \in (0, 1)$ and $\beta < 1$ then the density approaches ∞ at 0, while if $\beta > 1$ the density approaches 0 instead. If $a - \beta > 1$ then if $\beta > 2$ the approaches 0, otherwise if $\beta < 2$ the density diverges at 0. At large u, the function $U(\alpha, a+1-\beta, cu) \sim u^{-\alpha}$ so the density can be approximated by:

$$f(u) \sim u^{a-\alpha-1} e^{-cu} \tag{A.32}$$

and ultimately has an exponential tail.

With the likelihood Suppose $J = 2$. We can write the observed-data likelihood in terms of u_j and v_j as follows:

$$\exp\left(-\textstyle\sum_j u_j \sum_i E_{ij}\right) \displaystyle\prod_{i=1}^{I} \frac{(E_{i1}u_1 + E_{i2}u_2)^{m_i}}{m_i!} \prod_{j=1}^{J} \frac{\exp\left(-v_j \sum_i E_{ij}\right) \left(\prod_{i=1}^{I} (E_{ij})^{x_{ij}}\right) v_j^{\sum_i x_{ij}}}{\prod_{i=1}^{I} x_{ij}!}$$

If we assume that, WLOG, group 2 is the majority group for all i, or, in other words, $E_{i1} \ll E_{i2}$ for all i, then using the approximation

$$\begin{aligned}(E_{i1}u_1 + E_{i2}u_2)^{m_i} &= (E_{i2}u_2)^{m_i} \exp m_i \log\left(1 + \frac{E_{i1}}{E_{i2}u_2} u_1\right) \\ &\approxeq (E_{i2}u_2)^{m_i} \exp m_i \left(u_1 \frac{E_{i1}}{E_{i2}u_2} - u_1^2 \frac{E_{i1}^2}{2E_{i2}^2 u_2^2}\right)\end{aligned}$$

Leading to the approximate likelihood:

$$\begin{aligned}&\exp\left(-\textstyle\sum_j u_j \sum_i E_{ij}\right) \left(\displaystyle\prod_{i=1}^{I} E_{i2}\right) u_2^{\sum_i m_i} \frac{\exp\left(u_1 \sum_i \frac{m_i E_{i1}}{E_{i2}u_2} - u_1^2 \sum_i \frac{m_i E_{i1}^2}{2E_{i2}^2 u_2^2}\right)}{\prod_{i=1}^{I} m_i!} \\ &\times \prod_{j=1}^{2} \frac{\exp\left(-v_j \sum_i E_{ij}\right) \left(\prod_{i=1}^{I} (E_{ij})^{x_{ij}}\right) v_j^{\sum_i x_{ij}}}{\prod_{i=1}^{I} x_{ij}!}\end{aligned}$$

Let $x_{+j} = \sum_i x_{ij}, E_{+j} = \sum_i E_{ij}, s_1 = \sum_i \frac{m_i E_{i1}}{E_{i2}}, s_2 = \sum_i \frac{m_i E_{i1}^2}{E_{i2}^2}$. We condition on u_2 and multiply by the joint prior above for u_1, v_1:

$$\pi_{v_1,u_1}(v,u|\text{Data}) \propto v^{\alpha+x_{+1}-1}u^{\beta-1}(v+u)^{a-(\alpha+\beta)}e^{-c(v+E_{+1})}e^{u(s_1/u_2-E_{+1}-c)-u^2s_2/2u_2^2}\mathbb{1}_{u>0}\mathbb{1}_{v>0}$$

$$= \int_0^\infty v^{\alpha+x_{+1}-1}(v+u)^{a-(\alpha+\beta)}e^{-v(c+E_{+1})}dv \tag{A.33}$$

$$= \int_0^\infty v^{\alpha+x_{+1}-1}v^{a-\alpha-\beta}(1+\frac{u}{v})^{a-(\alpha+\beta)}e^{-v(c+E_{+1})}dv \tag{A.34}$$

$$= \int_0^\infty v^{a+x_{+1}-(\beta+1)}(1+\frac{u}{v})^{a-(\alpha+\beta)}e^{-v(c+E_{+1})}dv \tag{A.35}$$

$$= \int_\infty^0 (\frac{u}{x})^{a+x_{+1}-(\beta+1)}(1+x)^{a-(\alpha+\beta)}e^{-\frac{u}{x}(c+E_{+1})}dx\frac{-u}{x^2} \tag{A.36}$$

$$= \int_0^\infty x^{\beta-(a+x_{+1}+1)}(1+x)^{a-(\alpha+\beta)}e^{-\frac{u}{x}(c+E_{+1})}dxu^{a+x_{+1}-\beta} \tag{A.37}$$

$$= \int_\infty^0 y^{(a+x_{+1}+1)-\beta}(1+y^{-1})^{a-(\alpha+\beta)}e^{-(c+E_{+1})uy} - dyy^{-2}u^{a+x_{+1}-\beta} \tag{A.38}$$

$$= \int_0^\infty y^{(a+x_{+1}+1)-\beta}y^{\alpha+\beta-a}(y+1)^{a-(\alpha+\beta)}e^{-(c+E_{+1})uy} - dyy^{-2}u^{a+x_{+1}-\beta} \tag{A.39}$$

$$= u^{a+x_{+1}-\beta}\int_0^\infty y^{\alpha+x_{+1}-1}(y+1)^{a-(\alpha+\beta)}e^{-(c+E_{+1})uy}dy \tag{A.40}$$

$$= u^{a+x_{+1}-\beta}\int_0^\infty y^{\alpha+x_{+1}-1}(y+1)^{a+x_{+1}+1-\beta-\alpha-x_{+1}-1}e^{-(c+E_{+1})uy}dy \tag{A.41}$$

$$= u^{a+x_{+1}-\beta}\Gamma(\alpha)U(\alpha+x_{+1}, a+x_{+1}+1-\beta, (c+E_{+1})u) \tag{A.42}$$

where $U(a,b,z)$ is Tricomi's confluent hypergeometric function. Then the marginal posterior distribution for u is

$$f(u) \propto U(\alpha+x_{+1}, a+x_{+1}+1-\beta, (c+E_{+1})u)u^{a+x_{+1}-1}e^{u(s_1/u_2-E_{+1}-c)-u^2s_2/2u_2^2}\mathbb{1}_{u>0} \tag{A.43}$$

When $a=\alpha$ and $\beta=1$,

$$U(\alpha+x_{+1}, a+x_{+1}+1-\beta, (c+E_{+1})u) = e^{(c+E_{+1})u}\int_{(c+E_{+1})u}^\infty z^{-(\alpha+x_{+1})}e^{-z}dz$$

A.4.2.2 Hierarchy over u_1

Let $p_{1g} \mid \alpha_1, \beta_1 \sim \text{Beta}(\alpha_1, \beta_1)$, $\lambda_{1g} \mid \alpha_1, \beta_1, r_1 \sim \text{Gamma}(\alpha_1 + \beta_1, r_1)$ so $u_{1g} \mid \beta_1, r_1 \sim \text{Gamma}(\beta_1, r_1)$. Then using the approximate likelihood from above:

$$\pi_{u_{11},\ldots,u_{1G}}(u_{11},\ldots,u_{1G},r_1|\text{Data},u_{21},\ldots,u_{2G}) \propto r_1^{G\beta_1}\mathbb{1}_{r_1>0}\prod_g u_{1g}^{\beta_1-1}e^{u_{1g}(s_1/u_{2g}-E_{+1}-r_1)-u_{1g}^2 s_2/2u_{2g}^2}\mathbb{1}_{u_{1g}>0}$$

Assuming we know $\beta_1 = 2$, we can integrate out each u_{1g} in order to get an Empirical Bayes estimator:

$$\pi_{r_1}(r_1|\text{Data},u_{21},\ldots,u_{2G}) \propto r_1^{G\beta_1}\mathbb{1}_{r_1>0}\int\prod_g e^{u_{1g}(s_1/u_{2g}-E_{+1}-r_1)-u_{1g}^2 s_2/2u_{2g}^2}\mathbb{1}_{u_{1g}>0}du_{1g}$$

Using the fact that the normalizing constant of a Modified Half Normal distribution is $\Psi(\frac{\beta_1}{2}, \frac{s_1/u_{2g}-E_{+1}-r_1}{\sqrt{s_2/2u_{2g}^2}})\frac{1}{2(s_2/2u_{2g}^2)^{\beta_1/2}}$, and that $\Psi(\frac{\beta_1}{2}, x)$ has an analytic form for integer β_1, we can derive the integrated distribution. Let $x_g(r,u_{2g}) = \frac{s_1/u_{2g}-E_{+1}-r_1}{\sqrt{s_2/2u_{2g}^2}}$

$$\Psi(\frac{2}{2}, x_g(r,u_{2g})) = 1 + \sqrt{\pi}x_g(r,u_{2g})e^{x_g(r,u_{2g})^2/4}\Phi\left(x_g(r,u_{2g})/\sqrt{2}\right)$$

Then

$$\pi_{r_1}(r_1|\text{Data},u_{21},\ldots,u_{2G}) \propto r_1^{G\beta_1}\mathbb{1}_{r_1>0}\prod_g\left(1 + \sqrt{\pi}x_g(r_1,u_{2g})e^{x_g(r_1,u_{2g})^2/4}\Phi\left(x_g(r_1,u_{2g})/\sqrt{2}\right)\right)$$

which is more easily expressible as

$$\pi_{r_1}(r_1|\text{Data},u_{21},\ldots,u_{2G}) \propto r_1^{G\beta_1}\mathbb{1}_{r_1>0}\prod_g\left(1 + \frac{z_g(r_1,u_{2g})}{\phi(z_g(r_1,u_{2g}))}\Phi(z_g(r_1,u_{2g}))\right)$$

where $z_g(r,u_{2g}) = x_g(r,u_{2g})/\sqrt{2}$ and $\phi(x)$ is the standard normal density. This is not a proper posterior, as the integral of the kernel does not converge. Note that $r_1 > 0$, so we need to show that the expression above is bounded below on $z_g(r,u_{2g}) < 0$ by a nonintegrable function. We can bound $\Phi(x)$ above by the expression:

$$\frac{2\phi(x)}{-x + \sqrt{x^2 + 8/\pi}}$$

for $x < 0$. Then

$$1 + \frac{z_g(r_1, u_{2g})}{\phi(z_g(r_1, u_{2g}))}\Phi(z_g(r_1, u_{2g})) \geq \tag{A.44}$$

$$1 + \frac{2z_g(r_1, u_{2g})}{-z_g(r_1, u_{2g}) + \sqrt{z_g(r_1, u_{2g})^2 + 8/\pi}} \tag{A.45}$$

on $z_g(r_1, u_{2g}) < 0$. Let $u = -z_g(r_1, u_{2g})$, then the expression above is

$$1 - \frac{2u}{u + \sqrt{u^2 + 8/\pi}} = \frac{-u + \sqrt{u^2 + 8/\pi}}{u + \sqrt{u^2 + 8/\pi}} \tag{A.46}$$

This is asymptotic to $\frac{2}{\pi u^2}$. Putting this together with the fact that $\beta_1 = 2$,

$$r_1^{2G}\mathbb{1}_{r_1>0}\prod_g \left(1 + \frac{z_g(r_1, u_{2g})}{\phi(z_g(r_1, u_{2g}))}\Phi(z_g(r_1, u_{2g}))\right) \geq r_1^{2G}\mathbb{1}_{r_1>0}\prod_g \frac{2}{\pi z_g(r_1, u_{2g})^2} \tag{A.47}$$

$$= \mathbb{1}_{r_1>0}\prod_g \frac{2s_2 r_1^2}{\pi(s_1 - u_{2g}(E_{+1} + r_1))^2} \tag{A.48}$$

for r_1 large. Of course,

$$\int_0^\infty \prod_g \frac{2s_2 r_1^2}{\pi(s_1 - u_{2g}(E_{+1} + r_1))^2} dr_1 \to \infty \tag{A.49}$$

which implies that

$$\int_0^\infty r_1^{2G}\prod_g \left(1 + \frac{z_g(r_1, u_{2g})}{\phi(z_g(r_1, u_{2g}))}\Phi(z_g(r_1, u_{2g}))\right) dr_1 \to \infty \tag{A.50}$$

We can, however, numerically maximize the expression to yield an MLE for r_1, or we can put a proper prior on r_1 and integrate over the uncertainty.

A.4.3 DCT lemma

Lemma 9. *Let $p_{\eta(\theta)}(X = k)$ be defined*

$$\frac{1}{k!}\exp\left(\eta(\theta)k - e^{\eta(\theta)}\right),$$

where $\eta(\theta)$ is a univariate differentiable function of θ, $\theta \in \mathbb{R}^d$. Let $g(\eta(\theta)) = \int f(x)p_{\eta(\theta)}(x)d\mu(x)$ where μ is the counting measure on $[0, 1, 2, \dots]$. If we define the set $\theta \in \Theta_f$ as the set for which

$\int |f(x)|\, p_{\eta(\theta)}(x) d\mu(x) < \infty$, *then*

$$\frac{\partial}{\partial \theta_j} g(\eta(\theta)) = \int f(x) \frac{\partial}{\partial \theta_j} p_{\eta(\theta)}(x) d\mu(x)$$

Proof. By the chain rule, the $\frac{\partial}{\partial \theta_j} g(\eta(\theta)) = \frac{dg(\eta(\theta))}{d\eta(\theta)} \frac{\partial \eta(\theta)}{\partial \theta_j}$. By Theorem 2.4 in Keener (2010), $\frac{dg(\eta(\theta))}{d\eta(\theta)}$ exists and can be obtained via differentiating under the integral sign:

$$\frac{d}{d\eta(\theta)} g(\eta(\theta)) = \int f(x) \frac{dp_{\eta(\theta)}(x)}{d\eta(\theta)} d\mu(x).$$

Using this result and the chain rule again yields

$$\begin{aligned} \frac{\partial}{\partial \theta_j} g(\eta(\theta)) &= \frac{dg(\eta(\theta))}{d\eta(\theta)} \frac{\partial \eta(\theta)}{\partial \theta_j} && \text{(A.51)} \\ &= \int f(x) \frac{dp_{\eta(\theta)}(x)}{d\eta(\theta)} \frac{\partial \eta(\theta)}{\partial \theta_j} d\mu(x) && \text{(A.52)} \\ &= \int f(x) \frac{\partial}{\partial \theta_j} dp_{\eta(\theta)}(x) && \text{(A.53)} \end{aligned}$$

□

A.4.4 Lemmas and theorems in service of Fisher Info

Our local identifiability result is based on the following theorem that is referenced, though not proven, in Mukerjee and Sutradhar (2002). To our knowledge an explicit proof has not been given, though it follows directly from the proof of the Cramér-Rao lower bound in Rao (2002). This proof provides a slightly different route to showing local identifiability compared to that of Catchpole (1997). We give a proof below where we use the same notation as used in Rao (2002) for clarity's sake.

Theorem 10. *Suppose we have observations X_n and let $\mathbf{x} \in \mathbb{R}^N$ be the collection of all observations, with the n-th element equal to X_n, where $\pi(\mathbf{x}, \boldsymbol{\theta})$, parameterized by $\boldsymbol{\theta} \in \mathbb{R}^d$ with i^{th} element θ_i is the joint density of the observations. Let $f_1(\mathbf{x}), \ldots, f_r(\mathbf{x})$ be r statistics for which $\mathbb{E}[f_i(\mathbf{x})] = g_i(\boldsymbol{\theta})$. Further, assume that $\frac{\partial}{\partial \theta_j} \int f_i(\mathbf{x}) \pi(\mathbf{x}, \boldsymbol{\theta}) d\mathbf{x} = \int f_i(\mathbf{x}) \frac{\partial}{\partial \theta_j} \pi(\mathbf{x}, \boldsymbol{\theta}) d\mathbf{x} = \frac{\partial g_i(\boldsymbol{\theta})}{\partial \theta_j}$. Let Δ be a matrix in $\mathbb{R}^{r \times d}$ with elements $\Delta_{ij} = \frac{\partial g_i(\boldsymbol{\theta})}{\partial \theta_j}$. Let $\mathcal{I}$ be a matrix in $\mathbb{R}^{d \times d}$ where the (i,j)-th element is defined as $\mathcal{I}_{ij} = Cov\left(\frac{\partial \log P(\mathbf{x}, \boldsymbol{\theta})}{\partial \theta_i}, \frac{\partial \log P(\mathbf{x}, \boldsymbol{\theta})}{\partial \theta_j}\right)$. Let $\mathbf{V}$ be the matrix in $\mathbb{R}^{r \times r}$ with (i,j) elements $\mathbf{V}_{ij} = Cov(f_i(\mathbf{x}), f_j(\mathbf{x}))$. If $\mathbf{V}$ is positive definite, and Δ is full-rank, then the Fisher information matrix $\mathcal{I}$ is positive definite.*

Proof. Let $\mathbf{f}$ be the ordered collection of elements $f_i(\mathbf{x})$, and let $\nabla \log \pi(\mathbf{x}, \boldsymbol{\theta})$ be the score vector.

Then let the random vector $\rho = (\mathbf{f}, \nabla \log \pi(\mathbf{x}, \boldsymbol{\theta}))$. The covariance matrix associated with ρ, Σ, is a block matrix. Under $\pi(\mathbf{x}, \boldsymbol{\theta})$, $\text{Cov}\left(f_i, \frac{\partial \log \pi(\mathbf{x}, \boldsymbol{\theta})}{\partial \theta_j}\right) = \frac{\partial g_i(\boldsymbol{\theta})}{\partial \theta_j}$, which is element (i, j) of the matrix Δ. The block covariance matrix for ρ is:

$$\Sigma = \begin{bmatrix} \mathbf{V} & \Delta \\ \Delta^T & \mathcal{I} \end{bmatrix}$$

Suppose $\mathbf{V}$ is positive definite. We know $\mathcal{I} - \Delta^T \mathbf{V}^{-1} \Delta \succeq 0$, because Σ is a covariance matrix which ensures it is positive semi definite. Furthermore, $\mathcal{I} \succeq \Delta^T \mathbf{V}^{-1} \Delta$. If Δ is full-rank then $\mathcal{I} \succ 0$. □

We need the following two lemmas to prove that the Fisher information is positive definite. First, the following lemma is stated in Tian (2004):

Lemma 11. *Let $\mathbf{A}^\dagger$ be the Moore-Penrose inverse of a matrix $\mathbf{A} \in \mathbb{R}^{L \times M}$. Let $\mathbf{B} \in \mathbb{R}^{L \times T}$. The rank of a block matrix $[\mathbf{A}\ \mathbf{B}]$ is the rank of $\mathbf{A}$ plus the rank of $\mathbf{B} - \mathbf{A}\mathbf{A}^\dagger\mathbf{B}$.*

Next, we will need this lemma later on:

Lemma 12. *Suppose $\mathbf{W} \in \mathbb{R}^{I \times K}$ and $\mathbf{E} \in \mathbb{R}^{I \times J}$. Let $\mathbf{E}_{[:,j]}$ be an I-vector of the j^{th} column of matrix $\mathbf{E}$, let $\mathbf{F}_j = \text{diag}(\mathbf{E}_{[:,j]})$.*

(L.a) $I \geq J + K$

(L.b) $\text{rank}\,(\mathbf{E}) = J$

(L.c) $\text{rank}\,(\mathbf{W}) = K$

(L.d) $\forall i \in 1, \ldots, I \sum_{j=1}^{J} E_{ij} > 0$

(L.e) $\lambda_j > 0 \forall j \in 1, \ldots, J$

(L.f) $\text{rank}\left(\begin{bmatrix} \mathbf{F}_1\mathbf{W} & \ldots & \mathbf{F}_J\mathbf{W} & \mathbf{E}_{[:,1]} & \mathbf{E}_{[:,2]} & \ldots & \mathbf{E}_{[:,J-1]} & \mathbf{E}_{[:,J]} \end{bmatrix}\right) > J + K$

Then the matrix

$$\begin{bmatrix} \left(\sum_j \lambda_j \mathbf{F}_j\right)\mathbf{W} & \mathbf{E}_{[:,1]} & \mathbf{E}_{[:,2]} & \ldots & \mathbf{E}_{[:,J-1]} & \mathbf{E}_{[:,J]} \end{bmatrix}$$

is rank $J + K$.

Proof. Given conditions (L.a) to (L.e) above , the matrix $\left(\sum_j \lambda_j \mathbf{F}_j\right)\mathbf{W}$ is rank K. Using lemma 11 we proceed sequentially, showing first that

$$\text{rank}\left(\begin{bmatrix} \left(\sum_j \lambda_j \mathbf{F}_j\right)\mathbf{W} & \mathbf{E}_{[:,1]} \end{bmatrix}\right) = K + 1$$

By 11,

$$\text{rank}\left(\left[\left(\sum_j \lambda_j \mathbf{F}_j\right)\mathbf{W} \quad \mathbf{E}_{[:,1]}\right]\right) = \text{rank}\left(\left(\sum_j \lambda_j \mathbf{F}_j\right)\mathbf{W}\right) + \text{rank}\left(\mathbf{E}_{[:,1]} - \left(\sum_j \lambda_j \mathbf{F}_j\right)\mathbf{W}\left(\left(\sum_j \lambda_j \mathbf{F}_j\right)\mathbf{W}\right)^{\dagger}\mathbf{E}_{[:,1]}\right)$$

We can show that rank $\left(\mathbf{E}_{[:,1]} - \left(\sum_j \lambda_j \mathbf{F}_j\right)\mathbf{W}\left(\left(\sum_j \lambda_j \mathbf{F}_j\right)\mathbf{W}\right)^{\dagger}\mathbf{E}_{[:,1]}\right) = 1$. We show by contradiction that given conditions on the coefficient matrix, no solution to the equation in a_k

$$\mathbf{E}_{[:,1]} = \sum_{k=1}^{K} a_k \left(\sum_{j=1}^{J} \lambda_j \mathbf{E}_{[:,j]} \odot \mathbf{W}_k\right) \tag{A.54}$$

can be found. This follows from examining the matrix of coefficients of the equation, shown here in block form:

$$\begin{bmatrix}\mathbf{E}_{[:,1]} & \lambda_1\mathbf{F}_1\mathbf{W} & \dots & \lambda_J\mathbf{F}_J\mathbf{W}\end{bmatrix} \tag{A.55}$$

Given condition (L.e), $\lambda_j > 0$ for all j, for a fixed set of parameters λ_j the matrix in eq. (A.55) has the same column space as

$$\begin{bmatrix}\mathbf{E}_{[:,1]} & \mathbf{F}_1\mathbf{W} & \dots & \mathbf{F}_J\mathbf{W}\end{bmatrix} \tag{A.56}$$

Given condition (L.f), matrix A.56 has rank greater than K, so the system of equations in A.54 will not have a solution in $a_k, k \in [1, \dots, K]$. Thus the rank of matrix

$$\mathbf{E}_{[:,1]} - \left(\sum_j \lambda_j \mathbf{F}_j\right)\mathbf{W}\left(\left(\sum_j \lambda_j \mathbf{F}_j\right)\mathbf{W}\right)^{\dagger}\mathbf{E}_{[:,1]} \tag{A.57}$$

is 1 so the rank of .

$$\left[\left(\sum_j \lambda_j \mathbf{F}_j\right)\mathbf{W} \quad \mathbf{E}_{[:,1]}\right]$$

is $K + 1$.

Now for the induction step: Suppose that

$$\left[\left(\sum_j \lambda_j \mathbf{F}_j\right)\mathbf{W} \quad \mathbf{E}_{[:,1:M]}\right]$$

is rank $M + K$ and suppose we want to determine the rank of

$$\left[\left(\sum_j \lambda_j \mathbf{F}_j\right)\mathbf{W} \quad \mathbf{E}_{[:,1:M]} \quad \mathbf{E}_{[:,M+1]}\right].$$

By lemma 11, the rank of the above matrix is

$$\text{rank}\left[\left(\sum_j \lambda_j \mathbf{F}_j\right)\mathbf{W} \quad \mathbf{E}_{[:,1:M]}\right] \tag{A.58}$$

$$+\,\text{rank}\left(\mathbf{E}_{[:,M+1]} - \left[\left(\sum_j \lambda_j \mathbf{F}_j\right)\mathbf{W} \quad \mathbf{E}_{[:,1:M]}\right]\left[\left(\sum_j \lambda_j \mathbf{F}_j\right)\mathbf{W} \quad \mathbf{E}_{[:,1:M]}\right]^{\dagger}\mathbf{E}_{[:,M+1]}\right) \tag{A.59}$$

By the induction hypothesis, we know that eq. (A.58) equals $K + M$ and we look to see whether eq. (A.59) equals 1 by determining whether the system of equations below:

$$\mathbf{E}_{[:,M+1]} = \sum_{k=1}^{K} a_k \left(\sum_{j=1}^{J} \lambda_j \mathbf{E}_j \odot \mathbf{W}_k\right) + \sum_{m=1}^{M} d_m \mathbf{E}_{[:,m]} \tag{A.60}$$

has a solution in the variables $a_k, k \in [1, \ldots, K]$, $d_m, m \in [1, \ldots, M], M < J$. Then for a fixed $\{\lambda_j, j \in [1, \ldots, J]\}$ the coefficient matrix for the system of equations in eq. (A.60) has the same column space as

$$\left[\mathbf{E}_{[:,M+1]} \quad \mathbf{F}_1\mathbf{W} \quad \ldots \quad \mathbf{F}_J\mathbf{W} \quad \mathbf{E}_{[:,1:M]}\right] \tag{A.61}$$

By condition (L.f), the rank of matrix (A.61) is greater than $M+K$, which precludes $\mathbf{E}_{[:,M+1]}$ from lying in the column space of

$$\left[\left(\sum_j \lambda_j \mathbf{F}_j\right)\mathbf{W} \quad \mathbf{E}_{[:,1:M]}\right]$$

. Thus the line (A.59) is equal to 1 and summing with line (A.58) shows that the rank of

$$\left[\left(\sum_j \lambda_j \mathbf{F}_j\right)\mathbf{W} \quad \mathbf{E}_{[:,1:M]} \quad \mathbf{E}_{[:,M+1]}\right]$$

is $M + 1 + K$. Therefore by induction,

$$\left[\left(\sum_j \lambda_j \mathbf{F}_j\right)\mathbf{W} \quad \mathbf{E}_{[:,1]} \quad \mathbf{E}_{[:,2]} \quad \ldots \quad \mathbf{E}_{[:,J-1]} \quad \mathbf{E}_{[:,J]}\right]$$

is rank $J + K$. □

A.5 Full model local identifiability proof

Proof. In order to draw on the results of 10, we must ensure $\frac{\partial}{\partial\theta_j}\int f_i\pi(\mathbf{x},\theta)d\nu = \int f_i\frac{\partial}{\partial\theta_j}\pi(\mathbf{x},\theta)d\nu$ holds for model (2.10). This condition does indeed hold by lemma 9 because our observational density is an exponential family density. Now we look for moment estimators, $\mathbf{f} = (f_1,\ldots,f_r)$ with full-rank Δ and positive definite covariance matrices. As shown in section A.2, conditionally on unknown parameters, known covariates $\mathbf{z}_i$ and population counts $\mathbf{e}_i$, X_{ij} is independent of M_i for all i and j. Let $\mathbf{x}_j = (X_{1j}, X_{2j},\ldots,X_{Ij})$, and set

$$\mathbf{f} = (\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_J, M_1, M_2,\ldots,M_I).$$

From the independence of the elements of the vector $(X_{i1}, X_{i2},\ldots,X_{iJ},, M_i)^T$, and the conditional independence between observations i, $\mathrm{Cov}(\mathbf{f})$ is diagonal, and is positive definite because conditions (S.d) to (S.f) are assumed to hold. The final condition for theorem 10 to hold, which establishes a lower bound on the positive definiteness of the Fisher information matrix is to ensure that Δ has full column-rank. To that end, let us calculate Δ, beginning with $\mathbb{E}[\mathbf{f}]$: The expected value of $\mathbf{f}$ is the vector:

$$\begin{bmatrix}
E_{11}p_{11}e^{\mathbf{z}_1^T\boldsymbol{\beta}}\lambda_1 \\
E_{21}p_{21}e^{\mathbf{z}_2^T\boldsymbol{\beta}}\lambda_1 \\
\vdots \\
E_{I1}p_{I1}e^{\mathbf{z}_I^T\boldsymbol{\beta}}\lambda_1 \\
E_{12}p_{12}e^{\mathbf{z}_1^T\boldsymbol{\beta}}\lambda_2 \\
E_{22}p_{22}e^{\mathbf{z}_2^T\boldsymbol{\beta}}\lambda_2 \\
\vdots \\
E_{I2}p_{I2}e^{\mathbf{z}_I^T\boldsymbol{\beta}}\lambda_2 \\
\vdots \\
E_{1J}p_{1J}e^{\mathbf{z}_1^T\boldsymbol{\beta}}\lambda_J \\
E_{2J}p_{2J}e^{\mathbf{z}_2^T\boldsymbol{\beta}}\lambda_J \\
\vdots \\
E_{IJ}p_{IJ}e^{\mathbf{z}_I^T\boldsymbol{\beta}}\lambda_J \\
\sum_j E_{1j}(1-p_{1j})e^{\mathbf{z}_1^T\boldsymbol{\beta}}\lambda_j \\
\vdots \\
\sum_j E_{Ij}(1-p_{Ij})e^{\mathbf{z}_I^T\boldsymbol{\beta}}\lambda_j
\end{bmatrix}$$

For efficiency of notation, let $c_{ij} = p_{ij}\lambda_j$ $p'_{ij} = p_{ij}(1-p_{ij})$, and $q_{ij} = (1-p_{ij})$ Let $\mathbf{H}$ by the $(J+1)I \times 2K$ matrix of partial derivatives of $\mathbb{E}[\mathbf{f}]$ with respect to the vector $(\boldsymbol{\beta}, \boldsymbol{\gamma})$:

$$\mathbf{H} = \begin{bmatrix} \mathbf{z}_1^T E_{11} c_{11} e^{\mathbf{z}_1^T \boldsymbol{\beta}} & \mathbf{z}_1^T E_{11} p'_{11} \lambda_1 e^{\mathbf{z}_1^T \boldsymbol{\beta}} \\ \mathbf{z}_2^T E_{21} c_{21} e^{\mathbf{z}_2^T \boldsymbol{\beta}} & \mathbf{z}_2^T E_{21} p'_{21} \lambda_1 e^{\mathbf{z}_2^T \boldsymbol{\beta}} \\ \vdots & \vdots \\ \mathbf{z}_I^T E_{I1} c_{I1} e^{\mathbf{z}_I^T \boldsymbol{\beta}} & \mathbf{z}_I^T E_{I1} p'_{I1} \lambda_1 e^{\mathbf{z}_I^T \boldsymbol{\beta}} \\ \mathbf{z}_1^T E_{12} c_{12} e^{\mathbf{z}_1^T \boldsymbol{\beta}} & \mathbf{z}_1^T E_{12} p'_{12} \lambda_2 e^{\mathbf{z}_1^T \boldsymbol{\beta}} \\ \mathbf{z}_2^T E_{22} c_{22} e^{\mathbf{z}_2^T \boldsymbol{\beta}} & \mathbf{z}_2^T E_{21} p'_{22} \lambda_2 e^{\mathbf{z}_2^T \boldsymbol{\beta}} \\ \vdots & \vdots \\ \mathbf{z}_I^T E_{I2} c_{I2} e^{\mathbf{z}_I^T \boldsymbol{\beta}} & \mathbf{z}_I^T E_{I2} p'_{I2} \lambda_2 e^{\mathbf{z}_I^T \boldsymbol{\beta}} \\ \vdots & \vdots \\ \mathbf{z}_1^T E_{1J} c_{1J} e^{\mathbf{z}_1^T \boldsymbol{\beta}} & \mathbf{z}_1^T E_{1J} p'_{1J} \lambda_J e^{\mathbf{z}_1^T \boldsymbol{\beta}} \\ \mathbf{z}_2^T E_{2J} c_{2J} e^{\mathbf{z}_2^T \boldsymbol{\beta}} & \mathbf{z}_2^T E_{2J} p'_{2J} \lambda_J e^{\mathbf{z}_2^T \boldsymbol{\beta}} \\ \vdots & \vdots \\ \mathbf{z}_I^T E_{IJ} c_{IJ} e^{\mathbf{z}_I^T \boldsymbol{\beta}} & \mathbf{z}_I^T E_{IJ} p'_{IJ} \lambda_J e^{\mathbf{z}_I^T \boldsymbol{\beta}} \\ \mathbf{z}_1^T e^{\mathbf{z}_1^T \boldsymbol{\beta}} \sum_j E_{1j} q_{1j} \lambda_j & -\mathbf{z}_1^T e^{\mathbf{z}_1^T \boldsymbol{\beta}} \sum_j E_{1j} p'_{1j} \lambda_j \\ \vdots & \vdots \\ \mathbf{z}_I^T e^{\mathbf{z}_I^T \boldsymbol{\beta}} \sum_j E_{Ij} q_{Ij} \lambda_j & -\mathbf{z}_I^T e^{\mathbf{z}_I^T \boldsymbol{\beta}} \sum_j E_{Ij} p'_{Ij} \lambda_j \end{bmatrix}$$

Let $\mathbf{T}$ be the $(J+1)I \times 2J$ matrix of partial derivatives with respect to $(\lambda_1, \lambda_2, \ldots, \lambda_J, \eta_1, \eta_2, \ldots, \eta_J)$,

$$\mathbf{T} = \begin{bmatrix} E_{11} p_{11} e^{\mathbf{z}_1^T \boldsymbol{\beta}} & 0 & \ldots & 0 & E_{11} p'_{11} \lambda_1 e^{\mathbf{z}_1^T \boldsymbol{\beta}} & 0 & \ldots & 0 \\ E_{21} p_{21} e^{\mathbf{z}_2^T \boldsymbol{\beta}} & 0 & \ldots & 0 & E_{21} p'_{21} \lambda_1 e^{\mathbf{z}_2^T \boldsymbol{\beta}} & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ E_{I1} p_{I1} e^{\mathbf{z}_I^T \boldsymbol{\beta}} & 0 & \ldots & 0 & E_{I1} p'_{I1} \lambda_1 e^{\mathbf{z}_I^T \boldsymbol{\beta}} & 0 & \ldots & 0 \\ 0 & E_{12} p_{12} e^{\mathbf{z}_1^T \boldsymbol{\beta}} & \ldots & 0 & 0 & E_{12} p'_{12} \lambda_2 e^{\mathbf{z}_1^T \boldsymbol{\beta}} & \ldots & 0 \\ 0 & E_{22} p_{22} e^{\mathbf{z}_2^T \boldsymbol{\beta}} & \ldots & 0 & 0 & E_{22} p'_{22} \lambda_2 e^{\mathbf{z}_2^T \boldsymbol{\beta}} & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & E_{I2} p_{I2} e^{\mathbf{z}_I^T \boldsymbol{\beta}} & \ldots & 0 & 0 & E_{I2} p'_{I2} \lambda_2 e^{\mathbf{z}_I^T \boldsymbol{\beta}} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & E_{1J} p_{1J} e^{\mathbf{z}_1^T \boldsymbol{\beta}} & 0 & 0 & \ldots & E_{1J} p'_{1J} \lambda_J e^{\mathbf{z}_1^T \boldsymbol{\beta}} \\ 0 & 0 & \ldots & E_{2J} p_{2J} e^{\mathbf{z}_2^T \boldsymbol{\beta}} & 0 & 0 & \ldots & E_{2J} p'_{2J} \lambda_J e^{\mathbf{z}_2^T \boldsymbol{\beta}} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & E_{IJ} p_{IJ} e^{\mathbf{z}_I^T \boldsymbol{\beta}} & 0 & 0 & \ldots & E_{IJ} p'_{IJ} \lambda_J e^{\mathbf{z}_I^T \boldsymbol{\beta}} \\ E_{11} q_{11} e^{\mathbf{z}_1^T \boldsymbol{\beta}} & E_{12} q_{12} e^{\mathbf{z}_1^T \boldsymbol{\beta}} & \ldots & E_{1J} q_{1J} e^{\mathbf{z}_1^T \boldsymbol{\beta}} & -E_{11} p'_{11} \lambda_1 e^{\mathbf{z}_1^T \boldsymbol{\beta}} & -E_{12} p'_{12} \lambda_2 e^{\mathbf{z}_1^T \boldsymbol{\beta}} & \ldots & -E_{1J} p'_{1J} \lambda_J e^{\mathbf{z}_1^T \boldsymbol{\beta}} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ E_{I1} q_{I1} e^{\mathbf{z}_I^T \boldsymbol{\beta}} & E_{I2} q_{I2} e^{\mathbf{z}_I^T \boldsymbol{\beta}} & \ldots & E_{IJ} q_{IJ} e^{\mathbf{z}_I^T \boldsymbol{\beta}} & -E_{I1} p'_{I1} \lambda_1 e^{\mathbf{z}_I^T \boldsymbol{\beta}} & -E_{I2} p'_{I2} \lambda_2 e^{\mathbf{z}_I^T \boldsymbol{\beta}} & \ldots & -E_{IJ} p'_{IJ} \lambda_J e^{\mathbf{z}_I^T \boldsymbol{\beta}} \end{bmatrix}$$

Then

$$\Delta = \begin{bmatrix} \mathbf{T} & \mathbf{H} \end{bmatrix}$$

Let the matrix $\mathbf{R}_{i,j}(m)$ be the elementary row-addition matrix. When a matrix $\mathbf{A} \in \mathbb{R}^{M \times N}$ is left-multiplied by $\mathbf{R}_{ij}(m)$, $\tilde{\mathbf{A}} = \mathbf{E}_{ij}(m)\mathbf{A}$, all rows of $\tilde{\mathbf{A}}$ equal that of $\mathbf{A}$ excepting $\tilde{\mathbf{A}}$'s i-th row, which is $\tilde{\mathbf{A}}_{[i,:]} = \mathbf{A}_{[i,:]} + m\mathbf{A}_{[j,:]}$. Let $\tilde{\mathbf{H}}$ and $\tilde{\mathbf{T}}$ be the result of left-multiplying $\mathbf{H}$ and $\mathbf{T}$ by the

same product of elementary row-addition matrices, namely:

$$\prod_{i=1}^{I}\prod_{j=1}^{J}\mathbf{R}_{JI+j,(j-1)I+i}(1)$$

Then let $\tilde{\Delta}$ be the matrix Δ after applying the product of elementary row-addition matrices:

$$\begin{aligned}\tilde{\Delta} &= \prod_{i=1}^{I}\prod_{j=1}^{J}\mathbf{R}_{JI+j,(j-1)I+i}(1)\begin{bmatrix}\mathbf{T} & \mathbf{H}\end{bmatrix}\\ &= \begin{bmatrix}\prod_{i=1}^{I}\prod_{j=1}^{J}\mathbf{R}_{JI+j,(j-1)I+i}(1)\mathbf{T} & \prod_{i=1}^{I}\prod_{j=1}^{J}\mathbf{R}_{JI+j,(j-1)I+i}(1)\mathbf{H}\end{bmatrix}\\ &= \begin{bmatrix}\tilde{\mathbf{T}} & \tilde{\mathbf{H}}\end{bmatrix}\end{aligned}$$

where

$$\tilde{\mathbf{H}} = \begin{bmatrix}
\mathbf{z}_1^T E_{11} c_{11} e^{\mathbf{z}_1^T\boldsymbol{\beta}} & \mathbf{z}_1^T E_{11} p'_{11}\lambda_1 e^{\mathbf{z}_1^T\boldsymbol{\beta}}\\
\mathbf{z}_2^T E_{21} c_{21} e^{\mathbf{z}_2^T\boldsymbol{\beta}} & \mathbf{z}_2^T E_{21} p'_{21}\lambda_1 e^{\mathbf{z}_2^T\boldsymbol{\beta}}\\
\vdots & \vdots\\
\mathbf{z}_I^T E_{I1} c_{I1} e^{\mathbf{z}_I^T\boldsymbol{\beta}} & \mathbf{z}_I^T E_{I1} p'_{I1}\lambda_1 e^{\mathbf{z}_I^T\boldsymbol{\beta}}\\
\mathbf{z}_1^T E_{12} c_{12} e^{\mathbf{z}_1^T\boldsymbol{\beta}} & \mathbf{z}_1^T E_{12} p'_{12}\lambda_2 e^{\mathbf{z}_1^T\boldsymbol{\beta}}\\
\mathbf{z}_2^T E_{22} c_{22} e^{\mathbf{z}_2^T\boldsymbol{\beta}} & \mathbf{z}_2^T E_{21} p'_{22}\lambda_2 e^{\mathbf{z}_2^T\boldsymbol{\beta}}\\
\vdots & \vdots\\
\mathbf{z}_I^T E_{I2} c_{I2} e^{\mathbf{z}_I^T\boldsymbol{\beta}} & \mathbf{z}_I^T E_{I2} p'_{I2}\lambda_2 e^{\mathbf{z}_I^T\boldsymbol{\beta}}\\
\vdots & \vdots\\
\mathbf{z}_1^T E_{1J} c_{1J} e^{\mathbf{z}_1^T\boldsymbol{\beta}} & \mathbf{z}_1^T E_{1J} p'_{1J}\lambda_J e^{\mathbf{z}_1^T\boldsymbol{\beta}}\\
\mathbf{z}_2^T E_{2J} c_{2J} e^{\mathbf{z}_2^T\boldsymbol{\beta}} & \mathbf{z}_2^T E_{2J} p'_{2J}\lambda_J e^{\mathbf{z}_2^T\boldsymbol{\beta}}\\
\vdots & \vdots\\
\mathbf{z}_I^T E_{IJ} c_{IJ} e^{\mathbf{z}_I^T\boldsymbol{\beta}} & \mathbf{z}_I^T E_{IJ} p'_{IJ}\lambda_J e^{\mathbf{z}_I^T\boldsymbol{\beta}}\\
\mathbf{z}_1^T e^{\mathbf{z}_1^T\boldsymbol{\beta}}\sum_j E_{1j}\lambda_j & 0_{1\times K}\\
\vdots & \vdots\\
\mathbf{z}_I^T e^{\mathbf{z}_I^T\boldsymbol{\beta}}\sum_j E_{Ij}\lambda_j & 0_{1\times K}
\end{bmatrix},$$

and

$$\tilde{\mathbf{T}} = \begin{bmatrix}
E_{11}p_{11}e^{\mathbf{z}_1^T\boldsymbol{\beta}} & 0 & \dots & 0 & E_{11}p'_{11}\lambda_1 e^{\mathbf{z}_1^T\boldsymbol{\beta}} & 0 & \dots & 0 \\
E_{21}p_{21}e^{\mathbf{z}_2^T\boldsymbol{\beta}} & 0 & \dots & 0 & E_{21}p'_{21}\lambda_1 e^{\mathbf{z}_2^T\boldsymbol{\beta}} & 0 & \dots & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
E_{I1}p_{I1}e^{\mathbf{z}_I^T\boldsymbol{\beta}} & 0 & \dots & 0 & E_{I1}p'_{I1}\lambda_1 e^{\mathbf{z}_I^T\boldsymbol{\beta}} & 0 & \dots & 0 \\
0 & E_{12}p_{12}e^{\mathbf{z}_1^T\boldsymbol{\beta}} & \dots & 0 & 0 & E_{12}p'_{12}\lambda_2 e^{\mathbf{z}_1^T\boldsymbol{\beta}} & \dots & 0 \\
0 & E_{22}p_{22}e^{\mathbf{z}_2^T\boldsymbol{\beta}} & \dots & 0 & 0 & E_{22}p'_{22}\lambda_2 e^{\mathbf{z}_2^T\boldsymbol{\beta}} & \dots & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
0 & E_{I2}p_{I2}e^{\mathbf{z}_I^T\boldsymbol{\beta}} & \dots & 0 & 0 & E_{I2}p'_{I2}\lambda_2 e^{\mathbf{z}_I^T\boldsymbol{\beta}} & \dots & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & \dots & E_{1J}p_{1J}e^{\mathbf{z}_1^T\boldsymbol{\beta}} & 0 & 0 & \dots & E_{1J}p'_{1J}\lambda_J e^{\mathbf{z}_1^T\boldsymbol{\beta}} \\
0 & 0 & \dots & E_{2J}p_{2J}e^{\mathbf{z}_2^T\boldsymbol{\beta}} & 0 & 0 & \dots & E_{2J}p'_{2J}\lambda_J e^{\mathbf{z}_2^T\boldsymbol{\beta}} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
0 & 0 & \dots & E_{IJ}p_{IJ}e^{\mathbf{z}_I^T\boldsymbol{\beta}} & 0 & 0 & \dots & E_{IJ}p'_{IJ}\lambda_J e^{\mathbf{z}_I^T\boldsymbol{\beta}} \\
E_{11}e^{\mathbf{z}_1^T\boldsymbol{\beta}} & E_{12}e^{\mathbf{z}_1^T\boldsymbol{\beta}} & \dots & E_{1J}e^{\mathbf{z}_1^T\boldsymbol{\beta}} & 0 & 0 & \dots & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\
E_{I1}e^{\mathbf{z}_I^T\boldsymbol{\beta}} & E_{I2}e^{\mathbf{z}_I^T\boldsymbol{\beta}} & \dots & E_{IJ}e^{\mathbf{z}_I^T\boldsymbol{\beta}} & 0 & 0 & \dots & 0
\end{bmatrix}.$$

We can represent $\tilde{\mathbf{T}}$ and $\tilde{\mathbf{H}}$ as block matrices. We let $\mathbf{E}$ be the $I \times J$ matrix with $(i,j)^{\text{th}}$ elements E_{ij}, and we similarly define the matrix $\mathbf{C}$ to be in $\mathbb{R}^{I\times J}$ with its i,j element equal to c_{ij}. Furthermore, let $\mathbf{p}'_{ij} = p'_{ij}$. Let the matrix $\boldsymbol{\omega}$ be the diagonal matrix in $\mathbb{R}^{I\times I}$ with i,j elements $e^{\mathbf{z}_i^T\boldsymbol{\beta}}\mathbb{1}(i=j)$. Let $\mathbf{E}_{[:,j]} \odot \mathbf{C}_{[:,j]}$ be the element-wise multiplication between the two matrices $\mathbf{E}_{[:,j]}$ and $\mathbf{C}_{[:,j]}$. Let $\mathbf{Z} \in \mathbb{R}^{I\times K}$ with rows $\mathbf{Z}_{[i,:]} = \mathbf{z}_i^T$. Let $\mathbf{1}$ be the I-dimensional vector with each element equal to 1 and let

$$\begin{aligned}
\mathbf{D}_j &= \text{diag}(\mathbf{E}_{[:,j]} \odot \mathbf{C}_{[:,j]}) \\
\mathbf{D}'_j &= \text{diag}(\mathbf{E}_{[:,j]} \odot \mathbf{p}'_{[:,j]}) \\
\mathbf{F}_j &= \text{diag}(\mathbf{E}_{[:,j]}).
\end{aligned}$$

Let $\boldsymbol{\Omega}$ be the block matrix:

$$\boldsymbol{\Omega} = \begin{bmatrix}
\boldsymbol{\omega} & 0_{I\times I} & \dots & 0_{I\times I} & 0_{I\times I} & 0_{I\times I} \\
0_{I\times I} & \boldsymbol{\omega} & \dots & 0_{I\times I} & 0_{I\times I} & 0_{I\times I} \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\
0_{I\times I} & 0_{I\times I} & \dots & \boldsymbol{\omega} & 0_{I\times I} & 0_{I\times I} \\
0_{I\times I} & 0_{I\times I} & \dots & 0_{I\times I} & \boldsymbol{\omega} & 0_{I\times I} \\
0_{I\times I} & 0_{I\times I} & \dots & 0_{I\times I} & 0_{I\times I} & \boldsymbol{\omega}
\end{bmatrix} \tag{A.62}$$

Then we can write $\tilde{\mathbf{H}}$ as

$$\boldsymbol{\Omega}\begin{bmatrix} \mathbf{D}_1\mathbf{Z} & \mathbf{D}_1'\mathbf{Z}\lambda_1 \\ \mathbf{D}_2\mathbf{Z} & \mathbf{D}_2'\mathbf{Z}\lambda_2 \\ \vdots & \vdots \\ \mathbf{D}_{J-1}\mathbf{Z} & \mathbf{D}_{J-1}'\mathbf{Z}\lambda_{J-1} \\ \mathbf{D}_J\mathbf{Z} & \mathbf{D}_J'\mathbf{Z}\lambda_J \\ \left(\sum_j \lambda_j \mathbf{F}_j\right)\mathbf{Z} & 0_{I\times K} \end{bmatrix} \tag{A.63}$$

and $\tilde{\mathbf{T}}$ as

$$\boldsymbol{\Omega}\begin{bmatrix} \mathbf{D}_1\mathbf{1} & 0_{I\times 1} & \dots & 0_{I\times 1} & 0_{I\times 1} & \lambda_1\mathbf{D}_1'\mathbf{1} & 0_{I\times 1} & \dots & 0_{I\times 1} & 0_{I\times 1} \\ 0_{I\times 1} & \mathbf{D}_2\mathbf{1} & \dots & 0_{I\times 1} & 0_{I\times 1} & 0_{I\times 1} & \lambda_2\mathbf{D}_2'\mathbf{1} & \dots & 0_{I\times 1} & 0_{I\times 1} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0_{I\times 1} & 0_{I\times 1} & \dots & \mathbf{D}_{J-1}\mathbf{1} & 0_{I\times 1} & 0_{I\times 1} & 0_{I\times 1} & \dots & \lambda_{J-1}\mathbf{D}_{J-1}'\mathbf{1} & 0_{I\times 1} \\ 0_{I\times 1} & 0_{I\times 1} & \dots & 0_{I\times 1} & \mathbf{D}_J\mathbf{1} & 0_{I\times 1} & 0_{I\times 1} & \dots & 0_{I\times 1} & \lambda_J\mathbf{D}_J'\mathbf{1} \\ \mathbf{E}_{[:,1]} & \mathbf{E}_{[:,2]} & \dots & \mathbf{E}_{[:,J-1]} & \mathbf{E}_{[:,J]} & 0_{I\times 1} & 0_{I\times 1} & \dots & 0_{I\times 1} & 0_{I\times 1} \end{bmatrix} \tag{A.64}$$

Rearranging the columns of $\tilde{\Delta}$ to form $\tilde{\Delta}'$ does not change the rank of the matrix, and clarifies the conditions needed for the matrix to be full-rank:

$$\tilde{\Delta}' = \boldsymbol{\Omega}\begin{bmatrix} \mathbf{D}_1\mathbf{Z} & \mathbf{D}_1\mathbf{1} & 0_{I\times 1} & \dots & 0_{I\times 1} & 0_{I\times 1} & \lambda_1\mathbf{D}_1'\mathbf{Z} & 0_{I\times 1} & \lambda_1\mathbf{D}_1'\mathbf{1} & 0_{I\times 1} & \dots & 0_{I\times 1} \\ \mathbf{D}_2\mathbf{Z} & 0_{I\times 1} & \mathbf{D}_2\mathbf{1} & \dots & 0_{I\times 1} & 0_{I\times 1} & \lambda_2\mathbf{D}_2'\mathbf{Z} & 0_{I\times 1} & 0_{I\times 1} & \lambda_2\mathbf{D}_2'\mathbf{1} & \dots & 0_{I\times 1} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{D}_{J-1}\mathbf{Z} & 0_{I\times 1} & 0_{I\times 1} & \dots & \mathbf{D}_{J-1}\mathbf{1} & 0_{I\times 1} & \lambda_{J-1}\mathbf{D}_{J-1}'\mathbf{Z} & 0_{I\times 1} & 0_{I\times 1} & 0_{I\times 1} & \dots & \lambda_{J-1}\mathbf{D}_{J-1}'\mathbf{1} \\ \mathbf{D}_J\mathbf{Z} & 0_{I\times 1} & 0_{I\times 1} & \dots & 0_{I\times 1} & \mathbf{D}_J\mathbf{1} & \lambda_J\mathbf{D}_J'\mathbf{Z} & \lambda_J\mathbf{D}_J'\mathbf{1} & 0_{I\times 1} & 0_{I\times 1} & \dots & 0_{I\times 1} \\ \left(\sum_j \lambda_j \mathbf{F}_j\right)\mathbf{Z} & \mathbf{E}_{[:,1]} & \mathbf{E}_{[:,2]} & \dots & \mathbf{E}_{[:,J-1]} & \mathbf{E}_{[:,J]} & 0_{I\times K} & 0_{I\times 1} & 0_{I\times 1} & 0_{I\times 1} & \dots & 0_{I\times 1} \end{bmatrix} \tag{A.65}$$

is $\in \mathbb{R}^{(I(J+1))\times(2K+2J)}$. The first matrix $\boldsymbol{\Omega}$ is a $I(J+1)\times I(J+1)$ block diagonal matrix of diagonal matrices and by condition (S.f) is rank $I(J+1)$. Then the product in equation (A.65) is rank $2K+2J$ if the second matrix above is rank $2K+2J$ by Sylvester's rank inequality. The second matrix above is rank $2K+2J$ if the following three sub-blocks are full column rank:

$$\mathbf{L}_1 = \begin{bmatrix} \lambda_1\mathbf{D}_1'\mathbf{1} & 0_{I\times 1} & \dots & 0_{I\times 1} \\ 0_{I\times 1} & \lambda_2\mathbf{D}_2'\mathbf{1} & \dots & 0_{I\times 1} \\ \vdots & \vdots & \ddots & \vdots \\ 0_{I\times 1} & 0_{I\times 1} & \dots & \lambda_{J-1}\mathbf{D}_{J-1}'\mathbf{1} \end{bmatrix} \tag{A.66}$$

and

$$\mathbf{L}_2 = \begin{bmatrix} \left(\sum_j \lambda_j \mathbf{F}_j\right)\mathbf{Z} & \mathbf{E}_{[:,1]} & \mathbf{E}_{[:,2]} & \dots & \mathbf{E}_{[:,J-1]} & \mathbf{E}_{[:,J]} \end{bmatrix} \tag{A.67}$$

and

$$\mathbf{L}_3 = \begin{bmatrix} \lambda_J\mathbf{D}_J'\mathbf{Z} & \lambda_J\mathbf{D}_J'\mathbf{1} \end{bmatrix} \tag{A.68}$$

Sufficient conditions for $\mathbf{L}_1$ to be full column rank is

1. $I \geq (J-1)S$
2. $\text{diag}(\mathbf{E}_{[:,j]})\mathbf{1}$ is full column rank for all j

3. $\lambda_j > 0 \forall j \leq J$

4. $\eta_j \in (-\infty, \infty) \forall j \leq J$

5. $\beta_k, \gamma_k \in (-\infty, \infty) \forall k \leq J$

Sufficient conditions for $\mathbf{L}_2$ to be full column rank is

1. $I \geq K + J$

2. $\forall i \in 1, \ldots, I \sum_{j=1}^{J} E_{ij} > 0$

3. rank $\left(\begin{bmatrix} \mathbf{Z} & \mathbf{E} \end{bmatrix} \right) = J + K$

4. rank $\left(\begin{bmatrix} \mathbf{F}_1 \mathbf{Z} & \ldots & \mathbf{F}_J \mathbf{Z} & \mathbf{E}_{[:,1]} & \mathbf{E}_{[:,2]} & \ldots & \mathbf{E}_{[:,J-1]} & \mathbf{E}_{[:,J]} \end{bmatrix} \right) > J + K$

Sufficient conditions for $\mathbf{L}_3$ to be full column rank is

1. $I \geq K + S$

2. $\mathbf{F}_J \begin{bmatrix} \mathbf{Z} & \mathbf{1} \end{bmatrix}$ is full column rank

We recognize matrix $\mathbf{L}_2$ as the same matrix as in lemma 12 and that conditions (S.a) to (S.g) are a superset of the conditions (L.a) to (L.f) in lemma 12. Then by lemma 12, matrix $\mathbf{L}_2$ is rank $J + K$. Conditions (S.a) to (S.g) ensure that $\mathbf{L}_1$ and $\mathbf{L}_3$ are full rank as well, so rank$(\tilde{\Delta})$ is full rank.

Given that rank$(\tilde{\Delta}')$ = rank$(\tilde{\Delta})$ = rank$(\Delta) = 2J + 2K$, the column dimension of Δ, Cov$(\mathbf{f})$ is positive definite, and that the observational density, $\pi_\theta(\mathbf{f})$ is Poisson so $\frac{\partial}{\partial \theta_j} \int f_i \pi_\theta(\mathbf{x}) d\nu = \int f_i \frac{\partial}{\partial \theta_j} \pi_\theta(\mathbf{x}) d\nu$, we can apply lemma 10, which bounds the positive definiteness below by 0 for the Fisher information matrix $\mathcal{I}$. In other words, the Fisher Information matrix is positive definite. By Theorem 1 in Rothenberg (1971), model 2.10 is locally identifiable for any $(\mathbf{p}, \boldsymbol{\lambda}, \boldsymbol{\beta}, \boldsymbol{\gamma}) \in ((0,1)^J \times (0,\infty)^J \times \mathbb{R}^K \times \mathbb{R}^K)$, where $((a,b)^n)$ is the n-fold Cartesian product of the set (a,b). □

A.6 Further simulation study results

A.6.1 Root mean squared error plots

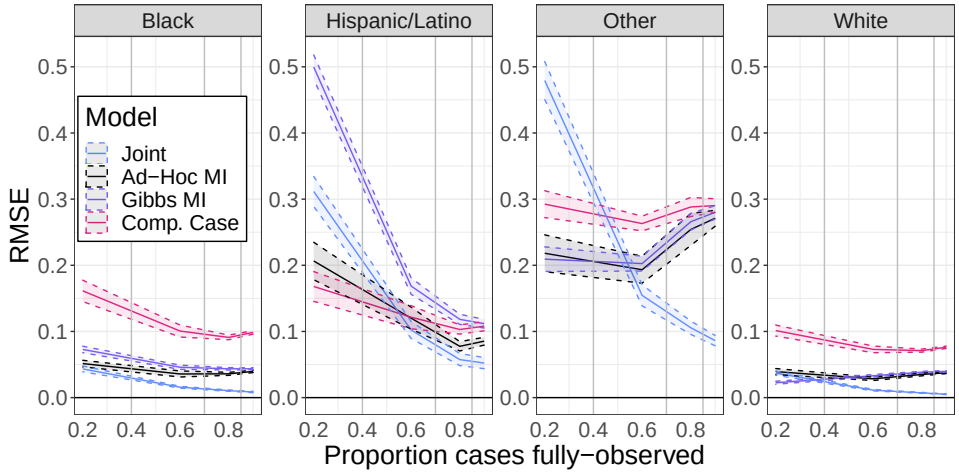


Figure A.2: Root mean squared error across simulated datasets for the standardized incidence ratio, or SIR_j for Blacks, Hispanic/Latinos, Others, and Whites plotted against the proportion of cases observed with race data. The blue color corresponds to the joint model in equation (2.11), while the red color corresponds to a the model defined in equation (2.15), or a complete-case analysis. Smaller magnitude is better.

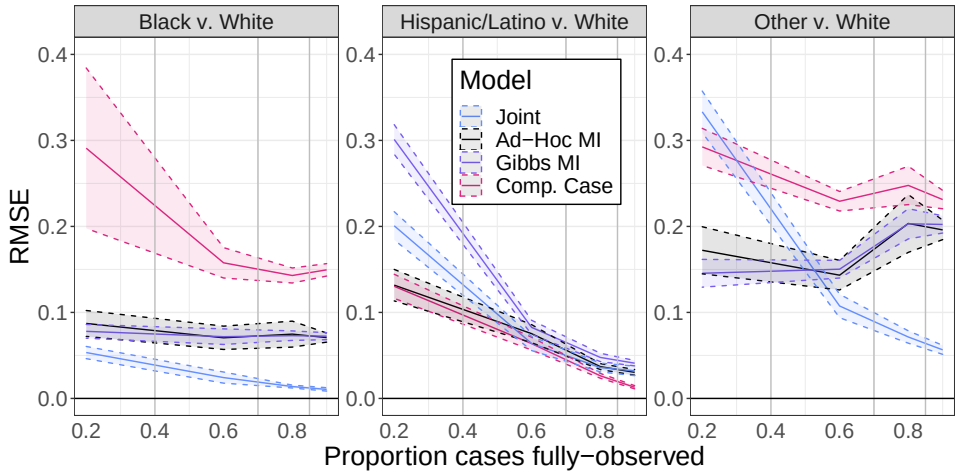


Figure A.3: Root mean squared error across simulated datasets for the relative risk ratio, or $\mathbb{I}_j/\mathbb{I}_J$ for Blacks, Hispanic/Latinos, Others relative to Whites plotted against the proportion of cases observed with race data. The blue color corresponds to the joint model in equation (2.11), while the red color corresponds to a the model defined in equation (2.15), or a complete-case analysis. Smaller magnitude is better.

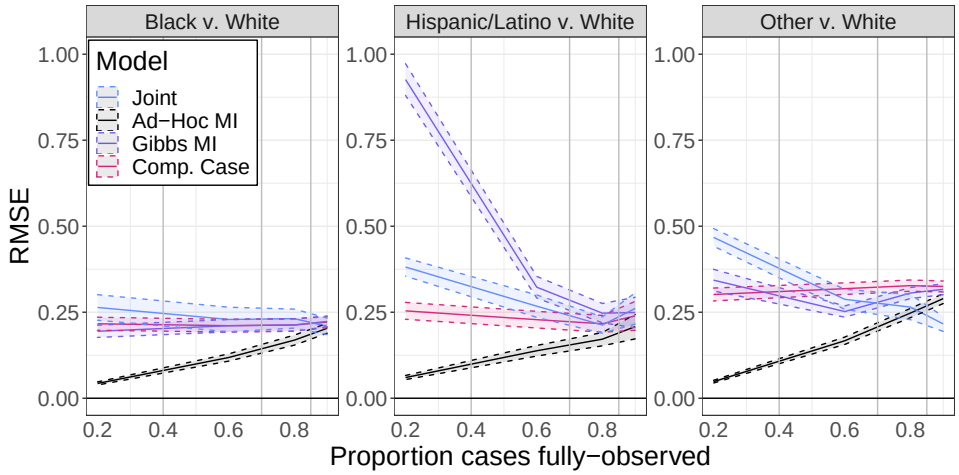


Figure A.4: Mean squared error across simulated datasets for the population relative risk ratio, or $\exp((\boldsymbol{\alpha}_\lambda)_j - (\boldsymbol{\alpha}_\lambda)_J)$ for Blacks, Hispanic/Latinos, Others relative to Whites plotted against the proportion of cases observed with race data. The blue color corresponds to the joint model in equation (2.11), while the red color corresponds to a the model defined in equation (2.15), or a complete-case analysis. Smaller magnitude is better.

A.6.2 80% posterior interval coverage

Table A.3: Table shows 80% posterior credible interval coverage and lengths for estimands of interest from the simulation study. Coverage proportion is calculated across 200 simulated datasets for each model/simulation scenario. Column headers for percentages (e.g. 20%) indicate the missing-data simulation scenario which corresponds to the statistic calculated in the table column; the simulation scenario corresponds to the proportion of cases observed with completely observed race covariates.

		80% interval coverage				80% mean interval length			
Parameter	Model	20%	60%	80%	90%	20%	60%	80%	90%
$\exp\left((\boldsymbol{\alpha}_\lambda)_{\text{Blacks}}\right.$	Complete Case	0.70	0.70	0.68	0.60	0.51	0.47	0.47	0.44
$\left.-(\boldsymbol{\alpha}_\lambda)_{\text{Whites}}\right)$	Joint	0.78	0.81	0.81	0.78	0.62	0.55	0.55	0.53
	MI-Ad-hoc	1.00	0.88	0.77	0.66	0.41	0.40	0.43	0.42
	MI-Gibbs	0.75	0.70	0.67	0.61	0.52	0.46	0.46	0.44
$\exp\left((\boldsymbol{\alpha}_\lambda)_{\text{Hispanics/Latinos}}\right.$	Complete Case	0.77	0.80	0.83	0.77	0.63	0.56	0.56	0.53
$\left.-(\boldsymbol{\alpha}_\lambda)_{\text{Whites}}\right)$	Joint	0.56	0.77	0.84	0.78	0.68	0.61	0.57	0.56
	MI-Ad-hoc	0.99	0.91	0.89	0.82	0.42	0.44	0.49	0.49
	MI-Gibbs	0.07	0.68	0.82	0.78	1.13	0.65	0.58	0.55
$\exp\left((\boldsymbol{\alpha}_\lambda)_{\text{Others}}\right.$	Complete Case	0.42	0.33	0.27	0.28	0.44	0.38	0.37	0.36
$\left.-(\boldsymbol{\alpha}_\lambda)_{\text{Whites}}\right)$	Joint	0.37	0.73	0.77	0.82	0.63	0.67	0.60	0.58
	MI-Ad-hoc	1.00	0.62	0.41	0.30	0.40	0.36	0.36	0.35

		80% interval coverage				80% mean interval length			
Parameter	Model	20%	60%	80%	90%	20%	60%	80%	90%
	MI-Gibbs	0.70	0.48	0.32	0.29	0.71	0.42	0.39	0.37
$\mathbb{SI}_{\text{Blacks}}$	Complete Case	0.06	0.04	0.01	0.00	0.03	0.02	0.01	0.01
	Joint	0.67	0.77	0.80	0.78	0.09	0.04	0.03	0.02
	MI-Ad-hoc	0.09	0.20	0.06	0.00	0.01	0.01	0.01	0.01
	MI-Gibbs	0.03	0.01	0.00	0.00	0.02	0.01	0.01	0.01
$\mathbb{SI}_{\text{Hispanics/Latinos}}$	Complete Case	0.34	0.24	0.12	0.01	0.13	0.07	0.06	0.06
	Joint	0.46	0.74	0.81	0.70	0.47	0.23	0.14	0.10
	MI-Ad-hoc	0.12	0.23	0.27	0.06	0.07	0.06	0.06	0.06
	MI-Gibbs	0.00	0.05	0.01	0.01	0.16	0.08	0.06	0.06
$\mathbb{SI}_{\text{Others}}$	Complete Case	0.03	0.00	0.00	0.00	0.11	0.06	0.05	0.05
	Joint	0.27	0.70	0.76	0.77	0.49	0.36	0.25	0.20
	MI-Ad-hoc	0.17	0.14	0.01	0.00	0.07	0.06	0.05	0.05
	MI-Gibbs	0.20	0.01	0.00	0.00	0.15	0.06	0.05	0.05
$\mathbb{SI}_{\text{Whites}}$	Complete Case	0.04	0.01	0.00	0.00	0.02	0.01	0.01	0.01
	Joint	0.50	0.80	0.82	0.84	0.06	0.03	0.02	0.01
	MI-Ad-hoc	0.14	0.17	0.00	0.00	0.01	0.01	0.01	0.01
	MI-Gibbs	0.21	0.01	0.00	0.00	0.02	0.01	0.01	0.01
$\mathbb{I}_{\text{Blacks}}/\mathbb{I}_{\text{Whites}}$	Complete Case	0.06	0.01	0.00	0.00	0.04	0.02	0.02	0.02
	Joint	0.73	0.79	0.81	0.81	0.12	0.05	0.04	0.03
	MI-Ad-hoc	0.10	0.20	0.02	0.00	0.02	0.02	0.02	0.02
	MI-Gibbs	0.12	0.01	0.00	0.00	0.03	0.02	0.02	0.02
$\mathbb{I}_{\text{Hispanics/Latinos}}/$	Complete Case	0.17	0.18	0.57	0.83	0.07	0.04	0.04	0.03
$\mathbb{I}_{\text{Whites}}$	Joint	0.43	0.73	0.81	0.73	0.28	0.14	0.09	0.06
	MI-Ad-hoc	0.12	0.21	0.32	0.41	0.04	0.04	0.04	0.03
	MI-Gibbs	0.00	0.14	0.24	0.17	0.10	0.05	0.04	0.04
$\mathbb{I}_{\text{Others}}/\mathbb{I}_{\text{Whites}}$	Complete Case	0.00	0.00	0.00	0.00	0.06	0.03	0.03	0.03
	Joint	0.26	0.72	0.74	0.76	0.32	0.24	0.17	0.13
	MI-Ad-hoc	0.14	0.11	0.01	0.00	0.05	0.04	0.03	0.03
	MI-Gibbs	0.21	0.00	0.00	0.00	0.10	0.04	0.03	0.03

Table A.4: Table shows 50% posterior credible interval coverage and lengths for estimands of interest from the simulation study. Coverage proportion is calculated across 200 simulated datasets for each model/simulation scenario. Column headers for percentages (e.g. 20%) indicate the missing-data simulation scenario which corresponds to the statistic calculated in the table column; the simulation scenario corresponds to the proportion of cases observed with completely observed race covariates.

		50% interval coverage				50% mean interval length			
Parameter	Model	20%	60%	80%	90%	20%	60%	80%	90%
$\exp\left((\boldsymbol{\alpha}_\lambda)_{\text{Blacks}}\right.$	Complete Case	0.39	0.35	0.41	0.32	0.27	0.24	0.24	0.23
$\left.-(\boldsymbol{\alpha}_\lambda)_{\text{Whites}}\right)$	Joint	0.52	0.53	0.49	0.47	0.32	0.28	0.29	0.27
	MI-Ad-hoc	0.96	0.56	0.47	0.30	0.21	0.21	0.22	0.22
	MI-Gibbs	0.48	0.38	0.38	0.29	0.27	0.24	0.24	0.23
$\exp\left((\boldsymbol{\alpha}_\lambda)_{\text{Hispanics/Latinos}}\right.$	Complete Case	0.48	0.47	0.47	0.44	0.33	0.29	0.29	0.28
$\left.-(\boldsymbol{\alpha}_\lambda)_{\text{Whites}}\right)$	Joint	0.27	0.48	0.47	0.43	0.35	0.31	0.30	0.29
	MI-Ad-hoc	0.96	0.62	0.53	0.47	0.22	0.23	0.25	0.25
	MI-Gibbs	0.01	0.39	0.46	0.45	0.58	0.33	0.30	0.28
$\exp\left((\boldsymbol{\alpha}_\lambda)_{\text{Others}}\right.$	Complete Case	0.20	0.10	0.12	0.09	0.23	0.20	0.19	0.19
$\left.-(\boldsymbol{\alpha}_\lambda)_{\text{Whites}}\right)$	Joint	0.17	0.41	0.47	0.51	0.32	0.35	0.31	0.30
	MI-Ad-hoc	0.91	0.29	0.15	0.10	0.21	0.18	0.19	0.18
	MI-Gibbs	0.38	0.23	0.12	0.10	0.36	0.22	0.20	0.19
$\mathbb{SI}_{\text{Blacks}}$	Complete Case	0.04	0.02	0.00	0.00	0.01	0.01	0.01	0.01
	Joint	0.39	0.49	0.48	0.49	0.04	0.02	0.01	0.01
	MI-Ad-hoc	0.04	0.09	0.03	0.00	0.01	0.01	0.01	0.01
	MI-Gibbs	0.03	0.01	0.00	0.00	0.01	0.01	0.01	0.01
$\mathbb{SI}_{\text{Hispanics/Latinos}}$	Complete Case	0.17	0.12	0.05	0.00	0.07	0.04	0.03	0.03
	Joint	0.27	0.47	0.51	0.41	0.24	0.12	0.07	0.05
	MI-Ad-hoc	0.07	0.15	0.18	0.01	0.04	0.03	0.03	0.03
	MI-Gibbs	0.00	0.01	0.01	0.00	0.08	0.04	0.03	0.03
$\mathbb{SI}_{\text{Others}}$	Complete Case	0.01	0.00	0.00	0.00	0.06	0.03	0.03	0.02
	Joint	0.12	0.45	0.48	0.41	0.25	0.18	0.13	0.10
	MI-Ad-hoc	0.08	0.07	0.01	0.00	0.04	0.03	0.03	0.03
	MI-Gibbs	0.09	0.00	0.00	0.00	0.08	0.03	0.03	0.02
$\mathbb{SI}_{\text{Whites}}$	Complete Case	0.03	0.01	0.00	0.00	0.01	0.01	0.00	0.00
	Joint	0.28	0.51	0.54	0.51	0.03	0.02	0.01	0.01
	MI-Ad-hoc	0.07	0.08	0.00	0.00	0.01	0.00	0.00	0.00
	MI-Gibbs	0.08	0.00	0.00	0.00	0.01	0.01	0.00	0.00

		50% interval coverage				50% mean interval length			
Parameter	Model	20%	60%	80%	90%	20%	60%	80%	90%
$\mathbb{I}_{\text{Blacks}}/\mathbb{I}_{\text{Whites}}$	Complete Case	0.03	0.01	0.00	0.00	0.02	0.01	0.01	0.01
	Joint	0.48	0.54	0.52	0.48	0.06	0.03	0.02	0.01
	MI-Ad-hoc	0.04	0.10	0.01	0.00	0.01	0.01	0.01	0.01
	MI-Gibbs	0.07	0.01	0.00	0.00	0.02	0.01	0.01	0.01
$\mathbb{I}_{\text{Hispanics/Latinos}}/$	Complete Case	0.09	0.08	0.33	0.53	0.03	0.02	0.02	0.02
$\mathbb{I}_{\text{Whites}}$	Joint	0.24	0.46	0.51	0.45	0.14	0.07	0.05	0.03
	MI-Ad-hoc	0.09	0.12	0.17	0.24	0.02	0.02	0.02	0.02
	MI-Gibbs	0.00	0.09	0.07	0.05	0.05	0.02	0.02	0.02
$\mathbb{I}_{\text{Others}}/\mathbb{I}_{\text{Whites}}$	Complete Case	0.00	0.00	0.00	0.00	0.03	0.02	0.02	0.01
	Joint	0.12	0.43	0.47	0.41	0.16	0.12	0.09	0.07
	MI-Ad-hoc	0.09	0.06	0.00	0.00	0.02	0.02	0.02	0.02
	MI-Gibbs	0.09	0.00	0.00	0.00	0.05	0.02	0.02	0.02

A.7 Prior sensitivity graphs

Graphs to support conclusions in 2.5.7

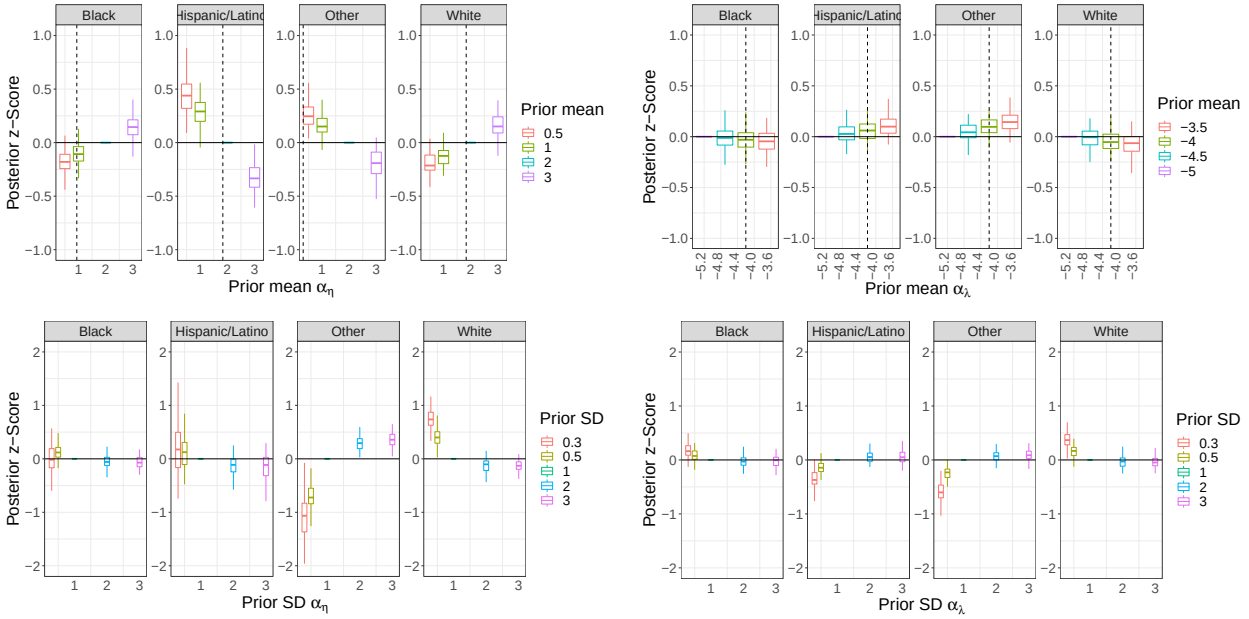


Figure A.5: Graphs above show box plots of posterior-standard-deviation-scaled differences in posterior mean incidences with respect to a baseline prior for various priors over population hyper-parameters, or $(\mathbb{E}_{\pi_a(\boldsymbol{\theta}|\text{Data})}[g(\boldsymbol{\theta})] - \mathbb{E}_{\pi_b(\boldsymbol{\theta}|\text{Data})}[g(\boldsymbol{\theta})])/\sqrt{\text{Var}_{\pi_b(\boldsymbol{\theta}|\text{Data})}(g(\boldsymbol{\theta}))}$. The graphs quantify how sensitive posterior mean incidence for each race/ethnicity group is to priors over population parameters $\boldsymbol{\alpha_\lambda}$ and $\boldsymbol{\alpha_\eta}$.

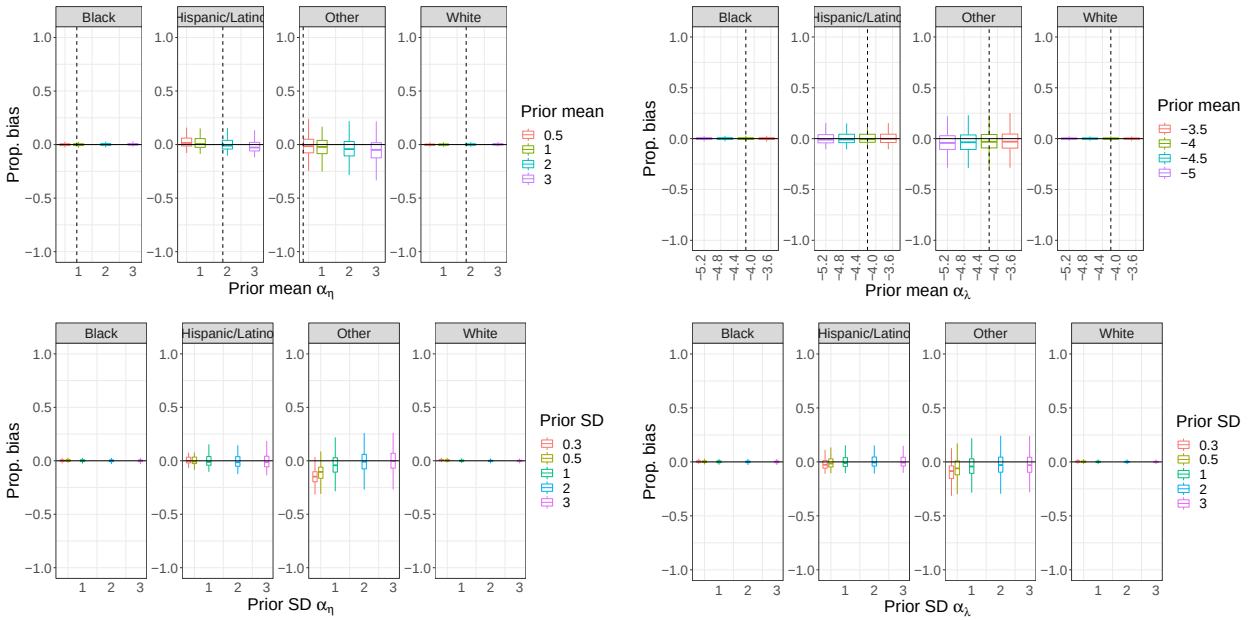


Figure A.6: Graphs above show box plots of scaled biases in the posterior mean for true incidences $g(\theta^\dagger)$, or $(\mathbb{E}_{\pi_a(\boldsymbol{\theta}|\text{Data})}[g(\boldsymbol{\theta})] - g(\boldsymbol{\theta}^\dagger))/g(\boldsymbol{\theta}^\dagger)$. The graphs quantify how priors over population parameters $\boldsymbol{\alpha_\lambda}$ and $\boldsymbol{\alpha_\eta}$ influence the bias of the posterior mean estimator.

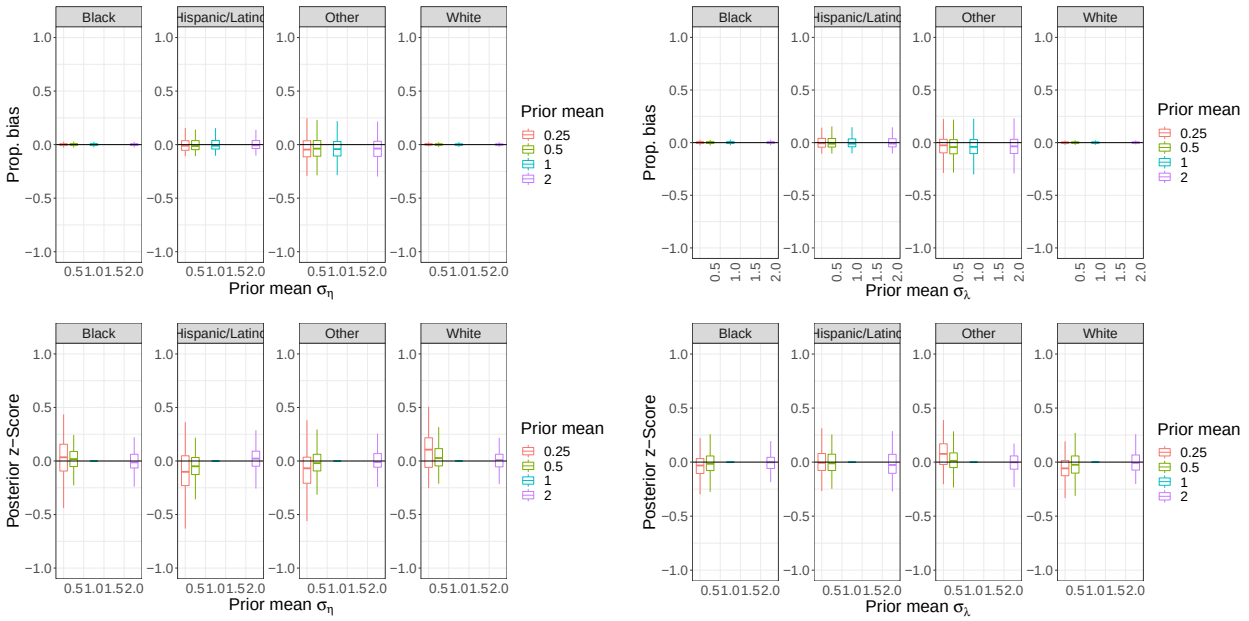


Figure A.7: The graphs above show the posterior bias (Equation (2.17)) and posterior z-scores (Equation (2.16)) for $\boldsymbol{\sigma_\lambda}$ and $\boldsymbol{\sigma_\eta}$

A.8 Further Wayne County applied data analysis results

A.8.1 Age-Race/Ethnicity posterior predictive checks

A.8.2 Rootogram

A.8.3 Tables for posterior summaries for estimands of interest

Table A.5: This table presents posterior summary statistics for the Wayne-County estimands of interest. Post. mean stands for Posterior Mean, and MCSE stands for Monte Carlo Standard Error, which is the standard error in the posterior estimator, which can be estimated assuming that the MCMC central limit theorem holds. See Betancourt and Girolami (2015) and Vehtari et al. (2020) for more details

		Post.		10% Post.		90% Post.	
Estimand	Model	Mean	MCSE	quant.	MCSE	quant.	MCSE
$\exp\left((\boldsymbol{\alpha}_\lambda)_{\text{Blacks}}\right.$	Joint	3.93	1.28e-02	3.16	1.40e-02	4.76	2.06e-02
$\left.-(\boldsymbol{\alpha}_\lambda)_{\text{Whites}}\right)$	Complete Case model	3.56	1.02e-02	2.89	1.09e-02	4.29	1.79e-02
	Ad-Hoc MI	2.96	1.80e-03	2.43	1.60e-03	3.52	2.15e-03
	Gibbs MI	3.49	1.75e-03	2.78	1.84e-03	4.24	2.59e-03
$\exp\left((\boldsymbol{\alpha}_\lambda)_{\text{Hispanics/Latinos}}\right.$	Joint	2.10	6.34e-03	1.65	6.39e-03	2.59	8.22e-03

Estimand	Model	Post. Mean	MCSE	10% Post. quant.	MCSE	90% Post. quant.	MCSE
$-(\boldsymbol{\alpha}_\lambda)_{\text{Whites}}\big)$	Complete Case model	1.93	4.35e-03	1.58	4.66e-03	2.29	8.52e-03
	Ad-Hoc MI	1.67	1.83e-03	1.35	1.72e-03	2.01	1.55e-03
	Gibbs MI	2.09	3.22e-03	1.65	2.62e-03	2.56	3.87e-03
$\exp\big((\boldsymbol{\alpha}_\lambda)_{\text{Others}}$	Joint	8.88	4.78e-02	6.09	5.51e-02	11.81	6.13e-02
$-(\boldsymbol{\alpha}_\lambda)_{\text{Whites}}\big)$	Complete Case model	5.21	1.30e-02	4.26	1.55e-02	6.22	1.91e-02
	Ad-Hoc MI	4.08	2.75e-03	3.32	2.52e-03	4.90	2.79e-03
	Gibbs MI	5.27	5.03e-03	4.18	4.14e-03	6.42	4.88e-03
$\exp\big((\boldsymbol{\alpha}_\lambda)_{\substack{\text{Asians/}\\\text{Pacific}\\\text{Islanders}}}$	Joint	1.35	4.79e-03	1.01	5.14e-03	1.73	7.37e-03
$-(\boldsymbol{\alpha}_\lambda)_{\text{Whites}}\big)$	Complete Case model	1.29	3.30e-03	1.00	3.59e-03	1.61	6.06e-03
	Ad-Hoc MI	1.21	1.98e-03	0.96	1.63e-03	1.48	2.43e-03
	Gibbs MI	1.66	4.44e-03	1.22	3.10e-03	2.15	6.17e-03
$\mathbb{SI}_{\text{Blacks}}$	Joint	1.59	6.62e-04	1.55	1.20e-03	1.64	7.33e-04
	Complete Case model	1.58	8.02e-05	1.57	1.44e-04	1.60	1.53e-04
	Ad-Hoc MI	1.56	1.47e-04	1.55	1.63e-04	1.57	1.37e-04
	Gibbs MI	1.60	1.65e-04	1.58	1.72e-04	1.61	1.65e-04
$\mathbb{SI}_{\text{Hispanics/Latinos}}$	Joint	1.16	1.59e-03	1.02	1.55e-03	1.30	2.20e-03
	Complete Case model	1.17	3.29e-04	1.12	6.04e-04	1.23	6.36e-04
	Ad-Hoc MI	1.15	7.02e-04	1.10	6.59e-04	1.20	7.41e-04
	Gibbs MI	1.23	9.89e-04	1.17	1.05e-03	1.28	1.01e-03
$\mathbb{SI}_{\text{Others}}$	Joint	4.64	1.81e-02	3.50	2.24e-02	5.64	1.38e-02
	Complete Case model	3.06	8.03e-04	2.93	1.33e-03	3.19	1.77e-03
	Ad-Hoc MI	2.68	1.21e-03	2.57	1.03e-03	2.80	1.38e-03
	Gibbs MI	3.15	2.40e-03	3.01	NA	3.28	2.69e-03
$\mathbb{SI}_{\substack{\text{Asians/}\\\text{Pacific}\\\text{Islanders}}}$	Joint	0.61	1.09e-03	0.53	4.95e-04	0.71	2.52e-03
	Complete Case model	0.65	3.27e-04	0.60	7.57e-04	0.71	7.64e-04
	Ad-Hoc MI	0.66	8.58e-04	0.61	8.63e-04	0.72	9.36e-04
	Gibbs MI	0.72	1.14e-03	0.66	NA	0.78	1.34e-03
$\mathbb{SI}_{\text{Whites}}$	Joint	0.46	4.25e-04	0.44	3.98e-04	0.49	4.65e-04
	Complete Case model	0.52	5.20e-05	0.52	7.63e-05	0.53	1.08e-04
	Ad-Hoc MI	0.55	1.00e-04	0.54	9.79e-05	0.56	1.02e-04
	Gibbs MI	0.50	9.26e-05	0.49	1.11e-04	0.51	8.00e-05
$\mathbb{I}_{\text{Blacks}}/\mathbb{I}_{\text{Whites}}$	Joint	3.01	2.86e-03	2.84	2.98e-03	3.19	3.31e-03

Estimand	Model	Post. Mean	MCSE	10% Post. quant.	MCSE	90% Post. quant.	MCSE
	Complete Case model	2.63	3.77e-04	2.57	5.53e-04	2.70	7.52e-04
	Ad-Hoc MI	2.49	6.53e-04	2.44	6.89e-04	2.55	6.51e-04
	Gibbs MI	2.79	7.12e-04	2.73	6.31e-04	2.85	8.37e-04
$\mathbb{I}_{\text{Hispanics/Latinos}}/$	Joint	1.69	3.26e-03	1.47	3.26e-03	1.92	4.36e-03
$\mathbb{I}_{\text{Whites}}$	Complete Case model	1.50	4.66e-04	1.42	9.17e-04	1.58	1.13e-03
	Ad-Hoc MI	1.42	9.31e-04	1.35	9.26e-04	1.49	9.86e-04
	Gibbs MI	1.66	1.43e-03	1.57	1.44e-03	1.74	1.50e-03
$\mathbb{I}_{\text{Others}}/\mathbb{I}_{\text{Whites}}$	Joint	6.55	2.96e-02	4.74	3.24e-02	8.12	2.06e-02
	Complete Case model	3.78	1.11e-03	3.60	2.05e-03	3.96	2.31e-03
	Ad-Hoc MI	3.20	1.67e-03	3.04	1.40e-03	3.35	1.90e-03
	Gibbs MI	4.10	3.36e-03	3.90	NA	4.30	4.10e-03
$\mathbb{I}_{\substack{\text{Asians/}\\\text{Pacific}\\\text{Islanders}}}/$	Joint	1.08	2.59e-03	0.92	1.60e-03	1.27	5.65e-03
$\mathbb{I}_{\text{Whites}}$	Complete Case model	1.01	5.26e-04	0.92	1.04e-03	1.10	1.02e-03
	Ad-Hoc MI	0.99	1.32e-03	0.91	NA	1.07	1.39e-03
	Gibbs MI	1.18	1.89e-03	1.08	NA	1.28	2.25e-03

Table A.6: The table shows sampling efficiency for population estimands of interest presented in table A.5. ESS stands for effective sample size; Bulk ESS and Tail ESS are measures of the equivalent number of independent samples generated from a MCMC procedure. See Vehtari et al. (2020) for more detail. Bulk and Tail ESS efficiency are the Bulk and Tail ESS figures divided by the total number of MCMC samples, which is $16,000$. As noted in Vehtari et al. (2020) MCMC samplers may generate Tail and Bulk ESS values greater than the total number of samples.

Estimand	Model	$\hat{R}$	Bulk ESS	Tail ESS	Bulk ESS eff.	Tail ESS eff.
$\exp\left((\boldsymbol{\alpha}_\lambda)_{\text{Blacks}}\right.$	Joint	1.00	2465	4753	0.15	0.30
$\left.-(\boldsymbol{\alpha}_\lambda)_{\text{Whites}}\right)$	Complete Case model	1.00	3037	5645	0.19	0.35
	Ad-Hoc MI	1.01	57036	223009	0.07	0.28
	Gibbs MI	1.00	114009	259368	0.14	0.32
$\exp\left((\boldsymbol{\alpha}_\lambda)_{\text{Hispanics/Latinos}}\right.$	Joint	1.00	3506	6994	0.22	0.44
$\left.-(\boldsymbol{\alpha}_\lambda)_{\text{Whites}}\right)$	Complete Case model	1.00	4440	6910	0.28	0.43
	Ad-Hoc MI	1.01	19965	91952	0.02	0.11
	Gibbs MI	1.02	12435	68983	0.02	0.09

Estimand	Model	$\hat{R}$	Bulk ESS	Tail ESS	Bulk ESS eff.	Tail ESS eff.
$\exp\left((\boldsymbol{\alpha}_\lambda)_{\text{Others}} - (\boldsymbol{\alpha}_\lambda)_{\text{Whites}}\right)$	Joint	1.00	2098	4315	0.13	0.27
	Complete Case model	1.00	3742	6528	0.23	0.41
	Ad-Hoc MI	1.01	51519	256045	0.06	0.32
	Gibbs MI	1.01	30320	220330	0.04	0.28
$\exp\left((\boldsymbol{\alpha}_\lambda)_{\substack{\text{Asians/}\\\text{Pacific}\\\text{Islanders}}} - (\boldsymbol{\alpha}_\lambda)_{\text{Whites}}\right)$	Joint	1.00	3850	7799	0.24	0.49
	Complete Case model	1.00	5511	8283	0.34	0.52
	Ad-Hoc MI	1.02	11078	45632	0.01	0.06
	Gibbs MI	1.03	7298	30332	0.01	0.04
$\mathbb{SI}_{\text{Blacks}}$	Joint	1.00	2531	4877	0.16	0.30
	Complete Case model	1.00	16204	14672	1.01	0.92
	Ad-Hoc MI	1.05	4322	15046	0.01	0.02
	Gibbs MI	1.06	3490	13738	0.00	0.02
$\mathbb{SI}_{\text{Hispanics/Latinos}}$	Joint	1.00	4523	9442	0.28	0.59
	Complete Case model	1.00	16596	12657	1.04	0.79
	Ad-Hoc MI	1.07	3392	14038	0.00	0.02
	Gibbs MI	1.12	1990		0.00	NA
$\mathbb{SI}_{\text{Others}}$	Joint	1.00	2034	4928	0.13	0.31
	Complete Case model	1.00	16338	11881	1.02	0.74
	Ad-Hoc MI	1.04	5584	18672	0.01	0.02
	Gibbs MI	1.12	1938		0.00	NA
$\mathbb{SI}_{\substack{\text{Asians/}\\\text{Pacific}\\\text{Islanders}}}$	Joint	1.00	6412	6513	0.40	0.41
	Complete Case model	1.00	17496	11891	1.09	0.74
	Ad-Hoc MI	1.10	2439		0.00	NA
	Gibbs MI	1.15	1673		0.00	NA
$\mathbb{SI}_{\text{Whites}}$	Joint	1.00	1722	5572	0.11	0.35
	Complete Case model	1.00	16277	13527	1.02	0.85
	Ad-Hoc MI	1.05	4083	17557	0.01	0.02
	Gibbs MI	1.05	4488		0.01	NA
$\mathbb{I}_{\text{Blacks}}/\mathbb{I}_{\text{Whites}}$	Joint	1.00	2197	5606	0.14	0.35
	Complete Case model	1.00	16184	13675	1.01	0.85
	Ad-Hoc MI	1.05	4225	16200	0.01	0.02
	Gibbs MI	1.05	4673		0.01	NA

Estimand	Model	$\hat{R}$	Bulk ESS	Tail ESS	Bulk ESS eff.	Tail ESS eff.
$\mathbb{I}_{\text{Hispanics/Latinos}}/$	Joint	1.00	2928	6147	0.18	0.38
$\mathbb{I}_{\text{Whites}}$	Complete Case model	1.00	16631	12370	1.04	0.77
	Ad-Hoc MI	1.06	3557	14586	0.00	0.02
	Gibbs MI	1.11	2144		0.00	NA
$\mathbb{I}_{\text{Others}}/\mathbb{I}_{\text{Whites}}$	Joint	1.00	1933	4788	0.12	0.30
	Complete Case model	1.00	16306	11978	1.02	0.75
	Ad-Hoc MI	1.04	5156		0.01	NA
	Gibbs MI	1.11	2114		0.00	NA
$\mathbb{I}_{\substack{\text{Asians/}\\ \text{Pacific}\\ \text{Islanders}}}/$	Joint	1.00	3827	6402	0.24	0.40
$\mathbb{I}_{\text{Whites}}$	Complete Case model	1.00	17423	11926	1.09	0.75
	Ad-Hoc MI	1.10	2447		0.00	NA
	Gibbs MI	1.14	1731		0.00	NA

Table A.7: The table shows the posterior means, 80% credible interval endpoints and the Monte Carlo standard errors of these estimates. CC stands for the complete-case model while J stands for the joint model.

	Post.		10% Post.		90% Post.	
Estimand	Mean	MCSE	quant.	MCSE	quant.	MCSE
$\mathbb{I}^{CC}_{Blacks}/\mathbb{I}^{J}_{Blacks}$	0.81	3.50e-04	0.79	3.55e-04	0.84	5.80e-04
$\mathbb{I}^{CC}_{Hispanics/Latinos}/\mathbb{I}^{J}_{Hispanics/Latinos}$	0.83	1.14e-03	0.73	1.14e-03	0.94	1.70e-03
$\mathbb{I}^{CC}_{Others}/\mathbb{I}^{J}_{Others}$	0.55	2.41e-03	0.44	1.36e-03	0.71	4.26e-03
$\mathbb{I}^{CC}_{\substack{Asians/\\Pacific\\Islanders}}/\mathbb{I}^{J}_{\substack{Asians/\\Pacific\\Islanders}}$	0.88	1.42e-03	0.73	2.50e-03	1.03	1.71e-03
$\mathbb{I}^{CC}_{Whites}/\mathbb{I}^{J}_{Whites}$	0.93	8.51e-04	0.88	7.85e-04	0.98	7.79e-04
$\mathbb{P}\left(\text{Race observed}\right)_{Blacks}$	0.85	7.55e-04	0.81	7.25e-04	0.91	1.35e-03
$\mathbb{P}\left(\text{Race observed}\right)_{Hispanics/Latinos}$	0.87	8.63e-04	0.78	1.43e-03	0.95	6.59e-04
$\mathbb{P}\left(\text{Race observed}\right)_{Others}$	0.58	2.93e-03	0.45	1.20e-03	0.77	5.31e-03
$\mathbb{P}\left(\text{Race observed}\right)_{\substack{Asians/\\Pacific\\Islanders}}$	0.90	8.25e-04	0.81	2.36e-03	0.97	3.80e-04
$\mathbb{P}\left(\text{Race observed}\right)_{Whites}$	0.94	7.09e-04	0.89	8.70e-04	0.98	4.33e-04

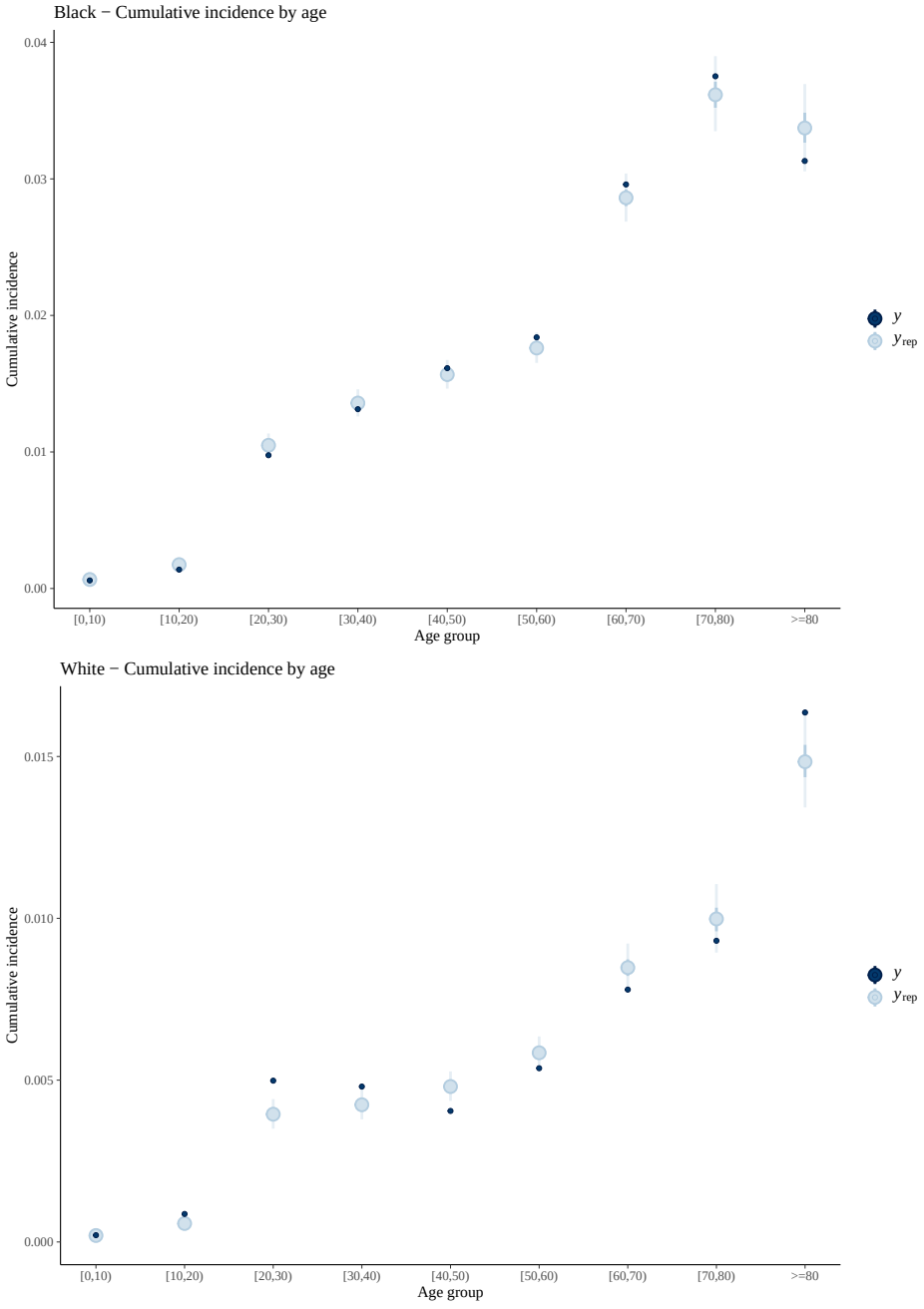


Figure A.8: Posterior predictive checks for cumulative incidence by age group by race for Blacks and Whites.

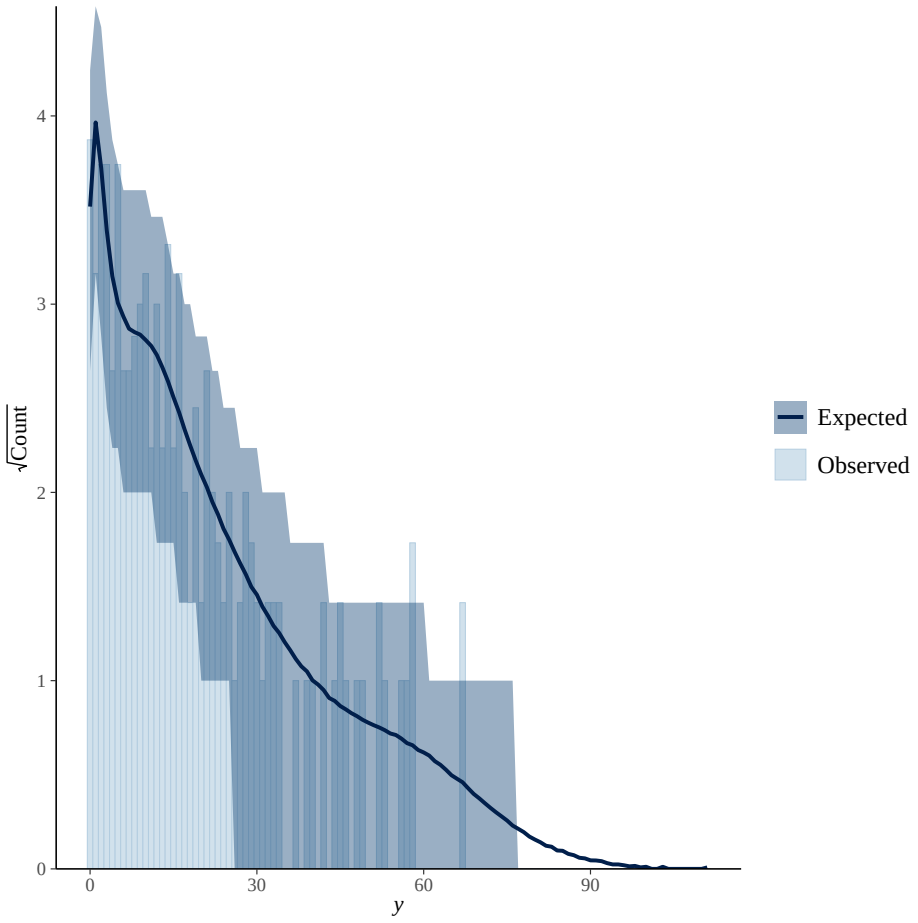


Figure A.9: Posterior predictive rootogram for missing case counts.

A.9 Stan code for binomial likelihood

The following Stan code computes the likelihood related to the following generative model using an efficient dynamic programming algorithm :

$$Y_{ij} \sim \text{Binomial}(E_{ij}, \theta_{ij})$$
$$X_{ij}|Y_{ij} \sim \text{Binomial}(Y_{ij}, p_{ij})$$

after marginalizing over all combinations of Y_{ij} such that $\sum_j Y_{ij} = T$ where T is the total identified cases of disease in stratum i, a known quantity.

The code was derived from Carpenter (2018).

```
functions {
  real binomial_2_lpmf(int y_obs, int y_miss,
                               real p, real theta, int E) {
    return binomial_lpmf(y_obs | y_miss, p)
           + binomial_lpmf(y_miss | E, theta);
  }
  real miss_lpmf(int[] y, int n_miss,
                 vector p, vector theta,
                 int[] E) {
    int N = rows(theta);
    real alpha[N + 1, n_miss + 1];

  // alpha[n + 1, tot + 1] = log p of tot missing cases
  // distributed among first n categories
    alpha[1, 1:(n_miss + 1)] = rep_array(0, n_miss+1);
    for (n in 1:N) {
      // tot = 0
      alpha[n + 1, 1] = alpha[n, 1]
          + binomial_2_lpmf(y[n]|y[n],p[n],theta[n], E[n]);

      // 0 < tot < n

        for (tot in 1:n_miss) {
          if (n > 1) {
            vector[tot + 1] vec;
```

```
                for (i in 1:(tot + 1)) {
                  vec[i] = alpha[n,i]
                  + binomial_2_lpmf(y[n] |
                      y[n] + tot - (i - 1),
                      p[n],theta[n], E[n]);
                }
                alpha[n + 1, tot + 1] = log_sum_exp(vec);
            } else {
                alpha[n + 1,tot + 1]
                 = binomial_2_lpmf(y[n]| y[n]
                 + tot,p[n],theta[n], E[n]);
            }
          }
      }
      return alpha[N + 1, n_miss + 1];
  }
}
```

A.10 Extensions to current model

A.10.1 Dynamic disease models

Suppose we collect information from multiple time periods, indexed by $1 \leq t \leq T$. Then our observations are of the form X_{itjg}, M_{itg}. Following the logic for dynamic infectious disease models outlined in Held and Paul (2012); Meyer and Held (2014); Bauer and Wakefield (2018); Wakefield et al. (2019), suppose that $w_{g,g'}$ are known weights, as in Bauer and Wakefield (2018) and let the observational model be of the form:

$$
\begin{aligned}
Y_{i1jg} &\sim \text{Poisson}(\lambda_{1jg} E_{ijg}) \\
X_{i1jg} | Y_{i1jg} &\sim \text{Binomial}(Y_{i1jg}, p_{1jg}) \\
Y_{itjg} | \mathbf{Y}_{t-1,g} &\sim \text{Poisson}(\lambda_{tjg} E_{ijg} + \beta_{tjg} \textstyle\sum_{j'} Y_{i,t-1,j',g} + \gamma_{tjg} \sum_{g'} w_{g,g'} \sum_{j',i'} Y_{i',t-1,j',g'}) \\
X_{itjg} | Y_{itjg} &\sim \text{Binomial}(Y_{itjg}, p_{tjg})
\end{aligned}
$$

which leads to an observational model of

$$\begin{aligned} X_{itjg}|\mathbf{Y}_{t-1,g} \sim \text{Poisson}(&p_{tjg}\lambda_{tjg}E_{ijg} + p_{tjg}\beta_{tjg}\textstyle\sum_{j'} Y_{i,t-1,j',g} \\ &+ p_{tjg}\gamma_{tjg}\textstyle\sum_{g'} w_{g,g'}\sum_{j',i'} Y_{i',t-1,j',g'}) \\ M_{itg}|\mathbf{Y}_{t-1,g} \sim \text{Poisson}(&\textstyle\sum_j(1-p_{tjg})\lambda_{tjg}E_{ijg} + (1-p_{tjg})\beta_{tjg}\sum_{j'} Y_{i,t-1,j',g} \\ &+ (1-p_{tjg})\gamma_{tjg}\textstyle\sum_{g'} w_{g,g'}\sum_{j',i'} Y_{i',t-1,j',g'}) \end{aligned}$$

Regressions of X_{itjg} on $E_{ijg}, \sum_{j'} Y_{i,t-1,j',g}$ stratified by j, t, g that include an intercept will yield unbiased estimators for $p_{tjg}\lambda_{tjg}, p_{tjg}\beta_{tjg}, p_{tjg}\gamma_{tjg}$ and regressions of M_{itg} on $\mathbf{E}_{itg}, \sum_{j'} Y_{i,t-1,j',g}$ with an intercept, will yield unbiased estimators of $(1-p_{t1g})\lambda_{t1g}, \ldots, (1-p_{t1g})\lambda_{tJg}, \sum_j(1-p_{tjg})\beta_{tjg}, \sum_j p_{tjg}\gamma_{tjg}$, provided all design matrices are full column rank. Using the same logic set out in theorem 1, the model is identifiable. Extensions to negative binomial likelihoods are also identifiable, and straightforward given the code in Appendix A.9. Note we can add stratum specific transmission effects as well provided the design matrix formed by combining all stratified regressions into a single regression model is full column rank.

A.10.2 Multivariate missing categorical data

Suppose we have observed the quintet $(U_n, C_n, R_n^C, D_n, R_n^D, S_n)$ for each person in a large population with total size E in a geographic area g, where U_n is an indicator for a positive test result for a given disease, C_n is a categorical variable with J levels coded $C_n \in \{1, \ldots, J\}$, R_n^C is a binary variable equal to 1 if C_n is observed, and 0 otherwise. D_n is a categorical variable with M levels coded $D_n \in \{1, \ldots, M\}$, R_n^D is a binary variable equal to 1 if D_n is observed, and 0 otherwise S_n is stratum information coded $S_n \in \{1, \ldots, I\}$. To make the problem more concrete, we will connect the notation to the problem of missing race and/or ethnicity data in COVID-19 case data. In the context of COVID-19 data, U_n is an indicator for a positive COVID-19 polymerase chain reaction (PCR) result, and C_n is a categorical variable encoding race information and D_n is a categorical variable encoding ethnicity information. S_n is additional information collected with each positive PCR test, like the patient sex at birth and age in years, while the geographic area could be a census tract, a zip code, or a larger area like a Public Use Microdata Area (hereafter referred to as PUMA).

Let the variable Y_{ijm} be the total cases within a stratum defined as all units for which $S_n = i$, $C_n = j$, and $D_n = m$, more explicitly,

$$Y_{ijm} = \sum_{\{n \,|\, S_n=i, C_n=j, D_n=m\}} U_n,$$

and assume that these cases are conditionally independent Poisson distributed random variables with rate μ_{ijm},

$$Y_{ijm}|\mu_{ijm} \sim \text{Poisson}(\mu_{ijm}).$$

Let the set of U_n be $\mathcal{U}$. Then we further assume that R_n^C are conditionally independent, Bernoulli distributed random variables:

$$R_n^C|p_{S_n,C_n,D_n}, \sim \text{Bernoulli}(p_{S_n,C_n,D_n}),$$

where p_{S_n,C_n,D_n} is equal to p_{ijm} if $S_n = i$ and $C_n = j$ and $D_n = m$.

Let $R_n^D|R_n^C$ be conditionally iid Bernoulli random variables:

$$R_n^D|R_n^C = r, q_{S_n,C_n,D_n}, \sim \text{Bernoulli}(q_{S_n,C_n,D_n,r}),$$

where $p_{S_n,C_n,D_n,r} = q_{ijmr}$

Then let the random vector $\vec{X}_{ijm}$ be the cases that are in stratum $S_n = i$, $C_n = j$ and $D_n = m$, summarized by their respective missingness pattern in (R_n^C, R_n^D):

$$\vec{X}_{ijm}|\mathcal{U} = \sum_{\{n \,|\, S_n=i, C_n=j, D_n=m, U_n=1\}} ((1-R_n^C)R_n^D, (1-R_n^C)(1-R_n^D), R_n^C R_n^D, R_n^C(1-R_n^D))$$

By the fact that R_n^C are conditionally independent Bernoulli distributed random variables and are identically distributed within a strata defined as $(S_n = i, C_n = j, D_n = m)$, and that $R_n^D|R_n^C = r$ are conditionally independent Bernoulli distributed random variables ID within a strata $(S_n = i, C_n = j, D_n = m, R_n^C = r)$ the following distributional equivalence holds:

$$\vec{X}_{ijm}|\mathcal{U} \stackrel{d}{=} \vec{X}_{ijm}|Y_{ijm},$$

and, furthermore,

$$\vec{X}_{ijm}|Y_{ijm}, p_{ijm}, q_{ijm0}, q_{ijm1} \sim \text{Multinomial}(Y_{ijm}, ((1-p_{ijm})q_{ijm0}, (1-p_{ijm})(1-q_{ijm0}), p_{ijm}q_{ijm1}, p_{ijm}(1-q_{ijm1})$$

Let W_{ij} be the cases for which we do not observe D_n, but for which we do observe C_n:

$$W_{ij}|\mathcal{U} = \sum_{m=1}^{M} (\vec{X}_{ijm})_4,$$

and let Z_{im} be the cases for which we do not observe C_n, but for which we do observe D_n:

$$Z_{im}|\mathcal{U} = \sum_{j=1}^{J} (\vec{X}_{ijm})_1,$$

and let T_i be the cases for which we observe neither C_n nor D_n:

$$T_i|\mathcal{U} = \sum_{j=1}^{J} (\vec{X}_{ijm})_2.$$

Given that Y_{ijm} are conditionally independent with means μ_{ijm}, then we can use results from Poisson process theory to show that

$$(\vec{X}_{ijm})_3 \sim \text{Poisson}(p_{ijm}q_{ijm1}\mu_{ijm}),$$

while

$$W_{ij} \sim \text{Poisson}(\sum_{m=1}^{M} p_{ijm}(1 - q_{ijm1})\mu_{ijm}),$$

and

$$Z_{im} \sim \text{Poisson}(\sum_{j=1}^{J} (1 - p_{ijm})q_{ijm0}\mu_{ijm}).$$

Finally,

$$T_i \sim \text{Poisson}(\sum_{j=1}^{J}\sum_{m=1}^{M} (1 - p_{ijm})(1 - q_{ijm0})\mu_{ijm}).$$

A.10.2.1 Modeling without stratum information

Suppose $M = 2$ and that we observe population counts

$$E_{ij1} = \sum_{n=1}^{E} \mathbb{1}_{S_n=i}\mathbb{1}_{C_n=j}\mathbb{1}_{D_n=1}$$

and

$$E_{i+2} = \sum_{j=1}^{J}\sum_{n=1}^{E} \mathbb{1}_{S_n=i}\mathbb{1}_{C_n=j}\mathbb{1}_{D_n=2}$$

If we assume that $\mu_{ij1} = \lambda_j E_{ij1}$, $\mu_{ij2} = \alpha E_{ij2} \forall j$ and that $p_{ij1} = p_j, \forall i$ and $p_{ij2} = \nu \forall j$, $q_{ij1r} = q_{jr}$ and $q_{ij2r} = \beta_r \forall j$, then the observed data model is:

$$\begin{aligned}
\left(\vec{X}_{ij1}\right)_3 &\sim \text{Poisson}\left(p_j q_{j1} \lambda_j E_{ij1}\right), \\
\sum_j \left(\vec{X}_{ij2}\right)_3 &\sim \text{Poisson}\left(\sum_j \nu \beta_1 \alpha E_{ij2}\right), \\
\sum_j W_{ij} &\sim \text{Poisson}\left(\sum_j p_j \left(1 - q_{j1}\right) \lambda_j E_{ij1} + \nu \left(1 - \beta_1\right) \alpha E_{ij2}\right), \\
Z_{i1} &\sim \text{Poisson}\left(\sum_{j=1}^{J} \left(1 - p_j\right) q_{j0} \lambda_j E_{ij1}\right), \\
Z_{i2} &\sim \text{Poisson}\left(\sum_{j=1}^{J} \left(1 - \nu\right) \beta_0 \alpha E_{ij2}\right),
\end{aligned} \tag{A.69}$$

which simplifies to

$$\begin{aligned}
\left(\vec{X}_{ij1}\right)_3 &\sim \text{Poisson}\left(p_j q_{j1} \lambda_j E_{ij1}\right), \\
\sum_j \left(\vec{X}_{ij2}\right)_3 &\sim \text{Poisson}\left(\nu \beta_1 \alpha E_{i+2}\right), \\
\sum_j W_{ij} &\sim \text{Poisson}\left(\sum_j p_j \left(1 - q_{j1}\right) \lambda_j E_{ij1} + \nu \left(1 - \beta_1\right) \alpha E_{i+2}\right), \\
Z_{i1} &\sim \text{Poisson}\left(\sum_{j=1}^{J} \left(1 - p_j\right) q_{j0} \lambda_j E_{ij1}\right), \\
Z_{i2} &\sim \text{Poisson}\left(\left(1 - \nu\right) \beta_0 \alpha E_{i+2}\right), \\
T_i &\sim \text{Poisson}\left(\sum_{j=1}^{J} \left(1 - p_j\right) \left(1 - q_{j0}\right) \lambda_j E_{ij1} + \left(1 - \nu\right) \left(1 - \beta_0\right) \alpha E_{i+2}\right).
\end{aligned} \tag{A.70}$$

This formulation explicitly throws away information, namely in the observed counts $(\vec{X}_{ij2})_3$ and W_{ij}, both of which are observed, but for which we do not have E_{ij2}.

The distinction between groups 1 and 2 is unnecessary, as can be seen by treating the group $M = 2$ as, say, group $j = J + 1$, and setting

$$\lambda_{J+1} = \alpha,\ p_{J+1} = \nu,\ q_{J+1,0} = \beta_0,\ q_{J+1,1} = \beta_1$$

and defining E_{ij} as $E_{ij} = E_{ij1}$ for $j \in [1, \ldots, J]$ and $E_{iJ+1} = E_{i+2}$. Finally, let $Z_i = Z_{i1} + Z_{i2}$

We can then rewrite the above equations more succinctly

$$\begin{aligned}
\left(\vec{X}_{ij1}\right)_3 &\sim \text{Poisson}\left(p_j q_{j1} \lambda_j E_{ij}\right), j \in [1, \ldots, J+1] \\
\sum_j W_{ij} &\sim \text{Poisson}\left(\sum_{j=1}^{J+1} p_j \left(1 - q_{j1}\right) \lambda_j E_{ij}\right), \\
Z_i &\sim \text{Poisson}\left(\sum_{j=1}^{J+1} \left(1 - p_j\right) q_{j0} \lambda_j E_{ij}\right), \\
T_i &\sim \text{Poisson}\left(\sum_{j=1}^{J+1} \left(1 - p_j\right) \left(1 - q_{j0}\right) \lambda_j E_{ij}\right).
\end{aligned} \tag{A.71}$$

We can design consistent estimators for λ_j for $j \in [1, \ldots, J+1]$, provided the $I \times J + 1$ matrix of population counts $\mathbf{E}$ with $(i,j)^{\text{th}}$ element E_{ij1} for all i and $j \in [1, \ldots, J]$ and $(i,j)^{\text{th}}$ element E_{i+2} for all i but $j = J+1$.

One wonders if the better model would be

$$\begin{aligned}
\vec{\mathcal{E}}_{i2} &\sim \text{Multinomial}(E_{i+2} | \vec{\phi}_i), \\
(\vec{X}_{ij1})_3 &\sim \text{Poisson}(p_j q_{j1} \lambda_j E_{ij1}), \\
(\vec{X}_{ij2})_3 &\sim \text{Poisson}(\nu \beta_1 \alpha (\vec{\mathcal{E}}_{i2})_j), \\
W_{ij} &\sim \text{Poisson}(p_j (1 - q_{j1}) \lambda_j E_{ij1} + \nu (1 - \beta_1) \alpha (\vec{\mathcal{E}}_{i2})_j), \\
Z_{i1} &\sim \text{Poisson}(\sum_{j=1}^{J} (1 - p_j) q_{j0} \lambda_j E_{ij1}), \\
Z_{i2} &\sim \text{Poisson}((1 - \nu) \beta_0 \alpha \sum_{j=1}^{J} E_{ij2}), \\
T_i &\sim \text{Poisson}(\sum_{j=1}^{J} (1 - p_j)(1 - q_{j0}) \lambda_j E_{ij1} + (1 - \nu)(1 - \beta_0) \alpha \sum_{j=1}^{J} E_{ij2}).
\end{aligned} \tag{A.72}$$

though the increase in dimensionality is quite staggering, as would be the integration over the space

of $\vec{\mathcal{E}}_{i2}$ Another option might be

$$\begin{aligned}
\vec{\mathbb{E}}_{i2} &\sim \text{Multinomial}(E_{i+2}|\vec{\phi}_i),\\
(\vec{X}_{ij1})_3 &\sim \text{Poisson}(p_j q_{j1}\lambda_j E_{ij1}),\\
(\vec{X}_{ij2})_3 &\sim \text{Poisson}(\nu\beta_1\alpha(\vec{\phi}_i)_j E_{i+2}),\\
W_{ij} &\sim \text{Poisson}(\sum_j p_j(1-q_{j1})\lambda_j E_{ij1} + \nu(1-\beta_1)\alpha(\vec{\phi}_i)_j E_{i+2}),\\
Z_{i1} &\sim \text{Poisson}(\sum_{j=1}^{J}(1-p_j)q_{j0}\lambda_j E_{ij1}),\\
Z_{i2} &\sim \text{Poisson}((1-\nu)\beta_0\alpha\sum_{j=1}^{J} E_{ij2}),\\
T_i &\sim \text{Poisson}(\sum_{j=1}^{J}(1-p_j)(1-q_{j0})\lambda_j E_{ij1} + (1-\nu)(1-\beta_0)\alpha\sum_{j=1}^{J} E_{ij2}).
\end{aligned} \tag{A.73}$$

where $\vec{\mathbb{E}}_{i2}$ are observed in national surveys subsequent to the Decennial census.

A.10.3 Multivariate Poisson model

A key assumption of all models is that latent disease cases Y_{ij} and $Y_{ij'}$ are conditionally independent given the data-generating parameters λ_j. However, this may not hold. To that end, we can induce positive dependence between the latent disease cases via an unobserved shared Poisson random variable.

A.10.3.1 Preliminaries

A useful expression to turn infinite series into finite series is Dobiński's formula Weisstein:

$$\sum_{\ell=1}^{n} S(n,\ell)\lambda^{\ell} = e^{-\lambda}\sum_{k=1}^{\infty}\frac{k^n}{k!}\lambda^k \tag{A.74}$$

where $S(n,k)$ is Stirling number of the second kind, which measures the number of ways of partitioning an n-element set into k non-empty subsets. The simple recurrence relation

$$S(n,k) = S(n-1,k-1) + kS(n-1,k)$$

when used with dynamic programming with cells can yield fast calculation.

A.10.3.2 Alternative generative model

Let $\lambda_j, \beta_j > 0$ for all j and $\lambda_0 > 0$.

$$\begin{aligned} Y_{i0} &\sim \text{Poisson}(\lambda_0) \\ Y_{ij}|Y_{i0} &\sim \text{Poisson}(E_{ij}\lambda_j + \beta_j Y_{i0}) \ \forall j \in \{1, \ldots, J\} \\ X_{ij}|Y_{ij} &\sim \text{Binomial}(Y_{ij}, p_j) \end{aligned}$$

Leads to the observed data model, with $M_i|Y_{i0} = \sum_j Y_{ij}|Y_{i0} - X_{ij}|Y_{i0}$,:

$$\begin{aligned} Y_{i0} &\sim \text{Poisson}(\lambda_0) \\ X_{ij}|Y_{i0} &\sim \text{Poisson}(p_j(E_{ij}\lambda_j + \beta_j Y_{i0})) \\ M_i|Y_{i0} &\sim \text{Poisson}(\sum_j E_{ij}(1-p_j)\lambda_j + \sum_j (1-p_j)\beta_j Y_{i0})) \end{aligned}$$

It turns out that the joint marginal distribution for $\{X_{ij}, M_i\}$ has an analytic, finite series representation: We derive the distribution for $J = 2$: Let $p_j\lambda_j = v_j$, $(1-p_j)\lambda_j = u_j$, and let $p_j\beta_j = b_j$, $(1-p_j)\beta_j = c_j$. Then the marginal distribution $p_{X_{i1},X_{i2},M_i}(x_1, x_2, m)$ is :

$$\frac{e^{-(E_{i1}(v_1+u_1)+E_{i2}(v_2+u_2)+\lambda_0)}}{x_1!x_2!m!} \sum_{k=0}^{\infty} (E_{i1}v_1 + b_1 k)^{x_1} (E_{i2}v_2 + b_2 k)^{x_2} (\sum_j E_{ij}u_j + k\sum_j c_j)^m \frac{(\lambda_0 e^{-(b_1+c_1+b_2+c_2)})^k}{k!}$$

$$= \frac{e^{-(E_{i1}(v_1+u_1)+E_{i2}(v_2+u_2)+\lambda_0)}}{x_1!x_2!m!} \sum_{k=0}^{\infty} \sum_{n=0}^{x_1} \binom{x_1}{n} (E_{i1}v_1)^{x_1-n} (b_1 k)^n \sum_{\ell=0}^{x_2} \binom{x_2}{\ell} (E_{i2}v_2)^{x_2-\ell} (b_2 k)^\ell$$
$$\sum_{q=0}^{m} \binom{m}{q} (\sum_j E_{ij}u_j)^{m-q} (k\sum_j c_j)^q \frac{(\lambda_0 e^{-(b_1+c_1+b_2+c_2)})^k}{k!}$$

$$= \frac{e^{-(E_{i1}(v_1+u_1)+E_{i2}(v_2+u_2)+\lambda_0)}}{x_1!x_2!m!} \sum_{n=0}^{x_1} \binom{x_1}{n} (E_{i1}v_1)^{x_1-n} (b_1)^n \sum_{\ell=0}^{x_2} \binom{x_2}{\ell} (E_{i2}v_2)^{x_2-\ell} (b_2)^\ell$$
$$\sum_{q=0}^{m} \binom{m}{q} (\sum_j E_{ij}u_j)^{m-q} (\sum_j c_j)^q \sum_{k=0}^{\infty} \frac{(\lambda_0 e^{-(b_1+c_1+b_2+c_2)})^k}{k!} k^{n+\ell+q}$$

$$= \frac{e^{-(E_{i1}(v_1+u_1)+E_{i2}(v_2+u_2)+\lambda_0)+\lambda_0 e^{-(b_1+c_1+b_2+c_2)}}}{x_1!x_2!m!} \sum_{n=0}^{x_1} \binom{x_1}{n} (E_{i1}v_1)^{x_1-n} (b_1)^n \sum_{\ell=0}^{x_2} \binom{x_2}{\ell} (E_{i2}v_2)^{x_2-\ell} (b_2)^\ell$$
$$\sum_{q=0}^{m} \binom{m}{q} (\sum_j E_{ij}u_j)^{m-q} (\sum_j c_j)^q \sum_{r=1}^{n+\ell+q} (\lambda_0 e^{-(b_1+c_1+b_2+c_2)})^r S(n+\ell+q, r)$$

A.10.3.3 Identifiability

In addition to the benefits of being amenable to computation, the parameters are identifiable because we can design consistent estimators for the parameters. The proof that consistency implies identifiability is shown on page 57 of Lehmann and Casella (1998), but the short version is that identifiability is a necessary condition for consistency.

Let our observations be $(x_{i1}, \ldots, x_{iJ}, m_i)$ for $i \in \{1, \ldots, I\}$, and let $\mathbf{E}_j$ be the matrix where the first column is an I-vector of 1s and the second column is $\mathbf{e}_j$, with the i^{th} element is E_{ij}. Define $\mathbf{x}_j$ the same way.

$$\mathbf{E}_j^\dagger \mathbf{x}_j = \hat{\boldsymbol{\theta}}_j \tag{A.75}$$

Let $\mathbf{E}$ be the $J+1$ column matrix where the first column is the I-vector of 1s and the last J columns are $\mathbf{e}_j$.

$$\mathbf{E}^\dagger \mathbf{m} = \hat{\boldsymbol{\phi}} \tag{A.76}$$

where $\mathbf{A}^\dagger$ is the Moore-Penrose inverse of a matrix $\mathbf{A}$.

By the consistency of least squares $\hat{\boldsymbol{\phi}}[j+1] \xrightarrow{p} u_j$ and $\hat{\boldsymbol{\theta}}_j[2] \xrightarrow{p} v_j$ so by the continuous mapping theorem

$$\hat{\boldsymbol{\theta}}_j[2] + \hat{\boldsymbol{\phi}}[j+1] \xrightarrow{p} \lambda_j \tag{A.77}$$

Then again by continuous mapping:

$$\frac{\hat{\boldsymbol{\theta}}_j[2]}{\hat{\boldsymbol{\theta}}_j[2] + \hat{\boldsymbol{\phi}}[j+1]} \xrightarrow{p} p_j \tag{A.78}$$

By the WLLN

$$\text{Cov}\left(\mathbf{x}_j, \mathbf{x}_k\right) \xrightarrow{p} b_j b_k \lambda_0 \tag{A.79}$$

and again by consistency of the least squares estimator

$$\hat{\boldsymbol{\theta}}_j[1] \xrightarrow{p} b_j \lambda_0 \tag{A.80}$$

so by the continuous mapping theorem

$$\frac{\text{Cov}\left(\mathbf{x}_j, \mathbf{x}_k\right)}{\hat{\boldsymbol{\theta}}_k[1] \frac{\hat{\boldsymbol{\theta}}_j[2]}{\hat{\boldsymbol{\theta}}_j[2] + \hat{\boldsymbol{\phi}}[j+1]}} \xrightarrow{p} \beta_j \tag{A.81}$$

and

$$\frac{\text{Cov}\left(\mathbf{x}_j, \mathbf{x}_k\right)}{\hat{\boldsymbol{\theta}}_j[1] \hat{\boldsymbol{\theta}}_k[1]} \xrightarrow{p} \frac{1}{\lambda_0} \tag{A.82}$$

In order to get incidence, we use the following estimator:

$$(\hat{\boldsymbol{\theta}}_j[2] + \hat{\boldsymbol{\phi}}[j+1]) \left(1 + \frac{\hat{\boldsymbol{\theta}}_j[1]/\hat{\boldsymbol{\theta}}_j[2]}{\sum_i E_{ij}}\right) \tag{A.83}$$

A.10.3.4 Properties of the multivariate Poisson

Let $E_{ij} = E_{ik} = 1$. The covariance for Y_{ij} and Y_{ik} is $\beta_j \beta_k \lambda_0$ while the variance is

$$\begin{aligned}
\text{Var}\left(Y_{ij}\right) &= \text{Var}\left(\mathbb{E}\left[Y_{ij}|Y_{i0}\right]\right) + \mathbb{E}\left[\text{Var}\left(Y_{ij}|Y_{i0}\right)\right] && \text{(A.84)}\\
&= \text{Var}\left(\lambda_j + \beta_j k\right) + \mathbb{E}\left[\lambda_j + \beta_j k\right] && \text{(A.85)}\\
&= \beta_j^2 \lambda_0 + \lambda_j + \beta_j \lambda_0 && \text{(A.86)}
\end{aligned}$$

Then the correlation between Y_{ij} and Y_{ik} is

$$c(\beta_j, \beta_k, \lambda_j, \lambda_k, \lambda_0) = \frac{\beta_j \beta_k \lambda_0}{\sqrt{\beta_j^2 \lambda_0 + \lambda_j + \beta_j \lambda_0}\sqrt{\beta_k^2 \lambda_0 + \lambda_k + \beta_k \lambda_0}} \tag{A.87}$$

When β_j or β_k is zero, the correlation is zero.

$$\lim_{\beta_j \to \infty} \lim_{\beta_k \to \infty} c(\beta_j, \beta_k, \lambda_j, \lambda_k, \lambda_0) \to 1 \tag{A.88}$$

A.10.4 Asymptotic bias when population observed with error

Suppose we have the following generative model:

Let $M_i|(E_{i1}, \ldots, E_{iJ}) \sim \text{Pois}(\sum_j E_{ij} u_j^\star)$ Suppose we cannot observe E_{ij} but instead observe:

$$\tilde{E}_{ij} \overset{\text{iid}}{\sim} \text{Poisson}(\mu_j)$$

such that

$$\mathbb{E}\left[E_{ij}|\tilde{E}_{ij}\right] = g_j \tilde{E}_{ij}$$

for $g_j > 0$ What happens when we fit the model

$$M_i \sim \text{Pois}(\textstyle\sum_j \tilde{E}_{ij} u_j)$$

The likelihood is:

$$\textstyle\sum_I m_i \log(\sum_j \tilde{E}_{ij} u_j) - \sum_j \tilde{E}_{ij} u_j \tag{A.89}$$

The MLEs, $\hat{u}_j^{(I)}$ will be the MLEs of the equation

$$\frac{1}{I}\sum_I m_i \log(\sum_j \tilde{E}_{ij} u_j) - \frac{1}{I}\sum_i \sum_j \tilde{E}_{ij} u_j \tag{A.90}$$

which converges a.s. as $I \to \infty$ to

$$\mathbb{E}\left[m_i \log(\sum_j \tilde{E}_{ij} u_j) - \sum_j \tilde{E}_{ij} u_j\right] \tag{A.91}$$

or

$$\mathbb{E}\left[\mathbb{E}\left[m_i | \tilde{\mathbf{e}}_i\right] \log(\sum_j \tilde{E}_{ij} u_j) - \sum_j \tilde{E}_{ij} u_j\right] \tag{A.92}$$

where $\tilde{\mathbf{e}}_i$ is the J-length vector with j^{th} element $\tilde{E}_{ij}$. Then using the generative model for M_i above:

$$\mathbb{E}\left[\sum_j \tilde{E}_{ij} g_j u_j^\star \log(\sum_j \tilde{E}_{ij} u_j) - \sum_j \tilde{E}_{ij} u_j\right] \tag{A.93}$$

Then the set of $\hat{u}_j^{(I)}$, the sequence of MLEs indexed by I for eq. (A.89), converges a.s. to the set of u_j that maximize eq. (A.93).

$$\sum_j \tilde{E}_{ij} g_j u_j^\star \log(\sum_j \tilde{E}_{ij} u_j) - \sum_j \tilde{E}_{ij} u_j \tag{A.94}$$

We note that the solution

$$u_j = g_j u_j^\star \tag{A.95}$$

maximizes eq. (A.97) for all realizations $\tilde{E}_{ij}$, namely:

$$\sum_j \tilde{E}_{ij} g_j u_j^\star \log(\sum_j \tilde{E}_{ij} u_j) - \sum_j \tilde{E}_{ij} u_j \geq \sum_j \tilde{E}_{ij} g_j u_j^\star \log(\sum_j \tilde{E}_{ij} u'_j) - \sum_j \tilde{E}_{ij} u'_j \tag{A.96}$$

for all $u'_j \neq g_j u_j^\star$. See Wooldridge (1999) for more information.: Monotonicity of the expectation operator shows that eq. (A.95) maximizes eq. (A.93) as well.

$$\mathbb{E}\left[\sum_j \tilde{E}_{ij} g_j u_j^\star \log(\sum_j \tilde{E}_{ij} u_j) - \sum_j \tilde{E}_{ij} u_j\right] \geq \mathbb{E}\left[\sum_j \tilde{E}_{ij} g_j u_j^\star \log(\sum_j \tilde{E}_{ij} u'_j) - \sum_j \tilde{E}_{ij} u'_j\right] \tag{A.97}$$

so

$$\hat{u}_j^{(I)} \overset{\text{a.s.}}{\to} g_j u_j^\star \quad \forall j \in \{1, \ldots, J\}$$

This implies the following relation:

$$\frac{\hat{u}_j}{\hat{u}_k} \xrightarrow{p} \frac{g_j u_j}{g_k u_k} \tag{A.98}$$

so

$$\frac{g_k \hat{u}_j}{g_j \hat{u}_k} \xrightarrow{p} \frac{u_j}{u_k} \tag{A.99}$$

APPENDIX B

VE appendix

We define our notation for principal stratification in vaccine efficacy (VE) in section B.1. In section B.5, we give general properties of the Kruskal rank, and extensions to Kruskal (1977) theorems that we derived. We apply these extensions in the context of principle stratification for VE in section B.2. The proof of our main result, Theorem 7, is given in section B.4. These proofs are based on results in Appendix B.5 and Appendix B.2.

B.1 Notation and definitions

In the following proofs, we have omitted the subscript i from random variables to simplify our notation. We have also elided conditioning on $X_i = x$; the proofs shown in Appendix B.4 are understood to be conditional on $X_i = x$. Let z be the N_z-category discrete variable taking values in the set $\{z_1, \ldots, z_{N_z}\}$ representing treatment, and let Z be treatment assignment. The principal stratum, S^{P_0} is defined as $(S(z_1), \ldots, S(z_{N_z}))$, $S(z_j) \in \{0, 1\}$. Let $\mathcal{S}$ be the set of principal strata, which is equal to $\{0, 1\}^{N_z}$ when there are no monotonicity assumptions; let $u \in \mathcal{S}$.

Let the set of treatments be $\{z_1, \ldots, z_{N_z}\}$, with $z \in \{z_1, \ldots, z_{N_z}\}$

Let A have N_a levels and take values in the set $\{1, \ldots, N_a\}$. Let $P(A \mid R)$ be the $N_a \times N_r$ matrix with $(i, j)^{\text{th}}$ element equal to $P(A = i \mid R = j)$. Let $P_{N_z}(A \mid S^{P_0})$ be the $N_a \times 2^{N_z}$ matrix with $(i, j)^{\text{th}}$ element equal to $P(A = i \mid S^{P_0} = \varpi_{N_z}(j - 1))$, and let $P_{N_z}(S^{P_0} \mid R)$ be the $2^{N_z} \times N_r$ matrix with $(i, j)^{\text{th}}$ element equal to $P(S^{P_0} = \varpi_{N_z}(i - 1) \mid R = j)$. Let $P(y \mid R, Z = z)$ be the $1 \times R$ matrix with element $(1, j)^{\text{th}}$ equal to $P(y \mid R = j, Z = z)$, and similarly let $P_{N_z}(y \mid S^{P_0}, Z = z)$ be the 1×2^{N_z} matrix with element $(1, j)^{\text{th}}$ equal to $P(y \mid S^{P_0} = \varpi_{N_z}(j - 1), Z = z)$. Let $P(y \mid R, Z = z, A = k)$ be the $1 \times R$ matrix with $(1, j)^{\text{th}}$ element equal to $P(y \mid R = j, Z = z, A = k)$, and similarly let $P_{N_z}(y \mid S^{P_0}, Z = z, A = k)$ be the 1×2^{N_z} matrix with element $(1, j)^{\text{th}}$ equal to $P(y \mid S^{P_0} = \varpi_{N_z}(j - 1), Z = z, A = k)$. Let the matrix $P_{N_z}(S \mid Z, S^{P_0})$ be in $\mathbb{R}^{2N_z \times 2^{N_z}}$ where column denotes principal stratum $S^{P_0} = \varpi_{N_z}(j - 1)$ and row represents a combination $(s, z) \in \{(1, 1), (1, 2), \ldots, (1, N_z), (0, 1), \ldots, (0, N_z)\}$, with $(i, j)^{\text{th}}$

element denoted $P_{N_z}(S \mid Z, S^{P_0})_{ij}$ defined as

$$P_{N_z}(S \mid Z, S^{P_0})_{ij} = \varpi_{N_z}(j-1)_i \mathbb{1}_{i \leq N_z} + (1 - \varpi_{N_z}(j-1)_{i-N_z}) \mathbb{1}_{i > N_z},$$

and let $P_{N_z}(\tilde{S} \mid Z, S^{P_0})$ be in $\mathbb{R}^{2N_z \times 2^{N_z}}$ with $(i,j)^{\text{th}}$ element denoted $P_{N_z}(\tilde{S} \mid Z, S^{P_0})_{ij}$ defined:

$$\begin{aligned} P_{N_z}(\tilde{S} \mid Z, S^{P_0})_{ij} = {} & \text{sn}_S^{\varpi_{N_z}(j-1)_i} (1 - \text{sp}_S)^{1 - \varpi_{N_z}(j-1)_i} \mathbb{1}_{i \leq N_z} \\ & + (1 - \text{sn}_S)^{\varpi_{N_z}(j-1)_{i-N_z}} \text{sp}_S^{1 - \varpi_{N_z}(j-1)_{i-N_z}} \mathbb{1}_{i > N_z}. \end{aligned}$$

Let B^+ be the Moore-Penrose inverse of the matrix B, $\mathbf{1}_m$ be the m-vector of 1s, $\mathbf{0}_m$ be the m-vector of 0s, and $\mathbf{I}_m$ be the $m \times m$ dimensional identity matrix.

B.2 Kruskal rank properties related to VE

In this section, we show that (a) the Kruskal rank of the matrix $P_{N_z}(\tilde{S} \mid Z, S^{P_0})$ is 3 for $N_z \geq 2$ when $\text{sn}_S + \text{sp}_S \neq 1$ and (b) the column domains of $P_{N_z}(\tilde{S} \mid Z, S^{P_0})$ are not invariant to column permutation when $\text{sn}_S, \text{sp}_S > 0.5$ or $\text{sn}_S, \text{sp}_S < 0.5$ for $N_z \geq 2$.

Lemma 13 (Kruskal rank $P_2(\tilde{S} \mid Z, S^{P_0})$)**.** *The Kruskal rank of*

$$\begin{array}{cccc} (0,0) & (1,0) & (0,1) & (1,1) & \\ \begin{bmatrix} 1-\text{sp}_S & \text{sn}_S & 1-\text{sp}_S & \text{sn}_S \\ 1-\text{sp}_S & 1-\text{sp}_S & \text{sn}_S & \text{sn}_S \\ \text{sp}_S & 1-\text{sn}_S & \text{sp}_S & 1-\text{sn}_S \\ \text{sp}_S & \text{sp}_S & 1-\text{sn}_S & 1-\text{sn}_S \end{bmatrix} & \begin{array}{l} (s=1,z=1) \\ (s=1,z=2) \\ (s=0,z=1) \\ (s=0,z=2) \end{array} \end{array} \tag{B.1}$$

is 3 *as long as* $\text{sn}_S + \text{sp}_S \neq 1$.

Proof. All subsets of 3 columns of the matrix $P_2(\tilde{S} \mid Z, S^{P_0})$ are of the form:

$$\begin{bmatrix} a & c & e \\ b & d & f \\ 1-a & 1-c & 1-e \\ 1-b & 1-d & 1-f \end{bmatrix}. \tag{B.2}$$

These submatrices have a common maximal minor of

$$a(d-f) - c(b-f) + e(b-d).$$

The quantities a, b, c, d, e, f are the elements of the 2×3 matrix

$$\begin{bmatrix} a & c & e \\ b & d & f \end{bmatrix} \tag{B.3}$$

in which $(a, b)^T, (c, d)^T, (e, f)^T$ are any 3 columns drawn without replacement from the 2×4 submatrix of Equation (B.1):

$$\begin{bmatrix} 1 - \mathrm{sp}_S & \mathrm{sn}_S & 1 - \mathrm{sp}_S & \mathrm{sn}_S \\ 1 - \mathrm{sp}_S & 1 - \mathrm{sp}_S & \mathrm{sn}_S & \mathrm{sn}_S \end{bmatrix}. \tag{B.4}$$

These minors are all equal to (up to a factor of -1):

$$(1 - \mathrm{sn}_S - \mathrm{sp}_S)^2,$$

which can be seen after a brute-force calculation. The minors are nonzero for all $\mathrm{sn}_S, \mathrm{sp}_S \in [0, 1]$ such that $\mathrm{sn}_S + \mathrm{sp}_S \neq 1$. Thus, by the determinantal rank definition, all 3 column matrices are rank 3. In contrast, the determinant of $P_2(\tilde{S} \mid Z, S^{P_0})$ is 0 for all values of $\mathrm{sn}_S, \mathrm{sp}_S$. Thus by the definition of Kruskal rank in Definition 3.2.6, $k_{P_2(\tilde{S}|Z,S^{P_0})} = 3$. □

Lemma 14 (Kruskal rank $P_{N_z}(\tilde{S} \mid Z, S^{P_0})$, $N_z \geq 2$)**.** *The Kruskal rank of* $P(\tilde{S} \mid Z, S^{P_0})$ *for* $N_z \geq 2$ *is* 3 *as long as* $\mathrm{sn}_S + \mathrm{sp}_S \neq 1$.

Proof. We proceed by induction. For $N_z = 2$, Lemma 13 shows that the Kruskal rank is 3. Let $N_z = n$ for $n > 2$. Recall that $P_n(\tilde{S} \mid Z, S^{P_0})$ is the $2n \times 2^n$ matrix with column j

$$\begin{bmatrix} s_j \\ \mathbf{1}_n - s_j \end{bmatrix}$$

with the i^{th} element of s_j denoted s_{ij} and defined as:

$$s_{ij} = \mathrm{sn}_S^{\varpi_n(j-1)_i}(1 - \mathrm{sp}_S)^{1-\varpi_n(j-1)_i}.$$

The induction hypothesis is that the Kruskal rank of $P_n(\tilde{S} \mid Z, S^{P_0})$ is 3. The columns of $P_{n+1}(\tilde{S} \mid Z, S^{P_0})$ are of the form

$$\begin{bmatrix} s_j \\ 1 - \mathrm{sp}_S \\ \mathbf{1}_n - s_j \\ \mathrm{sp}_S \end{bmatrix}$$

for $j \in \{1, \ldots, 2^n\}$, and

$$\begin{bmatrix} s_{j-2^n} \\ \mathrm{sn}_S \\ \mathbf{1}_n - s_{j-2^n} \\ 1 - \mathrm{sn}_S \end{bmatrix}$$

for $j \in \{2^n + 1, \ldots, 2^{n+1}\}$. The 3-column submatrices of $P_{N_z}(\tilde{S} \mid Z, S^{P_0})$ made from column j, ℓ, m indices fall into several classes. When $j, \ell, m \in \{1, \ldots, 2^n\}$, $j, \ell, m \in \{2^n + 1, \ldots, 2^{n+1}\}$ or $j, \ell \in \{1, \ldots, 2^n\}, m \in \{2^n + 1, \ldots, 2^{n+1}\} \setminus \{j + 2^n, \ell + 2^n\}$, $j \in \{1, \ldots, 2^n\}, m, \ell \in \{2^n + 1, \ldots, 2^{n+1}\} \setminus \{j + 2^n\}$ all matrices are rank 3 by the induction hypothesis. When $j, \ell \in \{1, \ldots, 2^n\}$ but $m \in \{j + 2^n, \ell + 2^n\}$ the submatrix is

$$\begin{bmatrix} s_j & s_\ell & s_{m-2^n} \\ 1 - \mathrm{sp}_S & 1 - \mathrm{sp}_S & \mathrm{sn}_S \\ \mathbf{1}_n - s_j & \mathbf{1}_n - s_\ell & \mathbf{1}_n - s_{m-2^n} \\ \mathrm{sp}_S & \mathrm{sp}_S & 1 - \mathrm{sn}_S \end{bmatrix}.$$

WLOG, let $m = j + 2^n$. This leads to the submatrix:

$$\begin{bmatrix} s_j & s_\ell & s_j \\ 1 - \mathrm{sp}_S & 1 - \mathrm{sp}_S & \mathrm{sn}_S \\ \mathbf{1}_n - s_j & \mathbf{1}_n - s_\ell & \mathbf{1}_n - s_j \\ \mathrm{sp}_S & \mathrm{sp}_S & 1 - \mathrm{sn}_S \end{bmatrix}.$$

The rank of this submatrix is

$$\begin{aligned}
\operatorname{rank}\begin{bmatrix} s_j & s_\ell & s_j \\ 1-\mathrm{sp}_S & 1-\mathrm{sp}_S & \mathrm{sn}_S \\ \mathbf{1}_n - s_j & \mathbf{1}_n - s_\ell & \mathbf{1}_n - s_j \\ \mathrm{sp}_S & \mathrm{sp}_S & 1-\mathrm{sn}_S \end{bmatrix} &= \operatorname{rank}\begin{bmatrix} s_j & s_\ell & s_j \\ 1-\mathrm{sp}_S & 1-\mathrm{sp}_S & \mathrm{sn}_S \\ \mathbf{1}_n - s_j & \mathbf{1}_n - s_\ell & \mathbf{1}_n - s_j \\ 1 & 1 & 1 \end{bmatrix} \\
&= \operatorname{rank}\begin{bmatrix} s_j & s_\ell & s_j \\ \mathbf{1}_n - s_j & \mathbf{1}_n - s_\ell & \mathbf{1}_n - s_j \\ 1-\mathrm{sp}_S & 1-\mathrm{sp}_S & \mathrm{sn}_S \\ 0 & 0 & 1-\frac{\mathrm{sn}_S}{1-\mathrm{sp}_S} \end{bmatrix} \\
&\geq \operatorname{rank}\begin{bmatrix} s_j & s_\ell \\ \mathbf{1}_n - s_j & \mathbf{1}_n - s_\ell \\ 1-\mathrm{sp}_S & 1-\mathrm{sp}_S \\ 0 & 0 \end{bmatrix} + \operatorname{rank}(1-\frac{\mathrm{sn}_S}{1-\mathrm{sp}_S}) \\
&= 3.
\end{aligned}$$

The inequality follows from Lemma 19. Other scenarios follow similarly. □

Lemma 15 (Domain restriction lemma)**.** *If* $\mathrm{sn}_S, \mathrm{sp}_S \in [0, 0.5)$ *or* $\mathrm{sn}_S, \mathrm{sp}_S \in (0.5, 1]$*, the matrix* $P_{N_z}(\tilde{S} \mid Z, S^{P_0}) \in \mathbb{R}^{2N_z \times 2^{N_z}}$ *has column domains that are not invariant to column permutation.*

Proof. We prove Lemma 15 by induction on N_z. The base case is $N_z = 2$. Let P be a 4×4 permutation matrix and let $P_2(\tilde{S} \mid Z, S^{P_0})$ be

$$\begin{array}{c} \begin{matrix} (0,0) & (1,0) & (0,1) & (1,1) \end{matrix} \\ \begin{bmatrix} 1-\mathrm{sp}_S & \mathrm{sn}_S & 1-\mathrm{sp}_S & \mathrm{sn}_S \\ 1-\mathrm{sp}_S & 1-\mathrm{sp}_S & \mathrm{sn}_S & \mathrm{sn}_S \\ \mathrm{sp}_S & 1-\mathrm{sn}_S & \mathrm{sp}_S & 1-\mathrm{sn}_S \\ \mathrm{sp}_S & \mathrm{sp}_S & 1-\mathrm{sn}_S & 1-\mathrm{sn}_S \end{bmatrix} \end{array} \begin{array}{l} \\ (s=1, z=1) \\ (s=1, z=2) \\ (s=0, z=1) \\ (s=0, z=2) \end{array} \tag{B.5}$$

Recall from the definition in Appendix B.1 that the column indices $\{1, 2, 3, 4\}$ of $P_2(\tilde{S} \mid Z, S^{P_0})$ map to the following principal strata S^{P_0}: $\varpi_2(0), \varpi_2(1), \varpi_2(2), \varpi_2(3)$. In other words, column index j is mapped to S^{P_0} via the relation $\varpi_2(j-1)$. We consider permutation matrix P without

loss of generality, and other cases are similarly shown,

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

Let $\mathcal{C} = [0, 1]$, and let $\mathcal{A}$ be one of two half intervals of $[0, 1]$: $[0, 0.5)$ or $(0.5, 1]$. Let $\mathcal{B} = \mathcal{C} \setminus \mathcal{A}$. Note that $P_2(\tilde{S} \mid Z, S^{P_0})$ maps $\mathcal{C} \times \mathcal{C}$ to a matrix with elements in $\mathcal{C}$. Let $1 - \mathrm{sp}_S \in \mathcal{A}$ and let $\mathrm{sn}_S \in \mathcal{B}$ and suppose that the column domains for $P_2(\tilde{S} \mid Z, S^{P_0})$ are not invariant after permutation by matrix P. Then we have the following domain for the map given by $P_2(\tilde{S} \mid Z, S^{P_0})$:

$$P_2(\tilde{S} \mid Z, S^{P_0}) \mid_{\mathcal{A}\times\mathcal{B}} \colon \mathcal{A} \times \mathcal{B} \to \begin{array}{c} \begin{matrix} (0,0) & (1,0) & (0,1) & (1,1) \end{matrix} \\ \begin{bmatrix} \mathcal{A} & \mathcal{A} & \mathcal{A} & \mathcal{A} \\ \mathcal{A} & \mathcal{A} & \mathcal{A} & \mathcal{A} \\ \mathcal{B} & \mathcal{B} & \mathcal{B} & \mathcal{B} \\ \mathcal{B} & \mathcal{B} & \mathcal{B} & \mathcal{B}) \end{bmatrix} \end{array}$$

However, we have,

$$\bar{P}_2(\tilde{S} \mid Z, S^{P_0}) \mid_{\mathcal{A}\times\mathcal{B}} = P_2(\tilde{S} \mid Z, S^{P_0}) \mid_{\mathcal{A}\times\mathcal{B}} P \tag{B.6}$$

$$= \begin{array}{c} \begin{matrix} (0,0) & (1,0) & (0,1) & (1,1) \end{matrix} \\ \begin{bmatrix} 1-\mathrm{sp}_S & \mathrm{sn}_S & 1-\mathrm{sp}_S & \mathrm{sn}_S \\ 1-\mathrm{sp}_S & 1-\mathrm{sp}_S & \mathrm{sn}_S & \mathrm{sn}_S \\ \mathrm{sp}_S & 1-\mathrm{sn}_S & \mathrm{sp}_S & 1-\mathrm{sn}_S \\ \mathrm{sp}_S & \mathrm{sp}_S & 1-\mathrm{sn}_S & 1-\mathrm{sn}_S \end{bmatrix} \end{array} \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \tag{B.7}$$

$$= \begin{array}{c} \begin{matrix} (0,0) & (1,0) & (0,1) & (1,1) \end{matrix} \\ \begin{bmatrix} \mathrm{sn}_S & 1-\mathrm{sp}_S & \mathrm{sn}_S & 1-\mathrm{sp}_S \\ \mathrm{sn}_S & 1-\mathrm{sp}_S & 1-\mathrm{sp}_S & \mathrm{sn}_S \\ 1-\mathrm{sn}_S & \mathrm{sp}_S & 1-\mathrm{sn}_S & \mathrm{sp}_S \\ 1-\mathrm{sn}_S & \mathrm{sp}_S & \mathrm{sp}_S & 1-\mathrm{sn}_S \end{bmatrix} \end{array}. \tag{B.8}$$

But we see that the column domains are invariant after column permutation:

$$\bar{P}_2(\tilde{S} \mid Z, S^{P_0}) \mid_{\mathcal{A}\times\mathcal{B}}: \mathcal{A}\times\mathcal{B} \to \begin{matrix} (0,0) & (1,0) & (0,1) & (1,1) & \\ \left[\begin{matrix} \mathcal{A} & \mathcal{A} & \mathcal{A} & \mathcal{A} \\ \mathcal{A} & \mathcal{A} & \mathcal{A} & \mathcal{A} \\ \mathcal{B} & \mathcal{B} & \mathcal{B} & \mathcal{B} \\ \mathcal{B} & \mathcal{B} & \mathcal{B} & \mathcal{B} \end{matrix}\right. & \begin{matrix} (s=1,z=1) \\ (s=1,z=2) \\ (s=0,z=1) \\ \left.(s=0,z=2)\right) \end{matrix} \end{matrix}$$

In order for the columns $\bar{P}_2(\tilde{S} \mid Z, S^{P_0}) \mid_{\mathcal{A}\times\mathcal{B}}$ to be on the same domain as $P_2(\tilde{S} \mid Z, S^{P_0}) \mid_{\mathcal{A}\times\mathcal{B}}$, a necessary and sufficient condition is that sn_S and $1 - \mathrm{sp}_S$ are on the same domain. In other words, $\{\mathrm{sn}_S \in \mathcal{A}, \mathrm{sp}_S \in \mathcal{B}\}$ or $\{\mathrm{sn}_S \in \mathcal{B}, \mathrm{sp}_S \in \mathcal{A}\}$.

Thus $\bar{P}_2(\tilde{S} \mid Z, S^{P_0}) \mid_{\mathcal{A}\times\mathcal{B}}$ maps $(\mathrm{sn}_S, \mathrm{sp}_S)$ to the same space that $P_2(\tilde{S} \mid Z, S^{P_0}) \mid_{\mathcal{A}\times\mathcal{B}}$. We contradict our statement that the columns are not invariant to permutation.

The case for $N_z > 2$. Let $N_z = n > 2$ and let the column domains of $P_n(\tilde{S} \mid Z, S^{P_0})$ be not invariant to permutation. Furthermore suppose that $\mathrm{sn}_S, \mathrm{sp}_S \in \mathcal{A}$ or $\mathrm{sn}_S, \mathrm{sp}_S \in \mathcal{B}$. Then matrix $P_{n+1}(\tilde{S} \mid Z, S^{P_0})$ has columns

$$\begin{bmatrix} s_j \\ 1 - \mathrm{sp}_S \\ \mathbf{1}_n - s_j \\ \mathrm{sp}_S \end{bmatrix}$$

for $j \in \{1, \ldots, 2^n\}$ and

$$\begin{bmatrix} s_{j-2^n} \\ \mathrm{sn}_S \\ \mathbf{1}_n - s_{j-2^n} \\ 1 - \mathrm{sn}_S \end{bmatrix}$$

for $j \in \{2^n + 1, \ldots, 2^{n+1}\}$. Permuting any two columns $j, k \in \{1, \ldots, 2^n\}$ or $j, k \in \{2^n + 1, \ldots, 2^{n+1}\}$ yields different column domains given the induction hypothesis. If $j \in \{1, \ldots, 2^n\}$ and $k = j + 2^n$, then the columns are

$$\begin{bmatrix} s_j & s_j \\ 1 - \mathrm{sp}_S & \mathrm{sn}_S \\ \mathbf{1}_n - s_j & \mathbf{1}_n - s_j \\ \mathrm{sp}_S & 1 - \mathrm{sn}_S \end{bmatrix}$$

Let the domain of s_j be $\mathcal{D}$, and let $\mathcal{D}^{\mathsf{c}} = [0,1]^n \setminus \mathcal{D}$ be the domain of $\mathbf{1}_n - s_j$. Then the domains

are

$$\begin{bmatrix} \mathcal{D} & \mathcal{D} \\ \mathcal{A} & \mathcal{B} \\ \mathcal{D}^{\mathsf{c}} & \mathcal{D}^{\mathsf{c}} \\ \mathcal{B} & \mathcal{A} \end{bmatrix}$$

if $\mathrm{sp}_S, \mathrm{sn}_S \in \mathcal{B}$ and

$$\begin{bmatrix} \mathcal{D} & \mathcal{D} \\ \mathcal{B} & \mathcal{A} \\ \mathcal{D}^{\mathsf{c}} & \mathcal{D}^{\mathsf{c}} \\ \mathcal{A} & \mathcal{B} \end{bmatrix}$$

if $\mathrm{sp}_S, \mathrm{sn}_S \in \mathcal{A}$. These two columns are not invariant to permutation. Because no two columns may be interchanged without a change in domain, right multiplying $P_{n+1}(\tilde{S} \mid Z, S^{P_0})$ by any $2^{n+1} \times 2^{n+1}$ permutation matrix $P \neq \mathbf{I}_{n+1}$ to will yield a matrix with different column domains than $P_{n+1}(\tilde{S} \mid Z, S^{P_0})$. □

B.3 Rank properties related to VE

In this section we show that when $N_z \geq 2$ the rank of $P_{N_z}(\tilde{S} \mid Z, S^{P_0}) = N_z + 1$ when $\mathrm{sn}_S + \mathrm{sp}_S \neq 1$.

Lemma 16 (Rank $P_2(\tilde{S} \mid Z, S^{P_0})$)**.** *The rank of* $P_2(\tilde{S} \mid Z, S^{P_0})$*, defined in Equation* (B.1)*, is* 3 *as long as* $\mathrm{sn}_S + \mathrm{sp}_S \neq 1$.

Proof. The determinant of $P_2(\tilde{S} \mid Z, S^{P_0})$ is 0. The determinant of the 3-minor $M_{4,4}$ is $(1 - \mathrm{sn}_S - \mathrm{sp}_S)^2$ which is nonzero as long as $\mathrm{sn}_S + \mathrm{sp}_S \neq 1$. □

Lemma 17 (Rank $P_{N_z}(\tilde{S} \mid Z, S^{P_0}),\ N_z \geq 2$)**.** *The rank of* $P(\tilde{S} \mid Z, S^{P_0})$ *for* $N_z \geq 2$ *is* $N_z + 1$ *as long as* $\mathrm{sn}_S + \mathrm{sp}_S \neq 1$.

Proof. We proceed by induction. For $N_z = 2$, Lemma 16 shows that the rank is 3. Let $N_z = n$ for $n > 2$. Recall that $P_n(\tilde{S} \mid Z, S^{P_0})$ is the $2n \times 2^n$ matrix with column j

$$\begin{bmatrix} s_j \\ \mathbf{1}_n - s_j \end{bmatrix}$$

with the i^{th} element of s_j denoted s_{ij} and defined as:

$$s_{ij} = \mathrm{sn}_S^{\varpi_n(j-1)_i}(1 - \mathrm{sp}_S)^{1-\varpi_n(j-1)_i}.$$

The induction hypothesis is that the rank of $P_n(\tilde{S} \mid Z, S^{P_0})$ is $n+1$. The columns of $P_{n+1}(\tilde{S} \mid Z, S^{P_0})$ are of the form

$$\begin{bmatrix} s_j \\ 1 - \text{sp}_S \\ \mathbf{1}_n - s_j \\ \text{sp}_S \end{bmatrix}$$

for $j \in \{1, \ldots, 2^n\}$, and

$$\begin{bmatrix} s_{j-2^n} \\ \text{sn}_S \\ \mathbf{1}_n - s_{j-2^n} \\ 1 - \text{sn}_S \end{bmatrix}$$

for $j \in \{2^n + 1, \ldots, 2^{n+1}\}$. After a row permutation we can express $P_{n+1}(\tilde{S} \mid Z, S^{P_0})$ as a block matrix:

$$\begin{bmatrix} P_n(\tilde{S} \mid Z, S^{P_0}) & P_n(\tilde{S} \mid Z, S^{P_0}) \\ (1 - \text{sp}_S)\mathbf{1}_{2^n}^T & \text{sn}_S \mathbf{1}_{2^n}^T \\ \text{sp}_S \mathbf{1}_{2^n}^T & (1 - \text{sn}_S)\mathbf{1}_{2^n}^T \end{bmatrix}$$

Recall that by construction the sum of the i^{th} row with the $(i+n)^{\text{th}}$ row of $P_n(\tilde{S} \mid Z, S^{P_0})$ is $\mathbf{1}_{2^n}^T$ for $i \leq n$. Then by Lemma 20, $\text{rank}\left(P_{n+1}(S \mid Z, S^{P_0})\right)$ is

$$\text{rank}\left(P_{n+1}(S \mid Z, S^{P_0})\right) = \text{rank}\left(P_n(S \mid Z, S^{P_0})\right) + \text{rank}\left(\begin{bmatrix} \text{sn}_S \mathbf{1}_{2^n}^T \\ (1 - \text{sn}_S)\mathbf{1}_{2^n}^T \end{bmatrix} - \begin{bmatrix} (1 - \text{sp}_S)\mathbf{1}_{2^n}^T \\ \text{sp}_S \mathbf{1}_{2^n}^T \end{bmatrix}\right) \tag{B.9}$$

$$= n + 1 + 1 \tag{B.10}$$

given that $\text{sn}_S + \text{sp}_S \neq 1$. □

B.4 Main results

Proof. Proof of Theorem 7

Define the three way array L with dimensions $2N_z \times N_a \times N_r$ and $(i, j, r)^{\text{th}}$ element $P(\tilde{S} = \mathbb{1}_{i \leq N_z}, A = a \mid Z = z_{i - N_z \mathbb{1}_{i > N_z}}, R = r)$. Recall that the definition of matrix $P_{N_z}(\tilde{S} \mid Z, S^{P_0})$ requires that column j be

$$\begin{bmatrix} s_j \\ \mathbf{1}_{N_z} - s_j \end{bmatrix}$$

where the i^{th} element of s_j is denoted as s_{ij} and is defined as:

$$s_{ij} = \text{sn}_S^{\varpi_{N_z}(j-1)_i}(1-\text{sp}_S)^{1-\varpi_{N_z}(j-1)_i}$$

Let the matrices $P_{N_z}(S^{P_0} \mid R)^T, P_{N_z}(A \mid S^{P_0})$ be defined as in Appendix B.1. Then

$$P(\tilde{S} = \mathbb{1}_{i \leq N_z}, A = k \mid Z = z_{i-N_z \mathbb{1}_{i>N_z}}, R = r) = \sum_{j=1}^{2^{N_z}} P_{N_z}(\tilde{S} \mid Z, S^{P_0})_{i,j} P_{N_z}(S^{P_0} \mid R)^T_{r,j} P_{N_z}(A \mid S^{P_0})_{k,j}.$$

Given that $\text{sn}_S + \text{sp}_S \neq 1$, as shown in Lemma 14, $k_{P_{N_z}(\tilde{S}|Z,S^{P_0})} = 3$ and $\text{rank}(P_{N_z}(\tilde{S} \mid Z, S^{P_0})) = N_z + 1$. Furthermore, by assumptions stated in Theorem 7, $\text{rank}(P_{N_z}(S^{P_0} \mid R)^T) = 2^{N_z}$ and $P_{N_z}(S^{P_0} \mid R)^T \in \mathbb{R}^{N_r \times 2^{N_z}}$ so by Definition 3.2.6, $k_{P_{N_z}(S^{P_0}|R)^T} = 2^{N_z}$. Given that $k_{P_{N_z}(A|S^{P_0})} \geq 2^{N_z} - 1$ as stated in Theorem 7, the conditions in Lemma 4 hold:

$$\min(3, 2^{N_z}) + 2^{N_z} - 1 \geq 2^{N_z} + 2 \tag{B.11}$$

$$\min(3, 2^{N_z} - 1) + 2^{N_z} \geq 2^{N_z} + 2 \tag{B.12}$$

$$\tag{B.13}$$

and

$$\text{rank}(P_{N_z}(S \mid Z, S^{P_0})) + \text{rank}(P_{N_z}(S^{P_0} \mid R)) + \text{rank}(P_{N_z}(A \mid S^{P_0})) \tag{B.14}$$

$$\geq N_z + 1 + 2^{N_z} + 2^{N_z} - 1 \tag{B.15}$$

$$= N_z + 2^{N_z+1} \tag{B.16}$$

by the fact that $\text{rank}(P_{N_z}(A \mid S^{P_0})) \geq k_{(P_{N_z}(A|S^{P_0})}$. Also

$$N_z + 2^{N_z+1} - 2(2^{N_z} - 1) = N_z - 1 \tag{B.17}$$

$$\geq \begin{cases} \min(N_z - 2, \text{rank}(P_{N_z}(A \mid S^{P_0})) - k_{(P_{N_z}(A|S^{P_0})}) \\ \min(N_z - 2, 0) \end{cases}. \tag{B.18}$$

Given that $P_{N_z}(A \mid S^{P_0})$ has columns that sum to 1, and $P_{N_z}(S^{P_0} \mid R)^T$ has rows that sum to 1, we can apply Lemma 4 to the 3-way array L. Applying Lemma 4 yields that the triple-product decomposition $[P_{N_z}(\tilde{S} \mid Z, S^{P_0}), P_{N_z}(A \mid S^{P_0}), P_{N_z}(S^{P_0} \mid R)^T]$ is unique up to a common column permutation. However, Theorem 7 states the assumption that sn_S, sp_S lie in a common half-interval. By Lemma 15, the only permutation matrix consistent with the column domain

of $P_{N_z}(\tilde{S} \mid Z, S^{P_0})$ is the identity matrix. We conclude that the 3-way decomposition of L, $[P_{N_z}(\tilde{S} \mid Z, S^{P_0}), P_{N_z}(A \mid S^{P_0}), P_{N_z}(S^{P_0} \mid R)^T]$, is unique. It follows that two different decompositions $[P_{N_z}(\tilde{S} \mid Z, S^{P_0}), P_{N_z}(A \mid S^{P_0}), P_{N_z}(S^{P_0} \mid R)^T]$ and $[P_{N_z}(\tilde{S} \mid Z, S^{P_0})', P_{N_z}(A \mid S^{P_0})', (P_{N_z}(S^{P_0} \mid R)^T)']$ yield different Ls. By the fact that L is a complete characterization of the data distribution $P(\tilde{S} = s, A = a \mid Z = z_j, R = r)$ and Definition 3.2.4 the parameter set $[P_{N_z}(\tilde{S} \mid Z, S^{P_0}), P_{N_z}(A \mid S^{P_0}), P_{N_z}(S^{P_0} \mid R)^T]$ is strictly identifiable.

Define the matrix $P(\tilde{Y} \mid Z, R, A = k)$ with dimensions $N_z \times N_r$ with elements $P(\tilde{Y} = y \mid R = r, Z = z, A = k)$

$$P(\tilde{Y} \mid Z, R, A = k)_{i,r} = P(\tilde{Y} = 1 \mid Z = z_i, R = r, A = k).$$

Let the matrix $P_{N_z}(\tilde{Y} \mid Z, S^{P_0}, A = k)$ be in $\mathbb{R}^{N_z \times 2^{N_z}}$ for all $k \in \{1, \ldots, N_a\}$ with elements

$$\begin{aligned} P_{N_z}(\tilde{Y} \mid Z, S^{P_0}, A = k)_{i,j} = \varpi_{N_z}(j-1)_i r_Y P(Y = 1 \mid Z = z_i, S^{P_0} = \varpi_{N_z}(j-1), A = k) \\ + (1 - \mathrm{sp}_Y) \end{aligned} \tag{B.19}$$

where $r_Y = \mathrm{sp}_Y + \mathrm{sn}_Y - 1$. Then $P(\tilde{Y} = 1 \mid Z = z_i, A = k, R = r) = \sum_{j=1}^{2^{N_z}} P_{N_z}(\tilde{Y} \mid Z, S^{P_0}, A = a)_{i,j} P_{N_z}(S^{P_0} \mid R)_{j,r}$ which can be represented as matrix multiplication:

$$P(\tilde{Y} \mid Z, R, A = a) = P_{N_z}(\tilde{Y} \mid Z, S^{P_0}, A = a) P_{N_z}(S^{P_0} \mid R) \tag{B.20}$$

Given our assumption that $P_{N_z}(S^{P_0} \mid R)$ is full row rank,$P_{N_z}(S^{P_0} \mid R) P_{N_z}(S^{P_0} \mid R)^+ = \mathbf{I}_{2^{N_z}}$ and

$$P(\tilde{Y} \mid Z, R, A = a) P_{N_z}(S^{P_0} \mid R)^+ = P_{N_z}(\tilde{Y} \mid Z, A = a, S^{P_0}) \tag{B.21}$$

It then follows the definition of $P_{N_z}(\tilde{Y} \mid Z, S^{P_0}, A = a)$ in Equation (B.19) that sp_Y is identifiable, as are the parameters $r_Y P(Y = 1 \mid Z = z_j, S^{P_0} = \varpi_{N_z}(j-1), A = k)$ for all $z_j, j \in \{1, \ldots, 2^{N_z}\}$ and k.

Let any allowable post-infection outcome vaccine efficacy estimand, necessarily where $u_j \, u_l = 1$, be defined as

$$\mathrm{VE}^u_{I,jl}(k) = 1 - \frac{\mathbb{E}\left[Y(z_j) \mid S^{P_0} = u, A = k\right]}{\mathbb{E}\left[Y(z_l) \mid S^{P_0} = u, A = k\right]}.$$

By Assumptions 2 to 3 $P(Y = 1 \mid Z = z, S^{P_0} = u, A = k) = P(Y(z) = 1 \mid S^{P_0} = u, A = k)$ for all $z \in \{z_1, \ldots, z_{N_z}\}$ and $\mathbb{E}\left[Y(z) \mid S^{P_0} = u, A = k\right] = P(Y(z) = 1 \mid S^{P_0} = u, A = k)$. Note that $\mathrm{sp}_Y = 1 - P_{N_z}(\tilde{Y} \mid Z, S^{P_0}, A = k)_{1,1}$ by our definition of $P_{N_z}(\tilde{Y} \mid Z, S^{P_0}, A = k)$ in

Equation (B.19). Then

$$P(Y=1 \mid Z=z, S^{P_0}=u, A=k) = \frac{P_{N_z}(\tilde{Y} \mid Z, S^{P_0}, A=k)_{z,j} - P_{N_z}(\tilde{Y} \mid Z, S^{P_0}, A=k)_{1,1}}{r_Y}$$

where $j = \varpi_{N_z}^{-1}(u) + 1$, so $\mathrm{VE}^u_{I,jl}(k)$ is identifiable. □

Proof. Proof of Corollary 8
By the conditions set forth in Corollary 8 we have that

$$P(\tilde{S} = \mathbb{1}_{i \leq N_z}, \tilde{A} = k \mid Z = z_{i - N_z \mathbb{1}_{i > N_z}}, R = r) = \sum_{j=1}^{2^{N_z}} P_{N_z}(\tilde{S} \mid Z, S^{P_0})_{i,j} P_{N_z}(S^{P_0} \mid R)^T_{r,j} P_{N_z}(\tilde{A} \mid S^{P_0})_{k,j}.$$

This decomposition holds because of our nondifferential misclassification assumption, namely $\tilde{A} \perp\!\!\!\perp S^{P_0}, \tilde{S}, R, Z \mid A$, which allows for the following complete characterization of $\tilde{A} \mid S^{P_0}$:

$$P(\tilde{A}=k \mid S^{P_0}=u) = \sum_{\ell=1}^{N_z} P(\tilde{A}=k \mid A=\ell) P(A=\ell \mid S^{P_0}=u).$$

Recall that $\mathrm{sn}_S, \mathrm{sp}_S$ lie in the same half interval of $[0,1]$, so by the same logic as Appendix B.4, the distributions $P(\tilde{S}=1 \mid Z=z, S^{P_0}=u), P(\tilde{A}=k \mid S^{P_0}=u), P(S^{P_0}=u \mid R=r)$ are identifiable. Define the matrix $P(\tilde{Y} \mid Z, R, \tilde{A}=k)$ with dimensions $N_z \times N_r$ with elements $P(\tilde{Y}=y \mid R=r, Z=z, \tilde{A}=k)$

$$P(\tilde{Y} \mid Z, R, \tilde{A}=k)_{i,r} = P(\tilde{Y}=1 \mid Z=z_i, R=r, \tilde{A}=k).$$

Let the matrix $P_{N_z}(\tilde{Y} \mid Z, S^{P_0}, \tilde{A}=k)$ be defined in the same way as Equation (B.19). Then $P(\tilde{Y}=1 \mid Z=z_i, \tilde{A}=k, R=r) = \sum_{j=1}^{2^{N_z}} P_{N_z}(\tilde{Y} \mid Z, S^{P_0}, \tilde{A}=a)_{i,j} P_{N_z}(S^{P_0} \mid R)_{j,r}$ which can be represented as matrix multiplication:

$$P(\tilde{Y} \mid Z, R, \tilde{A}=k) = P_{N_z}(\tilde{Y} \mid Z, S^{P_0}, \tilde{A}=a) P_{N_z}(S^{P_0} \mid R) \tag{B.22}$$

Given our assumption that $P_{N_z}(S^{P_0} \mid R)$ is full row rank,$P_{N_z}(S^{P_0} \mid R) P_{N_z}(S^{P_0} \mid R)^+ = \mathbf{I}_{2^{N_z}}$ and

$$P(\tilde{Y} \mid Z, R, \tilde{A}=k) P_{N_z}(S^{P_0} \mid R)^+ = P_{N_z}(\tilde{Y} \mid Z, \tilde{A}=k, S^{P_0}) \tag{B.23}$$

It then follows the definition of $P_{N_z}(\tilde{Y} \mid Z, S^{P_0}, \tilde{A}=k)$ that sp_Y is identifiable, as are the parameters $r_Y P(Y=1 \mid Z=z_j, S^{P_0}=\varpi_{N_z}(j-1), \tilde{A}=k)$ for all $j \in \{1, \ldots, 2^{N_z}\}$ and

$k \in \{1, \ldots, N_a\}$.

Let any allowable post-infection outcome vaccine efficacy estimand, necessarily where $u_j\, u_l = 1$, be defined as

$$\mathrm{VE}^{u}_{I,jl} = 1 - \frac{\mathbb{E}\left[Y(z_j) \mid S^{P_0} = u\right]}{\mathbb{E}\left[Y(z_l) \mid S^{P_0} = u\right]}.$$

By Assumptions 2 to 3 $P(Y = 1 \mid Z = z, S^{P_0} = u, \tilde{A} = k) = P(Y(z) = 1 \mid S^{P_0} = u, \tilde{A} = k)$ and $\mathbb{E}\left[Y(z) \mid S^{P_0} = u, \tilde{A} = k\right] = P(Y(z) = 1 \mid S^{P_0} = u, \tilde{A} = k)$ for all $z \in \{z_1, \ldots, z_{N_z}\}$. Note that $\mathrm{sp}_Y = 1 - P_{N_z}(\tilde{Y} \mid Z, S^{P_0}, \tilde{A} = k)_{1,1}$ by our definition of $P_{N_z}(\tilde{Y} \mid Z, S^{P_0}, \tilde{A} = k)$ in Equation (B.19). Then

$$P(Y = 1 \mid Z = z, S^{P_0} = u, \tilde{A} = k) = \frac{P_{N_z}(\tilde{Y} \mid Z, S^{P_0}, \tilde{A} = k)_{z,j} - P_{N_z}(\tilde{Y} \mid Z, S^{P_0}, \tilde{A} = k)_{1,1}}{r_Y}$$

where $j = \varpi^{-1}_{N_z}(u) + 1$. Then

$$\mathrm{VE}^{u}_{I,jl} = 1 - \frac{\sum_k \left(P_{N_z}(\tilde{Y} \mid Z, S^{P_0}, \tilde{A} = k)_{z,j} - P_{N_z}(\tilde{Y} \mid Z, S^{P_0}, \tilde{A} = k)_{1,1}\right) P(\tilde{A} = k \mid S^{P_0} = u)}{\sum_k \left(P_{N_z}(\tilde{Y} \mid Z, S^{P_0}, \tilde{A} = k)_{z,l} - P_{N_z}(\tilde{Y} \mid Z, S^{P_0}, \tilde{A} = k)_{1,1}\right) P(\tilde{A} = k \mid S^{P_0} = u)}.$$

□

B.5 Kruskal rank properties

In the section that follows, we use properties and several theorems and lemmas that are proven in Kruskal (1977). Where appropriate we will indicate on which pages the proofs of the theorems and lemmas can be found.

Lemma 18 (Rank lemma)**.** *Let*

$$H_{AB}(n) = \min_{\mathrm{card}(A')=n} \{\mathrm{rank}(A') + \mathrm{rank}(B')\} - n$$

for an integer n where A' is an n-column subset of the matrix A and B' is the same column-index subset of a matrix B. For any diagonal matrix $D \in \mathbb{R}^{n \times n}$ with rank δ,

$$\mathrm{rank}(ADB^T) \geq H_{AB}(\delta).$$

See proof on p. 121 in Kruskal (1977).

B.5.1 Proof of Lemma 4

Proof. Suppose that $L = [A, B, C]$ and that $[\bar{A}, \bar{B}, \bar{C}]$ is another decomposition of L, where $\bar{B}, \bar{C}$ satisfy the respective row- and column-sum constraints. Let $r_{\bar{B}}, r_{\bar{C}}$ be the ranks of $\bar{B}$ and $\bar{C}$ respectively. Definition 3.2.5 implies that $A\text{diag}(xC)B^T = \bar{A}\text{diag}(x\bar{C})\bar{B}^T$ for all $x \in \mathbb{R}^{1\times I}$. If for any $y \in \mathbb{R}^{1\times K}$ such that $y\bar{C} = 0 \implies yC = 0$ then $\text{col}(C) \subset \text{col}(\bar{C})$, $\text{null}(C) \supset \text{null}(\bar{C})$, and $r_C \leq r_{\bar{C}}$. If $y\bar{C} = 0$ then

$$\bar{A}\text{diag}(y\bar{C})\bar{B}^T = 0 \implies A\text{diag}(yC)B^T = 0$$

Recall the definition of $H_{AB}(n)$ from Lemma 18. Kruskal (1977) shows that the condition on the ranks and Kruskal ranks above imply the following inequalities (proof omitted):

$$k_A \geq \max(R - r_B + 2, R - r_C + 2), \tag{B.24}$$
$$k_B \geq R - r_C + 2, \tag{B.25}$$
$$k_C \geq R - r_B + 2, \tag{B.26}$$
$$H_{AB}(n) \geq R - r_C + 2 \,\text{if}\, n \geq R - r_C + 2 \tag{B.27}$$
$$H_{AC}(n) \geq R - r_B + 2 \,\text{if}\, n \geq R - r_B + 2 \tag{B.28}$$
$$H_{BC}(n) \geq 1 \,\text{if}\, n \geq 1 \tag{B.29}$$

The inequality eq. (B.27) implies that when $H_{AB}(n) < R - r_C + 2$ then $n < R - r_C + 2$. When $n < R - r_C + 2$, the inequalities eqs. (B.24) to (B.26) and the definition of $H_{AB}(n)$ imply that $H_{AB}(n) = n$. Then

$$\begin{aligned} 0 &= \text{rank}(A\text{diag}(yC)B^T) \\ &\geq H_{AB}(\text{rank}(\text{diag}(yC))) \geq 0, \end{aligned}$$

where the second to last inequality comes from Lemma 18 and the last inequality comes from the definition of $H_{AB}(n)$. This implies $yC = 0$. Let the function $w(y)$ for a generic vector y return the number of nonzero entries in the vector y. Let v be any vector such that $w(v\bar{C}) \leq R - \bar{K}_0 + 1$. Then we'll show that $w(vC) \leq w(v\bar{C})$.

$$R - r_C + 1 \geq R - \bar{K}_0 + 1 \geq w(v\bar{C}) = \text{rank}(\text{diag}(v\bar{C}) \tag{B.30}$$
$$\geq \text{rank}(A\text{diag}(y\bar{C})\bar{B}^T) = \text{rank}(A\text{diag}(yC)B^T) \tag{B.31}$$
$$\geq H_{AB}(\text{rank}(\text{diag}(vC)) = H_{AB}(w(vC)). \tag{B.32}$$

The final line implies that $H_{AB}(w(vC)) = w(vC)$, which shows that $w(v\bar{C}) \geq w(vC)$ when $R - \bar{K}_0 + 1 \geq w(v\bar{C})$.

Given this condition, Kruskal's permutation lemma (proved on page 134 of Kruskal (1977)) shows that for any matrices C and $\bar{C}$ that satisfy the inequality, $\bar{C} = CP_CN$ where P_C is a permutation matrix and N is a diagonal nonsingular scaling matrix. If we have the stronger condition that every two columns of C are linearly independent then P_C and N are unique. Our matrices satisfy these conditions, so we have that $\bar{C} = CP_CN$, and a similar argument can be used to show $\bar{B} = BP_BM$

Given that we also have the condition that $\mathbf{1}_{1\times K}C = \mathbf{1}_{1\times R}$ and $\mathbf{1}_{1\times K}\bar{C} = \mathbf{1}_{1\times R}$, then this implies that $\bar{C} = CP_C$ because $\mathbf{1}_{1\times K}\bar{C} = \mathbf{1}_{1\times K}CP_CN = \mathbf{1}_{1\times R}N$ which only equals $\mathbf{1}_{1\times R}$ if $N = \mathbf{I}_{R\times R}$.

Furthermore, if $r_B = R$, the equation $B\nu = \mathbf{1}_{J\times 1}$ has a unique solution in $\nu \in \mathbb{R}^{R\times 1}$, namely $\nu = \mathbf{1}_{R\times 1}$. This implies that M is the identity matrix, as the condition $\bar{B}\mathbf{1}_{R\times 1} = \mathbf{1}_{J\times 1}$ results in:

$$\mathbf{1}_{J\times 1} = \bar{B}\mathbf{1}_{R\times 1} \tag{B.33}$$

$$= BP_BM\mathbf{1}_{R\times 1} \tag{B.34}$$

$$\implies P_BM\mathbf{1}_{R\times 1} = \mathbf{1}_{R\times 1}. \tag{B.35}$$

Given that M is a nonsingular diagonal matrix and P_B is a permutation matrix, M must be the identity to solve the equation $P_BM\mathbf{1}_{R\times 1} = \mathbf{1}_{R\times 1}$.

We now have $\bar{C} = CP_C$ and $\bar{B} = BP_B$. We can apply Kruskal's permutation matrix proof from pages 129-130 in Kruskal (1977) to show that $P_C = P_B = P$. The following two identities hold for any diagonal scaling matrices M, N, any permutation matrix P, and any vector v:

$$M\text{diag}(v)N = \text{diag}(vMN) \tag{B.36}$$

$$P\text{diag}(v)P^T = \text{diag}(vP^T). \tag{B.37}$$

Given Equations (B.36) to (B.37) and the condition that $L = [A, B, C] = [\bar{A}, \bar{B}, \bar{C}]$, then, for all vectors $v \in \mathbb{R}^{1\times J}$,

$$B\text{diag}(vA)C^T = \bar{B}\text{diag}(v\bar{A})\bar{C}^T \tag{B.38}$$

$$= BP\text{diag}(v\bar{A})P^TC^T \tag{B.39}$$

$$= B\text{diag}(v\bar{A}P^T)C^T. \tag{B.40}$$

The equality $B\text{diag}(vA)C^T = B\text{diag}(v\bar{A}P^T)C^T$ implies

$$B\text{diag}(v(A - \bar{A}P^T))C^T = 0 \tag{B.41}$$

for all v. Furthermore,

$$0 = \text{rank}(B\text{diag}(v(A - \bar{A}P^T))C^T) \tag{B.42}$$

$$\geq H_{BC}(\text{rank}(\text{diag}(v(A - \bar{A}P^T))) \geq 0. \tag{B.43}$$

The last line follows from Lemma 18. Then using the implication from eq. (B.29) that if $H_{BC}(n) < 1 \implies n = 0$, $\text{rank}(\text{diag}(v(A - \bar{A}P^T)) = 0$ or $v(A - \bar{A}P^T) = 0$ for all v. This further implies that

$$A = \bar{A}P^T$$

or

$$\bar{A} = AP.$$

□

B.6 Supporting lemmas and definitions from other work

Lemma 19 (Block rank lemmas Tian (2004))**.** *Let* $A \in \mathbb{R}^{m\times n}, B \in \mathbb{R}^{m\times k}, C \in \mathbb{R}^{l\times n}$.

$$\text{rank}\left(\begin{bmatrix} A & B \\ C & 0 \end{bmatrix}\right) = \text{rank}(B) + \text{rank}(C) + \text{rank}((I - BB^+)A(I - C^+C))$$

If $\text{range}(B) \subseteq \text{range}(A)$ *and* $\text{range}(C^T) \subseteq \text{range}(A^T)$

$$\text{rank}\left(\begin{bmatrix} A & B \\ C & D \end{bmatrix}\right) = \text{rank}(A) + \text{rank}(D - CA^+B)$$

Lemma 20 (Block rank lemma extension)**.** *Let* $A \in \mathbb{R}^{m\times n}, B \in \mathbb{R}^{m\times k}, C \in \mathbb{R}^{l\times n}$. *If* $\text{range}(C^T) \subseteq \text{range}(A^T)$

$$\text{rank}\left(\begin{bmatrix} A & A \\ C & D \end{bmatrix}\right) = \text{rank}(A) + \text{rank}(D - C)$$

Proof. Given that $\text{range}(A) \subseteq \text{range}(A)$, we can apply the second block rank lemma from

Lemma 19 with $B = A$.

$$\text{rank}\left(\begin{bmatrix} A & A \\ C & D \end{bmatrix}\right) = \text{rank}(A) + \text{rank}(D - CA^{+}A).$$

By supposition, $\text{range}(C^T) \subseteq \text{range}(A^T)$ and $A^{+}A$ is the projection matrix onto the column space of A^T. Then $CA^{+}A = C$, and the statement follows. □

B.7 Details behind numerical examples

We have three simulation scenarios where we vary the sample size to determine the power: a two-arm trial to determine vaccine efficacy against severe symptoms, a three-arm trial to determine relative vaccine efficacy against severe symptoms, and a two-arm trial to determine vaccine efficacy against transmission. All trials are designed such that the assumptions of Theorem 7 are satisfied, so the three-arm trial includes 8 study sites, and a categorical covariate with 7 levels, and both two-arm trials include 4 study sites, and a categorical covariate with 3 levels. Within each scenario, we allow for the categorical covariate, A, to be measured perfectly or with error. In addition, we assume a 3-level, pretreatment categorical covariate has been measured for each participant. We simulate from the parametric model defined in Section 3.3.2, which requires that we specify μ_u^r, or the log-odds of belonging to stratum u relative to base stratum u_0 for each study site r. Let the ordered collection of log-odds of being in stratum u relative to stratum $u_{2^{N_z}}$ for the reference covariate level $x = 1$ be $\mu^r = \left(\mu_{u_1}^r, \mu_{u_2}^r, \ldots, \mu_{u_{2^{N_z}-1}}^r, 0\right)$.

Let softmax be the function from $v \in \mathbb{R}^L$ to the $L+1$-dimensional probability simplex, defined elementwise for the i^{th} element as:

$$\text{softmax}(v)_i = \frac{e^{v_i}}{\sum_{l=1}^{L} e^{v_l^r}}$$

and let softmax^{-1} be the inverse function from $\theta \in$ the $L+1$-dimensional simplex to $\mathbb{R}^L$, where the i^{th} element, $i < L+1$ is defined as

$$\text{softmax}(\theta)_i^{-1} = \log(\theta_i) - \log(\theta_{L+1})$$

Let $\theta_u^{r,x} = P(S^{P_0} = u \mid R = r, X = x)$, and let $\theta^{r,x}$ be the ordered vector $(\theta_{u_1}^{r,x}, \theta_{u_1}^{r,x}, \ldots, \theta_{u_{2^{N_z}}}^{r,x})$. For the 2-arm trials, the population principal strata proportions are as follows:

$$\theta^{r,1} \overset{\text{iid}}{\sim} \text{Dirichlet}((91, 5, 0.5, 3.5)) \forall r$$

while for the 3-arm trials, the proportions are

$$\theta^{r,1} \overset{\text{iid}}{\sim} \text{Dirichlet}((91, 5, 0.1, 0.1, 0.1, 0.1, 0.1, 3.5))\forall r$$

These parameter settings roughly equate to a cumulative true incidence of 0.05. Recall from Section 3.3.2 that $\eta^x \in \mathbb{R}^{2^{N_z}}$, so η^x_u is the change in log-odds of belonging to principal stratum u vs. u_0 relative to $x = 1$. We set $\eta^x_{2^{N_z}} = 0$ for identifiability. Then let $\mu^r_u = \text{softmax}^{-1}(\theta^{r,1})$ and

$$\theta^{r,x} = \text{softmax}\left(\mu^r_u + \eta^x\right),$$

where for all $x > 1$

$$\eta^x_i \overset{\text{iid}}{\sim} \text{Normal}(0, 1),\ i < 2^{N_z},\ \eta^x_{2^{N_z}} = 0.$$

Let the N_a-vector $a^{u,x}$ be defined elementwise as $a^{u,x}_k$ where $a^{u,x}_k = P(A = k \mid S^{P_0} = u, X = x)$.

$$a^{u,1} \overset{\text{iid}}{\sim} \text{Dirichlet}(2\mathbf{1}_{N_a})\forall u,$$

and $\nu^u = \text{softmax}^{-1}(a^{u,1})$. Then recall that $\gamma^x \in \mathbb{R}^{N_a}$ such that γ^x_k is the change in log-odds of $A = k$ relative to $A = k_0$, and that $\gamma^x_{N_a} = 0$ for identifiability. Then

$$a^{u,x} = \text{softmax}\left(\nu^u + \gamma^x\right),$$

and for all $x > 1$

$$\gamma^x_i \overset{\text{iid}}{\sim} \text{Normal}(0, 1),\ i < N_a,\ \gamma^x_{N_a} = 0.$$

Finally, recall that

$$\log \frac{P(Y(z_j) = 1 \mid S^{P_0} = u, A = k, X = x)}{P(Y(z_j) = 0 \mid S^{P_0} = u, A = k, X = x)} = \alpha^u_j + \delta^u_{j,k} + \omega^x_j,$$

where ω^x_j is the change in log-odds of $Y(z_j) = 1$, all else being equal, compared to $x = 1$. In all of our simulations, $\omega^x_j = (x - 1)\log(1.1)$ for all j. For the 2-arm trial example, we let $\alpha^{(1,1)}_1 = \log(0.3/0.7), \alpha^{(1,1)}_2 = \log(0.3/0.7) + \log(0.4)$, and $\delta^{(1,1)}_{1,k} = (k - 1)\log(0.925), \delta^{(1,1)}_{2,k} = (k - 1)\log(0.825)$. Further, we let $\alpha^{(1,0)}_1 = \log(0.15/0.85)$, $\alpha^{(0,1)}_2 = \log(0.2/0.8)$, and $\delta^{(1,0)}_{1,k} = (k - 1)\log(0.925)$, and $\delta^{(0,1)}_{2,k} = 0$

For the 3-arm trial example, we let $\alpha^{(1,1,1)}_1 = \log(0.3/0.7), \alpha^{(1,1,1)}_2 = \log(0.3/0.7), \alpha^{(1,1,1)}_3 = \log(0.3/0.7) + \log(0.4)$, and $\delta^{(1,1,1)}_{j,k} = (k - 1)\log(0.925)$ for $j = 1, 2, 3$. Further, we let $\alpha^{(1,0,1)}_1 = \log(0.2/0.8)$, $\alpha^{(1,0,1)}_3 = \log(0.1/0.9)$,$\alpha^{(1,1,0)}_1 = \log(0.3/0.7)$, $\alpha^{(1,1,0)}_2 =$

$\log(0.15/0.85)$,$\alpha_2^{(0,1,1)} = \log(0.25/0.75)$, $\alpha_3^{(0,1,1)} = \log(0.08/0.92)$, $\alpha_3^{(0,0,1)} = \log(0.25/0.75)$, $\alpha_2^{(0,1,0)} = \log(0.25/0.75)$, $\alpha_1^{(1,0,0)} = \log(0.1/0.9)$ and $\delta_{j,k}^u = 0$ for all k, $u \in \{(1,0,0),(0,1,0),(1,1,0),(0,0,1),(1,0,1),(0,1,1)\}$, and all allowable j

In the 2- and 3-arm trial examples that pertain to inferring vaccine efficacy against severe symptoms, we set $\text{sn}_S = 0.8, \text{sp}_S = 0.99$ which reflects the sensitivity and specificity of a typical PCR collected via nasopharyngeal swab (Kissler et al., 2021), and $\text{sn}_Y = 0.99, \text{sp}_S = 0.9$ to reflect the fact that most severe illness caused by the pathogen of interest will be reported, but that there are many severe illness episodes that are reported that may be caused by other pathogens. These lead to a true rate of severe illness of 0.01 but a rate of reported severe illness of 0.1. For comparison Monto et al. (2009) symptom reporting data shows that 10% of participants reported at least one severe symptom, but the cumulative incidence was 0.07.

For the transmission study, we use the same settings for sn_S, sp_S and set $\text{sn}_Y = \text{sn}_S, \text{sp}_Y = \text{sp}_S$. In order to generate $\tilde{A} \mid A$ for each participant, we use the following error model for the three arm trial:

$$P(\tilde{A} = a \mid A = a) = 0.5,\ P(\tilde{A} = a+1 \mid A = a) = P(\tilde{A} = a-1 \mid A = a) = 0.25$$

for $a \in \{2, \ldots, 6\}$. When $a = 1$:

$$P(\tilde{A} = a \mid A = a) = 0.5,\ P(\tilde{A} = a+1 \mid A = a) = 0.5,$$

and when $a = 7$

$$P(\tilde{A} = a \mid A = a) = 0.5,\ P(\tilde{A} = a-1 \mid A = a) = 0.5, a = 7.$$

For the two-arm trials, we generate $\tilde{A} \mid A$ from the following probability model when $a = 2$:

$$P(\tilde{A} = a \mid A = a) = 0.875,\ P(\tilde{A} = a+1 \mid A = a) = P(\tilde{A} = a-1 \mid A = a) = 0.0625.$$

When $a = 1$

$$P(\tilde{A} = a \mid A = a) = 0.95,\ P(\tilde{A} = a+1 \mid A = a) = 0.05,$$

and when $a = 3$

$$P(\tilde{A} = a \mid A = a) = 0.95,\ P(\tilde{A} = a-1 \mid A = a) = 0.05.$$

These distributions were chosen to reflect the fact that detailed pre-season antibody titer measurements are typically available for participants in influenza vaccination trials. Further discretizing

the titer measurements reduces the misclassification probabilities in our model. Let the collection of these conditional probabilities be $p^a_{N_a}$

For each hypothetical participant in a study site $R = r$ in our study we draw data in the following manner

$$\begin{aligned}
Z_i &\overset{\text{iid}}{\sim} \text{Categorical}(\frac{1}{N_z}\mathbf{1}_{N_z}) \\
X_i &\overset{\text{iid}}{\sim} \text{Categorical}(\frac{1}{3}\mathbf{1}_3) \\
S_i^{P_0} \mid R = r, X = x &\overset{\text{iid}}{\sim} \text{Categorical}(\theta^{r,x}) \\
A_i \mid S^{P_0} = u, X = x &\overset{\text{iid}}{\sim} \text{Categorical}(a^{u,x}) \\
Y_i \mid S^{P_0} = u, A = k, X = x, Z = j &\overset{\text{iid}}{\sim} \text{Bernoulli}(\text{inv_logit}(\alpha^u_j + \delta^u_{j,k} + \omega_x)) \\
\tilde{Y}_i \mid Y = y &\overset{\text{iid}}{\sim} \text{Bernoulli}(y\text{sn}_Y + (1-y)(1-\text{sp}_Y)) \\
\tilde{S}_i \mid S^{P_0} = u, Z = j &\overset{\text{iid}}{\sim} \text{Bernoulli}(u_j\text{sn}_S + (1-u_j)(1-\text{sp}_S)) \\
\tilde{A}_i \mid A = a &\overset{\text{iid}}{\sim} \text{Categorical}(p^a_{N_a})
\end{aligned} \tag{B.44}$$

and we do this for all sites $R \in \{1, \ldots, N_r\}$.

We fit the model defined in Equation (3.11). Recall

$$A_i \mid S_i^{P_0} = u, X_i = x \sim \text{Categorical}(\boldsymbol{\pi}_{u,x}) \tag{B.45}$$

$$S_i^{P_0} \mid R_i = r, X_i = x \sim \text{Categorical}(\boldsymbol{\rho}_{r,x}) \tag{B.46}$$

$$Y_i(z_j) \mid S_i^{P_0} = u, A_i = k, X_i = x \sim \text{Bernoulli}(\beta^{u,x}_{j,k}) \tag{B.47}$$

We use the following priors:

$$\begin{aligned}
\text{sn}_S &\sim \text{Beta}(0.5, 1, 4, 2) \\
\text{sp}_S &\sim \text{Beta}(0.5, 1, 10, 2) \\
\text{sn}_Y &\sim \text{Beta}(0.5, 1, 4, 2) \\
\text{sp}_Y &\sim \text{Beta}(0.5, 1, 4, 2) \\
\boldsymbol{\pi}_{u,x} &\overset{\text{iid}}{\sim} \text{Dirichlet}(3\mathbf{1}_{N_a}),\ \forall u, x \\
\boldsymbol{\rho}_{r,x} &\overset{\text{iid}}{\sim} \text{Dirichlet}((80, 1.5, 0.5\mathbf{1}^T_{N_U-3}, 1)^T),\ \forall r, x \\
\beta^{u,x}_{j,k} &\overset{\text{iid}}{\sim} \text{Uniform}(0, 1),\ \forall u, k, x
\end{aligned} \tag{B.48}$$

where $\text{Beta}(0.5, 1, 4, 2)$ is the shifted, scaled Beta distribution in which the first two arguments define the support of the distribution, and the second two parameters are shape parameters. For

example, for sn_S this corresponds to $\chi \sim \text{Beta}(4, 2)$ and $\text{sn}_S = (1 - 0.5)\chi + 0.5$.

For fitting scenarios in which Covariate homogeneity does not hold, we used a parametric model for A_i shown in Equation (3.12). It is shown here below

$$\log \frac{P(A_i = k \mid R_i = r, S_i^{P_0} = u, X_i = x)}{P(A_i = k_0 \mid R_i = r, S_i^{P_0} = u, X_i = x)} = \nu_k^u + \gamma_k^x + \gamma_k^{u,x} + \epsilon_k^{u,r}, \quad \epsilon_k^{u,r} \sim \text{Normal}(0, (\tau_\epsilon^u)^2) \, \forall r.$$

We fix $\tau_\epsilon^u \to \infty$ for all u, resulting in a nonpooled model for ϵ. The following priors are used for inference:

$$\begin{aligned}
\nu_k^u &\sim \text{Normal}(0, 1.7^2), \forall u \in \{u_1, u_2, u_{2^{N_z}}\}, 1 \leq k < N_a \\
\nu_k^u &\sim \text{Normal}(0, 0.5^2), \forall u \notin \{u_1, u_2, u_{2^{N_z}}\}, 1 \leq k < N_a \\
\gamma_k^x &\sim \text{Normal}(0, 0.5^2), 2 \leq x \leq 3, \, 1 \leq k < N_a \\
\gamma_k^{u,x} &\sim \text{Normal}(0, 0.5^2), 2 \leq x \leq 3, \, 1 \leq k < N_a, \, u \in \mathcal{S} \setminus u_0 \\
\epsilon_k^{u,r} &\sim \text{Normal}(0, 1), 2 \leq r \leq N_r, \, 1 \leq k < N_a, \, u \in \mathcal{S} \setminus u_0.
\end{aligned} \tag{B.49}$$

We use Stan for inference (Team, 2021) using the `cmdstanr` package (Gabry and Češnovar, 2022) in `R` (R Core Team, 2022). All models were run for 3,000 warmup and 3,000 post-warmup iterations; all $\hat{R}$ statistics (Gelman et al., 2013) were below 1.01, as recommended by Vehtari et al. (2020). The bulk and tail effective sample sizes were greater than 9% of samples for all models.

Some misspecified models did not achieve $\hat{R} < 1.01$. This indicated multimodal posteriors rather than lack of convergence. The summary of these results is in Table B.1.

Table B.1: Number of simulations with $\hat{R} > 1.01$. Null and alternative scenarios are combined, resulting in 200 simulations for each scenario.

Model	Measurements	20,000	40,000	80,000
A Incorrect	A	0	3	58
	$\tilde{A}$	0	0	5
A Correct	A	0	0	1
	$\tilde{A}$	0	1	2

We did not use these models for inference, as we expect that stronger priors would need to be used to constrain the models to a single mode.

In order to ensure that our decision rule did not lead to high Type 1 error rates, we simulated data under the null hypothesis that vaccine efficacy against post-infection outcomes was 0. The results presented in Table B.2 show that the Type 1 error rates are less that 0.05 for each scenario.

Table B.2: Type 1 error rates for vaccine efficacy against severe illness designs

Trial	A meas.	4,000	20,000	40,000	80,000	120,000
3-arm	A	NA	NA	0.01	0.05	0.03
	$\tilde{A}$	NA	NA	0.00	0.02	0.00
2-arm	A	0.00	0.01	0.00	0.02	NA
	$\tilde{A}$	0.00	0.00	0.01	0.01	NA

The Type 1 error rates for vaccine efficacy against transmission are presented in Table B.3. None of the Type 1 error rates are statistically significantly greater than 0.05.

Table B.3: Type 1 error rates for vaccine efficacy against transmission designs

A	0.00	0.05	0.03	0.01
$\tilde{A}$	0.00	0.06	0.03	0.05

Table B.4: Power of the test,$\mathrm{VE}_S \approxeq 0.5$, $\mathrm{VE}_{I,21}^{(1,1)} \approxeq 0.6$ for $N_z = 2$ for sample sizes of 20,000 through 80,000. Scenarios in which A_i was measured with error denoted by $\tilde{A}$, A otherwise.

Model	Measurements	20,000	40,000	80,000
A Incorrect	A	0.44	0.79	0.99
	$\tilde{A}$	0.33	0.66	0.87
A Correct	A	0.42	0.64	0.86
	$\tilde{A}$	0.36	0.66	0.85

B.8 Alternative identifiability scenarios

Theorem 7 is a set of sufficient conditions for the identifiability of VE_I. There are alternative conditions for identifiability, however, that do not rely on Kruskal rank conditions. The following is such an example.

Theorem 21 (Two-arm trial, binary covariate)**.** *Let A_i be a binary covariate observed for each participant in a multi-site randomized trail. Suppose Assumptions 1 to 3 hold and that there are at least five study sites. Given conditions on the covariate distribution conditional on principal stratum, VE_I is identifiable.*

Proof. Let $a^u = P(A_i = 1 \mid S_i^{P_0} = u)$, $p_{s+kzr} = P(\tilde{S}_i = s, A_i = k \mid Z_i = z, R_i = r)$, and $r_S = \text{sn}_S + \text{sp}_S - 1$.

$$\begin{aligned} p_{1+10r} &= r_S(\theta^r_{(0,1)}a^{(0,1)} + \theta^r_{(1,1)}a^{(1,1)}) + (1 - \text{sp}_S)p_{++10r} \\ p_{1+11r} &= r_S(\theta^r_{(1,0)}a_{(1,0)} + \theta^r_{(1,1)}a^{(1,1)}) + (1 - \text{sp}_S)p_{++11r} \end{aligned} \tag{B.50}$$

and that

$$\begin{aligned} \theta^r_{(0,1)} &= \frac{p_{1++0r} - (1 - \text{sp}_S)}{r_S} - \theta^r_{(1,1)} \\ \theta^r_{(1,0)} &= \frac{p_{1++1r} - (1 - \text{sp}_S)}{r_S} - \theta^r_{(1,1)} \end{aligned} \tag{B.51}$$

then

$$\begin{aligned} p_{1+10r} &= r_S(\frac{p_{1++0r} - (1 - \text{sp}_S)}{r_S}a^{(0,1)} + \theta^r_{(1,1)}(a^{(1,1)} - a^{(0,1)})) + (1 - \text{sp}_S)p_{++10r} \\ p_{1+11r} &= r_S(\frac{p_{1++1r} - (1 - \text{sp}_S)}{r_S}a^{(1,0)} + \theta^r_{(1,1)}(a^{(1,1)} - a^{(1,0)})) + (1 - \text{sp}_S)p_{++11r} \end{aligned} \tag{B.52}$$

so

$$\theta^r_{(1,1)} = \frac{\frac{p_{1+11r} - (1-\text{sp}_S)p_{++11r}}{r_S} - \frac{p_{1++1r} - (1-\text{sp}_S)}{r_S}a^{(1,0)}}{(a^{(1,1)} - a^{(1,0)})} \tag{B.53}$$

and

$$\begin{aligned} p_{1+10r} = (p_{1++0r} - (1 - \text{sp}_S))a^{(0,1)} + (p_{1+11r} - (1 - \text{sp}_S)p_{++11r})\frac{a^{(1,1)} - a^{(0,1)}}{a^{(1,1)} - a^{(1,0)}} \\ - (p_{1++1r} - (1 - \text{sp}_S))a^{(1,0)}\frac{a^{(1,1)} - a^{(0,1)}}{a_{(1,1)} - a^{(1,0)}} + (1 - \text{sp}_S)p_{++10r} \end{aligned}$$

expanding terms

$$\begin{aligned} p_{1+10r} = p_{1++0r}a^{(0,1)} - (1 - \text{sp}_S)a^{(0,1)} + p_{1+11r}\frac{a^{(1,1)} - a^{(0,1)}}{a^{(1,1)} - a^{(1,0)}} - (1 - \text{sp}_S)p_{++11r}\frac{a^{(1,1)} - a^{(0,1)}}{a^{(1,1)} - a^{(1,0)}} \\ - p_{1++1r}a^{(1,0)}\frac{a^{(1,1)} - a^{(0,1)}}{a^{(1,1)} - a^{(1,0)}} + (1 - \text{sp}_S)a^{(1,0)}\frac{a^{(1,1)} - a^{(0,1)}}{a^{(1,1)} - a^{(1,0)}} + (1 - \text{sp}_S)p_{++10r} \end{aligned}$$

and collecting terms

$$\begin{aligned}
p_{1+10r} = p_{1++0r}a^{(0,1)} + p_{1+11r}\frac{a^{(1,1)}-a^{(0,1)}}{a^{(1,1)}-a^{(1,0)}} - p_{1++1r}a^{(1,0)}\frac{a^{(1,1)}-a^{(0,1)}}{a^{(1,1)}-a^{(1,0)}} \\
- (1-\mathrm{sp}_S)p_{++11r}\frac{a^{(1,1)}-a^{(0,1)}}{a^{(1,1)}-a^{(1,0)}} - (1-\mathrm{sp}_S)a^{(0,1)} + (1-\mathrm{sp}_S)a^{(1,0)}\frac{a^{(1,1)}-a^{(0,1)}}{a^{(1,1)}-a^{(1,0)}} \\
+ (1-\mathrm{sp}_S)p_{++10r}
\end{aligned}$$

Note that $p_{++10r} = p_{++11r}$ so

$$\begin{aligned}
p_{1+10r} = p_{1++0r}a^{(0,1)} + p_{1+11r}\frac{a^{(1,1)}-a^{(0,1)}}{a^{(1,1)}-a^{(1,0)}} - p_{1++1r}a^{(1,0)}\frac{a^{(1,1)}-a^{(0,1)}}{a^{(1,1)}-a^{(1,0)}} \\
(1-\mathrm{sp}_S)p_{++11r}\frac{a^{(0,1)}-a^{(1,0)}}{a^{(1,1)}-a^{(1,0)}} - (1-\mathrm{sp}_S)a^{(0,1)} + (1-\mathrm{sp}_S)a^{(1,0)}\frac{a^{(1,1)}-a^{(0,1)}}{a^{(1,1)}-a^{(1,0)}}
\end{aligned}$$

If specificity sp_S is unknown, we need at least 5 groups to solve the above equations

$$\begin{aligned}
p_{1+10r_1} - p_{1+10r_5} &= (p_{1++0r_1} - p_{1++0r_5})a^{(0,1)} + (p_{1+11r_1} - p_{1+11r_5})\frac{a^{(1,1)}-a^{(0,1)}}{a^{(1,1)}-a^{(1,0)}} \\
&\quad - (p_{1++1r_1} - p_{1++1r_5})a^{(1,0)}\frac{a^{(1,1)}-a^{(0,1)}}{a^{(1,1)}-a^{(1,0)}} + (p_{++11r_1} - p_{++11r_5})(1-\mathrm{sp}_S)\frac{a^{(0,1)}-a^{(1,0)}}{a^{(1,1)}-a^{(1,0)}} \\
p_{1+10r_2} - p_{1+10r_5} &= (p_{1++0r_2} - p_{1++0r_5})a^{(0,1)} + (p_{1+11r_2} - p_{1+11r_5})\frac{a^{(1,1)}-a^{(0,1)}}{a^{(1,1)}-a^{(1,0)}} \\
&\quad - (p_{1++1r_2} - p_{1++1r_5})a^{(1,0)}\frac{a^{(1,1)}-a^{(0,1)}}{a^{(1,1)}-a^{(1,0)}} + (p_{++11r_2} - p_{++11r_5})(1-\mathrm{sp}_S)\frac{a^{(0,1)}-a^{(1,0)}}{a^{(1,1)}-a^{(1,0)}} \\
p_{1+10r_3} - p_{1+10r_5} &= (p_{1++0r_3} - p_{1++0r_5})a^{(0,1)} + (p_{1+11r_3} - p_{1+11r_5})\frac{a^{(1,1)}-a^{(0,1)}}{a^{(1,1)}-a^{(1,0)}} \\
&\quad - (p_{1++1r_3} - p_{1++1r_5})a^{(1,0)}\frac{a^{(1,1)}-a^{(0,1)}}{a^{(1,1)}-a^{(1,0)}} + (p_{++11r_3} - p_{++11r_5})(1-\mathrm{sp}_S)\frac{a^{(0,1)}-a^{(1,0)}}{a^{(1,1)}-a^{(1,0)}} \\
p_{1+10r_4} - p_{1+10r_5} &= (p_{1++0r_4} - p_{1++0r_5})a^{(0,1)} + (p_{1+11r_4} - p_{1+11r_5})\frac{a^{(1,1)}-a^{(0,1)}}{a^{(1,1)}-a^{(1,0)}} \\
&\quad - (p_{1++1r_4} - p_{1++1r_5})a^{(1,0)}\frac{a^{(1,1)}-a^{(0,1)}}{a^{(1,1)}-a^{(1,0)}} + (p_{++11r_4} - p_{++11r_5})(1-\mathrm{sp}_S)\frac{a^{(0,1)}-a^{(1,0)}}{a^{(1,1)}-a^{(1,0)}}
\end{aligned}$$

Let $\Delta_{1+10r_1} = p_{1+10r_1} - p_{1+10r_5}$ and other differences similarly defined. This results in the linear equation

$$\begin{bmatrix} \Delta_{1+10r_1} \\ \Delta_{1+10r_2} \\ \Delta_{1+10r_3} \\ \Delta_{1+10r_4} \end{bmatrix} = \begin{bmatrix} \Delta_{1++0r_1} & \Delta_{1+11r_1} & -\Delta_{1++1r_1} & \Delta_{++11r_1} \\ \Delta_{1++0r_2} & \Delta_{1+11r_2} & -\Delta_{1++1r_2} & \Delta_{++11r_2} \\ \Delta_{1++0r_3} & \Delta_{1+11r_3} & -\Delta_{1++1r_3} & \Delta_{++11r_3} \\ \Delta_{1++0r_4} & \Delta_{1+11r_4} & -\Delta_{1++1r_4} & \Delta_{++11r_4} \end{bmatrix} \begin{bmatrix} a^{(0,1)} \\ \frac{a^{(1,1)}-a^{(0,1)}}{a^{(1,1)}-a^{(1,0)}} \\ a^{(1,0)}\frac{a^{(1,1)}-a^{(0,1)}}{a^{(1,1)}-a^{(1,0)}} \\ (1-\mathrm{sp}_S)\frac{a^{(0,1)}-a^{(1,0)}}{a^{(1,1)}-a^{(1,0)}} \end{bmatrix}.$$

Provided that the coefficient matrix is invertible, we can solve the equation:

$$\begin{bmatrix} \Delta_{1++0r_1} & \Delta_{1+11r_1} & -\Delta_{1++1r_1} & \Delta_{++11r_1} \\ \Delta_{1++0r_2} & \Delta_{1+11r_2} & -\Delta_{1++1r_2} & \Delta_{++11r_2} \\ \Delta_{1++0r_3} & \Delta_{1+11r_3} & -\Delta_{1++1r_3} & \Delta_{++11r_3} \\ \Delta_{1++0r_4} & \Delta_{1+11r_4} & -\Delta_{1++1r_4} & \Delta_{++11r_4} \end{bmatrix}^{-1} \begin{bmatrix} \Delta_{1+10r_1} \\ \Delta_{1+10r_2} \\ \Delta_{1+10r_3} \\ \Delta_{1+10r_4} \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix}.$$

This yields solutions $\hat{a}^{(1,0)}, \hat{a}^{(1,1)}, \hat{a}^{(0,1)}, \hat{\text{sp}}_S$ in terms of observable probabilities:

$$\hat{a}^{(0,1)} = x_1 \tag{B.54}$$

$$\hat{a}^{(1,0)} = \frac{x_3}{x_2} \tag{B.55}$$

$$\hat{a}^{(1,1)} = \frac{x_1 - x_3/x_2}{1 - x_2} \tag{B.56}$$

$$\hat{\text{sp}}_S = 1 - x_4 \frac{x_1 + x_3 - 2\frac{x_3}{x_2}}{(1 - x_2)(x_1 - \frac{x_3}{x_2})} \tag{B.57}$$

provided that $x_2 \neq 0, x_2 \neq 1, x_1 \neq \frac{x_3}{x_2}$. These conditions imply that $x_1 + x_3 - 2\frac{x_3}{x_2} \neq 0$.

With these solutions, we may solve for $\widehat{r_s\theta^r}_{(1,1)}, \widehat{r_s\theta^r}_{(0,1)}, \widehat{r_s\theta^r}_{(1,0)}$ for all r:

$$\widehat{r_S\theta^r}_{(1,1)} = \frac{(p_{1+11r} - p_{1++1r}\frac{x_3}{x_2})(1 - x_2)}{x_1 + x_3 - 2\frac{x_3}{x_2}} - x_4 \frac{p_{++11r} + \frac{x_3}{x_2}}{x_1 - \frac{x_3}{x_2}} \tag{B.58}$$

$$\begin{aligned} \widehat{r_S\theta^r}_{(0,1)} &= p_{1++0r} - x_4 \frac{x_1 + x_3 - 2\frac{x_3}{x_2}}{(1 - x_2)(x_1 - \frac{x_3}{x_2})} - \frac{(p_{1+11r} - p_{1++1r}\frac{x_3}{x_2})(1 - x_2)}{x_1 + x_3 - 2\frac{x_3}{x_2}} + x_4 \frac{p_{++11r} + \frac{x_3}{x_2}}{x_1 - \frac{x_3}{x_2}} \\ \widehat{r_S\theta^r}_{(1,0)} &= p_{1++1r} - x_4 \frac{x_1 + x_3 - 2\frac{x_3}{x_2}}{(1 - x_2)(x_1 - \frac{x_3}{x_2})} - \frac{(p_{1+11r} - p_{1++1r}\frac{x_3}{x_2})(1 - x_2)}{x_1 + x_3 - 2\frac{x_3}{x_2}} + x_4 \frac{p_{++11r} + \frac{x_3}{x_2}}{x_1 - \frac{x_3}{x_2}} \end{aligned} \tag{B.59}$$

and

$$\begin{aligned} r_S - \widehat{r_S\theta^r}_{(0,0)} = p_{1++1r} + p_{1++0r} - 2x_4 &\frac{x_1 + x_3 - 2\frac{x_3}{x_2}}{(1 - x_2)(x_1 - \frac{x_3}{x_2})} \\ - &\frac{(p_{1+11r} - p_{1++1r}\frac{x_3}{x_2})(1 - x_2)}{x_1 + x_3 - 2\frac{x_3}{x_2}} + x_4 \frac{p_{++11r} + \frac{x_3}{x_2}}{x_1 - \frac{x_3}{x_2}} \end{aligned} \tag{B.60}$$

Finally, let

$$p_{0+10r} = (1-\mathrm{sn}_S)(\theta^r_{(0,1)}a^{(0,1)} + \theta^r_{(1,1)}a^{(1,1)}) + \mathrm{sp}_S(\theta^r_{(1,0)}a^{(1,0)} + \theta^r_{(0,0)}a^{(0,0)}) \tag{B.61}$$

$$= (1-\mathrm{sn}_S)p_{++10r} + r_S(\theta^r_{(1,0)}a^{(1,0)} + \theta^r_{(0,0)}a^{(0,0)}) \tag{B.62}$$

$$= (1-\mathrm{sn}_S)p_{++10r} + r_S(\theta^r_{(1,0)}a^{(1,0)} + (1-\theta^r_{(1,0)} - \theta^r_{(1,1)} - \theta^r_{(0,1)})a^{(0,0)}) \tag{B.63}$$

where we take advantage of $\theta^r_{(0,0)} = (1-\theta^r_{(1,1)} - \theta^r_{(0,1)} - \theta^r_{(1,0)})$. Then we can solve system of equations

$$\begin{bmatrix} p_{0+10r_1} - \widehat{r_S\theta^{r_1}}_{(1,0)}\hat{a}^{(1,0)} \\ p_{0+10r_2} - \widehat{r_S\theta^{r_2}}_{(1,0)}\hat{a}^{(1,0)} \\ p_{0+10r_3} - \widehat{r_S\theta^{r_3}}_{(1,0)}\hat{a}^{(1,0)} \end{bmatrix} = \begin{bmatrix} p_{++10r_1} & 1 & -\widehat{r_S\theta^{r_1}}_{(1,0)} - \widehat{r_S\theta^{r_1}}_{(0,1)} - \widehat{r_S\theta^{r_1}}_{(1,1)} \\ p_{++10r_2} & 1 & -\widehat{r_S\theta^{r_2}}_{(1,0)} - \widehat{r_S\theta^{r_2}}_{(0,1)} - \widehat{r_S\theta^{r_2}}_{(1,1)} \\ p_{++10r_3} & 1 & -\widehat{r_S\theta^{r_3}}_{(1,0)} - \widehat{r_S\theta^{r_3}}_{(0,1)} - \widehat{r_S\theta^{r_3}}_{(1,1)} \end{bmatrix} \begin{bmatrix} (1-\mathrm{sn}_S) \\ r_s a^{(0,0)} \\ a^{(0,0)} \end{bmatrix} \tag{B.64}$$

as long as

$$\det\left(\begin{bmatrix} p_{++10r_1} & 1 & -\widehat{r_S\theta^{r_1}}_{(1,0)} - \widehat{r_S\theta^{r_1}}_{(0,1)} - \widehat{r_S\theta^{r_1}}_{(1,1)} \\ p_{++10r_2} & 1 & -\widehat{r_S\theta^{r_2}}_{(1,0)} - \widehat{r_S\theta^{r_2}}_{(0,1)} - \widehat{r_S\theta^{r_2}}_{(1,1)} \\ p_{++10r_3} & 1 & -\widehat{r_S\theta^{r_3}}_{(1,0)} - \widehat{r_S\theta^{r_3}}_{(0,1)} - \widehat{r_S\theta^{r_3}}_{(1,1)} \end{bmatrix}\right) \neq 0$$

$$\begin{bmatrix} p_{++10r_1} & 1 & -\widehat{r_S\theta^{r_1}}_{(1,0)} - \widehat{r_S\theta^{r_1}}_{(0,1)} - \widehat{r_S\theta^{r_1}}_{(1,1)} \\ p_{++10r_2} & 1 & -\widehat{r_S\theta^{r_2}}_{(1,0)} - \widehat{r_S\theta^{r_2}}_{(0,1)} - \widehat{r_S\theta^{r_2}}_{(1,1)} \\ p_{++10r_3} & 1 & -\widehat{r_S\theta^{r_3}}_{(1,0)} - \widehat{r_S\theta^{r_3}}_{(0,1)} - \widehat{r_S\theta^{r_3}}_{(1,1)} \end{bmatrix}^{-1} \begin{bmatrix} p_{0+10r_1} - \widehat{r_S\theta^{r_1}}_{(1,0)}\hat{a}^{(1,0)} \\ p_{0+10r_2} - \widehat{r_S\theta^{r_2}}_{(1,0)}\hat{a}^{(1,0)} \\ p_{0+10r_3} - \widehat{r_S\theta^{r_3}}_{(1,0)}\hat{a}^{(1,0)} \end{bmatrix} = \begin{bmatrix} x_5 \\ x_6 \\ x_7 \end{bmatrix} \tag{B.65}$$

Then

$$\hat{a}^{(0,0)} = x_7 \tag{B.66}$$

$$\hat{r}_s = x_6/x_7 \mid x_7 \neq 0 \tag{B.67}$$

$$\hat{\mathrm{sn}}_S = 1 - x_5 \tag{B.68}$$

With these solutions, we can back out $\theta^r_{(1,1)}, \theta^r_{(0,1)}, \theta^r_{(1,0)}$. With solutions for $\theta^r_{(1,1)}, \theta^r_{(0,1)}, \theta^r_{(1,0)}$ in terms of observable probabilities, we can use the technique in Appendix B.4 to infer VE_I. □

B.9 Derivation of transparent reparameterization prior

Let $r \mid \text{sp}_Y \sim \text{Uniform}(\text{sp}_Y - 0.5, \text{sp}_Y)$, and let $\beta_{j,k}^{u,x} \sim \text{Uniform}(0, 1)$. Let r be the dummy variable for r_Y, let p_i be the dummy variable for a single $\beta_{j,k}^{u,x}$.

Suppose we have only p_1, p_2. We wish to compute the distribution of $c_1 = rp_1$ and $c_2 = rp_2$. Thus we have the transformation: $r = v, p_1 = c_1/v$ and $p_2 = c_2/v$. The Jacobian:

$$J_2 = \begin{array}{c} \begin{matrix} \frac{\partial}{\partial c_1} & \frac{\partial}{\partial c_2} & \frac{\partial}{\partial v} \end{matrix} \\ \begin{bmatrix} 0 & 0 & 1 \\ \frac{1}{v} & 0 & -\frac{c_1}{v^2} \\ -\frac{c_c v}{c_1^2} & \frac{v}{c_1} & \frac{c_2}{c_1} \end{bmatrix} \end{array} \begin{matrix} p_1 \\ r \\ p_2 \end{matrix}$$

Then $|\det J_2| = \frac{1}{v^2}$.

Suppose $i \in \{1, \ldots, n\}$, and $|\det J_n| = \frac{1}{v^n}$. Let $c_{n+1} = rp_{n+1}$. Then $p_{n+1} = \frac{c_{n+1}}{v}$. The partial derivatives of p_{n+1} are:

$$\frac{\partial p_{n+1}}{\partial c_i} = 0 \forall i \neq n+1 \tag{B.69}$$

$$\frac{\partial p_{n+1}}{\partial v} = \frac{-c_{n+1}}{v^2} \tag{B.70}$$

Collect the partial derivatives of p_{n+1} with respect to the first n parameters and v into a $n+1$ length vector $\nu = (\mathbf{0}_n^T, -\frac{c_{n+1}}{v^2})^T$ The partials of the other parameters with respect to c_{n+1} are all zero. Then J_{n+1}

$$J_{n+1} = \begin{bmatrix} J_n & \mathbf{0}_{n+1} \\ \nu & \frac{1}{v} \end{bmatrix}$$

so $|\det J_{n+1}| = \left|\det J_n \frac{1}{v}\right| = \frac{1}{v^{n+1}}$. Thus the absolute determinant of the Jacobian for the transformation with n p_i is $\frac{1}{v^n}$.

The constraints are simple.

$$\text{sp} - 1/2 \leq r \leq \text{sp} \tag{B.71}$$

$$0 \leq p_i \leq 1 \tag{B.72}$$

become

$$\mathrm{sp} - 1/2 \leq v \leq \mathrm{sp} \tag{B.73}$$

$$0 \leq \frac{c_i}{v} \leq 1. \tag{B.74}$$

This yields the combined bounds on v:

$$\max(\mathrm{sp} - 1/2, \max_n c_i) \leq v \leq \mathrm{sp} \tag{B.75}$$

The joint prior is:

$$f(c_1, \ldots, c_n) \propto \int_{\max(\mathrm{sp}-1/2, \max_n c_i)}^{\mathrm{sp}} v^{-n} dv \tag{B.76}$$

$$\propto \max\left(\mathrm{sp}_Y - 1/2, \max_n c_i\right)^{1-n_\beta} - \mathrm{sp}_Y^{1-n_\beta} \tag{B.77}$$

APPENDIX C

Cumulative exposure to environmental hazards appendix

C.1 Approximate integral proof

Lemma 22. *Approximation of log-Gaussian integral Let* $\mathcal{K} : \mathbb{R}^+ \to (0, 1]$ *be a continuously differentiable function, and let* $Z(c, \tau)$ *be a GP with domain* $(\mathcal{C} \times \mathbb{R}^+)$*, with a valid covariance function* σ *(validity as defined in (Møller et al., 1998, p. 453)). Let the target integral over domain* $\mathcal{C} \times t$ *be*

$$\int_{\mathcal{C}\times t} \mathcal{K}\left(\frac{\|\ell(c)-s_i\|_2}{\rho}\right) e^{Z(c,\tau)} d\mathcal{C}d\tau$$

Let $(C_m, t_l), m \in [1, \ldots, M], l \in [1, \ldots, L]$ *with volumes* $\Delta(C_m)\Delta(t_l)$ *be an equi-spaced partition of* $(\mathcal{C} \times t)$*, and let* $\bar{C}_m$*,* $\bar{t}_l$ *be the coordinates of the centroids of* C_m*, and* t_l*, respectively. The approximate integral for* M, L *is be*

$$\sum_{l=1}^{L}\sum_{m=1}^{M} \mathcal{K}\left(\frac{\|\ell(\bar{C}_m)-s_i\|_2}{\rho}\right) e^{Z(\bar{C}_m,\bar{t}_l)}\Delta(C_m)\Delta(t_l)$$

Then

$$\lim_{\substack{M\to\infty \\ L\to\infty}} \Bigg| \int_{\mathcal{C}\times t} \mathcal{K}\left(\frac{\|\ell(c)-s_i\|_2}{\rho}\right) e^{Z(c,\tau)} d\mathcal{C}d\tau - \sum_{m=1}^{M}\sum_{l=1}^{L} \mathcal{K}\left(\frac{\|\ell(\bar{C}_m)-s_i\|_2}{\rho}\right) e^{Z(\bar{C}_m,\bar{t}_l)})\Delta(C_m)\Delta(t_l)\Bigg| \to 0$$

Proof. First, we may use the integrability of $Z(c, \tau)$, as guaranteed for valid covariance functions (Møller et al., 1998), and the fact that $0 < \mathcal{K} \leq 1$ to show that the integral is well-defined.

Next, we show that the approximation error is almost surely bounded by a constant and terms

that depend on our approximation error:

$$
\begin{aligned}
&\left| \int_{\mathcal{C}\times t} \mathcal{K}\left(\tfrac{\|\ell(c)-s_i\|_2}{\rho}\right) e^{Z(c,\tau)} d\mathcal{C}d\tau - \sum_{m=1}^{M}\sum_{l=1}^{L} \mathcal{K}\left(\tfrac{\|\ell(\bar{C}_m)-s_i\|_2}{\rho}\right) e^{Z(\bar{C}_m,\bar{t}_l)})\Delta(C_m)\Delta(t_l) \right| \\
&= \left| \int_{C_m\times t_l} \mathcal{K}\left(\tfrac{\|\ell(c)-s_i\|_2}{\rho}\right) e^{Z(c,\tau)} d\mathcal{C}d\tau - \sum_{m=1}^{M}\sum_{l=1}^{L} \mathcal{K}\left(\tfrac{\|\ell(\bar{C}_m)-s_i\|_2}{\rho}\right) e^{Z(\bar{C}_m,\bar{t}_l)})\Delta(C_m)\Delta(t_l) \right| \\
&\leq \sum_{m=1}^{M}\sum_{l=1}^{L} \left| \int_{C_m\times t_l} \mathcal{K}\left(\tfrac{\|\ell(c)-s_i\|_2}{\rho}\right) e^{Z(c,\tau)} d\mathcal{C}d\tau - \mathcal{K}\left(\tfrac{\|\ell(\bar{C}_m)-s_i\|_2}{\rho}\right) e^{Z(\bar{C}_m,\bar{t}_l)})\Delta(C_m)\Delta(t_l) \right| \\
&\leq \sum_{m=1}^{M}\sum_{l=1}^{L} \Delta(C_m)\Delta(t_l) \left| \sup_{\{(c,\tau)\}\in C_m\times t_l} \mathcal{K}\left(\tfrac{\|\ell(c)-s_i\|_2}{\rho}\right) e^{Z(c,\tau)} - \mathcal{K}\left(\tfrac{\|\ell(\bar{C}_m)-s_i\|_2}{\rho}\right) e^{Z(\bar{C}_m,\bar{t}_l)}) \right| \\
&\leq \sum_{m=1}^{M}\sum_{l=1}^{L} \Delta(C_m)\Delta(t_l) \left| \sup_{c\in C_m} \mathcal{K}\left(\tfrac{\|\ell(c)-s_i\|_2}{\rho}\right) \sup_{\{(c,\tau)\}\in C_m\times t_l} e^{Z(c,\tau)} - \mathcal{K}\left(\tfrac{\|\ell(\bar{C}_m)-s_i\|_2}{\rho}\right) e^{Z(\bar{C}_m,\bar{t}_l)}) \right| \\
&\leq \sum_{m=1}^{M}\sum_{l=1}^{L} \Delta(C_m)\Delta(t_l) \Bigg(\sup_{\{(c,\tau)\}\in C_m\times t_l} e^{Z(c,\tau)} \left| \sup_{c\in C_m} \mathcal{K}\left(\tfrac{\|\ell(c)-s_i\|_2}{\rho}\right) - \mathcal{K}\left(\tfrac{\|\ell(\bar{C}_m)-s_i\|_2}{\rho}\right) \right| \\
&+ \mathcal{K}\left(\tfrac{\|\ell(c)-s_i\|_2}{\rho}\right) \left| \sup_{\{(c,\tau)\}\in C_m\times t_l} e^{Z(c,\tau)} - e^{Z(\bar{C}_m,\bar{t}_l)}) \right| \Bigg) \\
&\leq \sum_{m=1}^{M}\sum_{l=1}^{L} \Delta(C_m)\Delta(t_l) \Bigg(\sup_{\{(c,\tau)\}\in C_m\times t_l} e^{Z(c,\tau)} \sup_{c,c'\in C_m} \left| \mathcal{K}\left(\tfrac{\|\ell(c)-s_i\|_2}{\rho}\right) - \mathcal{K}\left(\tfrac{\|\ell(c')-s_i\|_2}{\rho}\right) \right| \\
&+ \mathcal{K}\left(\tfrac{\|\ell(c)-s_i\|_2}{\rho}\right) \sup_{\{(c,\tau),(c',\tau')\}\in C_m\times t_l} \left| e^{Z(c,\tau)} - e^{Z(c',\tau')} \right| \Bigg)
\end{aligned}
$$

Given that our partitions are equi-spaced, let $\Delta(C_m) = \Delta(\mathcal{C})/M$ and $\Delta(t_l) = (t_2 - t_1)/L$ for all m and l. Given the integrability condition on $e^{Z(c,\tau)}$, namely that $\int_{C\times t} e^{Z(c,\tau)} < \infty$ a.s. for any bounded region $(C \times t)$, $\sup_{\{(c,\tau)\}\in C_m\times t_l} e^{Z(c,\tau)} < K$ for some finite number K. Given that $\mathcal{K}$ is continuously differentiable, there exists a $B_{\mathcal{K}} < \infty$ such that:

$$
\sup_{c,c'\in C_m} \left| \mathcal{K}\left(\tfrac{\|\ell(c)-s_i\|_2}{\rho}\right) - \mathcal{K}\left(\tfrac{\|\ell(c')-s_i\|_2}{\rho}\right) \right| \leq B_{\mathcal{K}} \tfrac{\Delta\mathcal{C}}{M}.
$$

Finally, the term $\sup_{\{(c,\tau),(c',\tau')\}\in C_m\times t_l} \left| e^{Z(c,\tau)} - e^{Z(c',\tau')} \right|$ can be bounded as well. Given a sufficiently regular covariance function, sample functions are almost-surely s-Hölder continuous (Stuart, 2010), for a given compact interval $C \times t$, $s \in (0, 1)$, there exists a $B_{\exp Z} < \infty$ such that:

$$
\sup_{\{(c,\tau),(c',\tau')\}\in C\times t} \left| e^{Z(c,\tau)} - e^{Z(c',\tau')} \right| \leq B_{\exp Z} \Delta(\mathcal{C})(t_2 - t_1) \left(\sqrt{\left(\tfrac{\Delta(\mathcal{C})}{M}\right)^2 + \left(\tfrac{t_2-t_1}{L}\right)^2} \right)^s.
$$

Then

$$\sum_{m=1}^{M}\sum_{l=1}^{L}\Delta(C_m)\Delta(t_l)\Bigg(\sup_{\{(c,\tau)\}\in C_m\times t_l}e^{Z(c,\tau)}\sup_{c,c'\in C_m}\left|\mathcal{K}\left(\tfrac{\|\ell(c)-s_i\|_2}{\rho}\right)-\mathcal{K}\left(\tfrac{\|\ell(c')-s_i\|_2}{\rho}\right)\right|$$
$$+\mathcal{K}\left(\tfrac{\|\ell(c)-s_i\|_2}{\rho}\right)\sup_{\{(c,\tau),(c',\tau')\}\in C_m\times t_l}\left|e^{Z(c,\tau)}-e^{Z(c',\tau')}\right|\Bigg)$$
$$\leq\sum_{m=1}^{M}\sum_{l=1}^{L}\tfrac{\Delta(\mathcal{C})}{M}\tfrac{t_2-t_1}{L}\left(KB_{\mathcal{K}}\tfrac{\Delta\mathcal{C}}{M}+B_{\exp Z}\Delta(\mathcal{C})(t_2-t_1)\left(\sqrt{\left(\tfrac{\Delta(\mathcal{C})}{M}\right)^2+\left(\tfrac{t_2-t_1}{L}\right)^2}\right)^s\right)$$

Thus, as $M, L\to\infty$

$$\lim_{\substack{M\to\infty\\L\to\infty}}\left|\int_{\mathcal{C}\times t}\mathcal{K}\left(\tfrac{\|\ell(c)-s_i\|_2}{\rho}\right)e^{Z(c,\tau)}d\mathcal{C}d\tau-\sum_{m=1}^{M}\sum_{l=1}^{L}\mathcal{K}\left(\tfrac{\|\ell(\bar{C}_m)-s_i\|_2}{\rho}\right)e^{Z(\bar{C}_m,\bar{t}_l)})\Delta(C_m)\Delta(t_l)\right|\to 0$$

□

C.2 Error bounds

Let $\mathcal{K}_\rho(c)=\mathcal{K}\left(\tfrac{\|\ell(c)-s_j\|_2}{\rho}\right)$.

$$\left|\sum_{l=1}^{L}\sum_{m=1}^{M}\mathcal{K}_\rho(\bar{C}_m)e^{Z(\bar{C}_m,\bar{t}_l)}\Delta(C_m)\Delta(t_l)-\int_{\mathcal{C}\times t}\mathcal{K}_\rho(c)e^{Z(c,\tau)}d\mathcal{C}d\tau\right|$$
$$=\left|\sum_{l=1}^{L}\sum_{m=1}^{M}\int_{C_m\times t_l}\mathcal{K}_\rho(\bar{C}_m)e^{Z(\bar{C}_m,\bar{t}_l)}-\mathcal{K}_\rho(c)e^{Z(c,\tau)}d\mathcal{C}d\tau\right|$$
$$=\left|\sum_{l=1}^{L}\sum_{m=1}^{M}\int_{C_m\times t_l}\mathcal{K}_\rho(\bar{C}_m)\left(e^{Z(c,\tau)}-e^{Z(\bar{C}_m,\bar{t}_l)}\right)+e^{Z(c,\tau)}\left(\mathcal{K}_\rho(c)-\mathcal{K}_\rho(\bar{C}_m)\right)d\mathcal{C}d\tau\right|$$
$$=\left|\sum_{l=1}^{L}\sum_{m=1}^{M}\int_{C_m\times t_l}\mathcal{K}_\rho(\bar{C}_m)\left(e^{Z(c,\tau)}-e^{Z(\bar{C}_m,\bar{t}_l)}\right)d\mathcal{C}d\tau+\sum_{l=1}^{L}\sum_{m=1}^{M}\int_{C_m\times t_l}e^{Z(c,\tau)}\left(\mathcal{K}_\rho(c)-\mathcal{K}_\rho(\bar{C}_m)\right)d\mathcal{C}d\tau\right|$$
$$\leq\sum_{l=1}^{L}\sum_{m=1}^{M}\int_{C_m\times t_l}\mathcal{K}_\rho(\bar{C}_m)\left|e^{Z(c,\tau)}-e^{Z(\bar{C}_m,\bar{t}_l)}\right|d\mathcal{C}d\tau+\sum_{l=1}^{L}\sum_{m=1}^{M}\int_{C_m\times t_l}e^{Z(c,\tau)}\left|\mathcal{K}_\rho(c)-\mathcal{K}_\rho(\bar{C}_m)\right|d\mathcal{C}d\tau$$
$$\leq\sum_{l=1}^{L}\sum_{m=1}^{M}\int_{C_m\times t_l}\mathcal{K}_\rho(\bar{C}_m)\left|e^{Z(c,\tau)}-e^{Z(\bar{C}_m,\bar{t}_l)}\right|d\mathcal{C}d\tau$$
$$+\sum_{l=1}^{L}\sum_{m=1}^{M}\left(\mathcal{K}\left(\tfrac{\inf_c\|\ell(c)-s_j\|_2}{\rho}\right)-\mathcal{K}\left(\tfrac{\sup_c\|\ell(c)-s_j\|_2}{\rho}\right)\right)\int_{C_m\times t_l}e^{Z(c,\tau)}d\mathcal{C}d\tau$$

J	I	N	Distribution	No. sim.
500	10	5,000	Uniform	100
1,000	10	10,000	Uniform	100
2,000	10	20,000	Uniform	100
500	100	50,000	Uniform	100
1,000	100	100,000	Uniform	100
2,000	100	200,000	Uniform	100
500	10	5,000	Clustered	100
1,000	10	10,000	Clustered	100
2,000	10	20,000	Clustered	100
500	100	50,000	Clustered	100
1,000	100	100,000	Clustered	100
2,000	100	200,000	Clustered	100

Table C.1: Table of simulation study settings. J is the number of households, I is the number of observations per household, and $N = J \times I$. Distribution is the spatial distribution of

where in the last line we have used the fact that $\mathcal{K}$ is a monotonically decreasing function of distance $\|\ell(c) - s_j\|_2$. The lower bound, using the same decomposition of the error above

$$\left|\sum_{l=1}^{L}\sum_{m=1}^{M}\mathcal{K}_\rho(\bar{C}_m)e^{Z(\bar{C}_m,\bar{t}_l)}\Delta(C_m)\Delta(t_l) - \int_{\mathcal{C}\times t}\mathcal{K}_\rho(c)e^{Z(c,\tau)}d\mathcal{C}d\tau\right|$$
$$\geq \left|\left|\sum_{l=1}^{L}\sum_{m=1}^{M}\int_{C_m\times t_l} e^{Z(c,\tau)}\left(\mathcal{K}_\rho(c) - \mathcal{K}_\rho(\bar{C}_m)\right)d\mathcal{C}d\tau\right| - \left|\sum_{l=1}^{L}\sum_{m=1}^{M}\int_{C_m\times t_l}\mathcal{K}_\rho(\bar{C}_m)\left(e^{Z(c,\tau)} - e^{Z(\bar{C}_m,\bar{t}_l)}\right)d\mathcal{C}d\tau\right|\right|$$

C.3 Simulation scenarios

C.3.1 Comparison of different grid resolutions

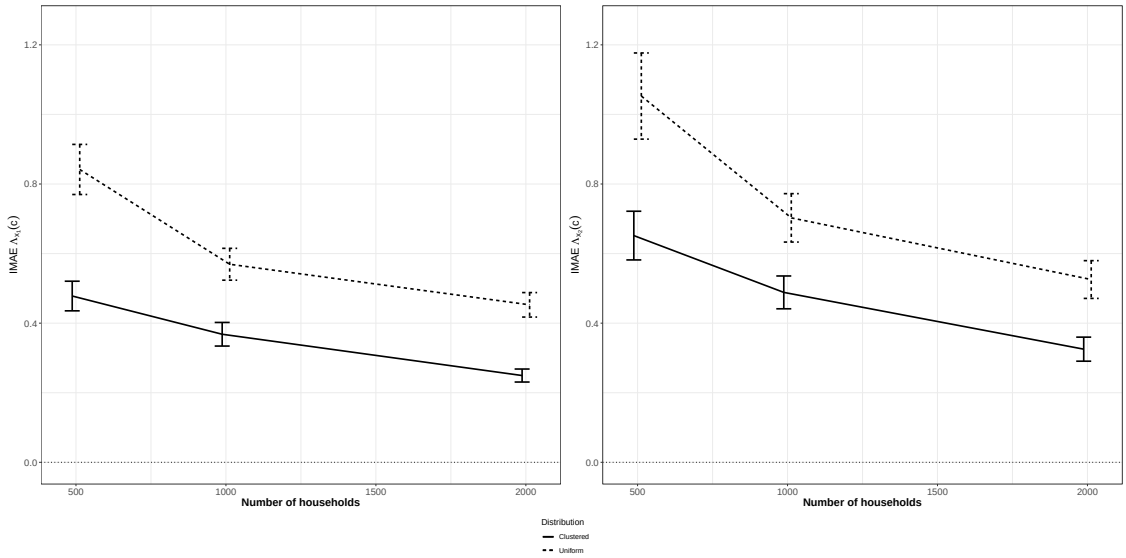


Figure C.1: Integrated mean absolute error for Λ_{x_1} and Λ_y with ± 1.96 standard errors plotted as black bars, 100 observations per household, $M = 160$

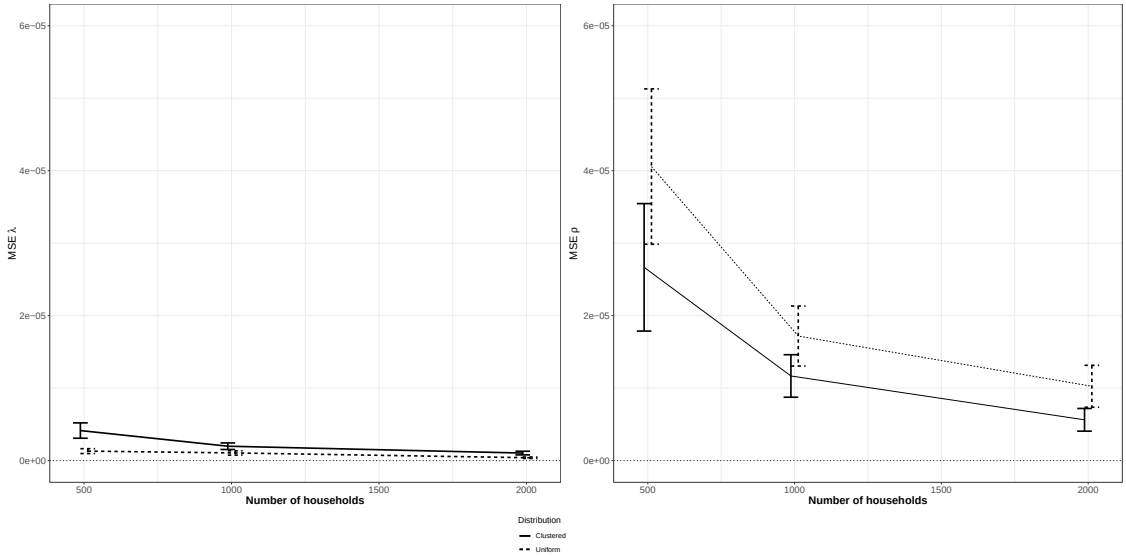


Figure C.2: MSE for ρ and λ with ± 2 standard errors plotted as black bars, x-jittered for clarity on the plot for ρ, 100 observations per household, $M = 160$

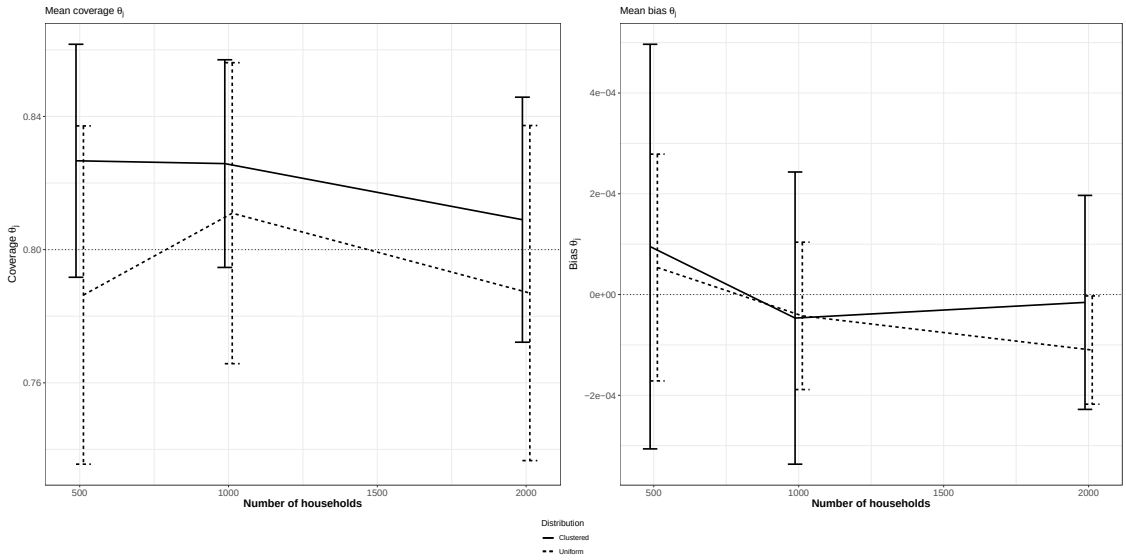


Figure C.3: Bias and 50% interval coverage for $\theta_j^{\text{environ}}$ ± 2 standard errors plotted as black bars. The horizontal dotted line in the left plot corresponds to the nominal coverage of 50%, while the horizontal dotted line in the right plot corresponds to zero bias., 100 observations per household, $M = 160$

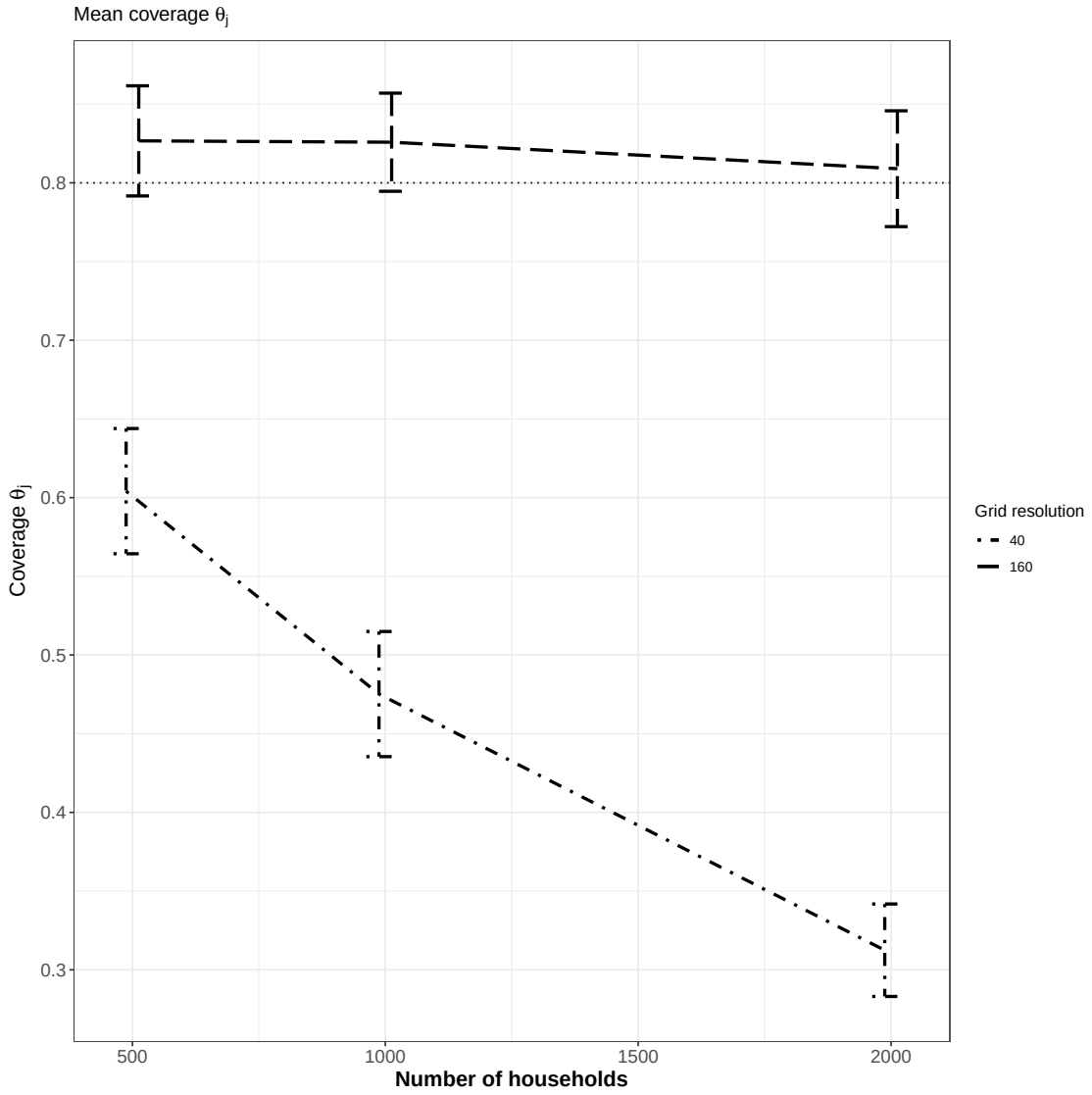


Figure C.4: Comparison of mean coverage rates of θ_j by number of households for $M = 40$ and $M = 160$ grid resolutions, 100 observations per household, clustered household distribution.

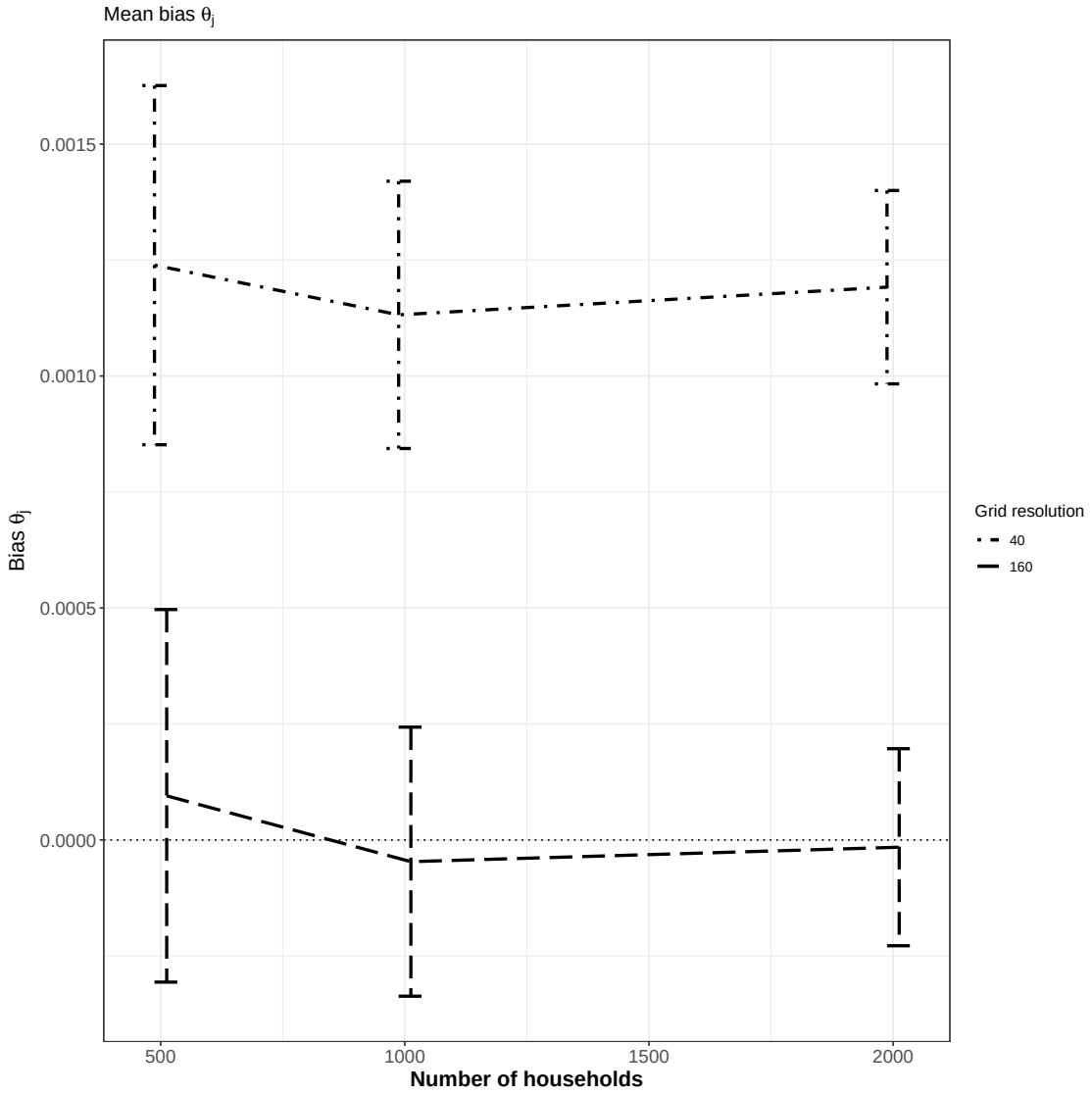


Figure C.5: Comparison of mean bias of posterior mean estimator for θ_j by number of households for $M = 40$ and $M = 160$ grid resolutions, 100 observations per household, clustered household distribution.

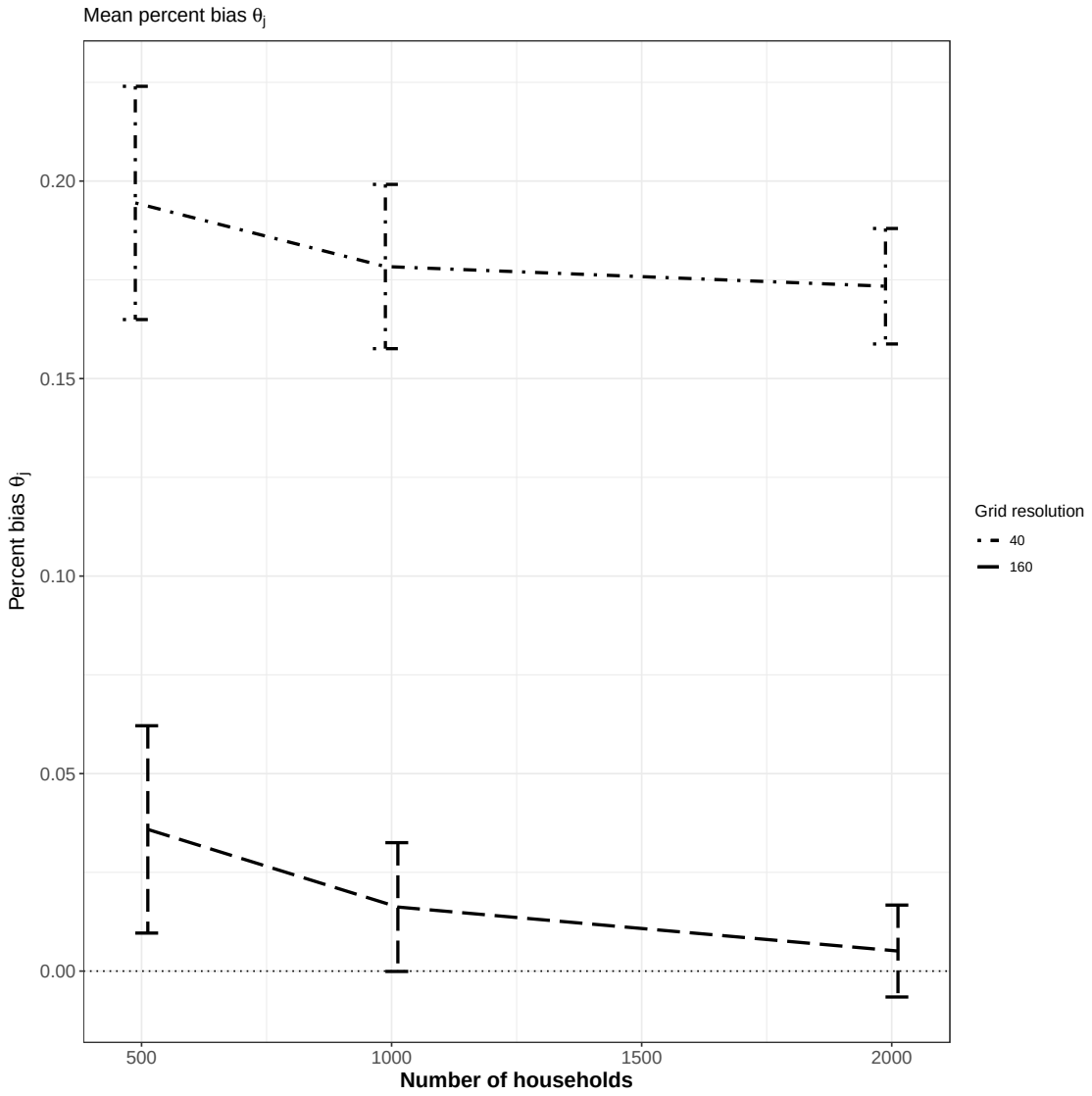


Figure C.6: Comparison of mean percent bias of posterior mean estimator for θ_j: $\frac{|\mathbb{E}[\theta_j|Y]-\theta_j|}{\theta_j}$ by number of households for $M = 40$ and $M = 160$ grid resolutions, 100 observations per household, clustered household distribution.

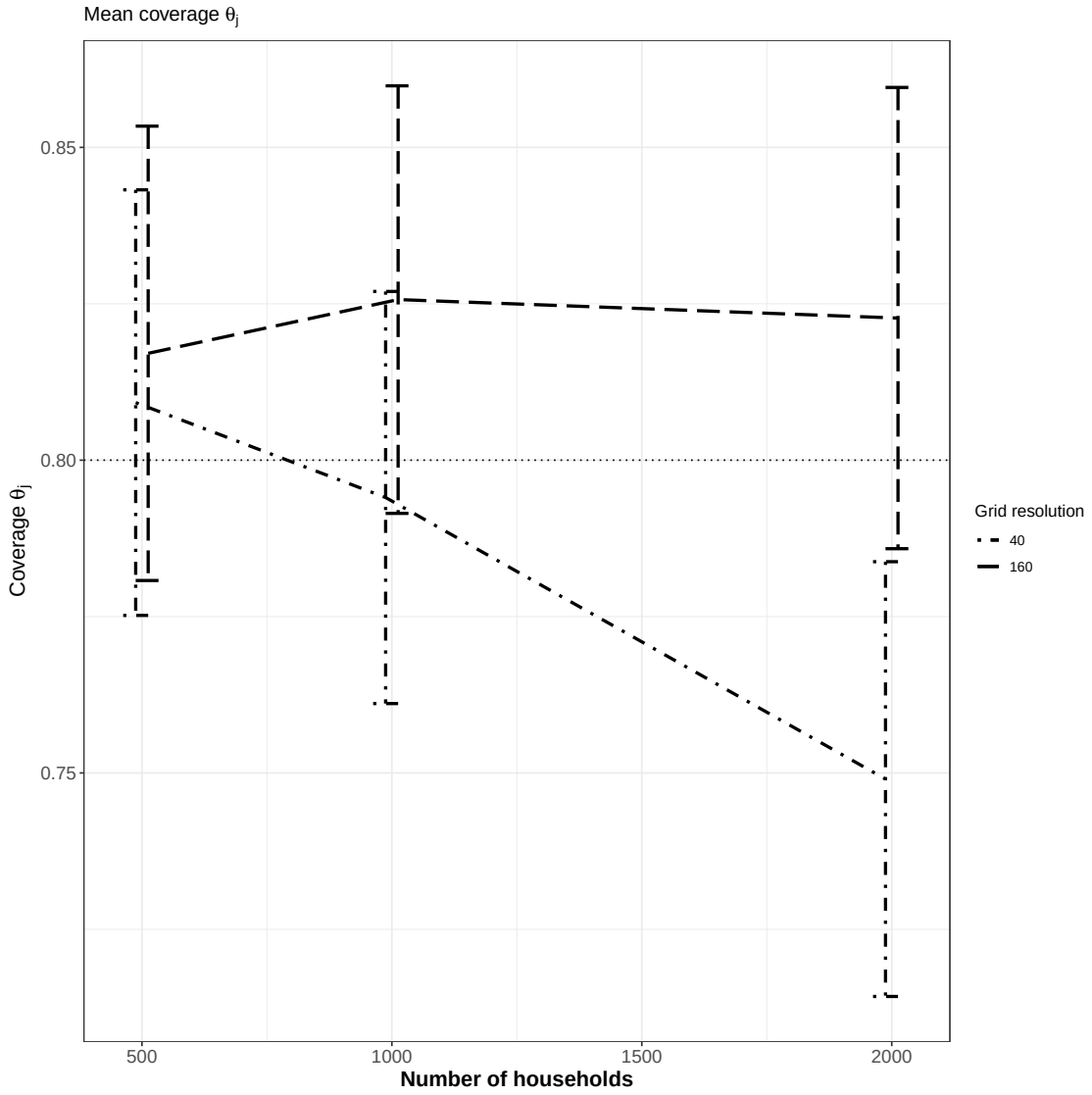


Figure C.7: Comparison of mean coverage rates of θ_j by number of households for $M = 40$ and $M = 160$ grid resolutions, 10 observations per household, clustered household distribution.

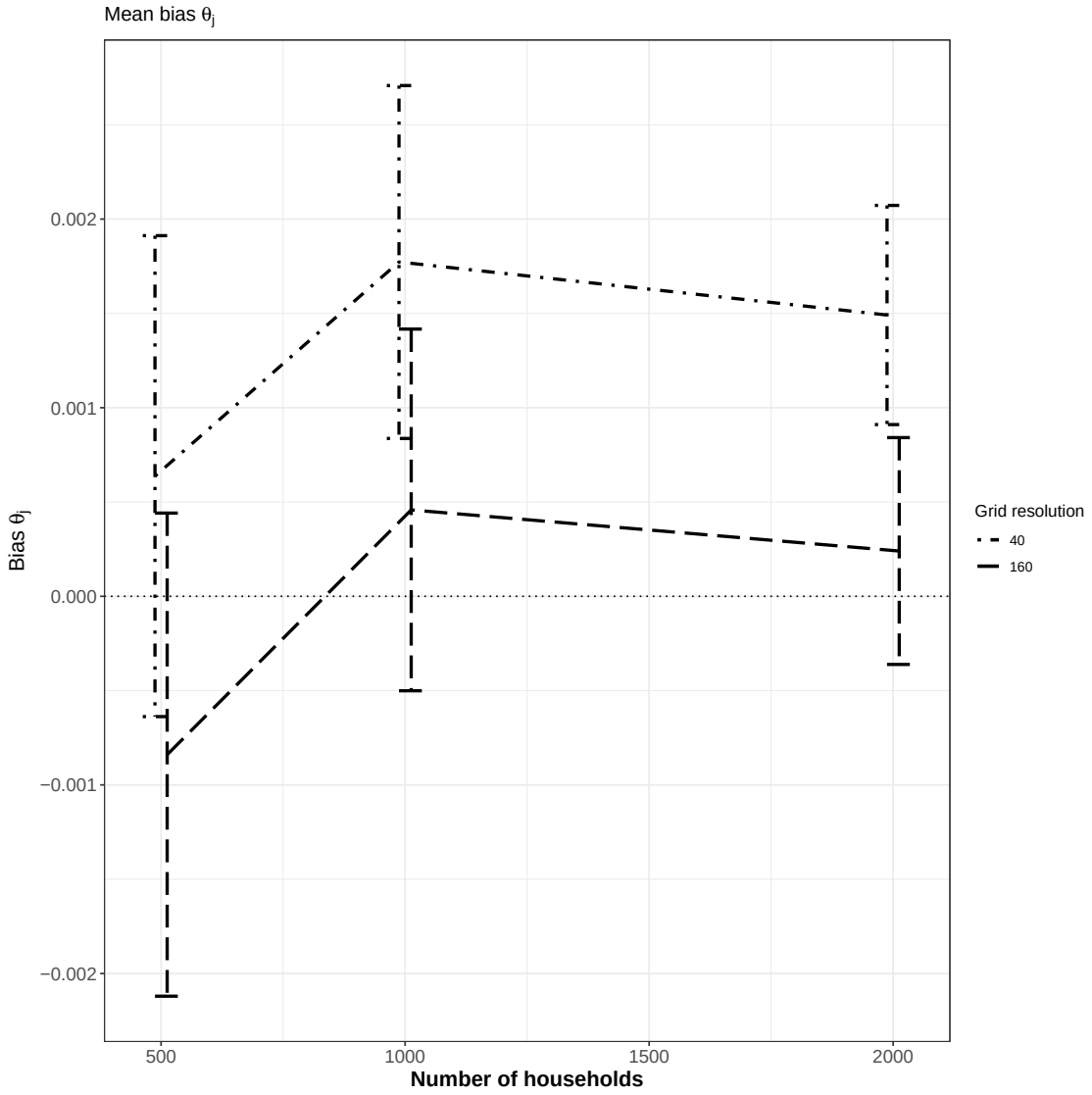


Figure C.8: Comparison of mean bias of posterior mean estimator for θ_j by number of households for $M = 40$ and $M = 160$ grid resolutions, 10 observations per household, clustered household distribution.

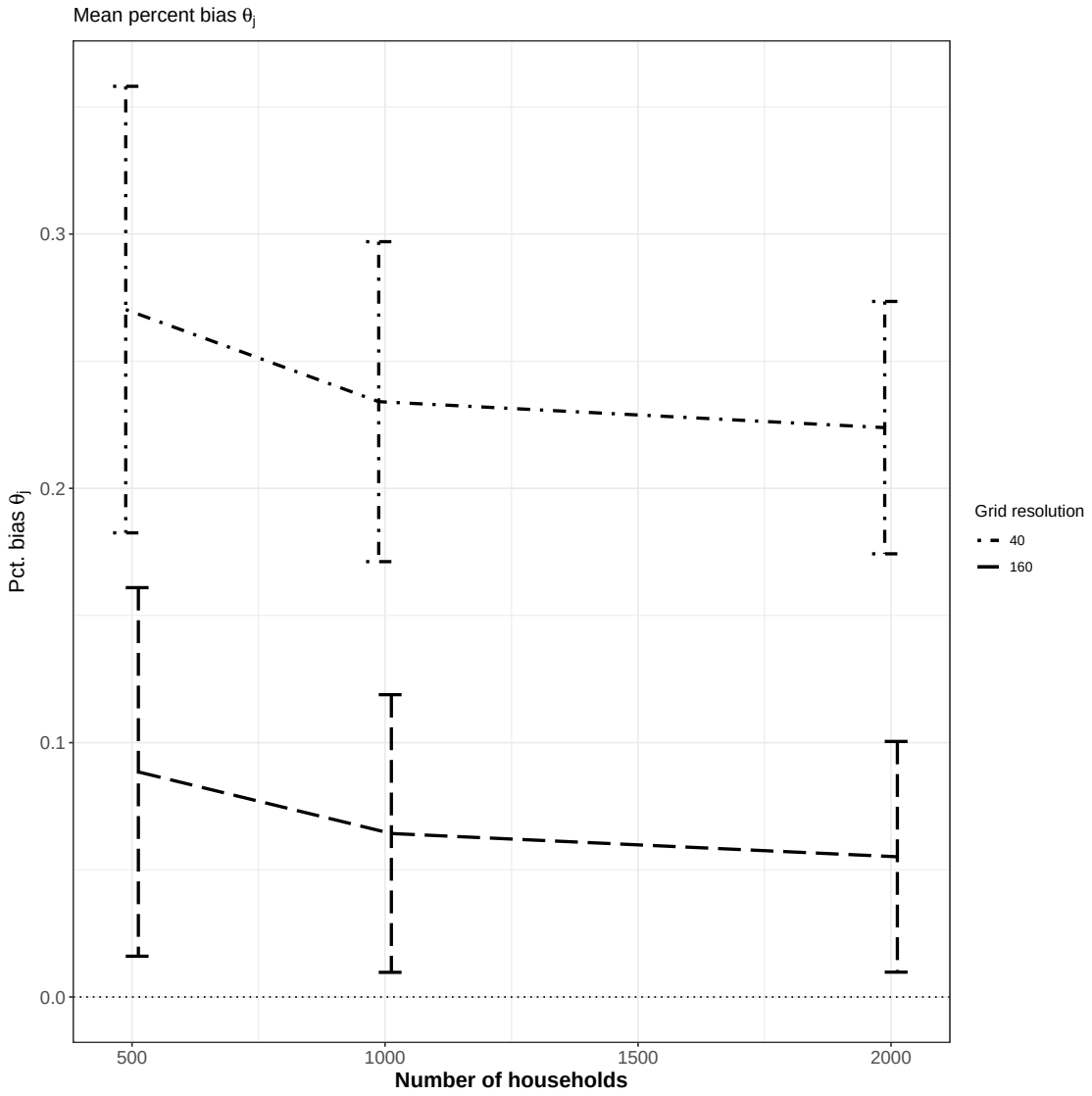


Figure C.9: Comparison of mean percent bias of posterior mean estimator for θ_j: $\frac{|\mathbb{E}[\theta_j|Y]-\theta_j|}{\theta_j}$ by number of households for $M = 40$ and $M = 160$ grid resolutions, 10 observations per household, clustered household distribution.

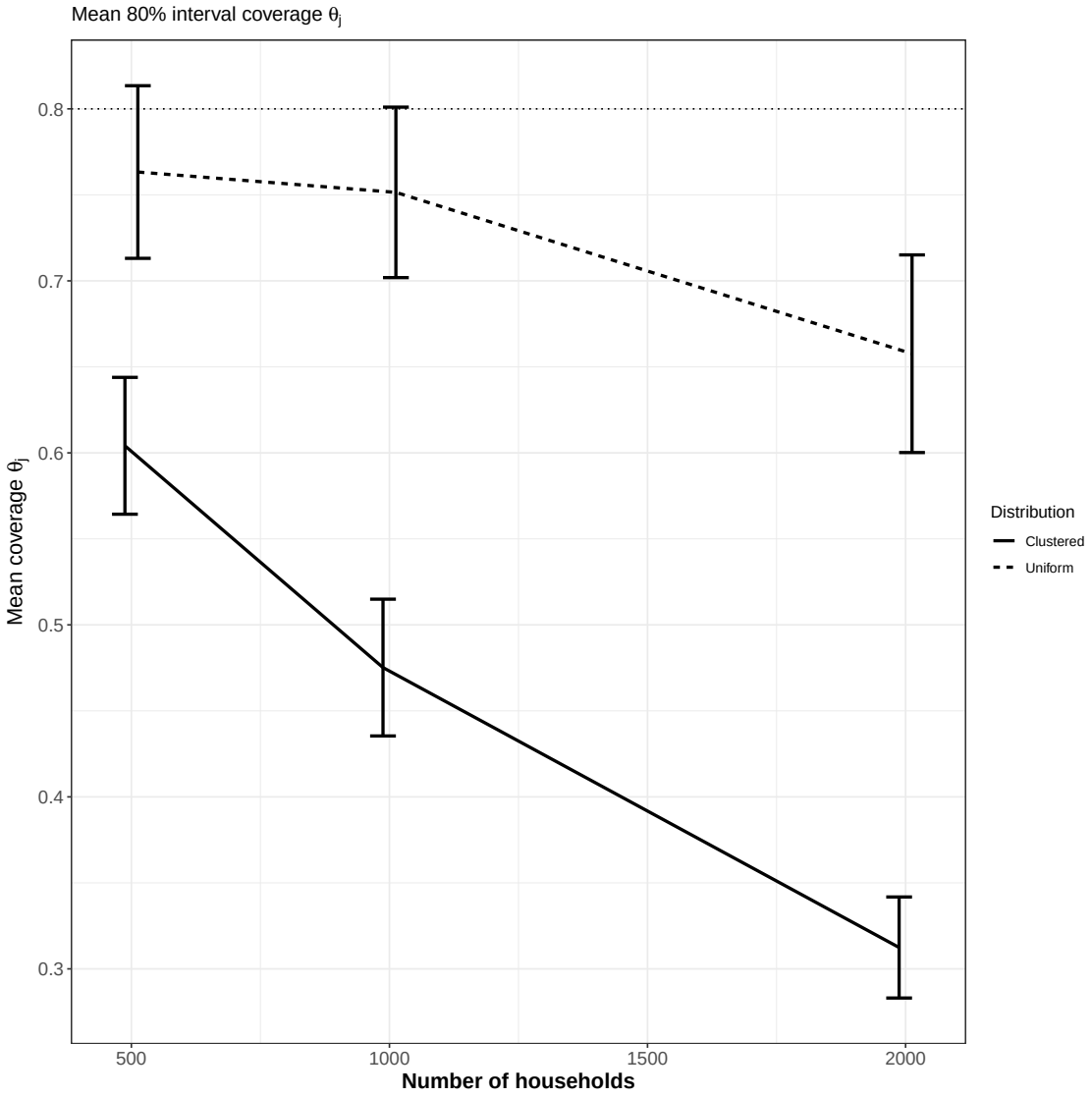


Figure C.10: Comparison of mean coverage of 80% posterior credible intervals for θ_j by number of households for $M = 40$ grid resolution, 100 observations per household, clustered household distribution.

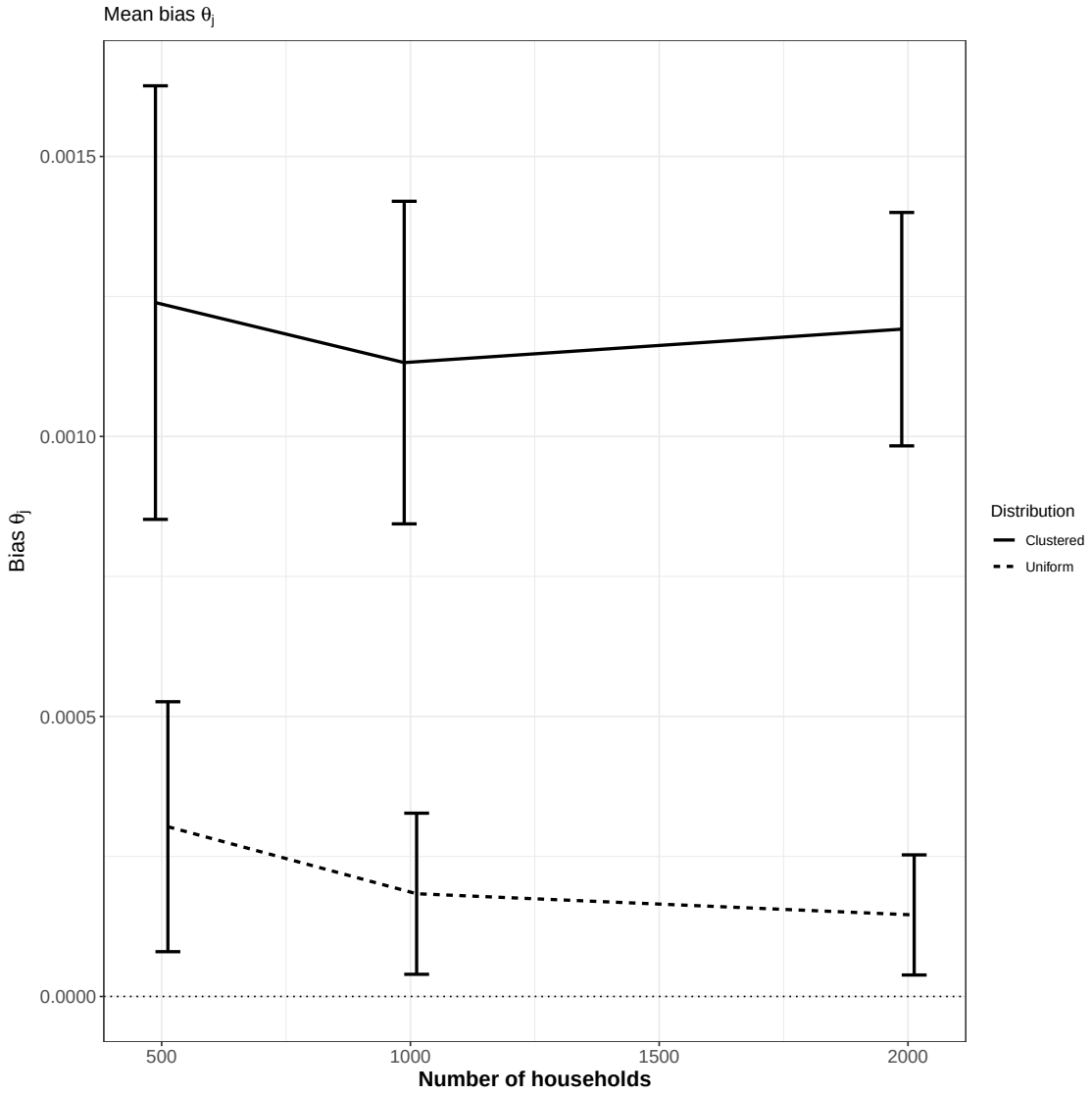


Figure C.11: Comparison of mean bias of posterior mean estimator for θ_j by number of households for $M = 40$ grid resolution, 100 observations per household, clustered household distribution.

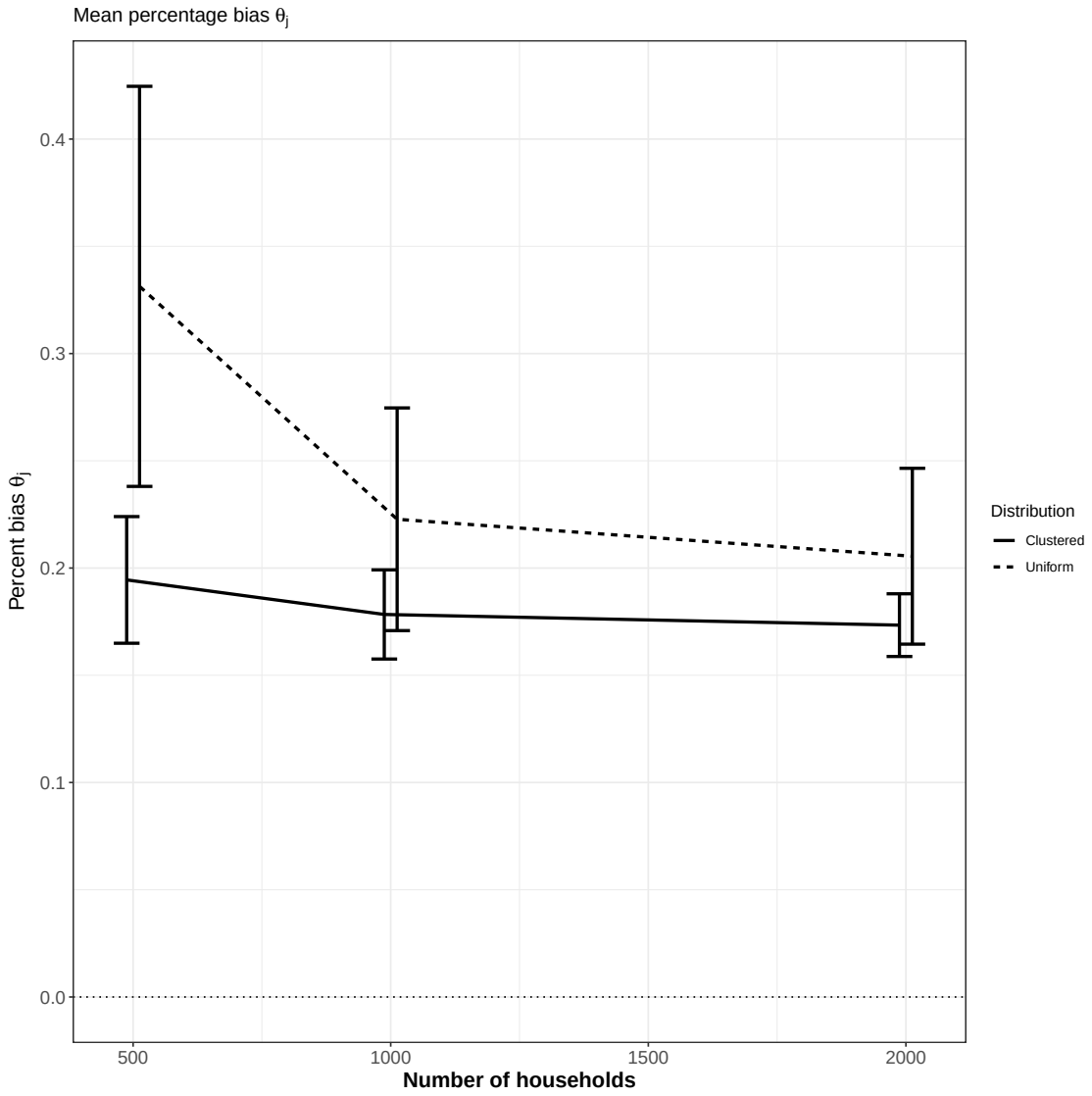


Figure C.12: Comparison of mean percent bias of posterior mean estimator for θ_j: $\frac{|\mathbb{E}[\theta_j|Y]-\theta_j|}{\theta_j}$ by number of households for $M = 40$ grid resolution, 100 observations per household, clustered household distribution.

BIBLIOGRAPHY

Gloria A. Aguayo, Anna Schritz, Maria Ruiz-Castell, Luis Villarroel, Gonzalo Valdivia, Guy Fagherazzi, Daniel R. Witte, and Andrew Lawson. Identifying hotspots of cardiometabolic outcomes based on a Bayesian approach: The example of Chile. *PLOS ONE*, 15(6), June 2020. ISSN 1932-6203. doi: 10.1371/journal.pone.0235009.

Elizabeth S. Allman, Catherine Matias, and John A. Rhodes. Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, 37(6A), December 2009. ISSN 0090-5364. doi: 10.1214/09-AOS689.

Isaiah Andrews and Anna Mikusheva. Maximum likelihood inference in weakly identified dynamic stochastic general equilibrium models: Maximum likelihood inference. *Quantitative Economics*, 6(1):123–152, 2015. ISSN 17597323. doi: 10.3982/QE331. URL http://doi.wiley.com/10.3982/QE331.

Eleni-Rosalina Andrinopoulou, D Rizopoulos, Johanna JM Takkenberg, and E Lesaffre. Combined dynamic predictions using joint models of two longitudinal outcomes and competing risk data. *Stat Methods Med Res*, 26(4):1787–1801, 2017. ISSN 0962-2802. doi: 10.1177/0962280215588340. URL https://doi.org/10.1177/0962280215588340.

Vincent Audigier, Ian R White, Shahab Jolani, Thomas PA Debray, Matteo Quartagno, James Carpenter, Stef Van Buuren, and Matthieu Resche-Rigon. Multiple imputation for multilevel data with continuous and binary variables. *Statistical Science*, 33(2):160–183, 2018.

Lindsey R. Baden, Hana M. El Sahly, Brandon Essink, Karen Kotloff, Sharon Frey, Rick Novak, David Diemert, Stephen A. Spector, Nadine Rouphael, C. Buddy Creech, John McGettigan, Shishir Khetan, Nathan Segall, Joel Solis, Adam Brosz, Carlos Fierro, Howard Schwartz, Kathleen Neuzil, Lawrence Corey, Peter Gilbert, Holly Janes, Dean Follmann, Mary Marovich, John Mascola, Laura Polakowski, Julie Ledgerwood, Barney S. Graham, Hamilton Bennett, Rolando Pajon, Conor Knightly, Brett Leav, Weiping Deng, Honghong Zhou, Shu Han, Melanie Ivarsson, Jacqueline Miller, and Tal Zaks. Efficacy and safety of the mRNA-1273 SARS-CoV-2 vaccine. *New England Journal of Medicine*, 384(5):403–416, 2021. ISSN 0028-4793, 1533-4406. doi: 10.1056/NEJMoa2035389. URL http://www.nejm.org/doi/10.1056/NEJMoa2035389.

Jannah Baker, Nicole White, and Kerrie Mengersen. Missing in space: An evaluation of imputation methods for missing data in spatial analysis of risk factors for type II diabetes. *International Journal of Health Geographics*, 13(1), 2014. ISSN 1476-072X. doi: 10.1186/1476-072X-13-47.

Alexander Balke and Judea Pearl. Counterfactual Probabilities: Computational Methods, Bounds and Applications. In *Uncertainty Proceedings 1994*, pages 46–54. Elsevier, 1994. ISBN 978-1-55860-332-5. doi: 10.1016/B978-1-55860-332-5.50011-0. URL https://linkinghub.elsevier.com/retrieve/pii/B9781558603325500110.

Cici Bauer and Jon Wakefield. Stratified space–time infectious disease modelling, with an application to hand, foot and mouth disease in China. *Journal of the Royal Statistical Society Series C*, 67(5), 2018.

Ralf Bender. Introduction to the Use of Regression Models in Epidemiology. In Mukesh Verma, editor, *Cancer Epidemiology*, Methods in Molecular Biology, pages 179–195. Humana Press, Totowa, NJ, 2009. ISBN 978-1-59745-416-2. doi: 10.1007/978-1-59745-416-2_9. URL https://doi.org/10.1007/978-1-59745-416-2_9.

Veronica J. Berrocal, Alan E. Gelfand, David M. Holland, Janet Burke, and Marie Lynn Miranda. On the use of a PM2.5 exposure simulator to explain birthweight. *Environmetrics*, 22(4):553–571, June 2011. ISSN 1180-4009. doi: 10.1002/env.1086. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3116241/.

Michael Betancourt. A Conceptual Introduction to Hamiltonian Monte Carlo. *arXiv:1701.02434 [stat]*, July 2018.

Michael Betancourt and Mark Girolami. Hamiltonian monte carlo for hierarchical models. *Current trends in Bayesian methodology with applications*, 79(30):2–4, 2015.

Annibale Biggeri and C Lagazio. Case-control analysis around putative sources. *Disease Mapping and Risk Assessment for Public Health. Lawson, Bertollini, Biggeri, Böhning, Lesaffre and Viel (eds). Wiley, London, UK*, pages 271–286, 1999.

Usama Bilal, Loni P Tabb, Sharrelle Barber, and Ana V Diez Roux. Spatial inequities in covid-19 testing, positivity, confirmed cases, and mortality in 3 us cities: An ecological study. *Annals of internal medicine*, 2021.

Björn Bornkamp, Kaspar Rufibach, Jianchang Lin, Yi Liu, Devan V. Mehrotra, Satrajit Roychoudhury, Heinz Schmidli, Yue Shentu, and Marcel Wolbers. Principal stratum strategy: Potential role in drug development. *Pharmaceutical Statistics*, 20(4):737–751, 2021. doi: https://doi.org/10.1002/pst.2104. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/pst.2104.

SL Brilleman, MJ Crowther, M Moreno-Betancur, J Buros Novik, and R Wolfe. Joint longitudinal and time-to-event models via Stan. URL https://github.com/stan-dev/stancon_talks/. StanCon 2018. 10-12 Jan 2018. Pacific Grove, CA, USA.

Andrew F. Brouwer, Marisa C. Eisenberg, Justin V. Remais, Philip A. Collender, Rafael Meza, and Joseph N. S. Eisenberg. Modeling Biphasic Environmental Decay of Pathogens and Implications for Risk Analysis. *Environ Sci Technol*, 51(4):2186–2196, February 2017a. ISSN 0013-936X. doi: 10.1021/acs.est.6b04030.

Andrew F. Brouwer, Mark H. Weir, Marisa C. Eisenberg, Rafael Meza, and Joseph N. S. Eisenberg. Dose-response relationships for environmentally mediated infectious disease transmission models. *PLOS Computational Biology*, 13(4):e1005481, April 2017b. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1005481.

Paul-Christian Bürkner, Jonah Gabry, Matthew Kay, and Aki Vehtari. posterior: Tools for working with posterior distributions, 2021. URL https://mc-stan.org/posterior/. R package version 1.0.1.

Crescenza Calculli, Alessio Pollice, and Lucia Bisceglia. Spatial variation of multiple diseases in relation to an environmental risk source. 38 WPG2010, June 2010. URL https://aisberg.unibg.it/handle/10446/950#.XbtLTEFKg5k.

Bob Carpenter. Stan implementation of poisson-binomial distribution. https://discourse.mc-stan.org/t/poisson-binomial-distribution-any-existing-stan-implementation/4220/7, 2018. Accessed: 2022-02-10.

Bob Carpenter, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan : A Probabilistic Programming Language. *Journal of Statistical Software*, 76(1), 2017. ISSN 1548-7660. doi: 10.18637/jss.v076.i01. URL https://www.osti.gov/pages/biblio/1430202-stan-probabilistic-programming-language.

Kelsie Cassell, Paul Gacek, Joshua L. Warren, Peter A. Raymond, Matthew Cartter, and Daniel M. Weinberger. Association between sporadic legionellosis and river systems in connecticut. *The Journal of Infectious Diseases*, 217(2):179–187, 2018. ISSN 1537-6613. doi: 10.1093/infdis/jix531.

E. Catchpole. Detecting parameter redundancy. *Biometrika*, 84(1):187–196, 1997. ISSN 0006-3444, 1464-3510. doi: 10.1093/biomet/84.1.187.

Jing Cheng and Dylan S. Small. Bounds on causal effects in three-arm trials with non-compliance. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(5), November 2006. ISSN 1369-7412, 1467-9868. doi: 10.1111/j.1467-9868.2006.00568.x.

Samuel J Clark and Brian Houle. Validation, replication, and sensitivity testing of heckman-type selection models to adjust estimates of hiv prevalence. *PloS one*, 9(11):e112563, 2014.

Jesse D Contreras, Rafael Meza, Christina Siebe, Sandra Rodríguez-Dozal, Yolanda A López-Vidal, Gonzalo Castillo-Rojas, Rosa I Amieva, Sandra G Solano-Gálvez, Marisa Mazari-Hiriart, Miguel A Silva-Magaña, et al. Health risks from exposure to untreated wastewater used for irrigation in the mezquital valley, mexico: A 25-year update. *Water Research*, 123:834–850, 2017.

Jesse D Contreras, Rob Trangucci, Eunice E Felix-Arellano, Sandra Rodríguez-Dozal, Christina Siebe, Horacio Riojas-Rodríguez, Rafael Meza, Jon Zelner, and Joseph NS Eisenberg. Modeling spatial risk of diarrheal disease associated with household proximity to untreated wastewater

used for irrigation in the mezquital valley, mexico. *Environmental Health Perspectives*, 128(7): 077002, 2020.

Samantha R Cook, Andrew Gelman, and Donald B Rubin. Validation of software for bayesian models using posterior quantiles. *Journal of Computational and Graphical Statistics*, 15(3): 675–692, 2006.

Simon L Cotter, Masoumeh Dashti, and Andrew M Stuart. Approximation of bayesian inverse problems for pdes. *SIAM journal on numerical analysis*, 48(1):322–345, 2010.

Forrest W Crawford, Florian M Marx, Jon Zelner, and Ted Cohen. Transmission modeling with regression adjustment for analyzing household-based studies of infectious disease: application to tuberculosis. *Epidemiology (Cambridge, Mass.)*, 2019.

Rhian M Daniel, Michael G Kenward, Simon N Cousens, and Bianca L De Stavola. Using causal diagrams to guide analysis in missing data problems. *Statistical Methods in Medical Research*, 21(3):243–256, June 2012. ISSN 0962-2802, 1477-0334. doi: 10.1177/0962280210394469. URL http://journals.sagepub.com/doi/10.1177/0962280210394469.

Peter Diggle and Michael G Kenward. Informative drop-out in longitudinal data analysis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 43(1):49–73, 1994.

Peter Diggle, Sara Morris, Paul Elliott, and Gavin Shaddick. Regression Modelling of Disease Risk in Relation to Point Sources. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 160(3):491–505, 1997. ISSN 1467-985X. doi: 10.1111/j.1467-985X.1997.00076.x. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-985X.1997.00076.x.

Peter J. Diggle. A point process modelling approach to raised incidence of a rare phenomenon in the vicinity of a prespecified point. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 153(3):349–362, 1990. ISSN 0964-1998. doi: 10.2307/2982977. URL http://www.jstor.org/stable/2982977.

Peter J. Diggle and Barry S. Rowlingson. A Conditional Approach to Point Process Modelling of Elevated Risk. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 157(3): 433–440, 1994. ISSN 0964-1998. doi: 10.2307/2983529.

Peng Ding and Fan Li. Causal inference. *Statistical Science*, 33(2):214–237, 2018.

Peng Ding and Jiannan Lu. Principal stratification analysis using principal scores. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3), June 2017. ISSN 13697412. doi: 10.1111/rssb.12191.

Peng Ding, Zhi Geng, Wei Yan, and Xiao-Hua Zhou. Identifiability and Estimation of Causal Effects by Principal Stratification With Outcomes Truncated by Death. *Journal of the American Statistical Association*, 106(496), December 2011. ISSN 0162-1459, 1537-274X. doi: 10.1198/jasa.2011.tm10265.

Iris Eekhout, R. Michiel de Boer, Jos W. R. Twisk, Henrica C. W. de Vet, and Martijn W. Heymans. Missing Data: A Systematic Review of How They Are Reported and Handled. *Epidemiology*, 23(5), 2012. ISSN 1044-3983. doi: 10.1097/EDE.0b013e3182576cdb.

Hana M. El Sahly, Lindsey R. Baden, Brandon Essink, Susanne Doblecki-Lewis, Judith M. Martin, Evan J. Anderson, Thomas B. Campbell, Jesse Clark, Lisa A. Jackson, Carl J. Fichtenbaum, Marcus Zervos, Bruce Rankin, Frank Eder, Gregory Feldman, Christina Kennelly, Laurie Han-Conrad, Michael Levin, Kathleen M. Neuzil, Lawrence Corey, Peter Gilbert, Holly Janes, Dean Follmann, Mary Marovich, Laura Polakowski, John R. Mascola, Julie E. Ledgerwood, Barney S. Graham, Allison August, Heather Clouting, Weiping Deng, Shu Han, Brett Leav, Deb Manzo, Rolando Pajon, Florian Schödel, Joanne E. Tomassini, Honghong Zhou, and Jacqueline Miller. Efficacy of the mRNA-1273 SARS-CoV-2 Vaccine at Completion of Blinded Phase. *New England Journal of Medicine*, 385(19):1774–1785, November 2021. ISSN 0028-4793, 1533-4406. doi: 10.1056/NEJMoa2113017.

Marc N. Elliott, Peter A. Morrison, Allen Fremont, Daniel F. McCaffrey, Philip Pantoja, and Nicole Lurie. Using the Census Bureau's surname list to improve estimates of race/ethnicity and associated disparities. *Health Services and Outcomes Research Methodology*, 9(2), 2009. ISSN 1387-3741, 1572-9400. doi: 10.1007/s10742-009-0047-1.

Felix Elwert and Christopher Winship. Endogenous Selection Bias: The Problem of Conditioning on a Collider Variable. *Annual Review of Sociology*, 40(1), 2014. ISSN 0360-0572, 1545-2115. doi: 10.1146/annurev-soc-071913-043455.

Donald P. Francis. The prevention of hepatitis b with vaccine: Report of the centers for disease control multi-center efficacy trial among homosexual men. *Annals of Internal Medicine*, 97(3): 362, 1982. ISSN 0003-4819. doi: 10.7326/0003-4819-97-3-362. URL http://annals.org/article.aspx?doi=10.7326/0003-4819-97-3-362.

Constantine E. Frangakis and Donald B. Rubin. Principal Stratification in Causal Inference. *Biometrics*, 58(1), March 2002. ISSN 0006341X. doi: 10.1111/j.0006-341X.2002.00021.x.

E. L. Frome. The Analysis of Rates Using Poisson Regression Models. *Biometrics*, 39(3), September 1983. ISSN 0006341X. doi: 10.2307/2531094.

Edward L. Frome and Harvey Checkoway. USE OF POISSON REGRESSION MODELS IN ESTIMATING INCIDENCE RATES AND RATIOS. *American Journal of Epidemiology*, 121 (2), February 1985. ISSN 1476-6256, 0002-9262. doi: 10.1093/oxfordjournals.aje.a114001.

Jonah Gabry and Tristan Mahr. bayesplot: Plotting for bayesian models, 2021. URL https://mc-stan.org/bayesplot/. R package version 1.8.1.

Jonah Gabry and Rok Češnovar. *cmdstanr: R Interface to 'CmdStan'*, 2021. https://mc-stan.org/cmdstanr, https://discourse.mc-stan.org.

Jonah Gabry and Rok Češnovar. *cmdstanr: R Interface to 'CmdStan'*, 2022. https://mc-stan.org/cmdstanr/, https://discourse.mc-stan.org.

Jonah Gabry, Daniel Simpson, Aki Vehtari, Michael Betancourt, and Andrew Gelman. Visualization in bayesian workflow. *J. R. Stat. Soc. A*, 182:389–402, 2019a. doi: 10.1111/rssa.12378.

Jonah Gabry, Daniel Simpson, Aki Vehtari, Michael Betancourt, and Andrew Gelman. Visualization in Bayesian workflow. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 182(2):389–402, 2019b. ISSN 1467-985X. doi: 10.1111/rssa.12378. URL http://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssa.12378.

Yuxiang Gao, Lauren Kennedy, Daniel Simpson, and Andrew Gelman. Improving Multilevel Regression and Poststratification with Structured Priors. *Bayesian Analysis*, -1(-1), January 2021. ISSN 1936-0975. doi: 10.1214/20-BA1223.

Andrew Gelman and Thomas C Little. Poststratification into many categories using hierarchical logistic regression. 1997.

Andrew Gelman and Donald B. Rubin. Inference from Iterative Simulation Using Multiple Sequences. *Statist. Sci.*, 7(4):457–472, November 1992. ISSN 0883-4237, 2168-8745. doi: 10.1214/ss/1177011136. URL https://projecteuclid.org/euclid.ss/1177011136.

Andrew Gelman, Aleks Jakulin, Maria Grazia Pittau, and Yu-Sung Su. A weakly informative default prior distribution for logistic and other regression models. *Ann. Appl. Stat.*, 2(4):1360–1383, December 2008. ISSN 1932-6157, 1941-7330. doi: 10.1214/08-AOAS191. URL https://projecteuclid.org/euclid.aoas/1231424214.

Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. Bayesian data analysis. 2013.

Andrew Gelman, Daniel Simpson, and Michael Betancourt. The Prior Can Often Only Be Understood in the Context of the Likelihood. *Entropy*, 19(10), October 2017. ISSN 1099-4300. doi: 10.3390/e19100555.

Andrew Gelman, Aki Vehtari, Daniel Simpson, Charles C Margossian, Bob Carpenter, Yuling Yao, Lauren Kennedy, Jonah Gabry, Paul-Christian Bürkner, and Martin Modrák. Bayesian workflow. *arXiv preprint arXiv:2011.01808*, 2020.

Subhashis Ghosal and Anindya Roy. Posterior consistency of Gaussian process prior for nonparametric binary regression. *The Annals of Statistics*, 34(5), October 2006. ISSN 0090-5364. doi: 10.1214/009053606000000795. URL https://projecteuclid.org/journals/annals-of-statistics/volume-34/issue-5/Posterior-consistency-of-Gaussian-process-prior-for-nonparametric-binary-10.1214/009053606000000795.full.

Peter B. Gilbert, Ronald J. Bosch, and Michael G. Hudgens. Sensitivity Analysis for the Assessment of Causal Vaccine Effects on Viral Load in HIV Vaccine Trials. *Biometrics*, 59(3), September 2003. ISSN 0006341X, 15410420. doi: 10.1111/1541-0420.00063.

Virgilio Gómez-Rubio, Michela Cameletti, and Marta Blangiardo. Missing data analysis and imputation via latent Gaussian Markov random fields. 2019.

Governor Whitmer Executive Order. Executive order 2020-55: Michigan coronavirus task force on racial disparities. https://www.michigan.gov/whitmer/0,9309,7-387-90499_90705-526476--,00.html, 2020. Accessed: 2022-02-10.

Leonardo Grilli and Fabrizia Mealli. Nonparametric Bounds on the Causal Effect of University Studies on Job Opportunities Using Principal Stratification. *Journal of Educational and Behavioral Statistics*, 33(1), March 2008. ISSN 1076-9986, 1935-1054. doi: 10.3102/1076998607302627.

Paul Gustafson. On Model Expansion, Model Contraction, Identifiability and Prior Information: Two Illustrative Scenarios Involving Mismeasured Variables. *Statistical Science*, 20(2):111–140, May 2005. ISSN 0883-4237, 2168-8745. doi: 10.1214/088342305000000098. URL https://projecteuclid.org/euclid.ss/1121347636. Publisher: Institute of Mathematical Statistics.

Paul Gustafson. Measurement error modelling with an approximate instrumental variable: Measurement Error Modelling. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(5):797–815, November 2007. ISSN 13697412. doi: 10.1111/j.1467-9868.2007.00611.x. URL http://doi.wiley.com/10.1111/j.1467-9868.2007.00611.x.

Paul Gustafson. *Bayesian inference for partially identified models: Exploring the limits of limited data*, volume 140. CRC Press, 2015.

M. Elizabeth Halloran and Claudio J. Struchiner. Causal Inference in Infectious Diseases:. *Epidemiology*, 6(2), March 1995. ISSN 1044-3983. doi: 10.1097/00001648-199503000-00010.

M. Elizabeth Halloran, Ira M. Longini, and Claudio J. Struchiner. *Design and Analysis of Vaccine Studies*. Statistics for Biology and Health. Springer New York, New York, NY, 2010. ISBN 978-0-387-68636-3. doi: 10.1007/978-0-387-68636-3. URL https://link.springer.com/10.1007/978-0-387-68636-3.

James Heckman. Varieties of selection bias. *The American Economic Review*, 80(2):313–318, 1990.

James J Heckman, Sergio Urzua, and Edward Vytlacil. Understanding Instrumental Variables in Models with Essential Heterogeneity. *The Review of Economics and Statistics*, 88(3):389–432, 2006.

Leonhard Held and Michaela Paul. Modeling seasonality in space-time infectious disease surveillance data: Modeling seasonality in space-time data. *Biometrical Journal*, 54(6), November 2012. ISSN 03233847. doi: 10.1002/bimj.201200037.

Leonhard Held, Niel Hens, Philip D O'Neill, and Jacco Wallinga. *Handbook of infectious disease data analysis*. CRC Press, 2019.

Graeme L. Hickey, Pete Philipson, Andrea Jorgensen, and Ruwanthi Kolamunnage-Dona. Joint modelling of time-to-event and multivariate longitudinal outcomes: Recent developments and issues. *BMC Medical Research Methodology*, 16(1):117, 2016. ISSN 1471-2288. doi: 10.1186/s12874-016-0212-5. URL https://doi.org/10.1186/s12874-016-0212-5.

Matthew D Hoffman, Andrew Gelman, et al. The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623, 2014.

Paul W. Holland. Statistics and Causal Inference. *Journal of the American Statistical Association*, 81(396):945–960, 1986. ISSN 0162-1459. doi: 10.2307/2289064. URL https://www.jstor.org/stable/2289064.

R. C. Holland, G. Jones, and J. Benschop. Spatio-temporal modelling of disease incidence with missing covariate values. *Epidemiology and Infection*, 143(8), June 2015. ISSN 0950-2688, 1469-4409. doi: 10.1017/S0950268814002854.

Michael G Hudgens and M. Elizabeth Halloran. Causal Vaccine Effects on Binary Postinfection Outcomes. *Journal of the American Statistical Association*, 101(473), March 2006. ISSN 0162-1459, 1537-274X. doi: 10.1198/016214505000000970.

Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

Katri Jalava, Hanna Rintala, Jukka Ollgren, Leena Maunula, Vicente Gomez-Alvarez, Joana Revez, Marja Palander, Jenni Antikainen, Ari Kauppinen, Pia Räsänen, Sallamaari Siponen, Outi Nyholm, Aino Kyyhkynen, Sirpa Hakkarainen, Juhani Merentie, Martti Pärnänen, Raisa Loginov, Hodon Ryu, Markku Kuusi, Anja Siitonen, Ilkka Miettinen, Jorge W. Santo Domingo, Marja-Liisa Hänninen, and Tarja Pitkänen. Novel Microbiological and Spatial Statistical Methods to Improve Strength of Epidemiological Evidence in a Community-Wide Waterborne Outbreak. *PLOS ONE*, 9(8):e104713, August 2014. ISSN 1932-6203. doi: 10.1371/journal.pone.0104713.

Yannis Jemiai, Andrea Rotnitzky, Bryan E. Shepherd, and Peter B. Gilbert. Semiparametric estimation of treatment effects given base-line covariates on an outcome measured after a post-randomization event occurs: Semiparametric Estimation of Treatment Effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(5), November 2007. ISSN 13697412. doi: 10.1111/j.1467-9868.2007.00615.x.

Zhichao Jiang and Peng Ding. Measurement errors in the binary instrumental variable model. *Biometrika*, 107(1), March 2020. ISSN 0006-3444, 1464-3510. doi: 10.1093/biomet/asz060.

Zhichao Jiang, Peng Ding, and Zhi Geng. Principal causal effect identification and surrogate end point evaluation by multiple trials. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(4), September 2016. ISSN 13697412. doi: 10.1111/rssb.12135.

Matt J Keeling and Pejman Rohani. *Modeling infectious diseases in humans and animals*. Princeton university press, 2011.

Robert W. Keener. *Theoretical Statistics*. Springer Texts in Statistics. Springer New York, New York, NY, 2010. ISBN 978-0-387-93838-7. doi: 10.1007/978-0-387-93839-4.

Lauren Kennedy, Katharine Khanna, Daniel Simpson, and Andrew Gelman. Using sex and gender in survey adjustment. *arXiv preprint arXiv:2009.14401*, 2020.

Stephen M. Kissler, Joseph R. Fauver, Christina Mack, Scott W. Olesen, Caroline Tai, Kristin Y. Shiue, Chaney C. Kalinich, Sarah Jednak, Isabel M. Ott, Chantal B. F. Vogels, Jay Wohlgemuth, James Weisberger, John DiFiori, Deverick J. Anderson, Jimmie Mancell, David D. Ho, Nathan D. Grubaugh, and Yonatan H. Grad. Viral dynamics of acute SARS-CoV-2 infection and applications to diagnostic and public health strategies. *PLOS Biology*, 19(7), July 2021. ISSN 1545-7885. doi: 10.1371/journal.pbio.3001333.

Joseph B. Kruskal. Three-way arrays: Rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear Algebra and its Applications*, 18(2), 1977. ISSN 00243795. doi: 10.1016/0024-3795(77)90069-6.

Katie Labgold, Sarah Hamid, Sarita Shah, Neel R. Gandhi, Allison Chamberlain, Fazle Khan, Shamimul Khan, Sasha Smith, Steve Williams, Timothy L. Lash, and Lindsay J. Collin. Estimating the Unknown: Greater Racial and Ethnic Disparities in COVID-19 Burden After Accounting for Missing Race and Ethnicity Data. *Epidemiology*, 32(2), 2021. ISSN 1044-3983. doi: 10.1097/EDE.0000000000001314.

Timothy L. Lash, Tyler J. VanderWeele, Sebastien Haneuse, and Kenneth J Rothman. *Modern Epidemiology. 4th Edition.* Lippincott Williams & Wilkins, 2021.

E. L. Lehmann and George Casella. *Theory of Point Estimation*. Springer Texts in Statistics. Springer, New York, 2nd ed edition, 1998. ISBN 978-0-387-98502-2.

Ye Li, Patrick Brown, Dionne C Gesink, and Håvard Rue. Log Gaussian Cox processes and spatially aggregated disease incidence data. *Statistical Methods in Medical Research*, 21(5), October 2012. ISSN 0962-2802, 1477-0334. doi: 10.1177/0962280212446326.

Ilya Lipkovich, Bohdana Ratitch, Yongming Qu, Xiang Zhang, Mingyang Shan, and Craig Mallinckrodt. Using principal stratification in analysis of clinical trials. *Statistics in Medicine*, 41(19):3837–3877, August 2022. ISSN 0277-6715, 1097-0258. doi: 10.1002/sim.9439. URL https://onlinelibrary.wiley.com/doi/10.1002/sim.9439.

Marc Lipsitch and Rebecca Kahn. Interpreting vaccine efficacy trial results for infection and transmission. *Vaccine*, 39(30), July 2021. ISSN 0264410X. doi: 10.1016/j.vaccine.2021.06.011.

Roderick Little. Selection and pattern-mixture models. 2008.

Roderick J. Little, Donald B. Rubin, and Sahar Z. Zangeneh. Conditions for Ignoring the Missing-Data Mechanism in Likelihood Inferences for Parameter Subsets. *Journal of the American Statistical Association*, 112(517), 2017. ISSN 0162-1459, 1537-274X. doi: 10.1080/01621459.2015.1136826.

Roderick J. A. Little and Donald B. Rubin. *Statistical Analysis with Missing Data*. Wiley Series in Probability and Statistics. Wiley, Hoboken, N.J, 2nd ed edition, 2002. ISBN 978-0-471-18386-0.

Roderick JA Little. Modeling the drop-out mechanism in repeated-measures studies. *Journal of the american statistical association*, 90(431):1112–1121, 1995.

Victoria Liublinska and Donald B. Rubin. Sensitivity analysis for a partially missing binary outcome in a two-arm randomized clinical trial. *Statistics in medicine*, 33(24), October 2014. ISSN 0277-6715. doi: 10.1002/sim.6197.

Sharon L. Lohr. *Sampling: Design and Analysis*. Chapman and Hall/CRC, 2 edition, April 2019. ISBN 978-0-429-29628-4. doi: 10.1201/9780429296284. URL https://www.taylorfrancis.com/books/9781000022087.

Dustin M. Long and Michael G. Hudgens. Sharpening Bounds on Principal Effects with Covariates: Principal Effect Bounds. *Biometrics*, 69(4), December 2013. ISSN 0006341X. doi: 10.1111/biom.12103.

Ira M Longini Jr, M Elizabeth Halloran, Azhar Nizam, Mark Wol, M Mendelman, Patricia E Fast, and Robert B Belshe. Estimation of the efficacy of live, attenuated inluenza vaccine from a two-year, multi-center vaccine trial: implications for in¯uenza epidemic control. *Vaccine*, 2000.

Haidong Lu, Stephen R. Cole, Chanelle J. Howe, and Daniel Westreich. Toward a Clearer Definition of Selection Bias When Estimating Causal Effects. *Epidemiology*, 33(5):699–706, 2022. ISSN 1044-3983. doi: 10.1097/EDE.0000000000001516. URL https://journals.lww.com/10.1097/EDE.0000000000001516.

Shanshan Luo, Wei Li, and Yangbo He. Causal inference with outcomes truncated by death in multiarm studies. *Biometrics*, 79(1):502–513, 2023.

Steven Manson, Jonathan Schroeder, David Van Riper, Tracy Kugler, and Steven Ruggles. Ipums national historical geographic information system: Version 16.0 [dataset], 2021. Minneapolis, MN: IPUMS, http://doi.org/10.18128/D050.V16.0.

Sebastian Meyer and Leonhard Held. Power-law models for infectious disease spread. *The Annals of Applied Statistics*, 8(3), September 2014. ISSN 1932-6157. doi: 10.1214/14-AOAS743.

Wang Miao, Zhi Geng, and Eric J Tchetgen Tchetgen. Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, 105(4), December 2018. ISSN 0006-3444, 1464-3510. doi: 10.1093/biomet/asy038.

Michigan Department of Health and Human Services. *Michigan state and local public health COVID-19 standard operating procedures.*, volume 41. Michigan Department of Health and Human Services, Lansing, MI, 2020.

Gregorio A. Millett, Austin T. Jones, David Benkeser, Stefan Baral, Laina Mercer, Chris Beyrer, Brian Honermann, Elise Lankiewicz, Leandro Mena, Jeffrey S. Crowley, Jennifer Sherwood, and Patrick S. Sullivan. Assessing differential impacts of COVID-19 on black communities. *Annals of Epidemiology*, 47, 2020. ISSN 10472797. doi: 10.1016/j.annepidem.2020.05.003.

Jesper Møller, Anne Randi Syversveen, and Rasmus Plenge Waagepetersen. Log Gaussian Cox Processes. *Scandinavian Journal of Statistics*, 25(3):451–482, 1998. ISSN 0303-6898.

Arnold S. Monto, Suzanne E. Ohmit, Joshua G. Petrie, Emileigh Johnson, Rachel Truscon, Esther Teich, Judy Rotthoff, Matthew Boulton, and John C. Victor. Comparative Efficacy of Inactivated and Live Attenuated Influenza Vaccines. *New England Journal of Medicine*, 361(13), September 2009. ISSN 0028-4793, 1533-4406. doi: 10.1056/NEJMoa0808652.

Rahul Mukerjee and Brajendra C. Sutradhar. On the Positive Definiteness of the Information Matrix Under the Binary and Poisson Mixed Models. *Annals of the Institute of Statistical Mathematics*, 54:355–366, 2002. ISSN 1572-9052. doi: 10.1023/A:1022478119885.

Jerzy Neyman. On the application of probability theory to agricultural experiments. essay on principles. *Ann. Agricultural Sciences*, pages 1–51, 1923.

Press Office Office of Michigan Governor. Governor whitmer creates the michigan coronavirus task force on racial disparities. https://www.michigan.gov/coronavirus/0,9753,7-406-98163-525224--,00.html, 2020. Accessed: 2022-02-10.

Keith B. Oldham, Jan C. Myland, and Jerome Spanier. *The Tricomi Function U(a,c,x)*. Springer US, New York, NY, 2008. ISBN 978-0-387-48807-3. doi: 10.1007/978-0-387-48807-3_49.

Jing Ouyang and Gongjun Xu. Identifiability of Latent Class Models with Covariates. *Psychometrika*, March 2022. ISSN 0033-3123, 1860-0980. doi: 10.1007/s11336-022-09852-y.

Laura Perez, Fred Lurmann, John Wilson, Manuel Pastor, Sylvia J. Brandt, Nino Künzli, and Rob McConnell. Near-Roadway Pollution and Childhood Asthma: Implications for Developing "Win–Win" Compact Urban Development and Clean Vehicle Strategies. *Environ Health Perspect*, 120(11):1619–1626, November 2012. ISSN 0091-6765. doi: 10.1289/ehp.1104785. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3556611/.

Neil J Perkins, Stephen R Cole, Ofer Harel, Eric J Tchetgen Tchetgen, BaoLuo Sun, Emily M Mitchell, and Enrique F Schisterman. Principled Approaches to Missing Data in Epidemiologic Studies. *American Journal of Epidemiology*, 187(3), 2018. ISSN 0002-9262, 1476-6256. doi: 10.1093/aje/kwx348.

Adam Peterson and Brisa Sanchez. rstap: An r package for spatial temporal aggregated predictor models. *arXiv:1812.10208 [stat]*, 2018. URL http://arxiv.org/abs/1812.10208.

Fernando P. Polack, Stephen J. Thomas, Nicholas Kitchin, Judith Absalon, Alejandra Gurtman, Stephen Lockhart, John L. Perez, Gonzalo Pérez Marc, Edson D. Moreira, Cristiano Zerbini, Ruth Bailey, Kena A. Swanson, Satrajit Roychoudhury, Kenneth Koury, Ping Li, Warren V. Kalina, David Cooper, Robert W. Frenck, Laura L. Hammitt, Özlem Türeci, Haylene Nell, Axel Schaefer, Serhat Ünal, Dina B. Tresnan, Susan Mather, Philip R. Dormitzer, Uğur Şahin, Kathrin U. Jansen, and William C. Gruber. Safety and Efficacy of the BNT162b2 mRNA Covid-19 Vaccine. *New England Journal of Medicine*, 383(27), December 2020. ISSN 0028-4793, 1533-4406. doi: 10.1056/NEJMoa2034577.

James L. Powell. Chapter 41 Estimation of semiparametric models. In *Handbook of Econometrics*, volume 4, pages 2443–2521. Elsevier, 1994. ISBN 978-0-444-88766-5. doi: 10.1016/S1573-4412(05)80010-8. URL https://linkinghub.elsevier.com/retrieve/pii/S1573441205800108.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021. URL https://www.R-project.org/.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022. URL https://www.R-project.org/.

Rebeca Ramis, Peter Diggle, Koldo Cambra, and Gonzalo López-Abente. Prostate cancer and industrial pollution: Risk around putative focus in a multi-source scenario. *Environment International*, 37(3):577–585, 2011. ISSN 0160-4120. doi: 10.1016/j.envint.2010.12.001. URL http://www.sciencedirect.com/science/article/pii/S0160412010002461.

Calyampudi Radhakrishna Rao. *Linear Statistical Inference and Its Applications*. Wiley, New York, 2. ed., paperback ed edition, 2002.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian processes for machine learning*. The MIT Press, 2006.

Thomas S Richardson, Robin J Evans, and James M Robins. Transparent parameterizations of models for potential outcomes. *Bayesian Statistics*, 9:569–610, 2011.

Thomas J. Rothenberg. Identification in Parametric Models. *Econometrica*, 39(3):577–591, 1971. ISSN 0012-9682. doi: 10.2307/1913267.

Jason Roy and Michael J Daniels. A general class of pattern mixture models for nonignorable dropout with many possible dropout times. *Biometrics*, 64(2):538–545, 2008.

Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, October 1974. ISSN 1939-2176, 0022-0663. doi: 10.1037/h0037350. URL http://doi.apa.org/getdoi.cfm?doi=10.1037/h0037350.

Donald B. Rubin. Inference and Missing Data. *Biometrika*, 63(3), 1976. ISSN 0006-3444. doi: 10.2307/2335739.

Donald B. Rubin. Bayesian Inference for Causal Effects: The Role of Randomization. *The Annals of Statistics*, 6(1), January 1978. ISSN 0090-5364. doi: 10.1214/aos/1176344064. URL https://projecteuclid.org/journals/annals-of-statistics/volume-6/issue-1/Bayesian-Inference-for-Causal-Effects-The-Role-of-Randomization/10.1214/aos/1176344064.full.

Donald B. Rubin. Causal Inference Through Potential Outcomes and Principal Stratification: Application to Studies with "Censoring" Due to Death. *Statistical Science*, 21(3), August 2006. ISSN 0883-4237. doi: 10.1214/088342306000000114.

Olli Saarela, David A. Stephens, and Erica E. M. Moodie. The Role of Exchangeability in Causal Inference. *Statistical Science*, pages 1–17, 2023. doi: 10.1214/22-STS879. URL https://doi.org/10.1214/22-STS879.

Yves-Laurent Kom Samo and Stephen Roberts. String and membrane gaussian processes. *arXiv:1507.06977 [stat]*, 2015. URL http://arxiv.org/abs/1507.06977.

Bryan E. Shepherd, Peter B. Gilbert, Yannis Jemiai, and Andrea Rotnitzky. Sensitivity Analyses Comparing Outcomes Only Existing in a Subset Selected Post-Randomization, Conditional on Covariates, with Application to HIV Vaccine Trials. *Biometrics*, 62(2), June 2006. ISSN 0006341X. doi: 10.1111/j.1541-0420.2005.00495.x.

Bryan E Shepherd, Peter B Gilbert, and Thomas Lumley. Sensitivity Analyses Comparing Time-to-Event Outcomes Existing Only in a Subset Selected Postrandomization. *Journal of the American Statistical Association*, 102(478), June 2007. ISSN 0162-1459, 1537-274X. doi: 10.1198/016214507000000130.

Xu Shi, Wang Miao, Jennifer C. Nelson, and Eric J. Tchetgen Tchetgen. Multiply robust causal inference with double-negative control adjustment for categorical unmeasured confounding. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(2), April 2020. ISSN 13697412. doi: 10.1111/rssb.12361.

Yulia Sidi and Ofer Harel. The treatment of incomplete data: Reporting, analysis, reproducibility, and replicability. *Social Science & Medicine*, 209, 2018. ISSN 02779536. doi: 10.1016/j.socscimed.2018.05.037.

D. Simpson, J. B. Illian, F. Lindgren, S. H. Sørbye, and H. Rue. Going off grid: Computationally efficient inference for log-Gaussian Cox processes. *Biometrika*, 103(1), March 2016. ISSN 0006-3444. doi: 10.1093/biomet/asv064.

Elizabeth A. Stasny. Hierarchical Models for the Probabilities of a Survey Classification and Nonresponse: An Example from the National Crime Survey. *Journal of the American Statistical Association*, 86(414), 1991. ISSN 0162-1459, 1537-274X. doi: 10.1080/01621459.1991.10475033.

Marianne Riksheim Stavseth, Thomas Clausen, and Jo Røislien. How handling missing data may impact conclusions: A comparison of six different imputation methods for categorical questionnaire data. *SAGE Open Medicine*, 7, 2019. ISSN 2050-3121, 2050-3121. doi: 10.1177/2050312118822912.

A. M. Stuart. Inverse problems: A Bayesian perspective. *Acta Numerica*, 19:451–559, May 2010. ISSN 0962-4929, 1474-0508. doi: 10.1017/S0962492910000061. URL https://www.cambridge.org/core/product/identifier/S0962492910000061/type/journal_article.

Jingchao Sun, Maiying Kong, and Subhadip Pal. The Modified-Half-Normal distribution: Properties and an efficient sampling scheme. *Communications in Statistics - Theory and Methods*, June 2021. ISSN 0361-0926, 1532-415X. doi: 10.1080/03610926.2021.1934700.

Eric J. Tchetgen Tchetgen. Identification and estimation of survivor average causal effects. *Statistics in Medicine*, 33(21), September 2014. ISSN 0277-6715, 1097-0258. doi: 10.1002/sim.6181.

Stan Development Team. *Stan Modeling Language Users Guide and Reference Manual, v2.27*, 2021.

The FUTURE II Study Group. Quadrivalent vaccine against human papillomavirus to prevent high-grade cervical lesions. *New England Journal of Medicine*, 356(19):1915–1927, 2007. ISSN 0028-4793, 1533-4406. doi: 10.1056/NEJMoa061741. URL http://www.nejm.org/doi/abs/10.1056/NEJMoa061741.

Corinne N. Thompson, Jonathan L. Zelner, Tran Do Hoang Nhu, My VT Phan, Phuc Hoang Le, Hung Nguyen Thanh, Duong Vu Thuy, Ngoc Minh Nguyen, Tuan Ha Manh, Tu Van Hoang Minh, Vi Lu Lan, Chau Nguyen Van Vinh, Hien Tran Tinh, Emmiliese von Clemm, Harry Storch, Guy Thwaites, Bryan T. Grenfell, and Stephen Baker. The impact of environmental and climatic variation on the spatiotemporal trends of hospitalized pediatric diarrhea in Ho Chi Minh City, Vietnam. *Health Place*, 35:147–154, September 2015. ISSN 1353-8292. doi: 10.1016/j.healthplace.2015.08.001. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4664115/.

Yongge Tian. Rank equalities for block matrices and their moore-penrose inverses. *Houston J. Math*, 30(4):483–510, 2004.

Rob Trangucci, Yang Chen, and Jon Zelner. Modeling racial/ethnic differences in COVID-19 incidence with covariates subject to non-random missingness. *Annals of Applied Statistics, Forthcoming*.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Tyler J. VanderWeele and Eric J. Tchetgen Tchetgen. Bounding the Infectiousness Effect in Vaccine Trials. *Epidemiology*, 22(5), September 2011. ISSN 1044-3983. doi: 10.1097/EDE.0b013e31822708d5.

Aki Vehtari, Andrew Gelman, Daniel Simpson, Bob Carpenter, and Paul-Christian Bürkner. Rank-normalization, folding, and localization: An improved rhat for assessing convergence of mcmc. *Bayesian Analysis*, 2020.

Jon Wakefield, Tracy Qi Dong, and Vladimir N. Minin. Spatio-temporal analysis of surveillance data. In Leonhard Held, Niel Hens, Philip D O'Neill, and Jacco Wallinga, editors, *Handbook of infectious disease data analysis*, chapter 23, pages 455–475. CRC Press, 2019.

Linbo Wang, Thomas S Richardson, and Xiao-Hua Zhou. Causal analysis of ordinal treatments and binary outcomes under truncation by death. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 79(3):719, 2017.

Wenling Wang, Yanli Xu, Ruqin Gao, Roujian Lu, Kai Han, Guizhen Wu, and Wenjie Tan. Detection of SARS-CoV-2 in Different Types of Clinical Specimens. *JAMA*, March 2020. ISSN 0098-7484. doi: 10.1001/jama.2020.3786.

Joshua L. Warren, Louis Grandjean, David A. J. Moore, Anna Lithgow, Jorge Coronel, Patricia Sheen, Jonathan L. Zelner, Jason R. Andrews, and Ted Cohen. Investigating spillover of

multidrug-resistant tuberculosis from a prison: A spatial and molecular epidemiological analysis. *BMC Medicine*, 16(1):122, 2018. ISSN 1741-7015. doi: 10.1186/s12916-018-1111-x. URL https://doi.org/10.1186/s12916-018-1111-x.

Sumio Watanabe. *Algebraic Geometry and Statistical Learning Theory*. Cambridge University Press, Cambridge, 2009. ISBN 978-0-511-80047-4. doi: 10.1017/CBO9780511800474.

Morris Weinberger, Eugene Z. Oddone, William G. Henderson, David M. Smith, James Huey, Anita Giobbie-Hurder, and John R. Feussner. Multisite Randomized Controlled Trials in Health Services Research: Scientific Challenges and Operational Issues:. *Medical Care*, 39(6):627–634, June 2001. ISSN 0025-7079. doi: 10.1097/00005650-200106000-00010. URL http://journals.lww.com/00005650-200106000-00010.

Eric W. Weisstein. Dobiński's formula. From MathWorld—A Wolfram Web Resource.

Jeffrey Wooldridge. Quasi-Likelihood Methods for Count Data. In M. Hashem Pesaran and Peter Schmidt, editors, *Handbook of Applied Econometrics Volume II: Microeconomics*, pages 202–245. Blackwell Publishing Ltd, Oxford, UK, 1999. ISBN 978-0-631-21633-9. doi: 10.1111/b.9780631216339.1999.00009.x.

Lo-Hua Yuan, Avi Feller, and Luke W. Miratrix. Identifying and estimating principal causal effects in a multi-site trial of Early College High Schools. *The Annals of Applied Statistics*, 13(3), September 2019. ISSN 1932-6157. doi: 10.1214/18-AOAS1235.

Sahar Z Zangeneh and Roderick J Little. Likelihood-based inference for the finite population mean with post-stratification information under non-ignorable non-response. *International Statistical Review*, 90:S17–S36, 2022.

Jon Zelner, Joshua G. Petrie, Rob Trangucci, Emily T. Martin, and Arnold S. Monto. Effects of Sequential Influenza A(H1N1)pdm09 Vaccination on Antibody Waning. *The Journal of Infectious Diseases*, 220(1):12–19, June 2019. ISSN 0022-1899. doi: 10.1093/infdis/jiz055. URL http://academic.oup.com/jid/article/220/1/12/5306487.

Jon Zelner, Rob Trangucci, Ramya Naraharisetti, Alex Cao, Ryan Malosh, Kelly Broen, Nina Masters, and Paul Delamater. Racial Disparities in Coronavirus Disease 2019 (COVID-19) Mortality Are Driven by Unequal Infection Risks. *Clinical Infectious Diseases*, 72(5), 03 2021. ISSN 1058-4838, 1537-6591. doi: 10.1093/cid/ciaa1723.

Guangyu Zhang, Charles E. Rose, Yujia Zhang, Rui Li, Florence C. Lee, Greta Massetti, and Laura E. Adams. Multiple Imputation of Missing Race and Ethnicity in CDC COVID-19 Case-Level Surveillance Data. *International Journal of Statistics in Medical Research*, 11, January 2022. ISSN 1929-6029. doi: 10.6000/1929-6029.2022.11.01.

Junni L. Zhang and Donald B. Rubin. Estimation of Causal Effects via Principal Stratification When Some Outcomes are Truncated by "Death". *Journal of Educational and Behavioral Statistics*, 28(4), December 2003. ISSN 1076-9986, 1935-1054. doi: 10.3102/10769986028004353.

Junni L. Zhang, Donald B. Rubin, and Fabrizia Mealli. Likelihood-Based Analysis of Causal Effects of Job-Training Programs Using Principal Stratification. *Journal of the American Statistical Association*, 104(485), March 2009. ISSN 0162-1459, 1537-274X. doi: 10.1198/jasa.2009.0012.

Jincheng Zhou, Haitao Chu, Michael G. Hudgens, and M. Elizabeth Halloran. A Bayesian approach to estimating causal vaccine effects on binary post-infection outcomes. *Statistics in Medicine*, 35(1):53–64, January 2016. ISSN 02776715. doi: 10.1002/sim.6573.

Xiang Zhou and Jerome P. Reiter. A Note on Bayesian Inference After Multiple Imputation. *The American Statistician*, 64(2), May 2010. ISSN 0003-1305, 1537-2731. doi: 10.1198/tast.2010.09109.