INVERSE PROBABILITY WEIGHTING AND OUTCOME REGRESSION APPROACHES IN CAUSAL INFERENCE AND SURVEY SAMPLING

Bonnie E. Shook-Sa

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Public Health in the Department of Biostatistics in the Gillings School of Global Public Health.

Chapel Hill
2020

Approved by:

Michael Hudgens

Stephen Cole

John Preisser

David Rosen

Donglin Zeng

ABSTRACT

Bonnie E. Shook-Sa: Inverse Probability Weighting and Outcome Regression Approaches in Causal Inference and Survey Sampling
(Under the direction of Michael G. Hudgens)

Survey sampling and causal inference share much of the same theoretical foundation. Both fields commonly use estimation methods that rely on randomization-based or prediction-based inferential paradigms, and inverse-probability weighting (IPW) and outcome regression methods are common in both fields (Lohr, 2010; Hernán and Robins, 2020). IPW estimators are used in conjunction with marginal structural models (MSMs) to estimate causal effects from observational studies by controlling for confounding. The parametric g-formula is an outcome regression approach utilized to make causal estimates in the presence of confounding by directly modeling the outcome as a function of the exposure and confounding variables and then integrating over the distribution of the confounders. IPW estimators are fundamental in survey sampling, as they appropriately account for each unit's probability of selection within a finite population and can be further adjusted to account for nonresponding units and undercoverage of the target population. Under a prediction-based inferential paradigm, outcome regression is used to impute outcomes for units not selected into the sample based on data from sampled units.

We develop and compare methods based on IPW and outcome regression with applications in survey sampling and causal inference. Our first paper develops methods to estimate the number of HIV-positive persons incarcerated in North Carolina jails. Study data are derived from record-linkage techniques and are incomplete. Survey sampling methods are used to adjust estimates from a portion of counties to make state-level estimates that are representative of all counties. An IPW estimator is compared with an estimator based on outcome regression in simulations and with preliminary study data.

A common technique for sample size determination for complex sample surveys is to make use of the design effect, the ratio of the variance of an estimator under a complex sample design to the variance of the estimator under a simple random sample (Kish, 1965). Design effects allow researchers to calculate sample sizes under the simpler design and then inflate them to account for the use of weights in the analysis. In our second paper, we extend the theory of design effects to causal inference. The design effect approximation can be used to design causal studies that will be analyzed using MSM with IPW to control for confounding.

MSMs, the parametric g-formula, and doubly robust estimators are commonly used to make causal estimates for observational studies when the outcome of interest is continuous, binary, or categorical. In our third paper, we provide a theoretical justification for the use of these methods when the outcome is a count. We consider methods to account for overdispersion, zero-inflation, and data heaping, a common type of measurement error for count data. We present estimators for causal rate ratios along with their properties and compare the three classes of estimators via simulations. We demonstrate these methods using data from the Women's Interagency HIV Study to assess the effect of incarceration on the number of sexual partners in the subsequent six-month period.

To Gustavo and Julia

ACKNOWLEDGEMENTS

I would like to thank my dissertation advisor, Dr. Michael Hudgens, for his guidance throughout the dissertation process and for empowering me to reach my full potential. He has inspired me to be a better statistician in both methods research and in application. Thank you also to my dissertation committee Dr. Stephen Cole, Dr. John Preisser, Dr. David Rosen, and Dr. Donglin Zeng for their valuable feedback and helpful suggestions that have strengthened this research.

The research in Chapter 2 was supported by NIH grant R01 AI129731. Dr. David Rosen and Andrew Kavee are co-authors on this paper. Thanks also to Dr. Phillip Kott at RTI International for his helpful suggestions.

The research in Chapters 3 and 4 was supported by NIH grant R01 AI085073. The research in Paper 3 was funded in part through Developmental funding from the University of North Carolina at Chapel Hill Center For AIDS Research (CFAR), an NIH funded program P30 AI050410. The authors thank Dr. Stephen Cole, Noah Greifer, Shaina Alexandria, Bryan Blette, Kayla Kilpatrick, and Dr. Jaffer Zaidi for their helpful suggestions.

Data in Chapter 4 were collected by the Women's Interagency HIV Study, now the MACS/WIHS Combined Cohort Study (MWCCS). The contents of this publication are solely the responsibility of the authors and do not represent the official views of the National Institutes of Health (NIH). MWCCS (Principal Investigators): Atlanta CRS (Ighovwerha Ofotokun, Anandi Sheth, and Gina Wingood), U01-HL146241; Baltimore CRS (Todd Brown and Joseph Margolick), U01-HL146201; Bronx CRS (Kathryn Anastos and Anjali Sharma), U01-HL146204; Brooklyn CRS (Deborah Gustafson and Tracey Wilson), U01-HL146202; Data Analysis and Coordination Center (Gypsyamber D'Souza, Stephen Gange and Elizabeth Golub), U01-HL146193; Chicago-Cook County CRS (Mardge Cohen and Audrey French), U01-HL146245; Chicago-Northwestern CRS (Steven Wolinsky), U01- HL146240; Connie Wofsy Women's HIV Study, Northern California CRS (Bradley

Aouizerat and Phyllis Tien), U01-HL146242; Los Angeles CRS (Roger Detels), U01-HL146333; Metropolitan Washington CRS (Seble Kassaye and Daniel Merenstein), U01-HL146205; Miami CRS (Maria Alcaide, Margaret Fischl, and Deborah Jones), U01-HL146203; Pittsburgh CRS (Jeremy Martinson and Charles Rinaldo), U01-HL146208; UAB-MS CRS (Mirjam-Colette Kempf and Deborah Konkle-Parker), U01-HL146192; UNC CRS (Adaora Adimora), U01-HL146194. The MWCCS is funded primarily by the National Heart, Lung, and Blood Institute (NHLBI), with additional co-funding from the Eunice Kennedy Shriver National Institute Of Child Health & Human Development (NICHD), National Human Genome Research Institute (NHGRI), National Institute On Aging (NIA), National Institute Of Dental & Craniofacial Research (NIDCR), National Institute Of Allergy And Infectious Diseases (NIAID), National Institute Of Neurological Disorders And Stroke (NINDS), National Institute Of Mental Health (NIMH), National Institute On Drug Abuse (NIDA), National Institute Of Nursing Research (NINR), National Cancer Institute (NCI), National Institute on Alcohol Abuse and Alcoholism (NIAAA), National Institute on Deafness and Other Communication Disorders (NIDCD), National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). MWCCS data collection is also supported by UL1-TR000004 (UCSF CTSA), P30-AI-050409 (Atlanta CFAR), P30-AI-050410 (UNC CFAR), and P30-AI-0277 67 (UAB CFAR). Thanks to Dr. Andrea Knittel, Dr. Andrew Edmonds, Catalina Ramirez, and Dr. Adaora Adimora for contributions to the work in Chapter 4.

Thank you to Gustavo, Julia, my parents, and the rest of my family and friends. Without their support I would not have made it here. Thank you to Kimberly Enders, Hillary Heiling, Nathan Bean, and Ethan Alt for your help and support along the way. And thank you to my dogs, my silent but supportive coauthors who never left my side.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION AND LITERATURE REVIEW

1.1 Introduction

Survey sampling and causal inference share much of the same theoretical foundation, and inverse-probability weighting (IPW) and outcome regression approaches are common in both fields. This dissertation focuses on applications and extensions of IPW and outcome regression approaches in the fields of causal inference and survey sampling. Chapter 2 proposes and compares survey sampling methods utilizing IPW and outcome regression for estimating the number of HIV-positive persons incarcerated in NC jails when study data are available only for a nonrandom subset of counties. Chapter 3 extends the theory of design effects from survey sampling to causal inference to allow for design of studies that will be analyzed using MSM with IPTWs. Chapter 4 provides a theoretical justification for the use of MSMs, the parametric g-formula, and doubly robust estimators for count outcome and proposes methods that account for overdispersion and data heaping.

This chapter provides background information and covers existing literature relevant to the methodology presented in Chapters 2, 3, and 4. We discuss the inferential paradigms commonly used in causal inference and survey sampling and review research in these fields related to IPW estimators and outcome regression approaches. We discuss the history of design effects and their applications in survey sampling and causal inference, and issues related to the analysis of self-reported count data.

1.2 Causal Inference

1.2.1 Inferential Paradigms

Causal inference aims to move beyond measures of association to estimate causal effects between variables. Neyman (1923) introduced the potential outcomes framework, and Rubin (1974; 1977; 1978) popularized this framework in the 1970s. This led to the Neyman-Rubin causal model, which serves as the foundation of causal inference. Under the potential outcomes framework, for a dichotomous treatment A the potential outcome Y^a represents the outcome that would be observed under treatment assignment $A = a$, $a \in \{0, 1\}$. In causal inference, the estimands are functions of the potential outcomes. For each participant, only the potential outcome associated with the realized treatment status is observed, and the other (counterfactual) outcome is unobserved. This leads to what Holland (1986) calls the fundamental problem of causal inference, which is a missing data problem.

Two of the inferential paradigms commonly employed in causal inference are randomization-based and prediction-based (or large sample frequentist) inference. Under randomization-based inference, the potential outcomes are commonly viewed as fixed characteristics of a finite population, and the only random component is the treatment assignment A for each participant. Expectations and variances are calculated based on all possible randomizations of treatment assignments. Neyman (1923) provides unbiased estimators for the average causal effect under the randomization-based paradigm along with their variance estimators.

The prediction-based paradigm relies on a large-sample frequentist perspective. Let A_i represent the binary treatment status for participant i ($A_i = 1$ means participant i received treatment, $A_i = 0$ means participant i did not receive treatment), $\boldsymbol{L_i}$ be a vector of baseline covariates measured prior to A_i or unaffected by treatment A_i, and Y_i be the measured outcome for participant i. Under the prediction-based paradigm, $(Y_1, \boldsymbol{L_1}, A_1), (Y_2, \boldsymbol{L_2}, A_2), ..., (Y_n, \boldsymbol{L_n}, A_n)$ are viewed as an independent and identically distributed sample from an infinite population. The causal methods discussed subsequently are all based on the prediction-based paradigm.

1.2.2 IPW and Outcome Regression Approaches

Standard statistical methods cannot be used to calculate causal effects from observational data when confounding is present, as measures of association differ from the causal effects of interest. Two methods commonly employed to calculate causal estimates in the presence of confounding are marginal structural models (MSMs) with IPW and the parametric g-formula, which is an outcome regression approach.

Both methods rely on three identifiability conditions: causal consistency, positivity, and conditional exchangeability. Causal consistency defines the observed outcome Y as a function of the potential outcomes and treatment assignment: $Y = AY^1 + (1 - A)Y^0$ (Rubin, 1980; Gibbard and William, 1981; Cole and Frangakis, 2009; Pearl, 2010). Positivity states that there is a non-zero probability of each level of treatment A for all combinations of A and $\boldsymbol{L}$ in the population (Cole and Frangakis, 2009; Hernán and Robins, 2020). Under conditional exchangeability, it is assumed that $Y^a \perp A \mid \boldsymbol{L}$, $a \in \{0, 1\}$. That is, the potential outcomes are independent of the treatment assignment given the set of measured confounding variables $\boldsymbol{L}$. This assumption is also referred to as the assumption of no unmeasured confounding (Hernán and Robins, 2020).

1.2.2.1 IPW Estimation

The propensity score for each participant is defined as $e_i = Pr(A_i = 1 \mid \boldsymbol{L_i})$, the probability that participant i received treatment $A = 1$ conditional on covariates $\boldsymbol{L_i}$ (Rosenbaum and Rubin, 1983). The inverse-probability of treatment weight (IPTW) is equal to $W_i = A_i e_i^{-1} + (1 - A_i)(1 - e_i)^{-1}$. Weighting individuals by W_i creates a pseudo-population in which confounding by variables L is not present, which allows for the estimation of causal effects (Robins et al., 2000).

Marginal structural models (MSMs) were introduced by Robins (1998) and further refined by Robins et al. (2000) and Hernán et al. (2000). To fit a MSM for a binary treatment, estimated propensity scores are first obtained using a method such as logistic regression to predict the observed treatment assignments $A = a$ based on confounding variables $\boldsymbol{L}$ (Cole and Hernán, 2008). For each participant i, the corresponding IPTW is estimated as a function of the estimated propensity

score and the observed treatment assignment: $\hat{W}_i = I(A_i = 1)\hat{e}_i^{-1} + I(A_i = 0)(1 - \hat{e}_i)^{-1}$, where $I(A_i = a)$ is a $\{0,1\}$ treatment indicator for participant i. Weighted estimating equations are then used to regress the observed outcome Y on treatment A with weights $\hat{W}$. Under the assumptions of causal consistency, positivity, and conditional exchangeability, causal effects are identifiable because within the weighted population $E(Y \mid A = a) = E(Y^a \mid A = a) = E(Y^a)$ for $a \in \{0, 1\}$. When the weight model is correctly specified the average causal effect, $E(Y^1) - E(Y^0)$, can be consistently estimated as a difference in Hájek estimators for the two causal means (Lunceford and Davidian, 2004):

$$\widehat{ACE} = \hat{\mu^1} - \hat{\mu^0} = \frac{\sum_{i=1}^{n} \hat{W}_i Y_i I\left[A_i = 1\right]}{\sum_{i=1}^{n} \hat{W}_i I\left[A_i = 1\right]} - \frac{\sum_{i=1}^{n} \hat{W}_i Y_i I\left[A_i = 0\right]}{\sum_{i=1}^{n} \hat{W}_i I\left[A_i = 0\right]}$$

Lunceford and Davidian (2004) show that the empirical sandwich variance estimator provides a consistent estimate for the asymptotic variance of the estimated average causal effect when the weight model is correctly specified and that the asymptotic variance of $\widehat{ACE}$ is $\Sigma = E\{(Y^1 - \mu^1)^2 e^{-1} + (Y^0 - \mu^0)^2 (1 - e)^{-1}\}$ when the weights are treated as fixed. Furthermore, Lunceford and Davidian (2004) show that treating the weights as fixed leads to a conservative variance estimate, as the asymptotic variance when the weights are appropriately treated as estimated is less than Σ.

MSMs with IPTWs can be used to estimate casual effects from observational studies by controlling for confounding variables. They do not require modeling the relationship between the confounding variables and the outcome of interest. One key advantage of MSMs is that they can adjust for confounding from time-varying covariates that are affected by prior exposures (Robins, 1998). The drawbacks of using MSM is that it requires correct specification of the weight model, and the IPW estimator can be unstable when weights are extreme (Cole and Hernán, 2008; Little and Rubin, 2019; Robins et al., 2000).

1.2.2.2 Outcome Regression Approaches

The parametric g-formula is an outcome regression approach used in causal inference that is an alternative to MSMs with IPWs. Standardization is a common analytic technique in epidemiology research (see, for example Rothman, 2012, pages 188-192). Robins (1986) introduced the parametric g-formula as a type of standardization that allows for the estimation of causal effects by directly modeling the outcome as a function of the exposure and confounding variables and then integrating over the distribution of the confounding variables.

More specifically, under the assumptions of conditional exchangeability and causal consistency, $E[Y^a \mid \boldsymbol{L} = \boldsymbol{l}] = E[Y^a \mid \boldsymbol{L} = \boldsymbol{l}, A = a] = E[Y \mid \boldsymbol{L} = \boldsymbol{l}, A = a]$. The final quantity is identifiable from the data and can be estimated using standard parametric models (e.g. linear regression, logistic regression) (Hernán and Robins, 2020). Causal means are then estimated using the law of total probability and integrating over the distribution of $\boldsymbol{L}$, $E[Y^a] = \int E[Y \mid \boldsymbol{L} = \boldsymbol{l}, A = a] dF_L(\boldsymbol{l})$. This is typically done empirically by taking the average of the $\hat{E}[Y \mid \boldsymbol{L} = \boldsymbol{l}, A = a]$ for the observed data (Hernán and Robins, 2020).

The parametric g-formula can lead to more stable and efficient estimates compared to MSM with IPW (Daniel et al., 2013). The drawbacks of the parametric g-formula are that it requires correct specification of the outcome regression model (i.e., correctly specifying the relationship between the outcome and the exposure and confounding variables) and that it can be problematic in longitudinal settings due to a phenomenon known as the "g-null paradox". Under the g-null paradox, when there is treatment-confounder feedback, the null hypothesis of no treatment effect will be rejected with probability approaching one under the null given enough data (Hernán and Robins, 2020; Robins, 1986).

1.2.2.3 Doubly Robust Estimation Approaches

Doubly robust estimators, also referred to as augmented IPW estimators, incorporate both MSM and parametric g-formula estimators to provide protection against incorrect model-specification for either the weight or outcome model (Bang and Robins, 2005; Hernán and Robins, 2020; Funk

et al., 2011). Because the relationships between the exposure, outcome, and confounding variables are typically unknown in observational settings, doubly robust estimators afford some protection against misspecification of these models (Funk et al., 2011). While in practice all models are at least partially misspecified, Bang and Robins (2005) note that doubly robust estimators allow for minimal bias when either the weight or outcome model is nearly correct, giving the researcher two chances to get close to correct specification. Lunceford and Davidian (2004) consider the doubly robust estimators below, proposed by Robins et al. (1994), for the causal means of two binary treatment groups.

$$\hat{E}[Y^1] = n^{-1} \sum_{i=1}^{n} \frac{A_i Y_i - \{A_i - \pi(\boldsymbol{L_i}, \hat{\boldsymbol{\gamma}})\} m_1(\boldsymbol{L_i}, \hat{\boldsymbol{\beta}}_{\mathbf{1}})}{\pi(\boldsymbol{L_i}, \hat{\boldsymbol{\gamma}})}$$

and

$$\hat{E}[Y^0] = n^{-1} \sum_{i=1}^{n} \frac{(1 - A_i) Y_i + \{A_i - \pi(\boldsymbol{L_i}, \hat{\boldsymbol{\gamma}})\} m_0(\boldsymbol{L_i}, \hat{\boldsymbol{\beta}}_{\mathbf{0}})}{1 - \pi(\boldsymbol{L_i}, \hat{\boldsymbol{\gamma}})}$$

where $\pi(\boldsymbol{L_i}, \hat{\boldsymbol{\gamma}})$ is the estimated propensity score for participant i from the weight model and $m_a(\boldsymbol{L_i}, \hat{\boldsymbol{\beta}}_{\boldsymbol{a}})$ is the predicted potential outcome for participant i from the parametric g-formula model. These estimators combine the MSM and parametric g-formula estimators and are consistent and asymptotically normal when either model is correctly specified (Lunceford and Davidian, 2004). An alternative doubly robust estimator is constructed by incorporating the reciprocal of the estimated propensity score, $\hat{e}^{-1}$ as a covariate in the outcome regression model (Scharfstein et al., 1999; Bang and Robins, 2005).

1.3 Survey Sampling

1.3.1 Inferential Paradigms

Survey sampling emerged in the late 1800s and early 1900s as an estimation approach for finite populations, with early practitioners making use of methods such as stratification, clustering, and multistage sampling (Kiaer, 1897). The classic paper by Neyman (1923) presented a finite population inferential approach based on a probability sample. This laid the foundation for the randomization-based (or design-based) inferential paradigm of survey sampling. The outcomes of interest are

treated as fixed quantities associated with the finite population and the only random component is which elements of the finite population are selected into the sample (Lohr, 2010, page 54). As in the randomization-based approach to causal inference described above, expectations and variances are calculated based on all possible randomizations. In survey sampling, these randomizations are the possible samples that could have been selected. These methods are sometimes referred to as unconditional, as they rely on averages over all possible samples and are not conditioned on the observed sample (Lavrakas, 2008).

An alternative approach to sample survey estimation is the prediction-based (or model-based) inferential paradigm. Royall (1970; 1976; 1978) was one of the early pioneers of the prediction-based paradigm within survey sampling, proposing models based on a superpopulation perspective and the use of auxiliary data for analyzing survey data. Under the prediction-based approach, the outcomes themselves are considered random variables that follow a model, and the finite population is thought of as a single realization of these random variables from a superpopulation (Särndal et al., 1978). Data from the observed sample is used to predict the unobserved values from the non sampled members of the finite population, and variances are estimated using standard parametric modeling theory (Särndal et al., 2003; Bolfarine and Zacks, 1992; Lohr, 2010, page 148). These methods can be more efficient than randomization-based methods when the model is correctly specified, but can lead to substantially biased estimates when the model is incorrectly specified (Hansen, 1987).

Model assisted survey sampling and randomization assisted model based inference are additional inferential paradigms that emerged to remedy the limitations of the randomization-based and predication-based approaches. Classic randomization-based inferential methods do not accommodate nonresponding units, undercoverage of the sampling frame, or measurement error. Under model-assisted survey sampling, models are incorporated into the randomization-based inferential framework as needed to account for these limitations of randomization-based methods, but models play a secondary role to design-based inference (Särndal et al., 2003; Särndal, 2010). The Generalized Regression Estimator (GREG) is a model-assisted survey sampling estimator with

the double-robustness property. The GREG remains design-unbiased even when the regression model is misspecified (Cassel et al., 1976; Särndal et al., 2003; Kang and Schafer, 2007). The Generalized Exponential Model (GEM) is a generalization of the GREG and is further discussed in Section 1.3.2.1. Proponents of randomization assisted model based inference argue that inference should be based on the realized sample rather than across all possible samples, so this alternative framework uses prediction-based theory but brings features of the sample design (e.g. stratification and clustering) into the modeling framework (Särndal, 2010; Kott, 2005).

1.3.2 IPW and Outcome Regression Approaches

IPW and outcome regression approaches are common in survey sampling. We focus on weight calibration estimators, specifically the GEM, under the randomization assisted model based inferential paradigm and the use of outcome regression to impute missing data for units that were not selected under the prediction-based paradigm. Let N be the finite population size and n represent the sample size.

1.3.2.1 IPW Estimation

Under the randomization-based paradigm, weights w_i are defined as the reciprocal of each unit i's probability of selection. These are commonly referred to as base weights (Valliant et al., 2013, pages 311-314). Base weights can be further adjusted using auxiliary data under the model assisted or randomization assisted model based paradigms to account for nonresponding units or sampling frame undercoverage of the target population (Kott, 2006; Kott and Day, 2014; Valliant et al., 2013, Ch 13-14).

Weight calibration models aim to balance an observed sample with respect to a set of calibration variables by adjusting the weights such that sample sums of the weighted calibration variables equal the population totals of the calibration variables (Deville and Särndal, 1992; Folsom and Singh, 2000). These models can reduce selection bias resulting from missing data under the assumption that each missed observation had some unknown (but positive) probability of participating in the

study (Kott and Liao, 2015). Calibration models also reduce the variance of an estimated total compared to estimation based on uncalibrated weights when the outcome of interest is correlated with calibration variables (Deville and Särndal, 1992).

A GEM is a type of weight calibration model used to calibrate the sample to known population totals (Folsom and Singh, 2000). Under this model, the weight adjustment for each participating unit i is defined as:

$$\theta_i(\boldsymbol{x}_i, \boldsymbol{\gamma}) = \left(\frac{l_i(u_i - c_i) + u_i(c_i - l_i)\exp(a_i \boldsymbol{x}_i \boldsymbol{\gamma})}{(u_i - c_i) + (c_i - l_i)\exp(a_i \boldsymbol{x}_i \boldsymbol{\gamma})} \right)$$

where $\boldsymbol{x}_i$ is a $1 \times p$ vector of calibration variables for participant i, where p is the number of calibration variables in the model; l_i and u_i are specified by the analyst and determine the lower and upper bounds of the adjustments, respectively; c_i is a centering constant for the model; and a_i is a function of l_i, u_i, and c_i (Folsom and Singh, 2000; RTI International, 2012). Let r_i be a response indicator for participant $i : r_i = 1$ if participant i responded, 0 otherwise. The specified lower and upper bounds put constraints on the assumed response model. Specifying a lower bound of 0 implicitly models each participant's probability of response as an exponential function of the calibration variables, while specifying a lower bound of 1 implicitly assumes a logistic relationship (Kott, 2006; Kott and Liao, 2012). Let $\boldsymbol{T}_x$ be a $p \times 1$ vector of population totals known for the finite population. That is, $\boldsymbol{T}_{\boldsymbol{x}} = \sum_{i=1}^{N} \boldsymbol{x}_{\boldsymbol{i}}^T$. Then, weight adjustments for each responding unit are obtained by solving the following set of calibration equations for $\boldsymbol{\gamma}$ using Newton-Raphson:

$$\boldsymbol{s}_p(\boldsymbol{\gamma}) = \sum_{i=1}^{N} \boldsymbol{x}_i^T w_i r_i \theta_i(\boldsymbol{x}_i, \boldsymbol{\gamma}) - \boldsymbol{T}_x = \boldsymbol{0}$$

From the model, $w_{ci} = w_i \theta_i(\boldsymbol{x}_i, \hat{\boldsymbol{\gamma}})$ is the calibration adjusted weight for responding participant i ($i = 1, ..., n$). Finite population totals are estimated by taking the weighted sum of observed outcomes across the n sampled and responding units. Folsom and Singh (2000) show that the calibration estimator is asymptotically consistent and derive the asymptotic variance. The large sample variance can be approximated using Taylor series linearization (Kott and Day, 2014; Singh

and Folsom, 2000; RTI International, 2012). Kott and Liao (2012) discuss the doubly robust properties of the GEM estimator, as it provides consistent estimates when either the implied response model or the linear predictor model is correctly specified.

Utilizing weight calibration provides the researcher control over the weight adjustments while adjusting for sample undercoverage or nonresponse. This method appropriately treats the weights as estimated when calculating variances instead of treating adjusted weights as fixed or known (Shook-Sa et al., 2017). The disadvantage of weight calibration models are that results are based on asymptotic theory, so unbiasedness and confidence interval coverage might not hold in small samples (Kott, 2006).

1.3.2.2 Outcome Regression Approaches

Under the prediction-based inferential paradigm, data from the observed sample of size n is used to predict the $N - n$ unobserved values from the non-sampled units of a finite population (Lohr, 2010, page 148; Royall, 1976; Särndal et al., 2003, pages 533-534). For example, for a continuous outcome Y, the following linear regression model can be fit for sampled and responding units: $Y_i = \boldsymbol{x}_i\boldsymbol{\beta} + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$. Predicted values from the model are used to estimate the outcome of interest for non-sampled units of the finite population, and finite population totals T are estimated as the sum of the observed and predicted values in the population. Assume that units are ordered such that $i = 1, ..., n$ were selected and units $i = n + 1, ..., N$ were not selected. Then the estimated population total $\hat{T}$ can be defined as follows, and variances can be computed using using standard linear model theory (Royall, 1976; Royall and Cumberland, 1978; Lohr, 2010, page 148; Särndal et al., 2003, pages 533-534; Bolfarine and Zacks, 1992).

$$\hat{T} = \sum_{i=1}^{n} Y_i + \sum_{j=n+1}^{N} \hat{Y}_j$$

Outcome regression approaches can be more efficient than randomization-based inferential approaches (Little, 2004) and do not depend on the underlying sampling scheme (Lohr, 2010, page

148). However, bias can be significant when the model is not correctly specified (Hansen, 1987; Särndal et al., 2003, page 535).

1.4 Design Effects

Weighted estimators are fundamental in survey sampling, and methods have been developed to quantify the effect of weighting on the precision of resulting estimates. Kish (1965, page 257) introduced the design effect, which is the ratio of the variance of an estimator under a complex sample design to the variance of the estimator under a simple random sample. When observations are independent, the design effect simplifies to the design effect due to weighting ($deff_w$), or the unequal weighting effect (Kish, 1992). Let w_i represent the sampling weight for the i^{th} participant and n be the sample size. The design effect due to weighting can be calculated as:

$$deff_w = \frac{n \sum_{i=1}^{n} {w_i}^2}{\{\sum_{i=1}^{n} w_i\}^2}$$

The design effect is interpreted as an estimator's increase in variance due to differential weights across participants. While this metric was theoretically motivated by a comparison of variance estimators under different stratification allocations, it is commonly applied to all types of complex sample designs in which sample members have different probabilities of selection (Valliant et al., 2013, page 375). Gabler et al. (1999) provided a justification for how Kish's design effect applies to model-based estimators.

The design effect is commonly utilized in sample size calculations through the use of the effective sample size. The effective sample size is equal to the observed sample size divided by the design effect. It can be interpreted as the sample size under simple random sampling that produces the same variance as the sample selected under the complex design (Valliant et al., 2013, page 5).

Bayesian importance sampling uses weighting methods when sampling from one distribution to estimate the properties of another distribution (Kong et al., 1994). Importance sampling also uses the effective sample size metric to compare the variance of the weighted estimator to the variance

that would be achieved if sampling had been conducted directly from the distribution of interest (Kong et al., 1994). Kong (1992) approximates the effective sample size as the observed sample size divided by $deff_w$.

Within causal inference, $deff_w$ has been used to quantify the loss of statistical precision due to weighting (McCaffrey et al., 2004, 2013), but the design effect has not previously been theoretically justified for causal methods and has not been used in power and sample size calculations. For the purposes of study design, the advantages of the design effect are that it is (1) outcome invariant and (2) allows the sample size under the complex design to be translated into an equivalent sample size under a simpler design. The former implies that the approximated design effect depends only on the participants' weights and is constant across outcomes. The latter means that once $deff_w$ is known or approximated, it can be used in power and sample size calculations along with the simpler assumptions needed to design a study without weights.

While the design effect has a simple computational form and is thus a useful tool in study design, it is not without its limitations. For many complex sample designs, and for the extension to causal inference presented in Chapter 3, Kish's $deff_w$ approximates the true ratio of the complex variance estimator to the simple or naïve estimator. In survey sampling, practitioners note that for nonresponse adjusted weights, $deff_w$ is a good approximation when the outcome is weakly associated with the adjustment cells (Little and Vartivarian, 2005). Kalton et al. (2005) note that for sample surveys, use of $deff_w$ relies on the assumptions of homogeneous variance across adjustment strata and that weights are unrelated to the outcome of interest and note that $deff_w$ should not be applied uncritically. However, $deff_w$ remains a useful rule of thumb in the design of sample surveys and performs well in practice (Little et al., 1997; Verma et al., 1980). Kong (1992) too notes the limitation of using the effective sample size in Bayesian importance sampling because of the approximation used in its derivation. He provides the form of the remainder term and says that the approximation can be off substantially when the remainder is large.

1.5 Issues Surrounding the Analysis of Count Data

Two common issues that can complicate the analysis of self-reported count data are the presence of overdispersion in the data and data heaping. Analytic approaches are needed to account for the presence of overdispersion or data heaping when the estimand is a count or a function of counts.

1.5.1 Overdispersion

The Poisson distribution takes on non-negative integer values and is often used to model the number of events that occur in a specific time or place (Weisberg, 2005, page 271). Count data are commonly modeled using Poisson generalized linear models (GLMs), in which the rate parameter is modeled as a linear function of predictors $\boldsymbol{X}$ through the link function (McCullagh and Nelder., 1989, Ch 6). That is, we assume that $Y \sim Poisson(\lambda)$ and model the canonical (log) of λ as a linear function of covariates: $log(\lambda) = \boldsymbol{X\beta}$.

Because the Poisson distribution has a single parameter λ such that $E(Y) = Var(Y) = \lambda$, count data commonly exhibit overdispersion. Overdispersion occurs when variation in the data exceeds the expected variation based on an assumed Poisson distribution (i.e., $Var(Y) > E(Y)$) (Agresti, 2002, page 130).

Negative binomial models are one alternative when data exhibit overdispersion. The negative binomial distribution has two parameters, with $E(Y) = \lambda$ and $Var(Y) = \lambda + \lambda^2\theta$. Because the negative binomial has an additional parameter θ (the dispersion parameter), it allows the variance to exceed the mean (Agresti, 2002, page 131).

Zero-inflated Poisson (ZIP) and negative binomial (ZINB) models are also used to account for overdispersion when the data exhibit an excess of zero values compared to those expected under the assumed parametric distribution. Parameter estimates have latent interpretations, as these models have separate components that predict susceptibility for the outcome and the count outcomes among those who are susceptible (Lambert, 1992; Long et al., 2014; Preisser et al., 2016). Marginalized versions of these models have been developed to yield parameter estimates with

similar interpretations to those in the Poisson and negative binomial GLMs (Long et al., 2014; Preisser et al., 2016).

1.5.2 Data Heaping

Self-reported count data frequently exhibit a form of measurement error known as data heaping, where reported counts are rounded to different levels of precision (Wang and Heitjan, 2008). This phenomenon is commonly observed when collecting self-reported retrospective counts or measures of duration, including cigarette usage (Klesges et al., 1995), duration of breastfeeding (Singh and Folsom, 2000), and number of sexual partners (Wiederman, 1997; Roberts and Brewer, 2001). Data heaping is often attributed to cognitive processes in respondents, including choosing round numbers or approximations (digit preference) or using estimation methods to aid in recall (Roberts and Brewer, 2001). Data heaping distorts the true underlying distribution of counts and can thus lead to biased inference and increased variance (Wang and Heitjan, 2008; Cummings et al., 2015).

Methods have been proposed both to detect and account for heaping. Roberts and Brewer (2001) propose a measure and test to quantify the amount of heaping in discrete data. Heitjan and Rubin (1990) propose multiple imputation methods to account for data heaping. Singh et al. (1994) present a smoothing method for heaped time-to-event data. Bayesian mixture models are another way to account for heaping in observed data (Wright and Bray, 2003; Wang and Heitjan, 2008). Cummings et al. (2015) present an interval-censored likelihood approach for accounting for heaped count data.

1.6 Motivating Examples

Chapters 2 and 4 were directly motivated by public health research applications at the University of North Carolina. Chapter 2 was developed to support the estimation of the number of HIV-positive persons incarcerated in North Carolina jails, and Chapter 4 was developed to estimate the effect of incarceration on the number of sexual partners in the subsequent six-month period using data from the Women's Interagency HIV Study.

1.6.1 Estimating the Number of HIV-Positive Persons in North Carolina Jails

The methods in Chapter 2 were developed to estimate the number of HIV-positive persons incarcerated in North Carolina (NC) jails overall and within each of the 100 counties over a fixed period of time. NC does not maintain a list of HIV-positive persons incarcerated in its jails. To estimate the size of this population, jail incarceration records, statewide court records, and confidential statewide health department records of persons living with HIV will be linked. This record linkage process will provide estimates for the number of HIV-positive persons incarcerated in the 26 counties with available jail incarceration data, but not for the remaining 74 counties. The characteristics of the counties with publicly-available data likely differ from those without publicly-available data, so appropriate statistical methods are needed generalize these results to the entire state of NC.

1.6.2 Women's Interagency HIV Study

The Women's Interagency HIV Study (WIHS) is a multicenter cohort study of women living with HIV or at risk of acquiring HIV (Adimora et al., 2018). At each biannual visit, the WIHS collects data regarding women's incarceration status and the sexual behavior since the prior visit. The methods in Chapter 4 allow for estimation of the effect of incarceration on the number of sexual partners in the subsequent six-month period. Because these are observational data and many variables can confound this effect (e.g. drug and alcohol use, unstable housing, sex exchange practices), statistical methods are needed to control for confounding, overdispersion, and data heaping in estimation.

CHAPTER 2: SURVEY SAMPLING APPROACHES TO ESTIMATE THE NUMBER OF HIV-POSITIVE PERSONS IN NORTH CAROLINA JAILS

2.1 Introduction

Calculating estimates for rare or dynamic populations is challenging when no single data source contains all information necessary for estimation. A lack of a viable sampling frame further precludes the use of conventional methods for estimating the size and characteristics of a rare or ever-changing population. Record linkage techniques allow for multiple data sources to be combined, which can facilitate indirect estimation of the target population (Harron et al., 2017; Qayad and Zhang, 2009; St. Sauver et al., 2011). However, when record linkage results in missing data, methods are needed to generalize findings based on linked records to the target population (Bohensky et al., 2010; Ford et al., 2006; Harron et al., 2014; Judson et al., 2013).

The goal of this study is to develop and evaluate methods to account for missing data following record linkage to estimate the number of HIV-positive persons incarcerated in the state of North Carolina (NC) jails overall and within each of the 100 counties over a fixed period of time. Estimates of HIV within and across county jails in the state could be used by jail administrators as well as by local and state public health officials in efforts to ensure the availability of adequate medical resources to treat HIV during periods of jail incarceration and to support incarcerated persons' health needs as they are released and return to the community.

NC does not maintain a list of HIV-positive persons incarcerated in jails. To calculate estimates for this rare population, the following individual-level datasets will be linked using incomplete personal identifiers: jail incarceration records derived from published inmate rosters available from 26 counties, statewide court records, and confidential statewide health department records of persons living with HIV. This record linkage process will provide estimates for the number of

HIV-positive persons incarcerated in jails in the 26 counties with available jail incarceration data. The characteristics of the counties with publicly-available data likely differ from those without publicly-available data, so care must be taken when generalizing these results to the entire state of NC.

This paper uses preliminary study data to develop methods for estimating the number of persons incarcerated in NC jails living with HIV. The findings from this methods evaluation will be used to determine the primary analytic methods that will be applied to the final dataset once data collection and linkage are complete. Two methods are considered, both of which provide estimates for the total number of persons incarcerated in NC jails living with HIV, and one of which also provides county-level estimates within the 74 counties without publicly-available jail rosters. Both methods use sampling inference approaches that leverage county-level characteristics, either via outcome regression or weight calibration modeling. In Section 2.2 both methods are presented. Section 2.3 describes a simulation study assessing and comparing these methods based on bias and precision, and Section 2.4 provides an illustration of the two approaches based on preliminary study data. Section 2.5 concludes with a discussion of the results and outlines the limitations of our study.

2.2 Methods

2.2.1 Record Linkage

To generate estimates of the number of people in NC jails living with HIV, the following datasets will be linked: jail incarceration records, state court records, and confidential NC health department records of persons in the state known to be living with HIV. A database of jail incarceration records is being created using a technique called webscraping, in which, several times a day, automated 'bots' collect individual-level incarceration data from jail rosters published on county jail websites. In NC, 26 of the 100 counties have such online rosters, which include the incarcerated persons' full names and ages (or dates of birth, DOBs). The 26 counties with available jail rosters are depicted in Figure 2.1. Over a three-month period, this process generates over 45,000 inmate-level records.

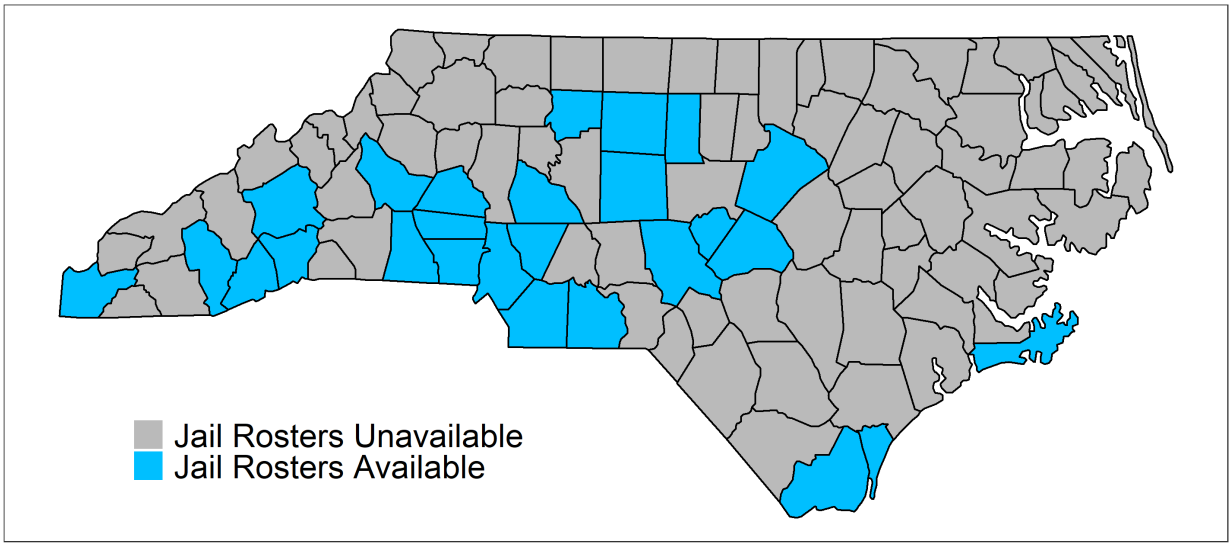

Figure 2.1: Availability of Jail Rosters by NC County

State criminal court records are being obtained from the NC Administrative Office of the Courts. These records include variables such as defendants' full names, DOBs, partial social security numbers (SSNs), and the county court system in which defendants' cases will be adjudicated. Over a three-month period, there are over 400,000 defendant-level records in the 26 counties with online inmate-level rosters and nearly 800,000 defendant-level records for the entire state of NC.

Finally, the NC State Health Department maintains a confidential database of all living persons who were diagnosed with HIV in the state. Among other variables, this database includes patients' full names, DOBs, SSNs, and dates of HIV diagnoses. This database tracks persons with positive HIV tests and persons accessing HIV services in NC.

A deterministic data linkage of jail and court records is being performed on an ongoing basis using the first and last names, DOBs or ages, and counties. These linked records will be provided to the state health department to be linked to the state's database of HIV records using probabilistic matching based on first and last name, DOB, and partial SSN. In return, the state health department will provide the estimated number of HIV-positive defendants in each of the 100 counties and the estimated number of HIV-positive persons in each of the 26 jails with roster data. The Internal

Review Board at the University of North Carolina at Chapel Hill reviewed and approved the study protocol (IRB # 17-0946).

2.2.2 Preliminary Data Description

To avoid using final study data to develop the statistical analysis plan, methods development is conducted using preliminary study data based on three months of linked jail-court records. Public health records have not yet been linked to the individual level jail-court records, so preliminary data includes the total number of defendants in each county over a three-month period but not the number of HIV-positive defendants or the number of HIV-positive persons incarcerated in the 26 counties with publicly-available jail rosters.

For the purposes of developing a preliminary data file for use in methods evaluation, the number of HIV-positive defendants in each county was approximated by multiplying the number of defendants by the estimated prevalence of HIV among the defendant population in each county. The estimated prevalence of HIV in each NC county was obtained by dividing the count of residents who were HIV-positive (North Carolina HIV/STD/Hepatitis Surveillance Unit, 2017) by the estimated number of county residents in 2016 from the US Census Bureau (U.S. Census Bureau Population Division, 2018). The assumed HIV prevalence among defendants was set equal to the county prevalence multiplied by five. This multiplier was chosen to reflect the approximate increased risk of HIV among persons entering the criminal justice system, based on the known prevalence of HIV in NC and the estimated prevalence among those entering the criminal justice system (Wohl et al., 2013).

The final linked dataset will include estimates of the number of HIV-positive defendants in each of the 100 counties and the number of HIV-positive persons incarcerated in 26 counties, so the methods below are developed based on the availability of these data. The simulations and preliminary analyses in Sections 2.3-2.4 treat the approximate numbers of HIV-positive defendants and incarcerated persons on the preliminary data file as estimates from record linkage.

2.2.3 Estimation

The estimands are the total number of HIV-positive persons incarcerated in jails over a fixed period of time (1) in NC ($n_{I,NC}$) and (2) within each of the 100 counties in NC ($n_{I,i},\ i = 1, .., 100$). Within each county *i*, the number of HIV-positive persons who were incarcerated equals the number of HIV-positive defendants ($n_{D,i}$) multiplied by the proportion of HIV-positive defendants in county *i* who were incarcerated ($P(I \mid D)_i$). That is: $n_{I,i} = n_{D,i}P(I \mid D)_i$.

The total number of HIV-positive persons incarcerated in NC jails is equal the sum of the 100 county-level totals: $n_{I,NC} = \sum_{i=1}^{100} n_{I,i}$. The number of HIV-positive defendants in each of the 100 counties in NC ($n_{D,i}$) will be estimated as described in Section 2.2.1 and the proportion of HIV-positive defendants who were incarcerated in the 26 counties with publicly-available jail rosters ($P(I \mid D)_i$) will be estimated through the jail-court record linkage process. Thus, $n_{I,i}$ will be estimated for the 26 counties that have publicly-available jail rosters. For the purposes of this paper, the estimates resulting from record linkage are treated as known quantities (i.e., error in the record linkage process is assumed to be negligible and is ignored). Furthermore, the 100 counties are assumed to be independent. Without loss of generality, the notation in the remainder of the paper assumes that counties are ordered such that $n_{I,i}$ is known for $i = 1, ..., 26$ and unknown for $i = 27, ..., 100$.

Two statistical methods for obtaining estimates of $n_{I,NC}$ are considered: outcome regression and weight calibration. The outcome regression approach aims to estimate $P(I \mid D)_i$ in the 74 counties where this quantity is unknown, and thus also provides estimates for $n_{I,i}$ within these counties. The weight calibration approach uses a weighting adjustment to estimate $n_{I,NC}$ directly and does not provide county-level estimates. Both approaches leverage county-level covariates thought to be associated with the number of HIV-positive persons incarcerated in each county and the proportion of HIV-positive defendants who were incarcerated. These county-level covariates, which were obtained from Vera Institute's publicly-available "Incarceration Trends" dataset for the year 2014 include the following: annual jail admissions, daily jail populations, index crime rates,

poverty levels, and urbanicity (Vera Institute of Justice, 2014). Additionally, the number of unique defendants obtained from court records is used as a county-level covariate.

2.2.3.1 Outcome Regression

Two inferential approaches that are commonly used to calculate sample estimates for a finite population are randomization-based and prediction-based inference (Lohr, 2010, pages 54-55). Under randomization-based inference, outcomes of the finite population are viewed as fixed, and the only random component is whether or not individual population members were selected into the sample. Under the prediction-based approach, the outcomes themselves are considered random variables that follow a model, and the finite population is thought of as a single realization of these random variables. Data from the observed sample is used to predict the unobserved values from the unsampled members of the finite population, and variability is estimated using standard linear model theory (Lohr, 2010, page 148).

The outcome regression approach follows the prediction-based paradigm by using $P(I \mid D)_i$ in the 26 counties with publicly available jail rosters to estimate the proportions in the remaining 74 counties without publicly available rosters, based on the set of predictors listed in Section 2.2.3. To estimate $P(I \mid D)_i$, the following linear regression model will be fit for the 26 counties with publicly-available rosters: $Y_i = \boldsymbol{x}_i\boldsymbol{\beta} + \epsilon_i$, where Y_i is the known proportion $P(I \mid D)_i$ in county i, $\boldsymbol{x}_i$ represents the $1 \times p$ vector of covariates for county i, and ϵ_i represents the random error. It is assumed that $\epsilon_i \sim N(0, \sigma^2)$ and that errors are independent across counties.

Predicted values $\hat{Y}_i^* = \hat{P}(I \mid D)_i$ and 95% prediction intervals ($\hat{P}(I \mid D)_{i,LCL}$, $\hat{P}(I \mid D)_{i,UCL}$) will be obtained from this model for each of the 74 counties without publicly-available jail rosters. Let $t_{0.025,25-p}$ be the 97.5^{th} percentile of the t-distribution with $25 - p$ degrees of freedom. Then $\hat{P}(I \mid D)_i = \boldsymbol{x}_i\hat{\boldsymbol{\beta}}$ and $\hat{P}(I \mid D)_{i,LCL}$ and $\hat{P}(I \mid D)_{i,UCL}$ are the lower and upper limits of

$$\hat{P}(I \mid D)_i \pm t_{0.025,25-p}\hat{\sigma}\sqrt{1 + \boldsymbol{x}_i(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i^T}$$

where $\hat{\boldsymbol{\beta}}$ are the maximum likelihood estimates for $\boldsymbol{\beta}$, $\boldsymbol{X}$ is the $26 \times p$ matrix of predictors for the regression model from the 26 counties with publicly-available jail rosters, and the mean squared error (MSE) $\hat{\sigma}^2$ is used to estimate σ^2.

To obtain the estimated number of HIV-positive persons incarcerated in jails in each of the 74 counties without publicly-available jail rosters along with 95% prediction intervals, the predicted proportions and prediction interval limits will be multiplied by the number of HIV-positive defendants in the county. That is, point estimates and the lower and upper endpoints for the 95% prediction intervals are defined as follows, respectively: $\hat{n}_{I,i} = n_{D,i}\hat{P}(I \mid D)_i$, $\hat{n}_{I,i,LCL} = n_{D,i}\hat{P}(I \mid D)_{i,LCL}$, and $\hat{n}_{I,i,UCL} = n_{D,i}\hat{P}(I \mid D)_{i,UCL}$.

An estimate for the total number of HIV-positive persons incarcerated in NC jails can be obtained by adding the known number of HIV-positive persons incarcerated in the 26 counties to the estimated number of HIV-positive persons incarcerated in the 74 counties without publicly-available jail rosters to obtain:

$$\hat{n}_{I,NC,OR} = \sum_{i=1}^{26} n_{I,i} + \sum_{j=27}^{100} n_{D,j}\hat{Y}_j^*$$

This estimator is unbiased and has minimum variance among the class of all unbiased estimators of $n_{I,NC}$ under the assumed linear regression model above (Bolfarine and Zacks, 1992, pages 31-32). Let $\boldsymbol{X}^*$ be the $74 \times p$ matrix of covariates for the 74 counties without publicly-available jail rosters, $\rho^T = (n_{D,27}, n_{D,28}, ..., n_{D,100})$, and $\boldsymbol{I}$ be the 74×74 identity matrix. Then, a 95% confidence interval for $n_{I,NC}$ is (Bolfarine and Zacks, 1992, Page 122): $\hat{n}_{I,NC,OR} \pm t_{0.025,25-p}\sqrt{\rho^T \hat{\boldsymbol{\Sigma}} \rho}$, where $\hat{\Sigma} = \hat{\sigma}^2 \left\{ \boldsymbol{X}^*(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^{*T} + \boldsymbol{I} \right\}$.

2.2.3.2 Weight Calibration

Weight calibration aims to balance an observed sample with respect to a set of calibration variables by constructing the weights such that sample sums of the weighted calibration variables equal the population totals of the calibration variables (Deville and Särndal, 1992; Folsom and Singh, 2000). Weight calibration can reduce selection bias resulting from missing data under the assumption that each missed observation had some unknown (but positive) probability of

participating in the study (Kott and Liao, 2015). Weight calibration can also reduce the variance of an estimated total compared to estimation based on uncalibrated weights when the outcome of interest is correlated with calibration variables (Deville and Särndal, 1992).

Weight calibration will be used to estimate the total number of HIV-positive persons incarcerated in NC jails. Calibrated weights will be computed for each of the 26 "responding" counties (i.e., the counties for which publicly-available jail rosters were available) via a generalized exponential model (Folsom and Singh, 2000; Kott, 2006), where the covariates listed above are calibrated to known state-level totals. In addition, because the weight calibration model is estimating $n_{I,NC}$ directly instead of the $P(I \mid D)_i$, the number of HIV-positive defendants will also be included as a calibration variable, as this variable should be highly predictive of $n_{I,i}$.

For this method, each of the 100 counties has an initial weight of 1 (i.e., $w_i = 1$, $i = 1, ..., 100$). That is, each county represents itself when combining county-level estimates to form an overall estimate for NC. The weight adjustment for county i is defined as:

$$\theta_i(\boldsymbol{x}_i, \boldsymbol{\gamma}) = \left(\frac{l_i(u_i - c_i) + u_i(c_i - l_i)\exp(a_i \boldsymbol{x}_i \boldsymbol{\gamma})}{(u_i - c_i) + (c_i - l_i)\exp(a_i \boldsymbol{x}_i \boldsymbol{\gamma})} \right)$$

where $\boldsymbol{x}_i$ is the $1 \times p$ vector of calibration variables for county i; l_i and u_i are specified by the analyst and determine the lower and upper bounds of the adjustments, respectively; c_i is a centering constant for the model; and a_i is a function of l_i, u_i, and c_i (Folsom and Singh, 2000; RTI International, 2012; Kott, 2006). For this application, no constraints are imposed on the adjustment factors, i.e., $l_i = 0,\ u_i = e^{20}$(essentially unbounded), and $c_i = 1$.

Let r_i be a response indicator for county i : $r_i = 1$ if county i responded, $r_i = 0$ otherwise. Let $\boldsymbol{T}_x$ be a $p \times 1$ vector of calibration variable totals for the finite population, i.e., all 100 counties in NC. That is, $\boldsymbol{T}_{\boldsymbol{x}} = \sum_{i=1}^{N} \boldsymbol{x}_{\boldsymbol{i}}^T$. Then, weight adjustments for each responding county are obtained by solving the following set of calibration equations for $\boldsymbol{\gamma}$ using Newton-Raphson:

$$\boldsymbol{s}_p(\boldsymbol{\gamma}) = \sum_{i=1}^{N} \boldsymbol{x}_i^T w_i r_i \theta_i(\boldsymbol{x}_i, \boldsymbol{\gamma}) - \boldsymbol{T}_x = \boldsymbol{0}$$

From the model, $w_{ci} = w_i\theta_i(\boldsymbol{x}_i, \hat{\boldsymbol{\gamma}})$ is the calibration adjusted weight for responding county i ($i = 1, ..., 26$). This method implicitly models each county's probability of response as an exponential function of the calibration variables (Kott, 2006). This estimator is doubly robust, in that it provides consistent estimates when either the implied response model or the linear predictor model is correctly specified (Kott and Liao, 2012).

The total number of HIV-positive persons incarcerated in NC jails will be estimated by the following calibration estimator: $\hat{n}_{I,NC,WC} = \sum_{i=1}^{n} w_{ci} n_{I,i}$, where $n = 26$ (the number of responding counties). Folsom and Singh (2000) show that the calibration estimator is asymptotically consistent and derive the asymptotic variance. Ignoring the finite population correction adjustment, the variance estimator is:

$$\widehat{Var}(\hat{n}_{I,NC,WC}) = \frac{n}{n-1}\left\{\sum_{i=1}^{n}(w_{ci}e_i)^2 - \frac{\left(\sum_{i=1}^{n} w_{ci}e_i\right)^2}{n}\right\}$$

where $e_i = n_{I,i} - \boldsymbol{x}_i\left(\sum_{j=1}^{n} w_{cj}\boldsymbol{x}_j^T\boldsymbol{x}_j\right)^{-1}\left(\sum_{j=1}^{n} w_{cj}\boldsymbol{x}_j^T n_{I,j}\right)$ (Kott, 2006; Shook-Sa et al., 2017). A finite population correction (fpc) adjustment can be made to this variance estimator to account for the large percentage of observed counties (26/100). The fpc adjustment is calculated as $(1 - n/N)$, where $N = 100$ is the total number of counties in the finite population (see, for example, Kish, 1965, page 43). The fpc-adjusted variance estimator is $\widehat{Var}(\hat{n}_{I,NC,WC})_{fpc} = (1 - n/N)\,\widehat{Var}(\hat{n}_{I,NC,WC})$. Then, a 95% confidence interval for $n_{I,NC}$ is (RTI International, 2012)

$$\hat{n}_{I,NC,WC} \pm t_{0.025,n-1}\sqrt{\widehat{Var}(\hat{n}_{I,NC,WC})}$$

where $\widehat{Var}(\hat{n}_{I,NC,WC})_{fpc}$ can be substituted in for $\widehat{Var}(\hat{n}_{I,NC,WC})$ to obtain an fpc-adjusted confidence interval.

Utilizing weight calibration provides the researcher control over the weight adjustments. When an intercept is included in the calibration model, this method also ensures that the calibration adjusted weights will sum to the number of counties, which is critical when the estimand of interest

is a total. Finally, this method appropriately treats the weights as estimated when calculating variances instead of treating adjusted weights as fixed or known (Shook-Sa et al., 2017).

2.3 Simulation Study

A simulation study was conducted to examine small and large sample properties of the two estimation approaches, provided that the models used for each method are correctly specified. True values of $P(I \mid D)_i$ and $n_{I,i}$ were generated for each of the $N = 100$ counties (as outlined below), as well as three sets of response probabilities that allow for the examination of three sample size scenarios: $n = 26$ (the observed number of counties), $n = 50$, and $n = 75$. In addition, larger finite populations of $N = 200$ and $N = 500$ were simulated, each with the same response probability distributions specified for the $N = 100$ population.

For the $N = 100$ population, the covariates and the number of defendants ($n_{D,i}$) from the preliminary data file were used. For the $N = 200$ and $N = 500$ populations, the the preliminary data file was duplicated, with each county appearing either twice (for $N = 200$) or five times (for $N = 500$). Normally-distributed random noise was added to each covariate and to the number of defendants in each county to obtain larger finite populations with counties that were similar, but not identical, to the 100 NC counties on the preliminary data file. Given the covariates for each county, a true $P\left(I \mid D\right)_i$ was generated for each of the counties in the finite population ($N = 100$, $N = 200$, or $N = 500$) as $P(I \mid D)_i = 0.155 - 0.03x_{i1} + 0.01x_{i2} - 0.001x_{i3} + \epsilon_i$, where x_{i1} was a rural/suburban indicator variable for county i, x_{i2} was the square root transformed index crime rate for county i, and x_{i3} was the percent below poverty value for county i. The error term $\epsilon_i \sim N(0, \sigma^2)$, where $\sigma = 0.015$. The resulting $P\left(I \mid D\right)_i$ ranged from 0.08 to 0.20 for $N = 100$ (median of 0.12), from 0.07 to 0.20 for $N = 200$ (median of 0.13), and from 0.06 to 0.22 for $N = 500$ (median of 0.13). The simulated distributions were fairly consistent with the distribution in the preliminary data. These $P\left(I \mid D\right)_i$ values were used to calculate a true number of HIV-positive persons incarcerated within each county under the simulation: $n_{I,i} = n_{D,i}P(I \mid D)_i$. This resulted in $n_{I,i}$ ranging from 0.1 to 314.8 for $N = 100$, from 0.1 to 309.7 for $N = 200$, and from 0 to 350.0 for $N = 500$, with a

sum of $n_{I,NC} = 1911.7$ (for $N = 100$), $n_{I,NC} = 3829.4$ (for $N = 200$), and $n_{I,NC} = 9752.4$ (for $N = 500$).

For each finite population size ($N = 100$, $N = 200$, and $N = 500$ counties), three sets of response probabilities were generated for each county, with $log\left(P(r_1 = 1)\right) = -\{\lambda_0 + 0.25x_{i1} - 0.02x_{i2} + 0.005x_{i3}\}$, where x_{i1}, x_{i2}, and x_{i3} are as defined above. Three values of λ_0 were chosen such that the mean response across the counties for the three simulation scenarios was 0.26 ($\lambda_0 = 1.12$), 0.50 ($\lambda_0 = 0.47$), and 0.75 ($\lambda_0 = 0.068$). This resulted in expected sample sizes for $N = 100$ of $n = 26$ (the number of responding counties in the preliminary data file), $n = 50$, and $n = 75$. For $N = 200$, this resulted in expected sample sizes of $n = 52$, $n = 100$, and $n = 150$. For $N = 500$, this resulted in expected sample sizes of $n = 130$, $n = 250$, and $n = 375$.

After generating true values under the specified models, $R = 1000$ simulated samples were generated for each of the nine finite population and sample size scenarios. Each simulated sample was obtained by assigning each county a binary response status based on a random draw from a Bernoulli random variable with mean $P(r_i = 1)$. After the respondent status was assigned, it was assumed that the true $P\left(I \mid D\right)_i$ and $n_{I,i}$ were observed only for the responding counties, and the outcome regression and weight calibration methods were implemented to estimate $n_{I,i}$ and/or $n_{I,NC}$ with correctly specified models. For each iteration of the simulation, the following statistics were obtained: $\hat{n}_{I,NC,OR}$, $\hat{n}_{I,NC,WC}$, their estimated standard errors, whether or not the 95% confidence interval for each method included $n_{I,NC}$, and the number of nonresponding counties in which the outcome regression 95% prediction intervals captured the true $n_{I,i}$.

The simulation results are summarized in Table 2.1 for each of the nine finite population size by sample size combinations. Density plots for the distribution of $\hat{n}_{I,NC,OR}$ and $\hat{n}_{I,NC,WC}$ across the $R = 1000$ simulated datasets for $N = 100$ are shown in Figure 2.2. Density plots for the $N = 200$ and $N = 500$ populations were similar and are not shown. Empirical bias was fairly small (3.17% or lower) for all sample sizes, finite population sizes, and methods. The largest empirical bias was for the weight calibration method when $N = 100$ and $n = 26$. Among the remaining scenarios, absolute mean percent bias ranged from 0.00% to 0.89% of the true value. For all population and

sample sizes, the distributions of both estimators are centered close to the true value of $n_{I,NC}$ (depicted with vertical dotted lines in Figure 2.2).

For outcome regression, the average estimated standard error tracked fairly closely with the empirical standard error. The outcome regression confidence intervals were slightly conservative for the smallest finite population $N = 100$, but but the empirical coverage was close to the nominal 95% for the larger finite populations. The weight calibration method exhibited undercoverage when $N = 100$, $n = 26$, regardless of whether or not the finite population correction (fpc) adjustment was made. The observed coverage rates were 77% and 81% with and without the fpc, respectively. For larger finite population sizes, weight calibration tended to be conservative when the fpc was ignored and provided close to nominal coverage when the fpc was applied.

The outcome regression method led to more precise estimates of $n_{I,NC}$, as the 95% confidence interval half-widths were much smaller than the weight calibration half-widths for all population and sample size scenarios. This is depicted in Figure 2.2, as there is more spread in the distributions of the estimates for the weight calibration method compared to the outcome regression method, regardless of the population size or the sample size. County-level prediction intervals for the outcome regression approach had the appropriate level of coverage of the true values, as 94-95% of prediction intervals included the true $n_{I,i}$ across the $R = 1000$ simulated samples for each scenario.

These simulations provide insight regarding the statistical methods specified in Section 2.2 under correctly-specified models. The outcome regression approach had good empirical properties, with the estimator empirically unbiased and the corresponding confidence and prediction intervals having empirical coverage rates approximately equal to the nominal level. Weight calibration led to anti-conservative confidence interval coverage for the small finite population and sample size scenario associated with the observed data for this study ($N = 100, n = 26$).

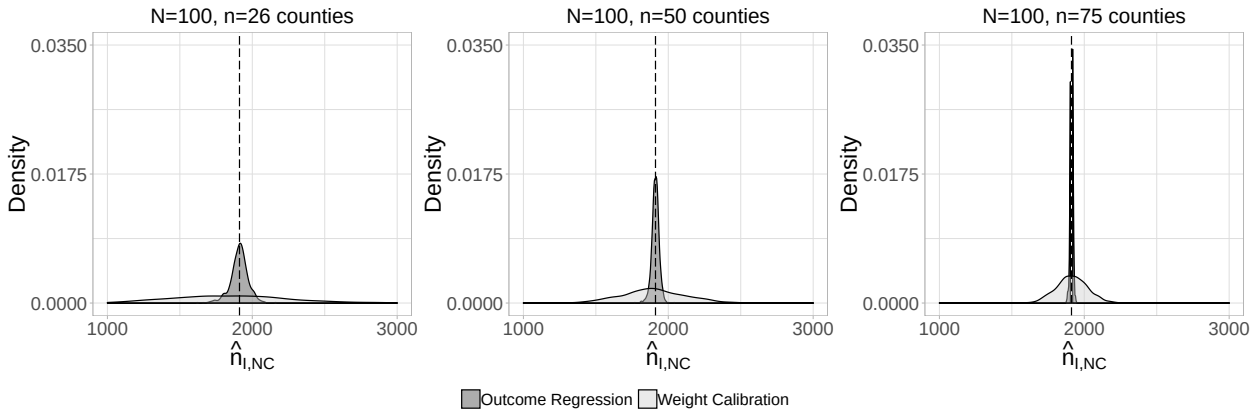

Figure 2.2: Density Plots for Outcome Regression vs. Weight Calibration, $R = 1000$ Simulations

Table 2.1: Simulation Summary Results, $R = 1000$ Simulations

N	*Avg n*	*Method*	*Avg* $\hat{n}_{I,NC}$	*Empirical Bias (%)*	*ASE*	*ESE*	*SER*	*95% CI Coverage (state-level estimate)*	*95% CI Half-Width (state-level estimate)*	*95% PI Coverage (county-level estimates)*
100	26	Outcome Regression	1909.9	-0.09%	65.8	64.5	1.02	0.98	137.0	0.94
		Weight Calibration, no fpc	1851.2	-3.17%	324.2	400.1	0.81	0.81	668.7	n/a
		Weight Calibration, fpc	1851.2	-3.17%	278.5	400.1	0.70	0.77	545.9	n/a
	50	Outcome Regression	1911.1	-0.03%	31.2	25.4	1.23	0.99	62.9	0.94
		Weight Calibration, no fpc	1894.6	-0.89%	275.3	221.1	1.25	0.95	553.5	n/a
		Weight Calibration, fpc	1894.6	-0.89%	194.6	221.1	0.88	0.89	381.4	n/a
	75	Outcome Regression	1912.4	0.04%	10.2	8.5	1.20	0.98	20.3	0.93
		Weight Calibration, no fpc	1914.4	0.14%	234.7	106.3	2.21	1.00	467.7	n/a
		Weight Calibration, fpc	1914.4	0.14%	117.2	106.3	1.10	0.97	229.6	n/a
200	52	Outcome Regression	3831.3	0.05%	84.3	78.9	1.07	0.97	169.7	0.94
		Weight Calibration, no fpc	3796.8	-0.85%	522.5	496.5	1.05	0.94	1049.4	n/a
		Weight Calibration, fpc	3796.8	-0.85%	449.5	496.5	0.91	0.90	881.1	n/a
	100	Outcome Regression	3827.0	-0.06%	43.8	43.6	1.00	0.96	86.9	0.94
		Weight Calibration, no fpc	3827.4	-0.05%	391.3	270.5	1.45	0.99	776.6	n/a
		Weight Calibration, fpc	3827.4	-0.05%	276.5	270.5	1.02	0.95	542.0	n/a
	150	Outcome Regression	3823.7	-0.15%	15.4	13.9	1.11	0.96	30.5	0.95
		Weight Calibration, no fpc	3827.7	-0.05%	322.0	130.6	2.46	1.00	636.2	n/a
		Weight Calibration, fpc	3827.7	-0.05%	161.8	130.6	1.24	0.98	317.1	n/a
500	130	Outcome Regression	9752.9	0.00%	123.1	132.2	0.93	0.94	243.6	0.94
		Weight Calibration, no fpc	9694.2	-0.60%	904.9	792.1	1.14	0.96	1790.4	n/a
		Weight Calibration, fpc	9694.2	-0.60%	777.7	792.1	0.98	0.93	1524.3	n/a
	250	Outcome Regression	9753.8	0.01%	66.4	68.9	0.96	0.94	130.9	0.94
		Weight Calibration, no fpc	9770.9	0.19%	664.8	435.1	1.53	1.00	1309.2	n/a
		Weight Calibration, fpc	9770.9	0.19%	469.2	435.1	1.08	0.96	919.7	n/a
	375	Outcome Regression	9751.2	-0.01%	24.0	25.5	0.94	0.95	47.1	0.94
		Weight Calibration, no fpc	9763.5	0.11%	543.4	230.0	2.36	1.00	1068.6	n/a
		Weight Calibration, fpc	9763.5	0.11%	271.9	230.0	1.18	0.98	532.9	n/a

fpc = finite-population correction; ASE=Average Estimated Standard Error; ESE=Empirical Standard Error; SER=Standard Error Ratio (ASE/ESE)

2.4 Preliminary Data Results

The outcome regression and weight calibration methods were implemented on the preliminary data, as outlined in Section 2.2. For both methods, annual jail admissions, daily jail populations, index crime rates, and the number of defendants per county were square-root transformed to reduce skewness of these variables and thus the over-influence of the largest counties on model fit.

2.4.1 Outcome Regression

Figure 2.3 compares the known $P(I \mid D)_i$ with the estimated $\hat{P}(I \mid D)_i$ in the 26 counties for which the outcome regression model was fit. These values align fairly well along the 45-degree line of equality ($R^2 = 0.656$), which is indicative of reasonable model prediction. Table 2.2 displays the estimated model parameters. When predicting $P(I \mid D)_i$ based on these covariates one at a time (single variable, or SV, models) or using all covariates simultaneously (the multivariable, or MV, model), the number of unique defendants is the strongest predictor of $P(I \mid D)_i$. This predictor is stronger when conditioning on the other predictors than it is marginally. Because of the high correlation among the set of predictor variables, a sensitivity analysis was conducted with subsets of predictor variables to ensure robustness of the findings. The resulting $\hat{n}_{I,NC,OR}$ estimates and 95% confidence intervals were similar for all models examined.

Based on the MV model, $\hat{P}(I \mid D)_i$ were computed and the number of HIV-positive persons incarcerated within the 74 counties without publicly-available jail rosters, $\hat{n}_{I,i}$, were estimated. Table 2.3 displays county-level estimates for the 10 largest and 10 smallest counties in NC along with 95% prediction intervals for the counties for which $\hat{P}(I \mid D)_i$ was estimated from the model rather than known. Because these results are based on preliminary data, counties are anonymized. Based on the outcome regression approach, $\hat{n}_{I,NC,OR} = 1297.6$. The 95% confidence interval for this estimate is (1198.5, 1396.7).

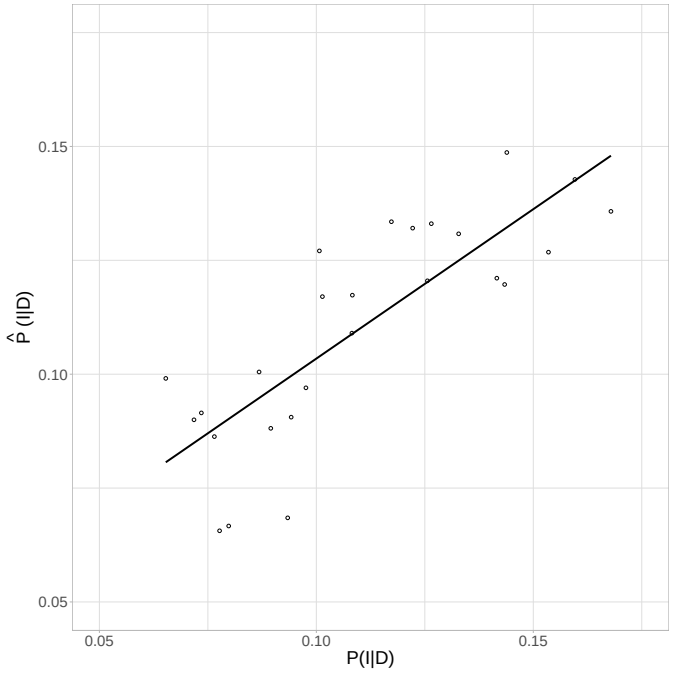

Figure 2.3: Predicted vs. Actual Proportion of Defendants who were Incarcerated, $n = 26$ Counties with Publicly-Available Jail Rosters

2.4.2 Weight Calibration

The association between county response status in the study (i.e., the availability of jail roster data) and the covariates of interest was explored. Table 2.4 presents the distribution of covariates by county response status. Responding counties tend to have more total defendants and HIV-positive defendants compared to nonresponding counties. Annual jail admissions, daily jail populations, and index crime rates are also higher in responding counties than nonresponding counties. Nonresponding counties have higher poverty rates and are more rural compared to responding counties.

The weight calibration model estimate was calculated as specified in Section 2.2.3.2, except that the daily jail population covariate was exclude due to collinearities with the other covariates that resulted in a lack of model convergence. Based on the simulation study, the sample size of $n = 26$ counties is not large enough to ensure appropriate coverage of 95% confidence intervals, and undercoverage was made worse when the fpc adjustment was applied. For this reason, the fpc

Table 2.2: Parameter Estimates for Multivariable and Single Variable Prediction Models, Outcome Regression

Variable Description	$\hat{\beta}$, *MV*	*Est SE, MV*	*95% Lower Conf Limit, MV*	*95% Upper Conf Limit, MV*	*p-value, MV*	$\hat{\beta}$, *SV*	*Est SE, SV*	*95% Lower Conf Limit, SV*	*95% Upper Conf Limit, SV*	*p-value, SV*
Percent below Poverty	0.0009	0.0015	-0.0022	0.0040	0.5543	0.0005	0.0019	-0.0035	0.0045	0.8075
Annual Jail Admissions (in thousands)[1]	0.0219	0.0135	-0.0065	0.0504	0.1226	0.0001	0.0044	-0.0091	0.0092	0.9863
Daily Jail Population[1]	0.0045	0.0031	-0.0021	0.0110	0.1686	-0.0002	0.0007	-0.0016	0.0012	0.7876
Index Crime Rate (in thousands)[1]	-0.0060	0.0158	-0.0392	0.0273	0.7109	-0.0037	0.0044	-0.0129	0.0054	0.4085
Number of Unique defendants in Court Records (in thousands)[1]	-0.0422	0.0096	-0.0625	-0.0220	0.0004	-0.0052	0.0035	-0.0125	0.0021	0.1534
Urbanicity Status: Rural	-0.0144	0.0120	-0.0396	0.0108	0.2462	-0.0007	0.0136	-0.0288	0.0275	0.9612
Urbanicity Status: Suburban	0.0063	0.0122	-0.0194	0.0320	0.6139	-0.0009	0.0164	-0.0348	0.0331	0.9587

MV=Multivariable; SV=Single Variable; Est SE=Estimated Standard Error;
[1] Square-root transformed

was excluded from the standard error calculation. The model parameters are presented in Table 2.5. County urbanicity status, the number of HIV-positive defendants, and annual jail admissions were all associated with county response status at the $\alpha = 0.1$ level in the multivariable calibration model.

The calibrated weights exhibited a fairly high amount of variation, ranging from 1.45×10^{-6} to 24.7 (with a median of 0.05). Based on the calibrated weights, $\hat{n}_{I,NC,WC} = 1090.6$, with a 95% confidence interval of $(969.1, 1212.1)$.

Figure 2.4 compares the estimates for the two methods. The confidence intervals overlap slightly, but the outcome regression method leads to a larger point estimate than the weight calibration method. The precision of the two methods differs greatly, with outcome regression providing the narrower interval (half-width of 99.1) and weight calibration providing a wider interval (half-width of 121.5).

Table 2.3: Estimated Number of HIV-positive Persons Incarcerated in Jails in the 10 Largest and 10 Smallest Counties, Outcome Regression

County Name	*Response*	$n_{D,i}$	$P(I\|D)_i$	$\hat{P}(I\|D)_i$	$\hat{n}_{I,i}$
Large County 1	Responding County	270	0.13	0.13 (0.09, 0.18)	34.1
Large County 2	Responding County	804	0.08	0.07 (0.02, 0.12)	62.5
Large County 3	Responding County	312	0.14	0.12 (0.07, 0.17)	44.7
Large County 4	Responding County	1104	0.09	0.07 (0.02, 0.12)	103.1
Large County 5	Responding County	1580	0.12	0.13 (0.08, 0.19)	185.4
Large County 6	Responding County	271	0.12	0.13 (0.09, 0.18)	33.2
Large County 7	Responding County	93	0.09	0.09 (0.04, 0.14)	8.3
Large County 8	Responding County	957	0.09	0.09 (0.04, 0.14)	90.2
Small County 1	Nonresponding County	1		0.12 (0.08, 0.17)	0.2 (0.1, 0.2)
Small County 2	Nonresponding County	3		0.11 (0.05, 0.17)	0.3 (0.1, 0.5)
Small County 3	Nonresponding County	5		0.13 (0.08, 0.18)	0.7 (0.4, 1.0)
Large County 9	Nonresponding County	626		0.09 (0.04, 0.14)	58.4 (26.5, 90.3)
Large County 10	Nonresponding County	420		0.14 (0.08, 0.20)	60.5 (35.3, 85.7)
Small County 4	Nonresponding County	3		0.12 (0.06, 0.17)	0.3 (0.2, 0.5)
Small County 5	Nonresponding County	1		0.12 (0.07, 0.17)	0.1 (0.1, 0.1)
Small County 6	Nonresponding County	4		0.12 (0.07, 0.18)	0.5 (0.3, 0.7)
Small County 7	Nonresponding County	33		0.11 (0.06, 0.16)	3.7 (2.0, 5.4)
Small County 8	Nonresponding County	11		0.15 (0.10, 0.21)	1.6 (1.0, 2.2)
Small County 9	Nonresponding County	21		0.09 (0.04, 0.14)	1.9 (0.9, 2.9)
Small County 10	Nonresponding County	27		0.10 (0.05, 0.15)	2.8 (1.4, 4.2)

Table 2.4: County Characteristics by Response Status

County Characteristics		*Responding Counties*	*Nonresponding Counties*
	n	26	74
Number of Unique defendants in Court Records (in thousands)	Median (Q1,Q3)	10.9 (6.5, 17.7)	4.0 (1.8, 6.4)
	Mean (SD)	15.6 (14.7)	5.2 (4.8)
	Min, Max	1.9, 54.2	0.4, 26.6
Number of HIV+ defendants	Median (Q1,Q3)	98 (55, 270)	36 (14, 84)
	Mean (SD)	263 (395)	73 (101)
	Min, Max	12, 1580	1, 626
Annual Jail Admissions (in thousands)	Median (Q1,Q3)	6.2 (3.5, 12.6)	1.6 (0.5, 3.1)
	Mean (SD)	9.5 (10.1)	2.5 (2.8)
	Min, Max	1.5, 40.3	0.0, 12.9
Daily Jail Population	Median (Q1,Q3)	250 (139, 457)	87 (36, 163)
	Mean (SD)	389 (422)	124 (125)
	Min, Max	46, 1881	0, 721
Index Crime Rate (in thousands)	Median (Q1,Q3)	3.8 (1.7, 6.6)	1.2 (0.3, 2.6)
	Mean (SD)	6.6 (8.9)	1.9 (2.7)
	Min, Max	0.6, 40.0	0.0, 17.3
Percent below Poverty	Median (Q1,Q3)	17.2 (15.2, 18.6)	20.5 (17.0, 24.3)
	Mean (SD)	16.8 (3.1)	20.6 (5.2)
	Min, Max	9.4, 24.2	6.0, 32.3
Urbanicity Status	Rural	9 (35%)	45 (61%)
	Suburban	5 (19%)	5 (7%)
	Small/Mid or Urban	12 (46%)	24 (32%)

Table 2.5: Parameter Estimates for Weight Calibration Model

Variable Description	$\hat{\gamma}$	*Est SE*	*95% Lower Conf Limit*	*95% Upper Conf Limit*	*P-Value*
Intercept	23.062	10.696	1.84	44.28	0.0335
Urbanicity Status: Rural or Suburban[1]	-6.908	3.504	-13.86	0.04	0.0514
Number of Unique defendants in Court Records (in thousands)[2]	-2.097	1.308	-4.69	0.50	0.1121
Number of HIV-positive Defendants	0.029	0.015	-0.00	0.06	0.0557
Annual Jail Admissions (in thousands) [2]	-7.648	3.386	-14.37	-0.93	0.0261
Index Crime Rate (in thousands) [2]	-0.806	2.036	-4.84	3.23	0.6930
Percent below Poverty	-0.009	0.151	-0.31	0.29	0.9523

[1] Small/Midsize, Urban is the reference level
[2] Square-root transformed

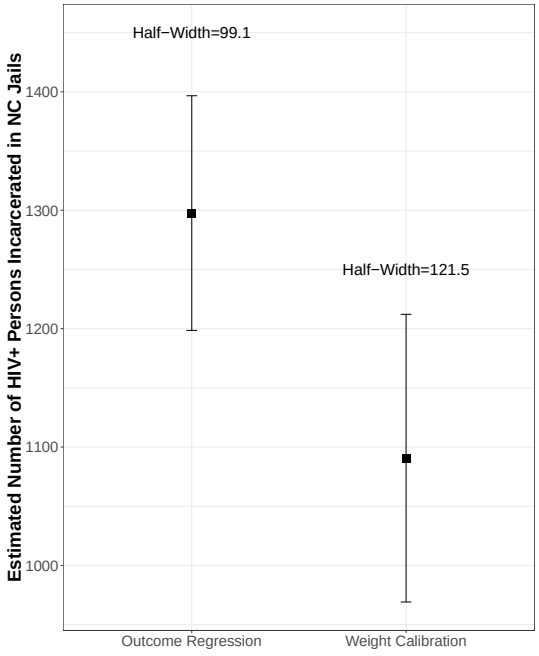

Figure 2.4: Estimated Number of HIV+ Persons Incarcerated in NC Jails ($\hat{n}_{I,NC}$) and 95% Confidence Intervals based on Outcome Regression and Weight Calibration

2.5 Discussion

Record linkage across three large NC databases, combined with outcome regression or weight calibration, will allow for indirect estimation of a rare population that cannot be directly measured given the current data collection practices and capabilities of NC jails. There are advantages and limitations of each method that are specific to the application at hand. Outcome regression has the advantage of producing county-level estimates, which will be useful for practitioners for targeting HIV interventions where they are most needed. However, outcome regression approaches rely on correct outcome model specification (Hansen, 1987). Weight calibration can be used to obtain an overall estimate for the entire state of NC, and the available covariates were predictive of county-level response status. The weight calibration model is doubly robust, providing consistent estimates if either the outcome model or the implied response model is correctly specified, but it unfortunately cannot provide county-level estimates. Furthermore, findings from the simulation called into question the small sample properties of its variance estimator for this population. In the simulations, the fpc was applied to the calibration model variance estimator without a formal justification. Use of the fpc led to more appropriate confidence interval coverage compared to ignoring the fpc adjustment for larger finite populations.

There are limitations associated with our findings. The small sample of $n = 26$ counties made model-fitting challenging and the validity of asymptotic properties questionable. The two evaluated approaches treat the number of HIV-positive defendants and HIV-positive persons in the 26 jails as known quantities, ignoring any error in the record linkage process. Despite these limitations, these findings demonstrate how outcome regression and weight calibration can be used to account for missing data following record linkage procedures in order to generalize results to a target population.

CHAPTER 3: DON'T LET CONFOUNDING CONFOUND YOU: POWER AND SAMPLE SIZE FOR MARGINAL STRUCTURAL MODELS

3.1 Introduction

Researchers often aim to estimate causal effects rather than just associations between variables. In settings where experimental designs are implausible, inference relies on observational data from which measured associations can be confounded. Marginal structural models (MSMs) are a commonly used method to estimate causal effects in the presence of confounding variables (Hernán et al., 2000; Robins et al., 2000; Cole and Hernán, 2008; Brumback et al., 2004). These models are fit using weighted estimating equations, where the weights are the inverse of each participant's probability of the observed treatment (or exposure). For a binary treatment, the estimand of interest is often the average causal effect, the difference in counterfactual means for the two treatment levels. With the assumptions of causal consistency, conditional exchangeability, and positivity, the inverse probability of treatment weight (IPTW) estimators are consistent for the MSM parameters for the causal means and the average causal effect (Lunceford and Davidian, 2004). Variance estimates are computed using standard estimating equation theory (Stefanski and Boos, 2002), with the empirical sandwich variance estimator providing a consistent estimator for the asymptotic variance of the estimated average causal effect.

While IPTW estimators provide researchers with an analytic tool for estimating causal effects in the presence of confounding variables, these estimators pose challenges during study design. The use of weights in the analysis affects the variance of the average causal effect estimator, making it challenging to determine the number of participants needed to achieve sufficient statistical power to detect a difference in causal means. Sample sizes cannot be calculated using standard methods that ignore weighting as in a randomized controlled trial (RCT) (e.g. as in Chow et al., 2017), as this

will tend to be anti-conservative. Numerous papers have examined the properties of IPTWs and have developed guidelines and diagnostics for specifying weight models and adjusting estimated weights (Austin, 2009; Austin and Stuart, 2015; Cole and Hernán, 2008; Lee et al., 2011). However, currently no methods exist for power and sample size calculations for studies that will be analyzed using MSMs fit with IPTWs.

Weighted estimators are common in survey sampling and for Bayesian methods that utilize importance sampling, and both fields have developed methods to quantify the effect of weighting on the precision of estimates. Kish (1965, page 257) introduced the *design effect* under the randomization-based inferential paradigm for survey sampling. The design effect is the ratio of the variance of an estimator under a complex sample design to the variance of the estimator under a simple random sample. When participants are selected directly from the finite population rather than from clusters of correlated observations, the design effect for a population mean estimator simplifies to the design effect due to weighting ($deff_w$), or the unequal weighting effect (Kish, 1992). Let n be the sample size and w_i represent the sampling weight for the i^{th} participant, i.e., the inverse of participant i's probability of selection. The design effect due to weighting is defined using either of the two equivalent forms:

$$deff_w = \frac{n \sum_{i=1}^{n} {w_i}^2}{(\sum_{i=1}^{n} w_i)^2} = 1 + \frac{S^2(w)}{(n^{-1} \sum_{i=1}^{n} w_i)^2} \tag{3.1}$$

where $S^2(w)$ is the finite sample variance of the weights. The design effect is interpreted as an estimator's increase in variance due to differential weights across participants. This metric is commonly applied to all types of complex sample designs in which individuals in the finite population have different probabilities of selection (Valliant et al., 2013, page 375). Gabler et al. (1999) provided a justification for how Kish's design effect also applies to model-based estimators. In practice, the design effect is used to calculate the *effective sample size*, which is equal to the observed sample size divided by the design effect. The effective sample size can be interpreted as

the sample size under simple random sampling that that would have produced the same variance as the sample selected under the complex design (Valliant et al., 2013, page 5).

Bayesian importance sampling uses weighting methods when sampling from one distribution to estimate the properties of another distribution (Kong et al., 1994). Importance sampling uses the effective sample size metric to compare the precision of the weighted estimator to the precision that would be achieved if sampling had been conducted directly from the distribution of interest (Kong et al., 1994). When the estimator of interest is a Hájek estimator, Kong (1992) provides an approximation for the effective sample size which is a function of (3.1).

Advantages of the approximated design effect are that it is outcome invariant and allows the sample size under a complex design to be translated into a sample size under a simpler design with the same variance. The former implies that the approximated design effect depends only on the participants' weights and is constant across outcomes. The latter means that once $deff_w$ is known or approximated, it can be used in power and sample size calculations along with the simpler assumptions needed to design a study without weights.

In this paper we consider design effects for planning observational studies to assess the effect of a treatment or exposure on an outcome of interest. In the analysis of observational data, McCaffrey et al. (2004, 2013) have used the effective sample size to quantify the loss of statistical precision following inference about causal effects using propensity score weighting. Here we describe the use of design effects for determining the sample size or power when designing an observational study. Section 3.2 introduces the design effect for causal inference and proves that it can be approximated with Kish's $deff_w$. Section 3.3 demonstrates how the design effect can be used to determine the sample size or power of an observational study that will be analyzed using MSM with IPTWs. Section 3.4 examines the accuracy of the design effect approximation for various exposure and outcome types via simulations, and Section 3.5 provides practical considerations regarding the use of design effects. Section 3.6 concludes with a discussion of the results and implications. Appendix A includes proofs of the propositions appearing in the main text.

3.2 The Design Effect

3.2.1 Preliminaries

Suppose an observational study is being planned where n independent and identically distributed copies of (A_i, L_i, Y_i) will be observed, where A_i is the binary treatment (exposure) status for participant i such that $A_i = 1$ if participant i received treatment and $A_i = 0$ otherwise, L_i is a vector of baseline covariates measured prior to A_i or unaffected by treatment A_i, and Y_i is the observed outcome for participant i.

The aim of the observational study will be to estimate the effect of treatment A on outcome Y. Specifically, let Y_{1i} denote the potential outcome if an individual i, possibly counter to fact, receives treatment. Similarly let Y_{0i} denote the potential outcome if individual i does not receive treatment, such that $Y_i = A_i Y_{1i} + (1 - A_i)Y_{0i}$. Inference from the observational study will focus on parameters of the MSM $E(Y_a) = \beta_0 + \beta_1 a$, with particular interest in the parameter β_1 which equals the average causal effect $ACE = E(Y_1) - E(Y_0) = \mu_1 - \mu_0$. Note the MSM is saturated and thus does not impose any restrictions on the assumed structure of the data.

Under certain assumptions, the parameters of the MSM can be consistently estimated using IPTW. In particular, assume conditional exchangeability holds, i.e., $Y_a \perp A \mid L$ for $a \in \{0, 1\}$. Also assume that positivity holds such that $Pr(A = a \mid L = l) > 0$ for all l such that $dF_L(l) > 0$ and $a \in \{0, 1\}$, where F_L is the cumulative distribution function of L. Estimating the average causal effect under the stated assumptions with the IPTW estimator first entails estimating the propensity score for each participant, defined as $p_i = Pr(A_i = 1 \mid L_i)$ (Rosenbaum and Rubin, 1983). A model is fit to obtain $\hat{p}_i$, each participant's estimated probability of treatment conditional on observed covariates L_i. The estimated IPTW is then equal to $\hat{W}_i = I(A_i = 1)\hat{p}_i^{-1} + I(A_i = 0)(1 - \hat{p}_i)^{-1}$, where $I(A_i = a)$ is a {0,1} treatment indicator for participant i. The estimated average causal effect $\hat{\beta}_1$ is obtained by regressing the observed outcome Y on treatment A with weights $\hat{W}$ using weighted least squares. The resulting IPTW estimator is a difference in Hájek estimators for the

two causal means (Hernán and Robins, 2020; Lunceford and Davidian, 2004):

$$\widehat{ACE} = \hat{\mu}_1 - \hat{\mu}_0 = \frac{\sum_{i=1}^{n} \hat{W}_i Y_i I\left(A_i = 1\right)}{\sum_{i=1}^{n} \hat{W}_i I\left(A_i = 1\right)} - \frac{\sum_{i=1}^{n} \hat{W}_i Y_i I\left(A_i = 0\right)}{\sum_{i=1}^{n} \hat{W}_i I\left(A_i = 0\right)} \tag{3.2}$$

Augmented IPW estimators, which incorporate both outcome and treatment models, may be used instead of (3.2) to estimate the ACE. Such estimators are doubly robust and will be more efficient than (3.2) if both the treatment and outcome models are correctly specified (Robins et al., 1994; Lunceford and Davidian, 2004). Thus, the power and sample size calculations derived below, which are based on (3.2), will be conservative for studies analyzed with augmented IPW estimators.

3.2.2 The Design Effect for a Single Causal Mean

Define the design effect to equal the ratio of the (finite sample) variance of $\hat{\mu}_a$ divided by the variance of a naïve causal mean estimator if, counter to fact, no confounding was present and weighting was not needed. That is,

$$deff_w^a = \frac{Var(\hat{\mu}_a)}{Var(\tilde{\mu}_a)} \tag{3.3}$$

where $\tilde{\mu}_a = \{\sum_{i=1}^{n} Y_i I(A_i = a)\}/\{\sum_{i=1}^{n} I(A_i = a)\}$. The derivation of the design effect estimator relies on the following proposition. The proposition assumes that the weights are known and are denoted by $W_a = P(A = a \mid L)^{-1}$ for $a \in \{0, 1\}$ with $W = AW_1 + (1-A)W_0$. Let $\sigma_a^2 = Var(Y_a)$ for $a \in \{0, 1\}$.

Proposition 3.1.

$$\sqrt{n}(\hat{\mu}_a - \mu_a) \xrightarrow{d} N(0, \Sigma_a)$$

where

$$\Sigma_a = \sigma_a^2 \left(\frac{E\left\{W^2 I(A = a)\right\}}{\left[E\left\{W I(A = a)\right\}\right]^2} \right) + R(L, Y_a)$$

and

$$R(L, Y_a) = E[\{W_a - E(W_a)\}(Y_a - \mu_a)^2]$$

with

$$|R(L, Y_a)| \leq \sqrt{Var(W_a)Var\{Y_a^2 - 2\mu_a Y_a\}}$$

for $a \in \{0, 1\}$

It follows from Proposition 3.1 that for large n the variance of $\hat{\mu}_a$ can be approximated as:

$$Var(\hat{\mu}_a) \approx \frac{\sigma_a^2}{n}\left(\frac{E\{W^2 I(A=a)\}}{[E\{WI(A=a)\}]^2}\right) + n^{-1}R(L, Y_a)$$

By similar arguments, for large n, $Var(\tilde{\mu}_a) \approx \sigma_a^2/\{nP(A=a)\}$. Therefore,

$$deff_w^a \approx \frac{P(A=a)E\{W^2 I(A=a)\}}{[E\{WI(A=a)\}]^2} + Er_a \tag{3.4}$$

where $Er_a = \{P(A=a)/\sigma_a^2\}R(L, Y_a)$, which by Proposition 3.1 is bounded by:

$$|Er_a| \leq \{P(A=a)/\sigma_a^2\}\sqrt{Var(W_a)Var(Y_a^2 - 2\mu_a Y_a)} \tag{3.5}$$

An approximation of (3.4) that does not depend on the potential outcome Y_a omits the remainder term Er_a:

$$\widetilde{deff}_w^a = \frac{P(A=a)E\{W^2 I(A=a)\}}{[E\{WI(A=a)\}]^2} \tag{3.6}$$

When planning an observational study, prior or pilot study data may be available to estimate (3.6). In particular, suppose based on a pilot study n_p copies of (L_i, A_i) are observed. Then replacing $P(A=a)$ with N_a/n_p where $N_a = \sum_{i=1}^{n_p} I(A_i = a)$, $E\{W^2 I(A=a)\}$ with $n_p^{-1}\sum_{i=1}^{n_p}\hat{W}_i^2 I(A_i = a)$, and $E\{WI(A=a)\}$ with $n_p^{-1}\sum_{i=1}^{n_p}\hat{W}_i I(A_i = a)$, a consistent estimator of (3.6) is:

$$\widehat{deff}_w^a = \frac{N_a \sum_{i=1}^{n_p}\hat{W}_i^2 I(A_i = a)}{\left\{\sum_{i=1}^{n_p}\hat{W}_i I(A_i = a)\right\}^2} \tag{3.7}$$

This estimator has the same form as Kish's design effect (3.1), applied to treatment group $A = a$. When prior data are not available, the design effect can be approximated using (3.6) based on an

assumed distribution for $A \mid L$ and the marginal distribution of L. The bias of (3.6) or (3.7) as an approximation to (3.4) in a given application depends on the value of Er_a. As further discussed in Section 3.6, Er_a is not guaranteed to be negligible. Bias of (3.6) and (3.7) for varying outcome types and confounding structures is evaluated empirically in simulation studies presented in Section 3.4.

3.3 Sample Size Calculations using the Design Effects

When the ACE is the focus of inference for the observational study being planned, the large sample distribution of $\widehat{ACE}$ can be used for power or sample size calculations. As $n \to \infty$, $\widehat{ACE}$ is consistent and asymptotically normal, i.e., $\sqrt{n}(\widehat{ACE} - ACE) \xrightarrow{d} N(0, \Sigma^*)$, where Σ^* is given by equation (13) in Lunceford and Davidian (2004). By the following proposition, Σ^* can be decomposed into the sum of asymptotic variances for the two causal mean estimators:

Proposition 3.2.

$$\Sigma^* = \Sigma_1 + \Sigma_0$$

Treating the weights as fixed or known leads to a larger asymptotic variance for $\widehat{ACE}$ compared to appropriately treating the weights as estimated, i.e., Σ^* is at least as large as the true asymptotic variance of $\widehat{ACE}$ (Lunceford and Davidian, 2004). Therefore, sample size formulae derived based on Σ^* would in general be expected to be conservative.

The results in Propositions 3.1 and 3.2 allow for sample size calculations for studies that will be analyzed using MSM with IPTW. Suppose the sample size for the observational study being planned is to be determined on the basis of the power to test $H_0 : ACE = 0$ versus $H_a : ACE \neq 0$ or equivalently $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$. Define the test statistic $t = \widehat{ACE}\{Var(\widehat{ACE})\}^{-1/2}$, where

$$Var(\widehat{ACE}) \approx Var(\hat{\mu}_1) + Var(\hat{\mu}_0) \approx \{nP(A=1)\}^{-1}\sigma^2_{1,adj} + \{nP(A=0)\}^{-1}\sigma^2_{0,adj} \qquad (3.8)$$

with $\sigma^2_{a,adj} = \sigma^2_a deff^a_w$ for $a \in \{0, 1\}$. Then,

$$\frac{\widehat{ACE} - ACE}{\sqrt{Var(\widehat{ACE})}} \tag{3.9}$$

is approximately standard normal for large n. Thus, H_0 is rejected when $|t| > z_{1-\frac{\alpha}{2}}$, where α is the type I error rate and z_q is the q^{th} quantile of the standard normal distribution.

Proposition 3.3. The sample size required to achieve power $1 - \beta$ for effect size $ACE = \delta$ and type I error rate α is approximately:

$$n_{deff} = \frac{(1 + k)(z_{1-\alpha/2} + z_{1-\beta})^2(\sigma^2_{1,adj}/k + \sigma^2_{0,adj})}{\delta^2} \tag{3.10}$$

where $k = P(A = 1)/P(A = 0)$ is the odds of treatment in the population.

The sample size formula (3.10) is the standard sample size equation commonly used to design RCTs, but with σ^2_a replaced by $\sigma^2_{a,adj}$ (Chow et al., 2017, page 48). Thus, Proposition 3.3 simplifies power and sample size calculations for observational studies by allowing researchers to design studies as if they were designing an RCT, but inflating the assumed variances by the approximated design effects. The researcher first assumes that no confounding is present, specifies the desired α and $1 - \beta$, and makes assumptions about σ^2_0, σ^2_1, δ, and k. The design effect is then approximated. When data from a pilot or prior study are available, $deff^1_w$ and $deff^0_w$ can be approximated based on (3.7) for each treatment group. When no prior study data are available, the distribution of the anticipated weights can be estimated based on assumptions about the distribution of L and $A \mid L$ and the design effect can be calculated based on (3.6). While these assumptions may not be easy to make, this approach notably requires no assumptions about the potential outcomes Y_0 and Y_1 and their associations with A and L. Further discussion about these practical considerations is included in Section 3.5. Once the design effects are approximated by $\widetilde{deff}^a_w$ or $\widehat{deff}^a_w$, adjusted variances $\sigma^2_{a,adj}$ can be estimated by $\tilde{\sigma}^2_{a,adj} = \sigma^2_a \widetilde{deff}^a_w$ or $\hat{\sigma}^2_{a,adj} = \sigma^2_a \widehat{deff}^a_w$, respectively, for $a \in \{0, 1\}$.

3.4 Simulation Study

3.4.1 Simulation Scenarios

Simulation studies were conducted to demonstrate use of the design effect in study design and estimate the bias of the approximation in (3.6) and (3.7) under a variety of confounding structures and outcome types. The scenarios in Table 3.1 were considered. For all scenarios, $\alpha = 0.05$ and $1 - \beta = 80\%$ were chosen.

Table 3.1: Five simulation scenarios. Scenarios 1-4 demonstrate use of the design effect when no prior study data are available, and Scenario 5 demonstrates use of the design effect with prior study data. $X \sim B(p)$ indicates that a random variable X follows the Bernoulli distribution with probability of success equal to p.

	Scenario	Treatment (A)	Confounders (L)	Outcome (Y)	δ
1	binary Y, small $deff_w^a$	$A \mid L = 0 \sim B(0.5)$ $A \mid L = 1 \sim B(0.75)$	$L \sim B(0.6)$	$Y_0 \mid L \sim B(0.85 - 0.2L)$ $Y_1 \mid L \sim B(0.70 - 0.2L)$	-0.15
2	binary Y, large $deff_w^a$	$A \mid L = 0 \sim B(0.1)$ $A \mid L = 1 \sim B(0.9)$	$L \sim B(0.5)$	$Y_0 \mid L \sim B(0.85 - 0.2L)$ $Y_1 \mid L \sim B(0.70 - 0.2L)$	-0.15
3	continuous Y, small $deff_w^a$	$A \mid L = 0 \sim B(0.5)$ $A \mid L = 1 \sim B(0.75)$	$L \sim B(0.6)$	$Y_0 \mid L \sim N(20 - 10L, 144)$ $Y_1 \mid L \sim N(25 - 10L, 256)$	5.0
4	continuous Y, large $deff_w^a$	$A \mid L = 0 \sim B(0.1)$ $A \mid L = 1 \sim B(0.9)$	$L \sim B(0.5)$	$Y_0 \mid L \sim N(20 - 10L, 144)$ $Y_1 \mid L \sim N(25 - 10L, 256)$	5.0
5	prior study data (NHEFS)	smoking cessation	9 baseline variables	weight gain	2.0

3.4.2 Sample Size Calculation

Two general approaches can be used to design a study with the design effect approximation: when prior study data are not available, as in Scenarios 1-4, and when prior study data are available, as in Scenario 5. One example from each general approach is presented in detail.

3.4.2.1 Example 1: No prior study data (Scenario 1)

Suppose no prior study data are available to design the study of interest. Then, the researcher must make the same assumptions and design choices as when designing an RCT, namely by specifying α, $1-\beta$, σ_0^2, σ_1^2, δ, and k. In general, σ_1^2 can be determined by deriving the marginal distribution of Y_1 based on the assumed distributions of $Y_1 \mid L$ and L. For Scenario 1, $P(Y_1 = 1) = \sum_{l=0}^{1} P(Y_1 = 1 \mid L = l)P(L = l) = 0.58$, and thus $\sigma_1^2 = 0.2436$. Similarly, $\sigma_0^2 = 0.1971$. Here, the average causal effect is assumed to be $\delta = -0.15$. The proportion of the population receiving treatment can be derived by integrating the distribution of $A \mid L$ over L. For Scenario 1, $P(A = 1) = \sum_{l=0}^{1} P(A = 1 \mid L = l)P(L = l) = 0.65$, and thus $k \approx 1.857$. When prior study data are not available, the distribution of the IPTWs must be assumed at the design phase. Based on the assumptions in Table 3.1, four possible values of W exist. These assumed values of W, along with the joint distribution of A and L, allow for the computation of the design effects using (3.6). This leads to $\widetilde{deff}_w^0 = 1.12$ and $\widetilde{deff}_w^1 = 1.04$, with approximated adjusted variances of $\tilde{\sigma}^2_{0,adj} = 0.2208$ and $\tilde{\sigma}^2_{1,adj} = 0.2533$.

Under the assumptions outlined in Table 3.1 for Scenario 1, to achieve 80% power to detect an average causal effect of -0.15 at the $\alpha = 0.05$ level, a sample size of approximately $n_{deff} = 356$ is required based on Proposition 3.3. The design effects and required sample sizes for Scenarios 2-4 can be determined similarly and are presented in Table 3.2. Note Scenarios 1 and 3 have the same design effects because in both instances the joint distribution of A and L is the same. Likewise, Scenarios 2 and 4 have the same design effects.

3.4.2.2 Example 2: Prior study data (Scenario 5)

Prior study or pilot data may allow for better informed assumptions about σ_0^2, σ_1^2, δ, and k. Because $\sigma_a^2 = E(Y_a^2) - \{E(Y_a)\}^2$, σ_a^2 can be estimated by obtaining $\hat{E}(Y_a^2)$ and $\hat{E}(Y_a)$ from fitted MSMs based on the prior study data. The estimate $\widehat{ACE}$ and prevalence of the exposure or treatment in the prior study can inform assumptions about δ and k.

As an example, consider designing a new study based on the National Health and Nutrition Examination Survey Data I Epidemiologic Follow-up Study (NHEFS) example presented in Chapter 12 of Hernán and Robins (2020). Hernán and Robins use MSM with IPTWs to estimate the average causal effect of smoking cessation (A) on weight gain after approximately 10 years of follow-up (Y) based on the NHEFS sample of smokers ($n = 1566$), assuming conditional exchangeability based on nine baseline confounders L: sex, age, race, education, smoking intensity, duration of smoking, physical activity, exercise, and weight.

Making the same assumptions as Hernán and Robins (2020), Scenario 5 considers the design of a new study to estimate the average causal effect of smoking cessation on 10-year weight gain. Based on the NHEFS data, assume that $\sigma_0^2 = 56.1$ and $\sigma_1^2 = 74.0$, obtained by fitting MSMs with IPTWs to estimate $E(Y_a^2)$ and $E(Y_a)$. In the Hernán and Robins example, $\widehat{ACE} = 3.441kg$. The new study will be designed to provide approximately 80% power to detect a difference in weight gain of $\delta = 2.0kg$. From the NHEFS sample, assume $k \approx 0.346$.

When prior study data are available, $deff_w^0$ and $deff_w^1$ can be estimated using (3.7). For the NHEFS data, $\widehat{deff}_w^0 = 1.03$ and $\widehat{deff}_w^1 = 1.24$. This leads to approximated adjusted variances of $\hat{\sigma}^2_{0,adj} = 57.78$ and $\hat{\sigma}^2_{1,adj} = 91.76$. Based on these assumptions, a sample size of $n_{deff} = 853$ is needed to achieve approximately 80% power to detect an average causal effect of $2.0kg$ at the $\alpha = 0.05$ level using MSM with IPTWs.

3.4.2.3 Naïve Sample Size Calculations

As a comparison, sample sizes n_{rct} were calculated naively under the assumptions of an RCT, ignoring the effect of weighting on the variances of the estimates. In other words, sample sizes were calculated as demonstrated above, except using σ_a^2 instead of $\tilde{\sigma}^2_{a,adj}$ or $\hat{\sigma}^2_{a,adj}$ from Table 3.2.

Table 3.2: Variances, approximated design effects, approximated adjusted variances, and required sample sizes for simulation scenarios by treatment.

	Scenario	a	σ_a^2	$\widetilde{deff}_w^a$ or $\widehat{deff}_w^a$	$\tilde{\sigma}_{a,adj}^2$ or $\hat{\sigma}_{a,adj}^2$	n_{deff}	n_{rct}
1	binary Y, small $deff_w^a$	0	0.1971	1.12	0.2208	356	327
		1	0.2436	1.04	0.2533		
2	binary Y, large $deff_w^a$	0	0.1875	2.78	0.5208	828	298
		1	0.2400	2.78	0.6667		
3	continuous Y, small $deff_w^a$	0	168.0	1.12	188.2	310	286
		1	280.0	1.04	291.2		
4	continuous Y, large $deff_w^a$	0	169.0	2.78	469.4	784	283
		1	281.0	2.78	780.6		
5	prior study data, (NHEFS)	0	56.10	1.03	57.78	853	713
		1	74.00	1.24	91.76		

3.4.3 Evaluation

For Scenarios 1-4, empirical power based on samples of size n_{deff} was evaluated via simulation by following these steps:

(i) Generate a superpopulation of size $N = 1,000,000$ based on distributions in Table 3.1.

(ii) Select a sample of size n_{deff} without replacement from the superpopulation, where n_{deff} is specified in Table 3.2.

(iii) Estimate $\hat{W}_i$ for each member of the sample based on the predicted values from the logistic regression of A on L.

(iv) Fit the MSM $E(Y_{ai}) = \beta_0 + \beta_1 a_i$ using weighted least squares, treating the weights as estimated by stacking the estimating equations from the weight model with the estimating equations for the causal means and difference in causal means using the geex package in R (Saul and Hudgens, 2020).

(v) Test $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$ using a Wald test, rejecting H_0 at the $\alpha = 0.05$ significance level.

(vi) Repeat steps (ii)-(v) $R = 2000$ times and calculate empirical power as the proportion of simulated samples where H_0 was rejected.

For Scenario 5, empirical power based on a sample of size n_{deff} was evaluated via simulation by following these steps:

(i) Estimate the propensity score for each of the 1566 NHEFS participants from a logistic regression model of A on L as $\hat{p}_i = \widehat{Pr}(A_i = 1 \mid L_i = l_i)$. As in Hernán and Robins (2020), the logistic regression model includes main effects for each of the nine baseline confounders and quadratic terms for the four continuous covariates.

(ii) For each participant, calculate $\hat{Y}_{ai}$, $a \in \{0, 1\}$, as the predicted value $\hat{E}(Y_{ai} \mid L_i = l_i)$ from the following linear regression model, fit only on participants with $A = a$: $E(Y_{ai} \mid L_i = l_i) = l_i\beta$, where l_i is a vector for participant i that includes an intercept term, the 9 previously defined covariates, and the four quadratic terms corresponding to continuous covariates. Also compute $\widehat{Var}(Y_{ai} \mid L_i = l_i) = MSE_a$, where MSE_a is the mean squared error from the model for $E(Y_{ai})$.

(iii) Add 1.441 to $\hat{Y}_{0i}$ for all participants, such that $ACE = 2.0$ in the simulated population instead of 3.441 as in the NHEFS sample.

(iv) Select a sample of size n_{deff} with replacement from the NHEFS dataset, where n_{deff} is specified in Table 3.2.

(v) Assign $A_i = a_i$ as a random draw from $A_i \sim Bernoulli(\hat{p}_i)$.

(vi) Let $Y_{ai} = \hat{Y}_{ai} + \epsilon_{ai}$, where $\epsilon_{ai} \sim N(0, \widehat{Var}(Y_{ai} \mid L_i = l_i))$.

(vii) Follow steps (iii)-(v) from the above list for Scenarios 1-4.

(viii) Repeat steps (iv)-(vii) $R = 2000$ times and estimate empirical power as the proportion of simulated samples where H_0 was rejected.

For each scenario, these steps were repeated to calculate empirical power based on the naïve sample sizes, replacing n_{deff} with n_{rct}.

The results of the simulation study are presented in Table 3.3. For all simulation scenarios, when the sample size was calculated using the design effect, empirical power was equal to or exceeded the nominal 80% level. That is, use of the design effects to calculate required sample sizes led to close to the intended level of statistical power. On the other hand, ignoring the effect of weighting and basing sample sizes on the naïve assumptions of an RCT led to empirical power that was lower than the nominal 80% level for all but one scenario. These results demonstrate that ignoring the weights in power and sample size calculations can lead to significantly underpowered studies, particularly when there are strong confounders that lead to high variability in the weights.

For all scenarios, the approximation errors Er_a from (3.4) for each sample and treatment were estimated by $\widehat{Er}_a = \{N_a/(n\sigma_a^2)\}\hat{E}\left[\{\hat{W}_a - \hat{E}(\hat{W}_a)\}\{Y_a - \hat{E}(Y_a)\}^2\right]$, where expected values were calculated empirically within each sample. Estimated approximation errors were then averaged across the $R = 2000$ simulated samples. Mean estimated approximation error was small for most scenarios (Table 3.3) and was in opposite directions for the two treatment groups, which tended to offset the effects of the errors. Approximation error was large for Scenario 2 (0.60 for $A = 0$ and -0.19 for $A = 1$), but empirical power still equaled the nominal level when the design effect was used to calculate the sample size. Note Scenario 2 is an extreme example, as it includes only a single and very strong confounding variable and only two possible and extreme values for W. For the binary outcome, this resulted in large approximation errors.

Table 3.3: Results of the simulation study by scenario across $R = 2000$ samples. Empirical power n_{deff} and n_{rct} are the proportions of simulated samples in which the p-values for testing $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$ were less than $\alpha = 0.05$ for the following MSM: $E(Y_{ai}) = \beta_0 + \beta_1 a_i$, based on sample sizes n_{deff} and n_{rct}, respectively, from Table 3.2

Scenario	Empirical Power n_{deff}	Empirical Power n_{rct}	Mean $\widehat{Er}_0$	Mean $\widehat{Er}_1$
1	0.81	0.76	0.08	-0.01
2	0.80	0.42	0.60	-0.19
3	0.85	0.81	-0.02	0.01
4	0.86	0.47	0.00	-0.01
5	0.82	0.76	0.02	-0.03

3.5 Practical Considerations

When prior study data are not available, specifying the design effects can be challenging. A few general guidelines are offered to help researchers determine reasonable assumptions to facilitate power and sample size calculations.

When only a few categorical covariates will be included in the weight model, researchers can use subject matter knowledge or prior study information to nonparametrically specify the joint distribution of A and L, or the marginal distribution of L and the conditional distribution of $A \mid L$ (as in Example 1). Based on these assumptions, the anticipated weights can be calculated nonparametrically and the design effects for each treatment group can be approximated.

When specification of these distributions is not feasible, researchers can forgo approximating the values of the weights and instead consider more generally how much variation is expected in the weights. The lower bound for $deff_w^a$ is 1, which implies that the weights within both treatment

groups are all equal and thus covariates are not predictive of the treatment. Design effects tend to increase when more covariates are added to the weight model. The presence of covariates that are strong predictors of treatment tends to increase the design effect. Care must be taken to identify the appropriate set of confounders to include in the weight model (Vansteelandt et al., 2012). Inclusion of instrumental variables, which are predictive of the exposure but which do not affect the outcome, inflate the variance of the ACE estimator without reducing bias (Rubin, 1997; Myers et al., 2011). The use of weight truncation will decrease the design effect.

Figure 3.1 provides a visual depiction of weight distributions within one treatment group for various values of the design effect to aid researchers in choosing a design effect consistent with the expected variation in the weights. These weight distributions were generated by taking the reciprocals of $N_a = 1000$ random draws from beta distributions with mean 0.5 and shape parameters set to achieve the desired design effect. As variation in the weights increases, so does the design effect approximation.

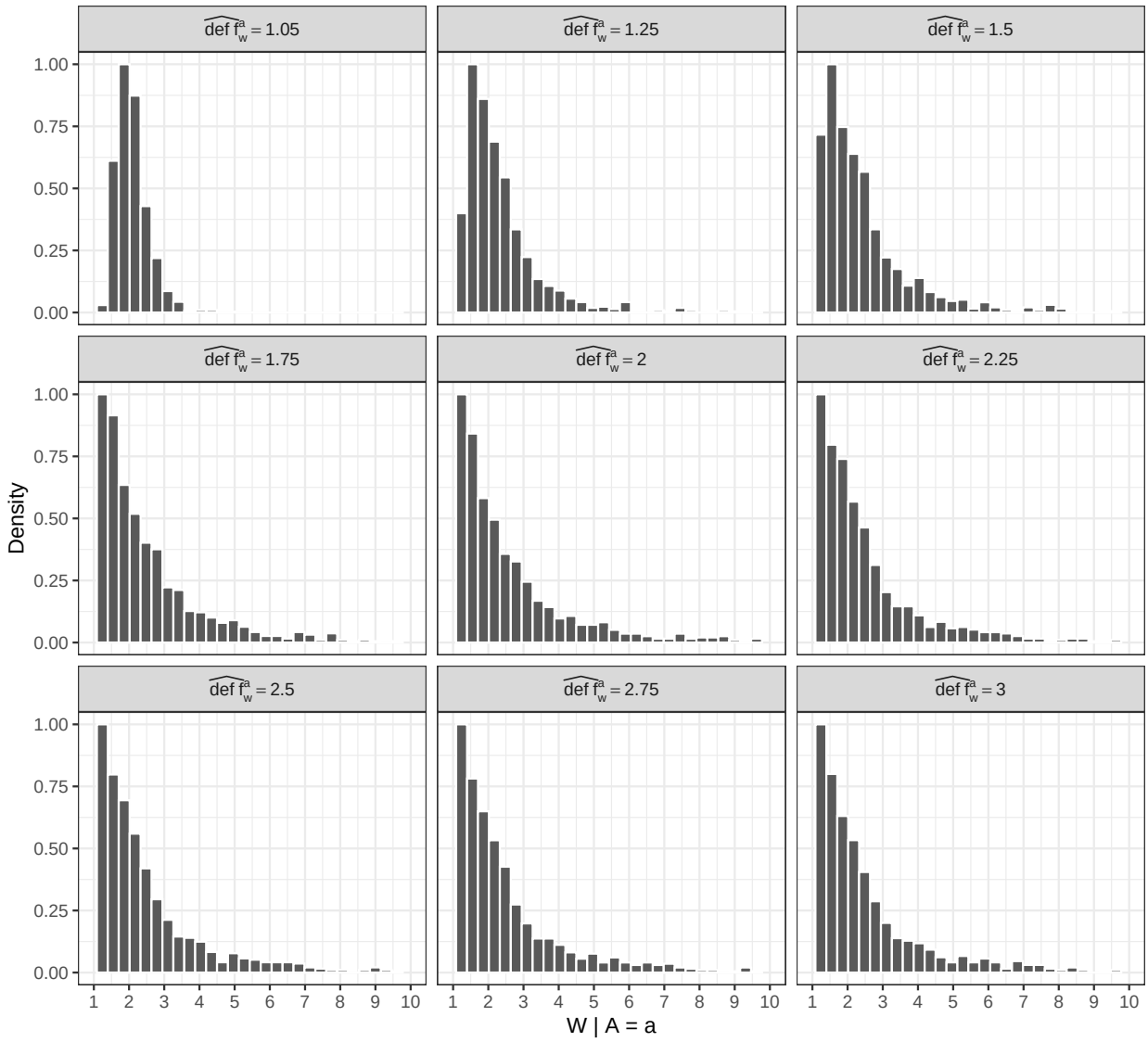

Figure 3.1: Examples of weight distributions for various approximated design effects. Distributions were generated by taking the reciprocals of $N_a = 1000$ random draws from beta distributions with mean 0.5 and shape parameters set to achieve the desired design effect.

3.6 Discussion

The design effect approximation simplifies power and sample size calculations of observational studies. Using the design effect allows researchers to utilize standard power and sample size software (e.g., nQuery, SAS Proc Power) for randomized trials, but with variances inflated by the approximate design effects. An additional advantage of using the design effect approximation is that no assumptions are required about the relationship between the potential outcomes and either the treatment or the confounders. Empirical results presented in Section 3.4 demonstrate the design effect approximation can yield the nominal level of power over a range of confounding and outcome structures.

Approximating the design effect when planning an observational study may be challenging. In survey sampling, it is common practice to report estimated design effects in analytic reports for better understanding of the precision of the estimates and to assist other researchers who are designing similar studies (see, for example Center for Behavioral Health Statistics and Quality, 2019). Reporting the estimated design effects corresponding to treatment or exposure effect estimates in observational studies may assist researchers with future study designs. In time, as more studies analyzed with MSMs start to report their design effects, rules of thumb and practical upper bounds for the design effects will likely emerge to aid in the design of future studies (see, for example, United Nations Statistical Division (2008, page 41), Daniel (2012, page 251), and Salganik (2006) from the survey sampling literature).

In the absence of knowledge of estimated design effects from prior studies, the design effect may be approximated either using (3.6) or, if pilot data are available, (3.7). In either case, the remainder term in (3.4) is ignored, which may introduce bias. The remainder may be large when individuals with extreme weight values tend to have potential outcomes that are also extreme relative to the mean. In the simulation studies in Section 3.4, the approximation error was small for all but one of the scenarios examined. Remainders were in opposite directions for the two treatment groups, which tended to offset the effects of the errors and thus use of the approximation did not result in

deviations from the nominal level of statistical power for any of the scenarios examined. However, there is no guarantee that approximation error will be negligible for a given study. When pilot or prior study data are available, approximation error Er_a can be estimated as in the simulations, but with Y_a replaced with $\hat{Y}_a$ for $a \in \{0, 1\}$ where $\hat{Y}_a$ is based on an assumed outcome regression model. Alternatively, an estimate for the upper bound of Er_a can be obtained by estimating the upper bound in (3.5).

Despite these limitations, the design effect approximation can be a useful tool for the design of studies that will be analyzed using MSM with IPTWs, as currently no power and sample size methods exist in this context. The design effect can also be used in precision calculations using approaches analogous to those described in this paper, i.e., basing calculations on the adjusted variances $\tilde{\sigma}^2_{a,adj}$ or $\hat{\sigma}^2_{a,adj}$ rather than σ^2_a.

CHAPTER 4: CAUSAL INFERENCE FROM OBSERVATIONAL DATA FOR COUNT OUTCOMES

4.1 Introduction

Researchers often seek to estimate the causal effect of an exposure or treatment on an outcome of interest. Experimental designs are infeasible for many exposures of interest, and thus inference often relies on observational data. Associations measured from observational data can be subject to confounding, so methods have been developed to estimate causal effects in the presence of confounding variables. Three commonly-used methods to estimate causal effects from observational data are marginal structural models (MSM) fit with inverse probability of treatment weights (IPTWs) (Robins, 1998; Robins et al., 2000; Hernán et al., 2000), the parametric g-formula (Robins, 1986), and doubly robust estimators that incorporate both exposure and outcome model estimators (Bang and Robins, 2005; Hernán and Robins, 2020; Funk et al., 2011; Kang and Schafer, 2007). These three methods provide consistent estimates of the average causal effect, the difference in counterfactual means for the two exposures, under the assumptions of causal consistency, conditional exchangeability, and positivity (Lunceford and Davidian, 2004). In practice, MSM with IPTWs, the parametric g-formula, and doubly robust estimators are commonly applied to observational data when the outcome of interest is continuous, binary, or categorical (Hernán et al., 2000; Bodnar et al., 2004; Cole and Hernán, 2008; Taubman et al., 2009; Young et al., 2011; Garcia-Aymerich et al., 2013; Funk et al., 2011; Waernbaum, 2012).

Count outcomes are common in observational studies, as researchers often seek to estimate measures over a fixed period of time such as the number of sexual partners (Wiederman, 1997; Knittel et al., 2020), pill counts to assess treatment adherence (Bangsberg et al., 2001), or the number of cigarettes smoked (Singh et al., 1994). To estimate the effect of a binary exposure on

a count outcome, one key estimand is the causal rate ratio, the ratio of the counterfactual mean under exposure to the counterfactual mean under no exposure. Applying causal methods to count outcomes poses challenges unique to count data that must be accounted for to yield valid inference. The Poisson distribution is commonly used to model count outcomes, but the observed variance of a count outcome often exceeds the variance assumed under the Poisson model. This phenomenon is known as overdispersion (Agresti, 2002, page 130). Zero-inflation is also common in count outcomes, where the number of observed zero counts exceeds the number expected under a Poisson distribution (Böhning et al., 1999).

In addition to overdispersion and zero-inflation, which pose challenges for correctly modeling count outcomes even when data are measured without error, count outcomes are also susceptible to a form of measurement error called data heaping. Data heaping occurs when reported counts are rounded to different levels of precision (Wang and Heitjan, 2008). This phenomenon is commonly observed when collecting self-reported retrospective counts or measures of duration, including cigarette usage (Klesges et al., 1995), duration of breastfeeding (Singh and Folsom, 2000), and number of sexual partners (Wiederman, 1997; Roberts and Brewer, 2001). Data heaping is often attributed to cognitive processes in respondents, including choosing round numbers or approximations (digit preference) or using estimation methods to aid in recall (Roberts and Brewer, 2001). Data coarsening is a type of data heaping where participants tend to round their reported outcomes (Heitjan, 1989; Cummings et al., 2015). When count outcomes are subject to data heaping, the true underlying distribution of counts is distorted which can lead to biased point and variance estimates (Wang and Heitjan, 2008; Cummings et al., 2015).

Causal methods have been applied to count data (Sato and Matsuyama, 2003), and methods have been proposed to estimate the causal rate ratio using the parametric g-formula for zero-inflated data (Albert et al., 2014). A general theoretical framework is needed to define causal estimands and estimators for count outcomes. Methods are needed to account for the unique features of count data, including overdispersion, zero-inflation, and data heaping. This paper develops and compares three estimators for the causal rate ratio for count data, each of which can accommodate

overdispersion and/or zero-inflation in the outcome. Methods are presented for estimating the causal rate ratio when the observed outcome data exhibit data heaping under a given set of assumptions. Section 4.2 presents the estimators in detail and describes their large sample properties. Section 4.3 demonstrates and compares the empirical properties of the estimators with a simulation study, and Section 4.4 applies the methods to Women's Interagency HIV Study (WIHS) data to estimate the effect of incarceration on the number of sexual partners in the subsequent six-month period. Section 4.5 concludes with a discussion of the results. Appendix B includes supplemental tables and proofs of the propositions appearing in the main text.

4.2 Methods

4.2.1 Preliminaries

Assume that the observed data $(Y_1, L_1, A_1), (Y_2, L_2, A_2), ..., (Y_n, L_n, A_n)$ are an independent and identically distributed sample from a superpopulation, where A_i is the binary exposure or treatment status for participant i ($A_i = 1$ if participant i was exposed, $A_i = 0$ if participant i was not exposed), L_i is a vector of baseline covariates measured prior to A_i or unaffected by exposure A_i, and Y_i is the observed count outcome for participant i. That is, $Y_i \in \mathbb{N}^0$ and $\mathbb{N}^0$ is the set of all non-negative integers. Let Y_i^1 denote the potential outcome if individual i, possibly counter to fact, is exposed. Similarly, let Y_i^0 denote the potential outcome if individual i is not exposed, such that $Y_i = A_i Y_i^1 + (1 - A_i) Y_i^0$. Assume that conditional exchangeability holds, i.e., $Y^a \perp A \mid L$, $a \in \{0, 1\}$. Also assume that positivity holds such that $Pr(A = a \mid L = l) > 0$ for all l such that $dF_L(l) > 0$ and $a \in \{0, 1\}$, where F_L is the cumulative distribution function of L. Let $E(Y^a) = \lambda^a$ for $a \in \{0, 1\}$. The estimand is the causal rate ratio, $CRR = \lambda^1/\lambda^0$.

4.2.2 MSM with IPTW

Marginal structural models were introduced by Robins (1998) and further refined by Robins et al. (2000) and Hernán et al. (2000). The parameters of these models are commonly estimated using

IPTW. Participant i's IPTW is $W_i = A_i e_i^{-1} + (1 - A_i)(1 - e_i)^{-1}$, where $e_i = Pr(A_i = 1 \mid L_i)$, the probability that participant i was exposed conditional on covariates L_i. Weighting by the IPTWs creates a pseudo-population in which confounding by L is not present, which allows causal estimands within a MSM to be identifiable and for the estimation of causal effects for observational data (Robins et al., 2000).

Consider the following MSM: $\log(\lambda^a) = \beta_0 + \beta_1 a$. Under the assumptions specified in Section 4.2.1, the parameters of the MSM can be consistently estimated using IPTW. To estimate the causal rate ratio, $\exp(\beta_1)$ from the above MSM, the propensity score for each participant is estimated based on the observed exposure A and covariates L using a finite dimensional parametric model. For example, A can be regressed on L using logistic regression, i.e., the model $logit(e_i) = l_i\alpha$ is fit, where l_i is the $1 \times c_w$ vector corresponding to the i^{th} row of the design matrix for the $A \mid L$ model, c_w is the number of columns in the design matrix, and α is the $c_w \times 1$ vector of regression coefficients from the weight model. Predicted propensity scores are calculated as $\hat{e}_i = \hat{e}_i(l_i, \hat{\alpha}) = \text{logit}^{-1}(l_i\hat{\alpha})$ where logit^{-1} represents the inverse logit function $(\exp(l_i\hat{\alpha})/\{1 + \exp(l_i\hat{\alpha})\})$ and $\hat{\alpha}$ are the maximum likelihood estimates for α from the logistic model. Predicted propensity scores are used to estimate the IPTWs as $\hat{W}_i = I(A_i = 1)\hat{e}_i^{-1} + I(A_i = 0)(1 - \hat{e}_i)^{-1}$, where $I(A_i = a)$ is a $\{0,1\}$ exposure indicator for participant i. The CRR is then estimated using weighted estimating equations by regressing the observed counts Y on exposure A with weights $\hat{W}$.

The following estimator for the CRR is proposed, which is equal to $\exp(\hat{\beta}_1)$ from the above MSM:

$$\widehat{CRR}_{MSM} = \frac{\sum_{i=1}^n \hat{W}_i Y_i I(A_i = 1)}{\sum_{i=1}^n \hat{W}_i I(A_i = 1)} \Bigg/ \frac{\sum_{i=1}^n \hat{W}_i Y_i I(A_i = 0)}{\sum_{i=1}^n \hat{W}_i I(A_i = 0)} \tag{4.1}$$

That is, the MSM estimator for the CRR is the ratio of two Hájek estimators, one for each causal mean λ^a. Next, the large sample properties of (4.1) are defined.

Proposition 4.1. When the weights are treated as a known function of A_i and L_i, (4.1) is a consistent and asymptotically normal estimator of the CRR.

Proposition 4.2. When the weights are treated as a known function of A_i and L_i,

$$\sqrt{n}\left(\widehat{CRR}_{MSM} - CRR\right) \xrightarrow{d} N\left(0, \Sigma_{MSM}\right)$$

where

$$\Sigma_{MSM} = E\left[e^{-1}\left(\frac{Y^1 - \lambda^1}{\lambda^0}\right)^2 + (1-e)^{-1}\left\{\frac{\lambda^1(Y^0 - \lambda^0)}{(\lambda^0)^2}\right\}^2\right]$$

Proposition 4.3. When the weights are estimated based on the observed values of A_i and L_i,

$$\sqrt{n}\left(\widehat{CRR}_{MSM} - CRR\right) \xrightarrow{d} N\left(0, \Sigma^*_{MSM}\right)$$

where

$$\Sigma^*_{MSM} \leq \Sigma_{MSM}$$

Note that the consistency and asymptotic normality of the MSM with IPTW estimator (4.1) requires no assumptions about the parametric distribution of the count outcome nor does it require special handling of overdispersion or zero-inflation.

4.2.3 Parametric g-formula

The parametric g-formula is an outcome regression approach used in causal inference and is an alternative to MSM with IPTW. Robins (1986) introduced the parametric g-formula as a type of standardization that allows for the estimation of causal effects by directly modeling the outcome as a function of the exposure and confounding variables and then integrating over the distribution of the confounding variables.

The parametric g-formula provides an alternative estimator for the CRR. Under the assumptions of conditional exchangeability and causal consistency, $E(Y^a \mid L = l) = E(Y^a \mid L = l, A = a) = E(Y \mid L = l, A = a)$. The final quantity is identifiable. By the law of total probability, the causal mean λ^a is then defined as $\lambda^a = \int E(Y \mid L = l, A = a) dF_L(l)$ for $a \in \{0, 1\}$. The parametric g-formula estimator of the CRR is specified below, where $\hat{E}(Y_i \mid L_i = l_i, A_i = a)$ is estimated for

$a \in \{0, 1\}$ from a parametric model:

$$\widehat{CRR}_{PG} = \frac{n^{-1} \sum_{i=1}^{n} \hat{E}(Y_i \mid L_i = l_i, A_i = 1)}{n^{-1} \sum_{i=1}^{n} \hat{E}(Y_i \mid L_i = l_i, A_i = 0)} \tag{4.2}$$

Four distributions are commonly used to model count outcomes:

1. Poisson:

$$p(Y_i = y_i \mid A_i = a_i, L_i = l_i) = \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!}$$

2. Negative Binomial (NB)(Lawless, 1987):

$$p(Y_i = y_i \mid A_i = a_i, L_i = l_i) = \frac{\Gamma(y_i + \theta^{-1})}{y_i! \Gamma(\theta^{-1})} \left(\frac{\theta \mu_i}{1 + \theta \mu_i} \right)^{y_i} \left(\frac{1}{1 + \theta \mu_i} \right)^{\theta^{-1}}$$

3. Zero-Inflated Poisson (ZIP) and Zero-Inflated Negative Binomial (ZINB) (Lambert, 1992; Long et al., 2014; Preisser et al., 2016):

$$p(Y_i = y_i \mid A_i = a_i, L_i = l_i) = \begin{cases} (1 - \nu_i) + \nu_i f(Y_i = y_i \mid A_i = a_i, L_i = l_i) & for\ y_i = 0 \\ \nu_i f(Y_i = y_i \mid A_i = a_i, L_i = l_i) & for\ y_i > 0 \end{cases}$$

where $f(Y_i = y_i \mid A_i = a_i, L_i = l_i)$ is the probability mass function for the Poisson distribution (for the ZIP) or the NB distribution (for the ZINB) and ν_i is the probability that individual i is in the susceptible population, i.e., the population of individuals who could have a count greater than zero.

For the Poisson distribution, $E(Y_i \mid A_i = a_i, L_i = l_i) = Var(Y_i \mid A_i = a_i, L_i = l_i) = \mu_i$ and for the NB distribution $E(Y_i \mid A_i = a_i, L_i = l_i) = \mu_i$ and $Var(Y_i \mid A_i = a_i, L_i = l_i) = \mu_i + \mu_i^2 \theta$, where θ represents the dispersion parameter. The Poisson distribution is reasonable when the variance is expected to equal the mean, while the NB distribution is useful in the presence of overdispersion (Agresti, 2002, page 131). For both distributions, a generalized linear model (GLM) is fit of the form: $log(\mu_i) = x_i \beta$ for $i = 1, ..., n$, where $x_i = g(l_i, a_i)$ is the $1 \times c$ vector

corresponding to the i^{th} row of the design matrix for the $Y \mid A, L$ model, c is the number of columns in the design matrix, and β is the $c \times 1$ vector of regression coefficients. The maximum likelihood estimates for β, $\hat{\beta}$, are obtained for the GLM, $\hat{E}(Y_i \mid L_i = l_i, A_i = a) = \exp\{g(l_i, a)\hat{\beta}\}$ are calculated for each participant for $a \in \{0, 1\}$, and $\widehat{CRR}_{PG}$ is derived as in (4.2).

The ZIP and ZINB distributions account for excess zeros in the count outcome without and with overdispersion, respectively. When a ZIP or ZINB distribution is assumed for modeling a count outcome, regression coefficients have latent interpretations corresponding to a subpopulation that is susceptible for an outcome greater than zero and a subpopulation that is non-susceptible and thus always has an outcome of zero (Long et al., 2014). When the outcome follows a ZIP or ZINB distribution, models for ν_i and the expected count for individual i within the susceptible population (η_i) are simultaneously fit: $logit(\nu_i) = x_{i1}\gamma$ and $log(\eta_i) = x_{i2}\xi$, where $x_{i1} = g_1(l_i, a_i)$ is the $1 \times c_1$ vector corresponding to the i^{th} row of the design matrix for the susceptibility model, $x_{i2} = g_2(l_i, a_i)$ is the $1 \times c_2$ vector corresponding to the i^{th} row of the design matrix for the count model, γ is the $c_1 \times 1$ vector of regression coefficients for the susceptibility model, and ξ is the $c_2 \times 1$ vector of regression coefficients for the count model. The exposure A is included in either or both models, as appropriate for correct model specification. The maximum likelihood estimates for γ and ξ, $\hat{\gamma}$ and $\hat{\xi}$ respectively, are obtained for the model and $\hat{E}(Y_i \mid L_i = l_i, A_i = a) = \text{logit}^{-1}\{g_1(l_i, a)\hat{\gamma}\} \exp\{g_2(l_i, a)\hat{\xi}\}$ are obtained for each participant for $a \in \{0, 1\}$. $\widehat{CRR}_{PG}$ is then calculated as in (4.2). Albert et al. (2014) demonstrates the use of the parametric g-formula to estimate the CRR for zero-inflated data.

Proposition 4.4. Under correctly specified models, (4.2) is a consistent and asymptotically normal estimator of the CRR.

Estimates of (4.2) and its variance can be computed using estimating equation theory (Stefanski and Boos, 2002), as demonstrated in Sections 4.3 and 4.4.

4.2.4 Doubly Robust Estimation

Doubly robust estimators, also referred to as augmented inverse probability weighted (AIPW) estimators, incorporate both MSM and parametric g-formula estimators to provide protection against

incorrect model specification for either the weight or outcome model (Bang and Robins, 2005; Hernán and Robins, 2020; Funk et al., 2011). Because the relationships between the exposure, outcome, and confounding variables are typically unknown in observational settings, doubly robust estimators afford some protection against misspecification of these models (Funk et al., 2011). While in practice all models are at least partially misspecified, Bang and Robins (2005) note that doubly robust estimators allow for minimal bias when either the weight or outcome model is nearly correct, giving the researcher two chances to get close to correct specification.

We propose the following doubly robust estimator for the CRR:

$$\widehat{CRR}_{DR} = \frac{\hat{\lambda}^1_{DR}}{\hat{\lambda}^0_{DR}} \tag{4.3}$$

where

$$\hat{\lambda}^1_{DR} = n^{-1} \sum_{i=1}^{n} \frac{A_i Y_i - \{A_i - \hat{e}_i\} m_1(L_i, \hat{\tau})}{\hat{e}_i}$$

and

$$\hat{\lambda}^0_{DR} = n^{-1} \sum_{i=1}^{n} \frac{(1 - A_i) Y_i + \{A_i - \hat{e}_i\} m_0(L_i, \hat{\tau})}{1 - \hat{e}_i}$$

Note $\hat{e}_i$ is the estimated propensity score for participant i from the weight model as described in Section 4.2.2 and $m_a(L_i, \hat{\tau}) = \hat{E}(Y_i \mid L_i = l_i, A_i = a)$ is the predicted potential outcome for participant i from the parametric g-formula model for $a \in \{0, 1\}$, based on one of the four distributions specified in Section 4.2.3. The causal mean estimators $\hat{\lambda}^a_{DR}$ for $a \in \{0, 1\}$ are of the form considered in Lunceford and Davidian (2004) that were originally proposed by Robins et al. (1994).

Proposition 4.5. When either the weight or outcome model is correctly specified, (4.3) is a consistent and asymptotically normal estimator of the CRR.

Estimates of (4.3) and its variance can be computed using estimating equation theory, as demonstrated in Sections 4.3 and 4.4.

4.2.5 Data Heaping

When data heaping is present, the estimators discussed in Sections 4.2.2-4.2.4 are biased in general. Define the potential outcomes for participant i following data heaping as Y_{hi}^a for $a \in \{0, 1\}$, with $Y_i^a \neq Y_{hi}^a$ for some i and thus $E(Y_{hi}^a) \neq \lambda^a$ in general. The extent of bias depends on the amount of heaping in the data and the mean of the underlying count distribution (Wang and Heitjan, 2008). One common type of data heaping is data coarsening, where some participants round their reported count outcome to the midpoint of a heaping interval. For example, a woman who has between 25 and 35 sexual partners might report her number of sexual partners rounded to the nearest ten and might report 30 sexual partners rather than the exact count.

Valid inference can be made when data are coarsened using a mixture of likelihoods approach, as outlined in Cummings et al. (2015). Under this approach, outcomes that are not multiples of the heaping interval length (H) are treated as not censored while observations that occur at multiples of H are treated as interval censored. Specifically, let $\Delta_i = I(Y_i \ mod \ H = 0) \ H/2$, where $I(Y_i \ mod \ H = 0)$ is a $\{0, 1\}$ indicator that is zero if Y_i is not a multiple of H and is one otherwise. The realized value of Δ_i based on the observed data is $\delta_i = I(y_i \ mod \ H = 0) \ H/2$. Define $Y_{li} = max(0, Y_i - \Delta_i)$ and $Y_{ri} = Y_i + \Delta_i$ as the left and right endpoints of the heaping interval for participant i, respectively, and $y_{li} = max(0, y_i - \delta_i)$ and $y_{ri} = y_i + \delta_i$ are the realized values based on the observed data. Here it is assumed that H is an even number and that the heaping interval is the same for all participants, but these endpoints can be defined more generally as in Cummings et al. (2015).

To estimate the CRR in the presence of data heaping using the interval censoring method, a parametric model is specified based on the assumed distribution of the underlying true counts and a regression model is fit to the observed heaped data based on the log-likelihood function

$$\mathcal{L} = \sum_{i=1}^{n} logP\{Y_i \in (y_{li}, y_{ri}) \mid Y_i \sim P(Y_i = y_i \mid Z_i = z_i)\} \tag{4.4}$$

where $P(Y_i = y_i \mid Z_i = z_i)$ is the probability mass function for the Poisson, NB, ZIP, or ZINB distribution as specified in Section 4.2.3 and Z_i is defined below for the MSM with IPTW, parametric g-formula, and doubly robust methods.

This approach relies on the assumption of noninformative interval censoring, i.e., that the underlying behavior that drives the censoring makes no contributions to the likelihood function. Stated more formally, noninformative interval censoring implies that $P(Y_i \leq y_i \mid Y_{li} = y_{li}, Y_{ri} = y_{ri}, Y_{li} \leq Y_i \leq Y_{ri}) = P(Y_i \leq y_i \mid y_{li} \leq Y_i \leq y_{ri})$ (Zhang and Sun, 2010). When interval censoring is noninformative, the likelihood can be factored and the portion of the likelihood that relies on censoring does not include the parameters and is thus ignored (Klein and Moeschberger, 2006, page 77).

4.2.5.1 MSM with IPTW

As discussed in Section 4.2.2, the parameters of the following MSM can be consistently estimated using IPTW: $\log(\lambda^a) = \beta_0 + \beta_1 a$. To account for data heaping using the interval censoring method, a parametric distribution is assumed for the marginal distribution of Y^a for $a \in \{0, 1\}$. For example, assume that Poisson distributions with means λ^a are specified. Then $log(\lambda_i^a) = x_i\beta$ for $i = 1, ..., n$, where x_i is the row vector corresponding to the i^{th} row of the design matrix for the Y^a model, and thus includes one (corresponding to the intercept term) and the observed exposure for participant i.

The parameters of the MSM are estimated by applying IPTWs to the score function associated with the log-likelihood (4.4), with $Z_i = A_i$. Define the $h \times 1$ vector of estimating equations for the heaped MSM to be

$$\sum_{i=1}^{n} W_i(\hat{\alpha}) \frac{\partial}{\partial \beta_j} \big\{ logP\{Y_i \in (y_{li}, y_{ri}) \mid Y_i \sim P(Y_i = y_i \mid A_i = a_i) \big\} = 0 \tag{4.5}$$

where $j = 1, .., h$ and h is the number of parameters in the heaped MSM, $W_i(\hat{\alpha})$ is the estimated IPTW as described in Section 4.2.2, and $P(Y_i = y_i \mid A_i = a_i)$ is the probability mass function

for the assumed parametric distribution of Y^a. Parameter estimates for the MSM are obtained by finding the maximum likelihood estimates (MLEs) for β from (4.5) and the CRR is estimated as $\widehat{CRR}_{MSM,heap} = \exp(\hat{\beta}_k)$, where k is the coefficient of the MSM corresponding to the exposure A. Note when Y^a is assumed to follow a ZIP or ZINB, a marginalized ZIP or ZINB can be used to model Y^a such that the CRR can be obtained across the susceptible and non-susceptible populations (Long et al., 2014; Preisser et al., 2016). Robust standard error estimates for $\widehat{CRR}_{MSM,heap}$ can be obtained using M-estimation.

4.2.5.2 Parametric g-formula

The parametric g-formula estimator of the CRR can be modified to accommodate data heaping using the interval censoring method by replacing the log-likelihood functions for the Poisson, NB, ZIP, or ZINB distribution with (4.4) where $P(Y_i = y_i \mid Z_i = z_i)$ is the probability mass function for the specified distribution and $Z_i = \{A_i, L_i\}$. Specifically, define the parametric g-formula estimator that accommodates data heaping as:

$$\widehat{CRR}_{PG,heap} = \frac{n^{-1}\sum_{i=1}^{n} \hat{E}_{heap}(Y_i \mid L_i = l_i, A_i = 1)}{n^{-1}\sum_{i=1}^{n} \hat{E}_{heap}(Y_i \mid L_i = l_i, A_i = 0)} \tag{4.6}$$

MLEs for the parameters in the heaping model are obtained and are used to calculate $\hat{E}_{heap}(Y_i \mid L_i = l_i, A_i = a)$ for $a \in \{0, 1\}$ and $i = 1...n$, as outlined in Section 4.2.3 and (4.6) is derived.

Corollary 4.1. Under correctly specified models, (4.6) is a consistent and asymptotically normal estimator of the CRR.

The proof of Corollary 4.1 follows from the proof of Proposition 4.4 after noting that the score function for the interval censored heaping model ψ_{τ_j} is unbiased based on maximum likelihood theory (McCullagh and Nelder., 1989, page 28), with solution $\hat{\tau}$.

Robust standard error estimates for $\widehat{CRR}_{PG,heap}$ can be obtained using M-estimation.

4.2.5.3 Doubly Robust Estimation

When observed count data are heaped, the doubly robust estimator proposed in Section 4.2.4 is biased, even when the predicted potential outcomes $m_a(L_i, \hat{\tau})$ for $a \in \{0, 1\}$ are derived as in Section 4.2.5.2. This can be shown by rewriting the estimating equation for $\hat{\lambda}^1_{DR}$, as derived in the proof of Proposition 4.5, in the presence of data heaping as $\psi_1(Y_i, A_i, L_i; \hat{\alpha}, \hat{\tau}, \lambda^1) = Y^1_{hi} + \{A_i - \hat{e}_i(L_i, \hat{\alpha})\}\{Y^1_{hi} - m_1(L_i, \hat{\tau})\}\{\hat{e}_i(L_i, \hat{\alpha})\}^{-1} - \lambda^1$. When a correctly specified heaping model is used to estimate the potential outcomes, $E\{Y^1_{hi} - m_1(L_i, \hat{\tau})\} = 0$. However, $E(Y^1_{hi}) - \lambda^1 \neq 0$ in general, so the doubly robust estimator (4.3) is generally biased in the presence of data heaping. Because there is no intuitive way to combine the MSM with IPTW estimator proposed in Section 4.2.5.1 with the parametric g-formula estimator proposed in Section 4.2.5.2, we propose a DR estimator of the form developed by Scharfstein et al. (1999) and further evaluated by Bang and Robins (2005):

$$\widehat{CRR}_{DR,heap} = \frac{n^{-1}\sum_{i=1}^{n} \hat{E}_{heap}(Y_i \mid L_i = l_i, A_i = 1, \hat{r}_i)}{n^{-1}\sum_{i=1}^{n} \hat{E}_{heap}(Y_i \mid L_i = l_i, A_i = 0, \hat{r}_i)} \tag{4.7}$$

where $\hat{r}_i = \hat{e}_i^{-1}$ is the inverse of the estimated propensity score for participant i as described in Section 4.2.2. The estimated potential outcomes $\hat{E}_{heap}(Y_i \mid L_i = l_i, A_i = a, \hat{r}_i)$ for $a \in \{0, 1\}$ are calculated from the parameter estimates for an interval censored outcome regression model, as specified in Section 4.2.5.2, but including $\hat{r}$ as a covariate in the outcome model. The DR estimator (4.7) is a consistent and asymptotically normal estimator of the CRR when either the weight or outcome model is correctly specified, based on a delta method argument to the proof in Bang and Robins (2005). Robust standard error estimates for $\widehat{CRR}_{DR,heap}$ can be obtained using M-estimation.

4.3 Simulation Study

A simulation study was conducted to examine and compare the empirical properties of the estimators. The goal of the simulation study was to estimate the CRR for a binary exposure A and a count outcome Y in the presence of three confounding variables L_1, L_2, and L_3 using the MSM,

parametric g-formula, and doubly robust estimators proposed in Section 4.2. Simulations were conducted both without data heaping, where the true outcome was observed (Section 4.3.1) and with data heaping, where the observed count was rounded to the nearest ten for some participants (Section 4.3.2).

4.3.1 Without Data Heaping

Data were simulated to approximate the distributions of WIHS variables, which are further discussed in Section 4.4. The sample size was $n = 800$, which is similar to the size of the WIHS analytic sample presented in Section 4.4. To examine the large-sample properties of the estimators, simulations were also conducted for $n = 2000$, with the results presented in Appendix B. Ignoring subscripts i for notational ease, let L_1 represent a participant's baseline age, where $L_1 \sim Uniform(20, 40)$. Let L_2 represent baseline binary drug use status, with $L_2 \sim Binomial(p_1)$, where $p_1 = \text{logit}^{-1}(-(l_1 - 0.5)/100 + \epsilon_1)$ and $\epsilon_1 \sim Uniform(-1, 1)$. The baseline binary sex exchange for money or drugs variable is represented by $L_3 \sim Binomial(p_2)$, where $p_2 = \text{logit}^{-1}(-3 - (l_1 - 0.5)/100 + 1.2l_2 + \epsilon_2)$, and $\epsilon_2 \sim Uniform(-0.5, 0.5)$. The exposure A represents the binary incarceration status at the visit following baseline, where $A \sim Binomial(p_3)$, and $p_3 = \text{logit}^{-1}(-0.5 - l_1/100 + 0.5l_2 + 0.5l_3)$.

The outcome of interest Y represents the number of total male sexual partners in the six-month period following measurement of the exposure, and it was generated under the four assumed parametric distributions: Poisson, NB, ZIP, and ZINB. Let the parameters of the four distributions from Section 4.2.3 equal $\mu^1 = \eta^1 = \exp(-1 - 0.005l_1 + 0.7l_2 + 3.5l_3 + 0.5)$, $\mu^0 = \eta^0 = \exp(-1 - 0.005l_1 + 0.7l_2 + 3.5l_3)$, $\theta = 0.5$, and $(1 - \nu^1) = (1 - \nu^0) = \text{logit}^{-1}(-2.5 + l_1/100 - 0.3l_2 - 2l_3)$.

For each scenario, $log(CRR) = 0.5$. The estimated causal rate ratios $\widehat{CRR}_{MSM}$, $\widehat{CRR}_{PG}$, and $\widehat{CRR}_{DR}$ and their estimated variances were calculated for each scenario as described in Section 4.2 (1) under correct model specification and (2) when the weight and/or outcome model were incorrectly specified by excluding L_2. Standard errors for $\widehat{CRR}_{MSM}$ were estimated both conservatively treating the weights as fixed or known and appropriately treating the weights as

estimated. Standard error estimates were computed using the geex package in R, which implements M-estimation (Stefanski and Boos, 2002; Saul and Hudgens, 2020).

The results of the simulation for $n = 800$ are presented in Table 4.1 when both models were correctly specified and Table 4.2 when one or both models were misspecified. These results demonstrate minimal empirical bias regardless of the method or underlying distribution of the data when models were correctly specified. Empirical bias was even smaller when the sample size was increased to $n = 2000$, as shown in Supplemental Tables B1 and B2. For MSM with IPTWs, standard error ratios were close to one and empirical coverage was close to the nominal 95% level when the weight model was correctly specified and weights were appropriately treated as estimated. When weights were treated as fixed, estimated standard errors were too large, leading to standard error ratios over 1.5 and confidence intervals with empirical coverage at or near 100%. The parametric g-formula and doubly robust estimators yielded more precise estimates than MSM with IPTW, with the parametric g-formula yielding the smallest estimated standard errors.

As anticipated, the doubly robust estimators yielded minimal empirical bias when the weight or outcome model was misspecified, while the MSM with IPTW and parametric g-formula estimators were empirically biased under weight and outcome model misspecification, respectively (see Table 4.2). The doubly robust estimators were empirically biased when both models were misspecified. For the doubly robust estimators, average estimated standard errors were smaller when the outcome model was correctly specified than when it was misspecified. However, misspecification of the weight models yielded similar doubly robust average standard errors as the average standard errors produced under correct specification of the weight model. These findings are consistent with the empirical results in Funk et al. (2011).

Table 4.1: Results of the simulation study by distribution and method across $R = 1000$ samples with correct model specification, $n = 800$. Empirical bias, ASE, ESE, SER, and empirical 95% confidence interval coverage calculated for the CRR.

Distribution	Method	Empirical Bias	ASE	ESE	SER	95% CI Coverage
Poisson	MSM, fixed	0.008	0.342	0.120	2.852	1.000
	MSM, estimated	0.008	0.117	0.120	0.977	0.947
	PG	0.004	0.082	0.085	0.957	0.946
	DR	0.004	0.082	0.086	0.954	0.946
NB	MSM, fixed	0.029	0.421	0.274	1.537	0.996
	MSM, estimated	0.029	0.268	0.274	0.978	0.941
	PG	0.003	0.162	0.160	1.016	0.953
	DR	0.023	0.253	0.257	0.987	0.941
ZIP	MSM, fixed	0.003	0.351	0.123	2.851	1.000
	MSM, estimated	0.003	0.125	0.123	1.020	0.956
	PG	0.000	0.083	0.087	0.956	0.929
	DR	0.000	0.090	0.095	0.952	0.937
ZINB	MSM, fixed	0.031	0.431	0.283	1.520	0.995
	MSM, estimated	0.031	0.275	0.283	0.971	0.928
	PG	0.006	0.177	0.174	1.022	0.945
	DR	0.030	0.266	0.271	0.983	0.921

Note: ZIP PG and DR results exclude 3.2% and 3.1% of simulations, respectively, where outcome models failed to converge in geex. ZINB PG and DR results exclude 3.9% and 3.8% of simulations, respectively, where outcome models failed to converge in geex. Abbreviations: NB=Negative Binomial; ZIP=Zero-Inflated Poisson; ZINB=Zero-Inflated Negative Binomial; MSM=Marginal Structural Modeling with IPTW; PG=Parametric g-formula; DR=Doubly Robust Estimator; ASE=Average Estimated Standard Error; ESE=Empirical Standard Error; SER=Standard Error Ratio (ASE/ESE); CI=Confidence Interval;

4.3.2 With Data Heaping

To demonstrate the empirical properties of the estimators that account for data heaping, data were simulated under a scenario where the estimators defined in Sections 4.2.2-4.2.4 were expected to be substantially biased. Let L_1, L_2, L_3, and A follow the same distributions as in Section 4.3.1. Let $Y^a \sim Poisson(\mu^a)$ for $a \in \{0, 1\}$, where $\mu^1 = \exp(1.2 - 0.005l_1 + 0.4l_2 + 0.4l_3 + 0.5)$ and $\mu^0 = \exp(1.2 - 0.005l_1 + 0.4l_2 + 0.4l_3)$. Thus, $log(CRR) = 0.5$. Let C represent a binary coarsening indicator that equals one if the participant rounded her reported count to the nearest ten and zero if the participant reported the exact count. Here, $C \sim Bernoulli(p_c)$ where $p_c = 0$ if $y < 5$ and $p_c = 0.5$ if $y \geq 5$. This resulted in approximately 20% of reported outcomes rounded to the nearest ten (Figure 4.1).

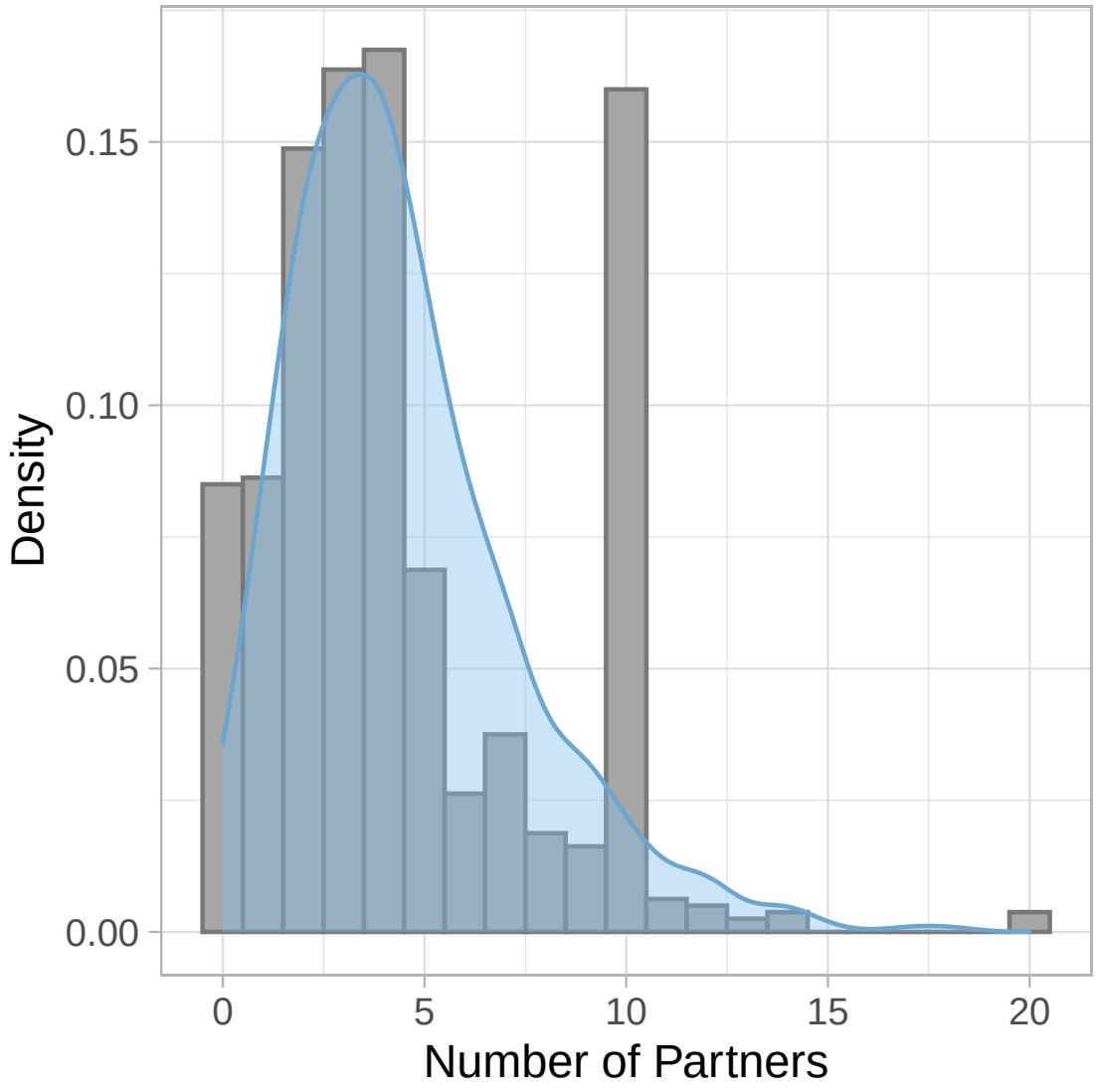

Figure 4.1: Density plot of the true distribution of partners with a histogram of the reported (heaped) number of partners for a single simulation, $n = 800$

The estimated causal rate ratios based on MSM with IPTW, the parametric g-formula, and the doubly robust estimator and their estimated variances were calculated for each scenario both ignoring data heaping, using the estimators described in Sections 4.2.2-4.2.4 (referred to as naïve in the results below), and accounting for data heaping using the interval censoring method, as described in Section 4.2.5. As with the simulations presented in Section 4.3.1, the CRR was estimated under correct model specification and when the weight and/or outcome model were incorrectly specified by excluding L_2. Standard errors for the MSM estimators were estimated both treating the weights as fixed or known and appropriately treating the weights as estimated. Standard error estimates were computed using M-estimation with the geex package in R.

The results of the data heaping simulations are presented in Table 4.3 under correct model specification and in Table 4.4 when the weight model, outcome model, or both models were incorrectly specified. When data heaping was ignored and the standard estimators from Sections 4.2.2-4.2.4 were applied to coarsened data, the estimates exhibit considerable bias and 95% CI coverage is below the nominal level. However, standard error ratios are close to one when the naïve estimators were applied to heaped data.

The heaping estimators demonstrate minimal empirical bias regardless of the method when models were correctly specified. For MSM with IPTWs, the standard error ratio was close to one when weights were appropriately treated as estimated, but exceeded one when weights were treated as fixed. The three estimators had similar average estimated standard errors.

As anticipated, the doubly robust estimators yielded minimal empirical bias when the weight or outcome model was misspecified, while the MSM with IPTW and parametric g-formula estimators were empirically biased under weight and outcome model misspecification, respectively (Table 4.4). Not surprisingly, the doubly robust estimators were empirically biased when both models were misspecified.

Table 4.2: Results of the simulation study by distribution and method across $R = 1000$ samples with one or both models misspecified, $n = 800$. Empirical bias, ASE, ESE, SER, and empirical 95% confidence interval coverage calculated for the CRR.

Distribution	Method	Empirical Bias	ASE	ESE	SER	95% CI Coverage
Poisson	MSM, fixed, MW	0.123	0.362	0.134	2.708	1.000
	MSM, estimated, MW	0.123	0.131	0.134	0.979	0.866
	DR, MW	0.004	0.082	0.085	0.955	0.946
NB	MSM, fixed, MW	0.144	0.447	0.293	1.525	1.000
	MSM, estimated, MW	0.144	0.288	0.293	0.982	0.945
	DR, MW	0.021	0.252	0.255	0.987	0.937
ZIP	MSM, fixed, MW	0.120	0.372	0.139	2.682	1.000
	MSM, estimated, MW	0.120	0.140	0.139	1.007	0.891
	DR, MW	0.000	0.090	0.094	0.953	0.939
ZINB	MSM, fixed, MW	0.148	0.457	0.305	1.497	0.997
	MSM, estimated, MW	0.148	0.296	0.305	0.968	0.950
	DR, MW	0.028	0.263	0.269	0.976	0.921
Poisson	PG, MO	0.123	0.130	0.134	0.966	0.860
	DR, MO	0.010	0.119	0.124	0.961	0.940
NB	PG, MO	0.143	0.183	0.180	1.018	0.922
	DR, MO	0.029	0.268	0.274	0.978	0.945
ZIP	PG, MO	0.119	0.134	0.134	1.000	0.880
	DR, MO	0.005	0.127	0.125	1.017	0.954
ZINB	PG, MO	0.146	0.196	0.196	1.003	0.925
	DR, MO	0.033	0.276	0.285	0.969	0.928
Poisson	DR, MB	0.123	0.130	0.135	0.965	0.861
NB	DR, MB	0.145	0.287	0.292	0.981	0.944
ZIP	DR, MB	0.121	0.139	0.138	1.004	0.893
ZINB	DR, MB	0.151	0.295	0.306	0.963	0.948

Note: ZIP PG and DR results exclude 0%-3.1% of simulations where outcome models failed to converge in geex. ZINB PG and DR results exclude 2.2%-3.8% of simulations where outcome models failed to converge in geex. Abbreviations: NB=Negative Binomial; ZIP=Zero-Inflated Poisson; ZINB=Zero-Inflated Negative Binomial; MSM=Marginal Structural Modeling with IPTW; PG=Parametric g-formula; DR=Doubly Robust Estimator; MW=Misspecified Weight Model; MO=Misspecified Outcome Model; MB=Both Weight and Outcome Models Misspecified; ASE=Average Estimated Standard Error; ESE=Empirical Standard Error; SER=Standard Error Ratio (ASE/ESE); CI=Confidence Interval

Table 4.3: Results of the data heaping simulation study by method across $R = 1000$ samples with correct model specification, $n = 800$. All heaping estimators and the naïve PG and DR estimators assume a Poisson distribution. Empirical bias, ASE, ESE, SER, and empirical 95% confidence interval coverage calculated for the CRR.

Method	Estimator	Empirical Bias	ASE	ESE	SER	95% CI Coverage
MSM, fixed	Naïve	0.072	0.086	0.079	1.087	0.900
	Heaping	-0.022	0.078	0.073	1.068	0.948
MSM, estimated	Naïve	0.072	0.079	0.079	0.991	0.864
	Heaping	-0.022	0.074	0.073	1.014	0.937
PG, Poisson	Naïve	0.072	0.079	0.079	0.991	0.865
	Heaping	0.006	0.072	0.074	0.978	0.949
DR, Poisson	Naïve	0.071	0.079	0.079	0.993	0.864
	Heaping	0.005	0.072	0.074	0.978	0.952

Abbreviations: MSM=Marginal Structural Modeling with IPTW; PG=Parametric g-formula; DR=Doubly Robust Estimator; ASE=Average Estimated Standard Error; ESE=Empirical Standard Error; SER=Standard Error Ratio (ASE/ESE); CI=Confidence Interval

Table 4.4: Results of the data heaping simulation study by method across $R = 1000$ samples with one or both models misspecified, $n = 800$. All estimators assume a Poisson distribution. Empirical bias, ASE, ESE, SER, and empirical 95% confidence interval coverage calculated for the CRR.

Method	Empirical Bias	ASE	ESE	SER	95% CI Coverage
MSM, fixed, MW	0.060	0.081	0.078	1.033	0.915
MSM, estimated, MW	0.060	0.079	0.078	1.007	0.901
DR, MW	0.005	0.074	0.074	1.006	0.951
PG, MO	0.068	0.078	0.078	0.996	0.873
DR, MO	0.009	0.073	0.074	0.980	0.944
DR, MB	0.068	0.082	0.079	1.039	0.877

Abbreviations: MSM=Marginal Structural Modeling with IPTW; PG=Parametric g-formula; DR=Doubly Robust Estimator; MW=Misspecified Weight Model; MO=Misspecified Outcome Model; MB=Both Weight and Outcome Models Misspecified; ASE=Average Estimated Standard Error; ESE=Empirical Standard Error; SER=Standard Error Ratio (ASE/ESE); CI=Confidence Interval

4.4 Example: Women's Interagency HIV Study

To demonstrate the application of these methods, we built on the analysis described in Knittel et al. (2020), which estimated the effect of incarceration on the total number of vaginal, oral, or anal male sex partners (subsequently referred to as partners) during subsequent six-month period using data from the Women's Interagency HIV Study (WIHS). The WIHS is a multicenter cohort study of women living with HIV or at risk of acquiring HIV (Adimora et al., 2018). At each biannual visit, the WIHS collects data regarding women's self-reported incarceration status and sexual behavior since the prior visit. Because of the complex relationships between incarceration, high risk sexual behavior, drug use, and sex exchange for money or drugs, the effect of incarceration on the number of partners is likely confounded. As shown in Figure 4.2, many participants reported no partners over a six-month period and the number of partners exhibits potential overdispersion, with some participants reporting large numbers of partners relative to the mean. Furthermore, there is evidence of data heaping as women with five or more partners tended to report counts at multiples of ten. Due to a lack of causal methods to accommodate these unique features of count outcomes, the authors categorized the number of partners in the original analysis described in Knittel et al. (2020).

To assess the effect of incarceration on the number of partners in the six-month period following incarceration, we estimated the causal rate ratio using the estimators presented in Section 4.2. Because the WIHS are observational data, many variables likely confound the effect of incarceration on the number of partners. As in Knittel et al. (2020), we assume age, educational attainment (high school or more versus less than high school), race (black, white, or other), WIHS site, HIV status, prior incarceration status, unstable housing (living in a rooming/boarding/halfway house), sex exchange practices (sex for drugs, money, or shelter), alcohol use (none, 1-7 drinks/week, or > 7 drinks/week), marijuana use, and hard drug use (crack cocaine, cocaine, heroin, methamphetamines, other opioids, or any injection use) confound the effect of incarceration on the number of partners. Let the study period be the visit in which the exposure, incarceration status, was measured. Then the visit prior to the study period represents the baseline visit. For the purposes of this application, we

assume that baseline values for the above set of covariates L provide conditional exchangeability, i.e., that $Y^a \perp A \mid L$, $a \in \{0, 1\}$, where A represents the binary incarceration status during the study period, Y^1 represents the total number of sexual partners in the six-month period following the study period if, possibly counter to fact, the participant was incarcerated during the study period. Y^0 represents the number of total sexual partners in the six-month period following the study period if, possibly counter to fact, the participant was not incarcerated during the study period. We also assume that positivity and causal consistency hold for this application. To facilitate fitting zero-inflated models to these data, we assume that baseline values for the set of covariates X_2 consisting of age, marital status (legally married/common-law married/living with a partner or widowed/divorced/marriage annulled/separated/never married/other), sex exchange practices, HIV status, and sexual orientation (lesbian/gay or heterosexual/straight/bisexual/other) allow for correct specification of the susceptibility model. These variable classifications were made to predict a woman's potential to have one or more male sexual partners in subsequent study visits.

The analytic sample was derived by restricting the longitudinal WIHS dataset of 4,982 women to women who attended at least one visit between 2007-2017, as 2007 is when incarceration questions were added to the WIHS questionnaire. As in Knittel et al. (2020), the dataset was further restricted to include only women without missing covariates L and X_2 following implementation of last value carried forward and previous value carried back imputation, excluding the history of incarceration covariate which was only asked at a single timepoint and thus could not be imputed using this method. For each woman who reported being incarcerated between 2007-2017, her first incarcerated visit following a non-incarcerated visit was selected as her study period visit. This allowed for an appropriate run-in period in which to measure covariates L at the visit preceding the study period visit. The outcome Y was measured at the visit following the study period visit. This resulted in $n = 294$ incarcerated women after excluding the 28 women missing outcome data at the visit following the study period visit. A sample of one visit from each of $n = 588$ women who did not report being incarcerated between 2007-2017 was randomly selected, ensuring the same distribution of study period visits as the incarcerated women and restricting sampling to women

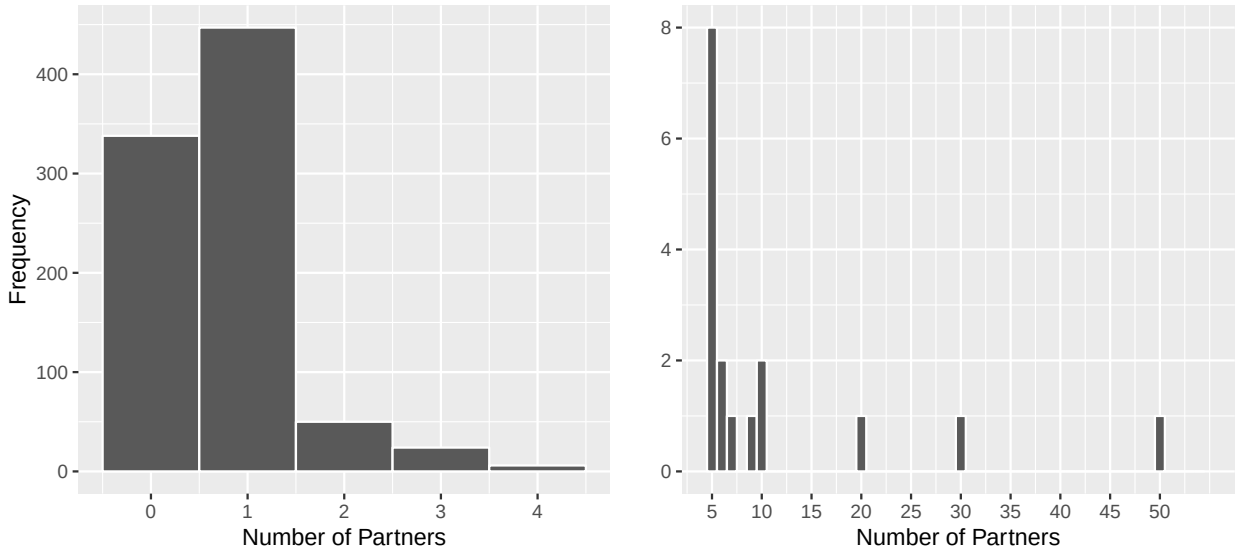

Figure 4.2: Distribution of partners during the six months following the study period reported by WIHS participants in the analytic sample: 0-4 partners (left, $n = 865$) and 5 or more partners (right, $n = 17$)

with non-missing outcome data at the visit following the sampled study period visit. Missing values for prior incarceration for $n = 14$ participants were imputed with the mode (no history of incarceration).

$\widehat{CRR}_{MSM}$, $\widehat{CRR}_{PG}$, and $\widehat{CRR}_{DR}$ were calculated as described in Section 4.2 and standard error estimates were computed using M-estimation with the geex package in R. For each estimate, 95% Wald confidence intervals were constructed. For the MSM with IPTW approach, the set of covariates L was included in the logistic regression model for computing the IPTWs. For the parametric g-formula estimators, the set of covariates $X_1 = \{L, A\}$ was included in the count outcome models, and for zero-inflated models, the covariates X_2 were included in the logistic susceptibility models. In all models, age was included as a continuous predictor and the remaining covariates were treated as categorical. Doubly robust estimators included the same weight and outcome model specifications as the MSM with IPTW and parametric g-formula models, respectively.

When calculating standard errors for $\widehat{CRR}_{MSM}$, the weights were appropriately treated as estimated in the computation of standard errors and confidence intervals. To compare the fit of

parametric models to these data, the Akaike information criterion (AIC) was computed for each parametric model (Agresti, 2002, pages 216-217).

Table 4.5 presents estimates of the CRR based on the three estimators. The estimators provide similar estimates of the CRR, regardless of the assumed underlying distribution of the outcome. The mean number of partners if everyone in the population were incarcerated is estimated to be about 1.3 times the mean number of partners if no one in the population were incarcerated. Precision estimates are also similar across estimates, with the parametric g-formula and an assumed NB distribution having the smallest estimated standard error and the parametric g-formula and an assumed ZIP distribution having the largest estimated standard error. AIC values for the Poisson, NB, ZIP, and ZINB were 2321, 2125, 2323, and 2132, respectively, indicating that the NB distribution provided the best fit for the WIHS data.

Table 4.5: Estimated causal rate ratios, estimated standard errors, and Wald 95% confidence intervals for the effect of incarceration on the number of male sexual partners in the subsequent six months by method and assumed parametric distribution, WIHS 2007-2017

Method	Distribution	$\widehat{CRR}$	$\widehat{SE}(\widehat{CRR})$	Wald 95% CI
MSM	n/a	1.27	0.30	(0.68, 1.85)
PG	Poisson	1.30	0.34	(0.64, 1.97)
	NB	1.35	0.21	(0.93, 1.76)
	ZIP	1.31	0.38	(0.55, 2.06)
	ZINB	1.35	0.22	(0.92, 1.78)
DR	Poisson	1.32	0.28	(0.77, 1.88)
	NB	1.33	0.30	(0.74, 1.93)
	ZIP	1.33	0.29	(0.76, 1.90)
	ZINB	1.34	0.31	(0.73, 1.94)

Abbreviations: NB=Negative Binomial; ZIP=Zero-Inflated Poisson; ZINB=Zero-Inflated Negative Binomial; MSM=Marginal Structural Modeling with IPTW; PG=Parametric g-formula; DR=Doubly Robust Estimator; SE=Standard Error; CI=Confidence Interval; n/a=Not Applicable

As a sensitivity analysis, the WIHS data were analyzed to account for data heaping at intervals of ten, as discussed in Section 4.2.5. Because over 98% of the analytic sample reported zero to four partners in the six months following the study period, data heaping had only a small effect on the estimates. Assuming that the number of partners follows a Poisson distribution, $\widehat{CRR}_{MSM,heap}$

was 1.27 with a 95% CI of (0.81, 1.72), $\widehat{CRR}_{PG,heap}$ was 1.26 with a 95% CI of (0.64, 1.88), and $\widehat{CRR}_{DR,heap}$ was 1.26 with a 95% CI of (0.64, 1.88). Both the MSM and parametric g-formula heaping estimates had smaller estimated standard errors than their non-heaping counterparts (0.23 versus 0.30 for the MSM with IPTW estimator and 0.32 versus 0.34 for the parametric g-formula estimator), while the doubly robust estimator's standard error was slightly larger when heaping was accounted for (0.31 versus 0.28).

4.5 Discussion

Count outcomes are of common interest in public health research. To estimate the causal effect of a binary exposure on a count outcome with observational data, methods are needed to control for confounding variables. This paper proposes estimators of the causal rate ratio based on marginal structural modeling with IPTWs, the parametric g-formula, and doubly robust estimation. All three estimators can accommodate overdispersion and/or zero-inflation. Under the assumptions of causal consistency, conditional exchangeability, and positivity, these estimators are consistent for the causal rate ratio. Consistency and asymptotic normality holds for the MSM and parametric g-formula estimators under correct exposure and outcome model specification, respectively, and for the doubly robust estimator when either the exposure or the outcome model is correctly specified. Modified estimators are proposed for the causal rate ratio when data are heaped using a mixture of likelihoods approach.

Simulations demonstrate that all estimators were empirically unbiased under correct model specification and led to appropriate standard error estimates when M-estimation was implemented. In simulations, the MSM with IPTW estimator exhibited very conservative 95% confidence interval coverage when weights were treated as fixed or known and thus we recommend appropriately treating weights as estimated by stacking the estimating equations using M-estimation. In simulations, the parametric g-formula and doubly robust estimators were more precise than the MSM with IPTW estimator but required specification of the parametric distribution of the counts. One notable advantage of the MSM with IPTW estimator is that it is robust to the distribution of the outcome

and does not require specification of a parametric model for the outcome when data heaping are not present.

APPENDIX A: TECHNICAL DETAILS FOR CHAPTER 3

This appendix contains proofs of the propositions in Chapter 3.

A.1 Proofs of Propositions

A.1.1 Proposition 3.1

Without loss of generality, consider $a = 1$. Let $X_i = W_{1i}Y_iA_i$ and $Z_i = W_{1i}A_i$. The asymptotic distribution of $\hat{\mu}_1 = \sum_{i=1}^n X_i / \sum_{i=1}^n Z_i$ can be derived using the multivariate delta method (Kong, 1992). Let

$$T_n = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n X_i \\ \frac{1}{n}\sum_{i=1}^n Z_i \end{pmatrix}, \quad \theta = \begin{pmatrix} \mu_x \\ \mu_z \end{pmatrix}, \quad g(\theta) = \frac{\mu_x}{\mu_z}, \quad \nabla g(\theta) = \begin{pmatrix} \frac{1}{\mu_z} \\ -\frac{\mu_x}{\mu_z^2} \end{pmatrix},$$

and

$$\Sigma = \begin{pmatrix} Var(X) & Cov(X,Z) \\ Cov(Z,X) & Var(Z) \end{pmatrix}$$

where $\mu_x = E(X_i)$, $\mu_z = E(Z_i)$, and $\nabla g(\theta)$ is the gradient vector for $g(\theta)$. From the bivariate central limit theorem, $\sqrt{n}(T_n - \theta) \xrightarrow{d} N_2(0, \Sigma)$. Applying the multivariate delta method, $\sqrt{n}\{g(T_n) - g(\theta)\} \xrightarrow{d} N(0, \Sigma_1)$ where $g(T_n) = \hat{\mu}_1$ and

$$\Sigma_1 = \nabla g(\theta)^T \Sigma \nabla g(\theta) = \left(\frac{\mu_x}{\mu_z}\right)^2 \left\{ \frac{Var(X)}{\mu_x^2} + \frac{Var(Z)}{\mu_z^2} - 2\frac{Cov(X,Z)}{\mu_x\mu_z} \right\}$$

Dropping subscripts i for notational ease, note that:

$$\mu_z = E(Z) = E(W_1A) = E_L\left\{ \frac{E_{A|L}A}{P(A=1 \mid L)} \right\} = 1$$

and from Hernán and Robins (2020) Technical Point 2.3, $\mu_x = E(X) = E(W_1AY_1) = \mu_1$. Then,

$$\Sigma_1 = Var(W_1AY_1) + \mu_1^2 Var(W_1A) - 2\mu_1 Cov(W_1AY_1, W_1A) \tag{A.1}$$

A simpler form for Σ_1 is derived by rewriting the components of (A.1) using the following results. First note that

$$\begin{aligned} Cov(W_1AY_1, W_1A) &= E\{(W_1AY_1)(W_1A)\} - E(W_1AY_1)E(W_1A) = E\left(W_1^2AY_1\right) - \mu_1 \\ &= E_L\left\{\frac{E_{A|L}A \;\; E_{Y_1|L}Y_1}{P(A=1 \mid L)^2}\right\} - \mu_1 = E_L\left\{\frac{E_{Y_1|L}Y_1}{P(A=1 \mid L)}\right\} - \mu_1 \\ &= E(W_1Y_1) - \mu_1 \end{aligned} \tag{A.2}$$

Also note that

$$\begin{aligned} Var(W_1AY_1) &= E\left(W_1^2AY_1^2\right) - \{E(W_1AY_1)\}^2 \\ &= E_L\left\{\frac{E_{A|L}A \;\; E_{Y_1|L}Y_1^2}{P(A=1 \mid L)^2}\right\} - \mu_1^2 = E_L\left\{\frac{E_{Y_1|L}Y_1^2}{P(A=1 \mid L)}\right\} - \mu_1^2 \\ &= E\left(W_1Y_1^2\right) - \mu_1^2 \end{aligned} \tag{A.3}$$

By the law of total variance:

$$Var(W_1A) = E\{Var(W_1A \mid L)\} + Var\{E(W_1A \mid L)\} = E(W_1) - 1 \tag{A.4}$$

Therefore, plugging (A.2), (A.3), and (A.4) into (A.1),

$$\begin{aligned} \Sigma_1 &= E(W_1Y_1^2) - \mu_1^2 + \mu_1^2\{E(W_1) - 1\} - 2\mu_1\{E(W_1Y_1) - \mu_1\} \\ &= E(W_1Y_1^2) + \mu_1^2 E(W_1) - 2\mu_1 E(W_1Y_1) \end{aligned} \tag{A.5}$$

Next define $R = E[\{W_1 - E(W_1)\}(Y_1 - \mu_1)^2]$ and note that

$$R = E(W_1 Y_1^2) - 2\mu_1 E(W_1 Y_1) - E(Y_1^2)E(W_1) + 2\mu_1^2 E(W_1) \tag{A.6}$$

From (A.5) and (A.6) it follows that

$$\begin{aligned}
\Sigma_1 &= E(W_1)E(Y_1^2) - E(W_1)\mu_1^2 + R = E(W_1)\sigma_1^2 + R \\
&= \sigma_1^2 E(W_1^2 A) + R = \sigma_1^2 \left[\frac{E(W_1^2 A)}{E(W_1 A)}\right] + R \\
&= \sigma_1^2 \left[\frac{E(W^2 A)}{\{E(WA)\}^2}\right] + R
\end{aligned}$$

Bounds for R follow from the Cauchy-Schwarz inequality:

$$|R| = |Cov(W_1, Y_1^2 - 2\mu_1 Y_1)| \leq \sqrt{Var(W_1)Var(Y_1^2 - 2\mu_1 Y_1)}$$

A.1.2 Proposition 3.2

From equation (13) in Lunceford and Davidian (2004),

$$\Sigma^* = E\{W_1(Y_1 - \mu_1)^2 + W_0(Y_0 - \mu_0)^2\}$$

Note

$$E\{W_1(Y_1 - \mu_1)^2\} = E(W_1 Y_1^2) - 2\mu_1 E(W_1 Y_1) + \mu_1^2 E(W_1)$$

which equals Σ_1 by (A.5). Similarly, $E\{W_0(Y_0 - \mu_0)^2\} = \Sigma_0$, proving the proposition.

A.1.3 Proposition 3.3

Let $1-\beta$ denote the power to detect a difference in causal means of size δ, i.e.,

$$\begin{aligned}
1-\beta &= P(|t| > z_{1-\alpha/2} \mid ACE = \delta) \\
&= P\left(\frac{\widehat{ACE} - \delta}{\sqrt{Var(\widehat{ACE})}} > z_{1-\alpha/2} - \frac{\delta}{\sqrt{Var(\widehat{ACE})}} \middle| ACE = \delta \right) \\
&\quad + P\left(\frac{\widehat{ACE} - \delta}{\sqrt{Var(\widehat{ACE})}} < z_{\alpha/2} - \frac{\delta}{\sqrt{Var(\widehat{ACE})}} \middle| ACE = \delta \right)
\end{aligned}$$

In large samples, (3.9) is approximately standard normal. Thus,

$$1-\beta \approx 1 - \Phi\left(z_{1-\alpha/2} - \frac{\delta}{\sqrt{Var(\widehat{ACE})}} \right) + \Phi\left(z_{\alpha/2} - \frac{\delta}{\sqrt{Var(\widehat{ACE})}} \right) \tag{A.7}$$

where $\Phi(*)$ represents the cumulative distribution function for the standard normal evaluated at $*$. Without loss of generality, assume $\delta > 0$. Then the second component on the right side of (A.7) will be less than $\alpha/2$ and often close to zero. Therefore,

$$z_\beta \approx z_{1-\alpha/2} - \frac{\delta}{\sqrt{Var(\widehat{ACE})}} \tag{A.8}$$

Define $k = P(A=1)/P(A=0)$. Given that $Var(\widehat{ACE}) \approx \{nP(A=1)\}^{-1}\sigma^2_{1,adj} + \{nP(A=0)\}^{-1}\sigma^2_{0,adj}$ and solving (A.8) for n yields (3.10).

APPENDIX B: TECHNICAL DETAILS FOR CHAPTER 4

This appendix contains supplemental tables and proofs of the propositions in Chapter 4.

B.1 Supplemental Tables

Table B1: Results of the simulation study by distribution and method across $R = 1000$ samples with correct model specification, $n = 2000$. Empirical bias, ASE, ESE, SER, and empirical 95% confidence interval coverage calculated for the CRR.

Distribution	Method	Empirical Bias	ASE	ESE	SER	95% CI Coverage
Poisson	MSM, fixed	0.003	0.215	0.078	2.748	1.000
	MSM, estimated	0.003	0.074	0.078	0.944	0.940
	PG	0.002	0.052	0.054	0.949	0.934
	DR	0.002	0.052	0.054	0.948	0.936
NB	MSM, fixed	0.004	0.264	0.176	1.498	0.994
	MSM, estimated	0.004	0.171	0.176	0.967	0.934
	PG	0.001	0.103	0.103	1.003	0.949
	DR	0.004	0.162	0.169	0.961	0.931
ZIP	MSM, fixed	0.005	0.22	0.080	2.767	1.000
	MSM, estimated	0.005	0.079	0.080	0.994	0.937
	PG	0.002	0.053	0.053	0.996	0.947
	DR	0.003	0.058	0.060	0.974	0.949
ZINB	MSM, fixed	0.014	0.272	0.177	1.534	0.996
	MSM, estimated	0.014	0.177	0.177	0.997	0.951
	PG	0.000	0.109	0.108	1.008	0.949
	DR	0.011	0.168	0.168	0.996	0.949

Note: ZIP PG and DR results exclude 1.2% and 1.1% of simulations, respectively, where outcome models failed to converge in geex. ZINB PG and DR results exclude 1.9% of simulations where outcome models failed to converge in geex. Abbreviations: NB=Negative Binomial; ZIP=Zero-Inflated Poisson; ZINB=Zero-Inflated Negative Binomial; MSM=Marginal Structural Modeling with IPTW; PG=Parametric g-formula; DR=Doubly Robust Estimator; ASE=Average Estimated Standard Error; ESE=Empirical Standard Error; SER=Standard Error Ratio (ASE/ESE); CI=Confidence Interval

Table B2: Results of the simulation study by distribution and method across $R = 1000$ samples with one or both models misspecified, $n = 2000$. Empirical bias, ASE, ESE, SER, and empirical 95% confidence interval coverage calculated for the CRR.

Distribution	Method	Empirical Bias	ASE	ESE	SER	95% CI Coverage
Poisson	MSM, fixed, MW	0.118	0.228	0.086	2.662	1.000
	MSM, estimated, MW	0.118	0.083	0.086	0.969	0.713
	DR, MW	0.002	0.052	0.054	0.948	0.934
NB	MSM, fixed, MW	0.120	0.280	0.189	1.483	0.997
	MSM, estimated, MW	0.120	0.183	0.189	0.969	0.924
	DR, MW	0.003	0.161	0.167	0.962	0.931
ZIP	MSM, fixed, MW	0.122	0.234	0.088	2.643	1.000
	MSM, estimated, MW	0.122	0.089	0.088	1.000	0.745
	DR, MW	0.003	0.058	0.060	0.976	0.948
ZINB	MSM, fixed, MW	0.131	0.289	0.190	1.520	0.996
	MSM, estimated, MW	0.131	0.190	0.190	1.000	0.922
	DR, MW	0.010	0.167	0.167	1.003	0.947
Poisson	PG, MO	0.118	0.082	0.085	0.968	0.714
	DR, MO	0.003	0.075	0.080	0.943	0.933
NB	PG, MO	0.139	0.116	0.115	1.004	0.822
	DR, MO	0.005	0.171	0.177	0.968	0.935
ZIP	PG, MO	0.120	0.085	0.085	1.005	0.741
	DR, MO	0.006	0.081	0.081	0.996	0.937
ZINB	PG, MO	0.143	0.123	0.124	0.988	0.812
	DR, MO	0.013	0.177	0.179	0.992	0.951
Poisson	DR, MB	0.118	0.082	0.085	0.967	0.713
NB	DR, MB	0.120	0.183	0.188	0.970	0.923
ZIP	DR, MB	0.122	0.088	0.088	1.000	0.745
ZINB	DR, MB	0.130	0.189	0.191	0.994	0.919

Note: ZIP PG and DR results exclude 0.1%-1.1% of simulations where outcome models failed to converge in geex. ZINB PG and DR results exclude 0.8%-1.9% of simulations where outcome models failed to converge in geex. Abbreviations: NB=Negative Binomial; ZIP=Zero-Inflated Poisson; ZINB=Zero-Inflated Negative Binomial; MSM=Marginal Structural Modeling with IPTW; PG=Parametric g-formula; DR=Doubly Robust Estimator; MW=Misspecified Weight Model; MO=Misspecified Outcome Model; MB=Both Weight and Outcome Models Misspecified; ASE=Average Estimated Standard Error; ESE=Empirical Standard Error; SER=Standard Error Ratio (ASE/ESE); CI=Confidence Interval

B.2 Proofs of Propositions

B.2.1 Proposition 4.1

Consider the following MSM: $\log(\lambda^a) = \beta_0 + \beta_1 a$. $\hat{\beta}_0$ and $\hat{\beta}_1$ are the solutions to the following estimating equations, where W_i is the true weight for participant i and is treated as known.

$$\sum_{i=1}^{n} W_i \begin{bmatrix} Y_i - \exp(\beta_0 + \beta_1 a_i) \\ A_i \left(Y_i - \exp(\beta_0 + \beta_1 a_i)\right) \end{bmatrix} = 0$$

Using the law of iterated expectation and assuming causal consistency and conditional exchangeability:

$$E\{W_i Y_i - W_i \exp(\beta_0 + \beta_1 a_i)\}$$

$$= E\{A_i W_i Y_i - A_i W_i \exp(\beta_0 + \beta_1) + (1 - A_i) W_i Y_i - (1 - A_i) W_i \exp(\beta_0)\}$$

$$= E_L\left\{W_i E_{A|L}(A_i \mid L) E_{Y^1|L}(Y_i^1 \mid L)\right\} - E_L\left\{W_i E_{A|L}(A_i \mid L) \exp(\beta_0 + \beta_1)\right\}$$

$$+ E_L\left\{W_i E_{A|L}\{(1 - A_i) \mid L\} E_{Y^0|L}(Y_i^0 \mid L)\right\} - E_L\left\{W_i E_{A|L}\{(1 - A_i) \mid L\} \exp(\beta_0)\right\}$$

$$= E(Y_i^1) - \exp(\beta_0 + \beta_1) + E(Y_i^0) - \exp(\beta_0) = 0$$

Similarly,

$$E\{W_i A_i Y_i - W_i A_i \exp(\beta_0 + \beta_1 a_i) = 0\}$$

Therefore, these estimating equations are unbiased. The solutions to the estimating equations are derived as follows:

$$\sum_{i=1}^{n} W_i \begin{bmatrix} Y_i - \exp(\hat{\beta}_0 + \hat{\beta}_1 a_i) \\ A_i \left\{Y_i - \exp(\hat{\beta}_0 + \hat{\beta}_1 a_i)\right\} \end{bmatrix} = 0$$

$\iff$

$$\sum_{i=1}^{n} W_i \exp(\hat{\beta}_0 + \hat{\beta}_1 a_i) = \sum_{i=1}^{n} W_i Y_i \tag{B.1}$$

and

$$\sum_{i=1}^{n} W_i A_i \exp(\hat{\beta}_0 + \hat{\beta}_1) = \sum_{i=1}^{n} W_i A_i Y_i \tag{B.2}$$

Solving (B.1) for $\hat{\beta}_0$:

$$\hat{\beta}_0 = log\left(\frac{\sum_{i=1}^{n} W_i Y_i}{\sum_{i=1}^{n} W_i \exp\{\hat{\beta}_1 I(A_i = 1)\}}\right)$$

Plugging $\hat{\beta}_0$ into (B.2):

$$\left[\frac{\sum_{i=1}^{n} W_i Y_i}{\sum_{i=1}^{n} W_i \exp\{\hat{\beta}_1 I(A_i = 1)\}}\right]\left\{\sum_{i=1}^{n} W_i \exp\{\hat{\beta}_1) I(A_i = 1)\right\} = \sum_{i=1}^{n} W_i Y_i I(A_i = 1)$$

$$\implies \hat{\beta}_1 = log\left(\frac{\sum_{i=1}^{n} W_i Y_i I(A_i = 1)}{\sum_{i=1}^{n} W_i I(A_i = 1)} \Big/ \frac{\sum_{i=1}^{n} W_i Y_i I(A_i = 0)}{\sum_{i=1}^{n} W_i I(A_i = 0)}\right)$$

Define $g(\beta_1) = \exp(\beta_1)$. Because $\widehat{CRR}_{MSM} = \exp(\hat{\beta}_1)$ is a delta method transformation of the solution to an unbiased estimating equation, $\widehat{CRR}_{MSM}$ is a consistent and asymptotically normal estimator of CRR when the weights are known (Stefanski and Boos, 2002).

B.2.2 Proposition 4.2

The proof of proposition 2 relies on standard estimating equation theory (see Stefanski and Boos, 2002). Define the set of estimating equations:

$$\sum_{i=1}^{n} \Psi(O_i, \Lambda) = \begin{bmatrix} \sum_{i=1}^{n} \Psi_1(O_i, \Lambda) \\ \sum_{i=1}^{n} \Psi_0(O_i, \Lambda) \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} e_i^{-1}(Y_i - \lambda^1) I(A_i = 1) \\ \sum_{i=1}^{n} (1 - e_i)^{-1}(Y_i - \lambda^0) I(A_i = 0) \end{bmatrix} = 0$$

Then,

$$\dot{\Psi}(O_i, \Lambda) = \frac{\partial \Psi(O_i, \Lambda)}{\partial \Lambda} = \begin{bmatrix} -e_i^{-1} I(A_i = 1) & 0 \\ 0 & -(1 - e_i)^{-1} I(A_i = 0) \end{bmatrix}$$

$$A(\Lambda) = E(-\dot{\Psi}) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

and

$$B(\Lambda) = E[\Psi(O_i, \Lambda)\Psi(O_i, \Lambda)^T] = E\begin{bmatrix} e_i^{-2}(Y_i - \lambda^1)^2 I(A_i = 1) & 0 \\ 0 & (1 - e_i)^{-2}(Y_i - \lambda^0)^2 I(A_i = 0) \end{bmatrix}$$

$$= E\begin{bmatrix} e_i^{-2}(Y_i^1 - \lambda^1)^2 I(A_i = 1) & 0 \\ 0 & (1 - e_i)^{-2}(Y_i^0 - \lambda^0)^2 I(A_i = 0) \end{bmatrix}$$

$$= \begin{bmatrix} E_L\left\{e_i^{-2} E_{Y^1|L}(Y_i^1 - \lambda^1)^2 E_{A|L} I(A_i = 1)\right\} & 0 \\ 0 & E_L\left\{(1 - e_i)^{-2} E_{Y^0|L}(Y_i^0 - \lambda^0)^2 E_{A|L} I(A_i = 0)\right\} \end{bmatrix}$$

$$= E\begin{bmatrix} e_i^{-1}(Y_i^1 - \lambda^1)^2 & 0 \\ 0 & (1 - e_i)^{-1}(Y_i^0 - \lambda^0)^2 \end{bmatrix}$$

by causal consistency, iterated expectation, and conditional exchangeability. Then as $n \to \infty$,

$$\sqrt{n}\left(\begin{bmatrix} \hat{\lambda}^1_{MSM} \\ \hat{\lambda}^0_{MSM} \end{bmatrix} - \begin{bmatrix} \lambda^1 \\ \lambda^0 \end{bmatrix}\right) \xrightarrow{d} N\left(0, V(\Lambda)\right)$$

where

$$V(\Lambda) = A(\Lambda)^{-1} B(\Lambda)\{A(\Lambda)^{-1}\}^T = B(\Lambda)$$

The delta method is applied to obtain the asymptotic variance of $\widehat{CRR}_{MSM} = \hat{\lambda}^1_{MSM}/\hat{\lambda}^0_{MSM}$, with $g(\Lambda) = \lambda^1/\lambda^0$. Thus,

$$G = \frac{\partial g}{\partial \Lambda} = \begin{bmatrix} \frac{1}{\lambda^0} & \frac{-\lambda^1}{(\lambda^0)^2} \end{bmatrix}$$

Then,

$$\sqrt{n}\left(\frac{\hat{\lambda}^1_{MSM}}{\hat{\lambda}^0_{MSM}} - \frac{\lambda^1}{\lambda^0}\right) \xrightarrow{d} N\left(0, \Sigma_{MSM}\right)$$

where

$$\Sigma_{MSM} = GV(\Lambda)G^T = E\left[e_i^{-1}\left(\frac{Y^1-\lambda^1}{\lambda^0}\right)^2 + (1-e_i)^{-1}\left\{\frac{\lambda^1(Y^0-\lambda^0)}{(\lambda^0)^2}\right\}^2\right]$$

B.2.3 Proposition 4.3

When the weights are treated as estimated rather than known, we solve the set of estimating equations:

$$\sum_{i=1}^{n}\Psi(O_i,\Lambda) = \begin{bmatrix}\sum_{i=1}^n \Psi_\alpha(O_i,\Lambda)\\ \sum_{i=1}^n \Psi_1(O_i,\Lambda)\\ \sum_{i=1}^n \Psi_0(O_i,\Lambda)\end{bmatrix} = \begin{bmatrix}\sum_{i=1}^n \Psi_\alpha(O_i,\alpha)\\ \sum_{i=1}^n W_i(\alpha)(Y_i-\lambda^1)I(A_i=1)\\ \sum_{i=1}^n W_i(\alpha)(Y_i-\lambda^0)I(A_i=0)\end{bmatrix} = 0$$

where $\mathbf{\Lambda}^T = (\alpha^T, \lambda^1, \lambda^0)$ are the p parameters from the logistic regression weight model (α) and the two causal means (λ^1 and λ^0), $O_i = (A_i, L_i, Y_i)$ are the observed data, and Ψ_α is the vector of score functions from the logistic regression weight model. Let the solutions to the estimating equations be denoted by $\hat{\Lambda}$, where

$$\hat{\Lambda} = \begin{bmatrix}\hat{\alpha}\\ \hat{\lambda}^1_{MSM}\\ \hat{\lambda}^0_{MSM}\end{bmatrix} = \begin{bmatrix}\hat{\alpha}\\ \sum_{i=1}^n W_i(\hat{\alpha})Y_iI(A_i=1)/\{\sum_{i=1}^n W_i(\hat{\alpha})I(A_i=1)\}\\ \sum_{i=1}^n W_i(\hat{\alpha})Y_iI(A_i=0)/\{\sum_{i=1}^n W_i(\hat{\alpha})I(A_i=0)\}\end{bmatrix}$$

When the weight model is correctly specified, $\hat{\Lambda}$ are solutions to an unbiased set of estimating equations. Thus, $\sqrt{n}(\hat{\Lambda}-\Lambda) \to N(0, V(\Lambda))$, where $V(\Lambda) = A(\Lambda)^{-1}B(\Lambda)\{A(\Lambda)^{-1}\}^T$. Note that $A(\Lambda) = E\{-\dot{\Psi}(O_i,\Lambda)\}$, $B(\Lambda) = E\{\Psi(O_i,\Lambda)\Psi(O_i,\Lambda)^T\}$, and $\dot{\Psi}(O_i,\Lambda) = \partial\Psi(O_i,\Lambda)/\partial\Lambda$ are

$(p+2) \times (p+2)$ matrices. Also note that:

$$\dot{\Psi}(O_i, \Lambda) = \begin{bmatrix} \partial\Psi_\alpha/\partial\alpha & \partial\Psi_\alpha/\partial\lambda^1 & \partial\Psi_\alpha/\partial\lambda^0 \\ \partial\Psi_1/\partial\alpha & \partial\Psi_1/\partial\lambda^1 & \partial\Psi_1/\partial\lambda^0 \\ \partial\Psi_0/\partial\alpha & \partial\Psi_0/\partial\lambda^1 & \partial\Psi_0/\partial\lambda^0 \end{bmatrix} = \begin{bmatrix} \partial\Psi_\alpha/\partial\alpha & 0_{p\times 1} & 0_{p\times 1} \\ \partial\Psi_1/\partial\alpha & -W_i(\alpha)I(A_i = 1) & 0 \\ \partial\Psi_0/\partial\alpha & 0 & -W_i(\alpha)I(A_i = 0) \end{bmatrix}$$

where $\partial\Psi_\alpha/\partial\alpha$ is the $p \times p$ Jacobian matrix of partial derivatives for Ψ_α, $\partial\Psi_a/\partial\alpha$ are gradient vectors for $a \in \{0, 1\}$, and $0_{p\times 1}$ are vectors of 0. Then,

$$A(\Lambda) = \begin{bmatrix} A_1 & 0_{p\times 2} \\ A_2 & I_{2\times 2} \end{bmatrix}$$

where $A_1 = E(-\partial\Psi_\alpha/\partial\alpha)$ and $I_{2\times 2}$ is the identity matrix. Note that

$$A_2 = \begin{bmatrix} E(-\partial\Psi_1/\partial\alpha) \\ E(-\partial\Psi_0/\partial\alpha) \end{bmatrix}$$

and thus

$$A(\Lambda)^{-1} = \begin{bmatrix} A_1^{-1} & 0_{p\times 2} \\ -A_2A_1^{-1} & I_{2\times 2} \end{bmatrix}$$

Let

$$B(\Lambda) = \begin{bmatrix} B_{11} & B_{21}^T \\ B_{21} & B_{22} \end{bmatrix}$$

where B_{11} is $(p \times p)$, B_{21} is $(2 \times p)$, and B_{22} is (2×2). By Lemma 7.3.11 in Casella and Berger (2002), $A_1 = B_{11}$. We claim that $A_2 = B_{21}$. Under this claim,

$$V(\Lambda) = \begin{bmatrix} A_1^{-1} & 0_{p\times 2} \\ -A_2A_1^{-1} & I_{2\times 2} \end{bmatrix} \begin{bmatrix} A_1 & A_2^T \\ A_2 & B_{22} \end{bmatrix} \begin{bmatrix} A_1^{-1} & -(A_1^{-1})^T A_2^T \\ 0_{2\times p} & I_{2\times 2} \end{bmatrix} = \begin{bmatrix} A_1^{-1} & 0_{p\times 2} \\ 0_{2\times p} & B_{22} - A_2A_1^{-1}A_2^T \end{bmatrix}$$

Thus, as $n \to \infty$,

$$\sqrt{n}(\hat{\Lambda} - \Lambda) \xrightarrow{d} N\left(0, V(\Lambda)\right)$$

The delta method is applied to obtain the asymptotic variance of $\widehat{CRR}_{MSM} = \hat{\lambda}^1_{MSM}/\hat{\lambda}^0_{MSM}$, with $g(\Lambda) = \lambda^1/\lambda^0$. Thus,

$$G = \frac{\partial g}{\partial \Lambda} = \begin{bmatrix} 0_{1\times p} & \frac{1}{\lambda^0} & \frac{-\lambda^1}{(\lambda^0)^2} \end{bmatrix}$$

Then,

$$\sqrt{n}\left(\frac{\hat{\lambda}^1_{MSM}}{\hat{\lambda}^0_{MSM}} - \frac{\lambda^1}{\lambda^0}\right) \xrightarrow{d} N\left(0, \Sigma^*_{MSM}\right)$$

where

$$\Sigma^*_{MSM} = GV(\Lambda)G^T = g^{*T}B_{22}g^* - g^{*T}(A_2A_1^{-1}A_2^T)g^* = \Sigma_{MSM} - g^{*T}(A_2A_1^{-1}A_2^T)g^*$$

and

$$g^* = \begin{bmatrix} \frac{1}{\lambda^0} & \frac{-\lambda^1}{(\lambda^0)^2} \end{bmatrix}$$

The final equality holds because $B_{22} = V(\Lambda)$ from the proof to Proposition 4.2. Note that $g^{*T}(A_2A_1^{-1}A_2^T)g^* = g^{*T}(A_2B_{11}^{-1}A_2^T)g^* \geq 0$ because B_{11} is positive semi-definite. This proves the proposition under the claim that $A_2 = B_{21}$.

Now, prove the claim that $A_2 = B_{21}$. Denote the propensity score as $e = \{1 + \exp(-\alpha L)\}^{-1}$. Let $k = 1, .., p$ represent the k^{th} column of the first row of A_2, which equals

$$-E\left(\frac{\partial \Psi_1}{\partial \alpha_k}\right) = E\left(e^{-2}(Y - \lambda^1)I(A_i = 1)\frac{\partial e}{\partial \alpha_k}\right) = E\left(e^{-1}(Y^1 - \lambda^1)\frac{\partial e}{\partial \alpha_k}\right)$$

The k^{th} column of the first row of B_{21} equals

$$E(\Psi_1\Psi_{\alpha_k}) = E\left(e^{-1}(Y - \lambda^1)I(A_i = 1)\frac{I(A_i = 1) - e}{e(1-e)}\frac{\partial e}{\partial \alpha_k}\right) = E\left(e^{-1}(Y^1 - \lambda^1)\frac{\partial e}{\partial \alpha_k}\right)$$

because $E_{A|L}\{I(A_i=1)^2 - eI(A_i=1)\} = E_{A|L}\{I(A_i=1) - eI(A_i=1)\} = (1-e)E_{A|L}\{I(A_i=1)\} = (1-e)e$.

Similarly, the k^{th} column of the second row of A_2 equals

$$-E\left(\frac{\partial \Psi_0}{\partial \alpha_k}\right) = -E\left(\frac{(Y-\lambda^0)I(A_i=0)}{(1-e)^2}\frac{\partial e}{\partial \alpha_k}\right) = -E\left(\frac{Y^0-\lambda^0}{1-e}\frac{\partial e}{\partial \alpha_k}\right)$$

and the k^{th} column of the second row of B_{21} equals

$$E(\Psi_0 \Psi_{\alpha_k}) = E\left(\frac{(Y-\lambda^0)I(A_i=0)}{1-e}\frac{I(A_i=0)-e}{e(1-e)}\frac{\partial e}{\partial \alpha_k}\right) = -E\left(\frac{Y^0-\lambda^0}{1-e}\frac{\partial e}{\partial \alpha_k}\right)$$

Thus, $\Sigma^*_{MSM} \leq \Sigma_{MSM}$.

B.2.4 Proposition 4.4

Assume that $\hat{E}(Y_i \mid L_i = l_i, A_i = a)$ is estimated based on one of the four models described in Section 4.2.3. Define the set of estimating equations:

$$\sum_{i=1}^{n} \psi(Y_i, A_i, L_i; \tau, \lambda) = \begin{bmatrix} \sum_{i=1}^n \psi_\tau(Y_i, A_i, L_i; \tau) \\ \sum_{i=1}^n \psi_1(Y_i, L_i; \hat{\tau}, \lambda^1) \\ \sum_{i=1}^n \psi_0(Y_i, L_i; \hat{\tau}, \lambda^0) \end{bmatrix} = 0$$

where ψ_{τ_j} is the derivative of the log-likelihood function for the model with respect to the j^{th} regression coefficient, for $j = 1, ..., p$. For the Poisson and NB distributions, $\tau = \beta$ and τ has dimension $p = c$, and for the ZIP and ZINB distributions, $\tau = [\gamma\ \xi]^T$ and has dimension $p = c_1 + c_2$. For example, for the Poisson distribution, $\psi_{\tau_j} = \{y_i - \exp(\sum_{j=1}^c x_{ij}\beta_j)\}x_{ij} = 0$, for $j = 1, ..., c$. When the model(s) are correctly specified, these estimating equations are unbiased based on maximum likelihood theory, with solutions $\hat{\tau}$. Now we define the estimating equations for the causal means. Define $\hat{\lambda}^a_{PG} = \hat{E}(Y^a) = \int \hat{E}(Y \mid L = l, A = a) dF_L(l) = n^{-1}\sum_{i=1}^n \hat{E}(Y_i \mid L_i = l_i, A_i = a)$. Then, $\sum_{i=1}^n \psi_a(Y_i, A_i, L_i; \hat{\tau}, \lambda^a) = \sum_{i=1}^n \{\hat{E}(Y_i \mid L_i = l_i, A_i = a) - \lambda^a\} = 0$ for $a \in \{0, 1\}$, where $\hat{E}(Y_i \mid L_i = l_i, A_i = a)$ is the predicted count for participant i based on the model. When the

model(s) are correctly specified and based on causal consistency and conditional exchangeability,

$$E\{\psi_a(Y_i, L_i; \hat{\tau}, \lambda^a)\} = E\{E(Y_i \mid L_i = l_i, A_i = a) - \lambda^a\} = E\{E(Y_i^a \mid L_i = l_i, A_i = a)\} - \lambda^a$$

$$= E\{E(Y_i^a \mid L_i = l_i)\} - \lambda^a = E(Y^a) - \lambda^a = 0$$

Thus, this represents an unbiased set of estimating equations. Under suitable regularity conditions (Stefanski and Boos, 2002),

$$\sqrt{n}\left(\begin{bmatrix}\hat{\tau}\\ \hat{\lambda}^1_{PG}\\ \hat{\lambda}^0_{PG}\end{bmatrix} - \begin{bmatrix}\tau\\ \lambda^1\\ \lambda^0\end{bmatrix}\right) \xrightarrow{d} N\left(0, \Sigma_{PG}\right)$$

where

$$\Sigma_{PG} = A(\Lambda)^{-1}B(\Lambda)\{A(\Lambda)^{-1}\}^T$$

where $\Lambda^T = (\tau^T, \lambda^1, \lambda^0)$, $A(\Lambda) = E\{-\dot{\psi}(Y_i, A_i, L_i, \Lambda)\}$, $B(\Lambda) = E\{\psi(Y_i, A_i, L_i, \Lambda)\psi(Y_i, A_i, L_i, \Lambda)^T\}$, and $\dot{\psi}(Y_i, A_i, L_i, \Lambda) = \partial\psi(Y_i, A_i, L_i, \Lambda)/\partial\Lambda^T$. The delta method is applied to obtain the asymptotic distribution of $\widehat{CRR}_{PG} = \hat{\lambda}^1_{PG}/\hat{\lambda}^0_{PG}$, with $g(\Lambda) = \lambda^1/\lambda^0$. Thus,

$$G_{PG} = \frac{\partial g(\Lambda)}{\partial(\Lambda)} = \begin{bmatrix}0_{1\times p} & \frac{1}{\lambda^0} & \frac{-\lambda^1}{(\lambda^0)^2}\end{bmatrix}$$

Then,

$$\sqrt{n}\left(\frac{\hat{\lambda}^1_{PG}}{\hat{\lambda}^0_{PG}} - \frac{\lambda^1}{\lambda^0}\right) \xrightarrow{d} N\left(0, \Sigma^*_{PG}\right)$$

where

$$\Sigma^*_{PG} = G_{PG}\Sigma_{PG}G^T_{PG}$$

Because $\widehat{CRR}_{PG}$ is a delta method transformation of solutions to an unbiased set of estimating equations, $\widehat{CRR}_{PG}$ is a consistent and asymptotically normal estimator of CRR when the model(s) are correctly specified.

B.2.5 Proposition 4.5

Define the set of estimating equations:

$$\sum_{i=1}^{n} \psi(Y_i, A_i, L_i; \alpha, \tau, \lambda) = \begin{bmatrix} \sum_{i=1}^{n} \psi_\alpha(A_i, L_i; \alpha) \\ \sum_{i=1}^{n} \psi_\tau(Y_i, A_i, L_i; \tau) \\ \sum_{i=1}^{n} \psi_1(Y_i, A_i, L_i; \hat{\alpha}, \hat{\tau}, \lambda^1) \\ \sum_{i=1}^{n} \psi_0(Y_i, A_i, L_i; \hat{\alpha}, \hat{\tau}, \lambda^0) \end{bmatrix} = 0$$

where ψ_{α_j} is the derivative of the log-likelihood function for the weight model with respect to the j^{th} regression coefficient, for $j = 1, .., c_w$, with c_w equal to the number of columns in the design matrix for the weight model; ψ_{τ_k} is the derivative of the log-likelihood function for the parametric outcome model with respect to the k^{th} regression coefficient, for $k = 1, ..., p$. For the Poisson and NB distributions, $\tau = \beta$ and τ has dimension $p = c$, and for the ZIP and ZINB distributions, $\tau = [\gamma\ \xi]^T$ and has dimension $p = c_1 + c_2$. For example, for the Poisson distribution, $\psi_{\tau_j} = \{y_i - \exp(\sum_{j=1}^{c} x_{ij}\beta_j)\}x_{ij} = 0$, for $j = 1, ..., c$.

Define $\hat{\alpha}$ such that $E\{\psi_\alpha(A_i, L_i; \hat{\alpha})\} = 0$, and $\hat{\tau}$ such that $E\{\psi_\tau(Y_i, A_i, L_i; \hat{\tau})\} = 0$. Then $\hat{e}_i(L_i, \hat{\alpha})$ is participant i's estimated propensity score from the weight model, i.e, $\hat{e}_i(L_i, \hat{\alpha}) = \hat{P}(A_i = 1 \mid L_i = l_i)$. Participant i's estimated potential outcomes for $a \in \{0, 1\}$ based on the outcome model(s) are defined as $m_a(L_i, \hat{\tau}) = \hat{E}(Y_i \mid L_i = l_i, A_i = a)$. Then, $\sum_{i=1}^{n} \psi_1(Y_i, A_i, L_i; \hat{\alpha}, \hat{\tau}, \lambda^1) = \sum_{i=1}^{n} ([A_iY_i - \{A_i - \hat{e}_i(L_i, \hat{\alpha})\}m_1(L_i, \hat{\tau})]\{\hat{e}_i(L_i, \hat{\alpha})\}^{-1} - \lambda^1) = 0$ is the estimating equation for $\hat{\lambda}^1_{DR}$ and $\sum_{i=1}^{n} \psi_0(Y_i, A_i, L_i; \hat{\alpha}, \hat{\tau}, \lambda^0) = \sum_{i=1}^{n} [\{(1 - A_i)Y_i + \{A_i - \hat{e}_i(L_i, \hat{\alpha})\}m_0(L_i, \hat{\tau})\}\{1 - \hat{e}_i(L_i, \hat{\alpha})\}^{-1} - \lambda^0] = 0$ is the estimating equation for $\hat{\lambda}^0_{DR}$.

We show that when either the weight model or the outcome model(s) are correctly specified, $\psi_1(Y_i, A_i, L_i; \hat{\alpha}, \hat{\tau}, \lambda^1)$ and $\psi_0(Y_i, A_i, L_i; \hat{\alpha}, \hat{\tau}, \lambda^0)$ are unbiased. Let α_0 and τ_0 represent the true values of the parameters from the weight and outcome models, respectively. When the weight model is correctly specified, $E(\hat{\alpha}) = \alpha_0$, and when the outcome model is correctly specified $E(\hat{\tau}) = \tau_0$. By causal consistency and algebraic manipulation, $\psi_1(Y_i, A_i, L_i; \hat{\alpha}, \hat{\tau}, \lambda^1) = Y_i^1 +$

$\{A_i - \hat{e}_i(L_i, \hat{\alpha})\}\{Y_i^1 - m_1(L_i, \hat{\tau})\}\{\hat{e}_i(L_i, \hat{\alpha})\}^{-1} - \lambda^1$. Note that by conditional exchangeability:

$$E[\{A_i - \hat{e}_i(L_i, \hat{\alpha})\}\{Y_i^1 - m_1(L_i, \hat{\tau})\}\{\hat{e}_i(L_i, \hat{\alpha})\}^{-1}]$$

$$= E_L\left(\{\hat{e}_i(L_i, \hat{\alpha})\}^{-1} E_{A|L}\{A_i - \hat{e}_i(L_i, \hat{\alpha})\} E_{Y^1|L}\{Y_i^1 - m_1(L_i, \hat{\tau})\}\right)$$

When the weight model is correctly specified, $E_{A|L}\{A_i - \hat{e}_i(L_i, \hat{\alpha})\} = E_{A|L}\{A_i\} - e_i(L_i, \alpha_0) = 0$ and when the outcome model(s) are correctly specified $E_{Y^1|L}\{Y_i^1 - m_1(L_i, \hat{\tau})\} = E_{Y^1|L}\{Y_i^1\} - m_1(L_i, \tau_0) = 0$. Then, $E\{\psi_1(Y_i, A_i, L_i; \hat{\alpha}, \hat{\tau}, \lambda^1)\} = E(Y^1) - \lambda^1 = 0$. Thus, $\psi_1(Y_i, A_i, L_i; \hat{\alpha}, \hat{\tau}, \lambda^1)$ is unbiased when the weight or outcome model is correctly specified. Similarly, $\psi_0(Y_i, A_i, L_i; \hat{\alpha}, \hat{\tau}, \lambda^0)$ is unbiased when either model is correctly specified.

Then, under suitable regularity conditions (Stefanski and Boos, 2002),

$$\sqrt{n}\left(\begin{bmatrix}\hat{\alpha}\\ \hat{\tau}\\ \hat{\lambda}^1_{DR}\\ \hat{\lambda}^0_{DR}\end{bmatrix} - \begin{bmatrix}\alpha\\ \tau\\ \lambda^1\\ \lambda^0\end{bmatrix}\right) \xrightarrow{d} N\left(0, \Sigma_{DR}\right)$$

where

$$\Sigma_{DR} = A(\Lambda)^{-1} B(\Lambda)\{A(\Lambda)^{-1}\}^T$$

where $\Lambda^T = (\alpha^T, \tau^T, \lambda^1, \lambda^0)$, $A(\Lambda) = E\{-\dot{\psi}(Y_i, A_i, L_i, \Lambda)\}$, $B(\Lambda) = E\{\psi(Y_i, A_i, L_i, \Lambda)\psi(Y_i, A_i, L_i, \Lambda)^T\}$, and $\dot{\psi}(Y_i, A_i, L_i, \Lambda) = \partial\psi(Y_i, A_i, L_i, \Lambda)/\partial\Lambda^T$. The delta method is applied to obtain the asymptotic distribution of $\widehat{CRR}_{DR} = \hat{\lambda}^1_{DR}/\hat{\lambda}^0_{DR}$, with $g(\Lambda) = \lambda^1/\lambda^0$. Thus,

$$G_{DR} = \frac{\partial g(\Lambda)}{\partial(\Lambda)} = \begin{bmatrix} 0_{1\times c_w} & 0_{1\times p} & \frac{1}{\lambda^0} & \frac{-\lambda^1}{(\lambda^0)^2}\end{bmatrix}$$

Then,

$$\sqrt{n}\left(\frac{\hat{\lambda}^1_{DR}}{\hat{\lambda}^0_{DR}} - \frac{\lambda^1}{\lambda^0}\right) \xrightarrow{d} N\left(0, \Sigma^*_{DR}\right)$$

where

$$\Sigma^*_{DR} = G_{DR}\Sigma_{DR}G^T_{DR}$$

Because $\widehat{CRR}_{DR}$ is a delta method transformation of solutions to an unbiased set of estimating equations, $\widehat{CRR}_{DR}$ is a consistent and asymptotically normal estimator of CRR when the weight model or the outcome model(s) are correctly specified.

BIBLIOGRAPHY

Adimora, A. A., Ramirez, C., Benning, L., Greenblatt, R. M., Kempf, M.-C., Tien, P. C., Kassaye, S. G., Anastos, K., Cohen, M., Minkoff, H., et al. (2018). Cohort profile: the Women's Interagency HIV Study (WIHS). *International Journal of Epidemiology*, 47(2):393–394i.

Agresti, A. (2002). *Categorical Data Analysis, 2nd Edition*. John Wiley & Sons, Inc.

Albert, J. M., Wang, W., and Nelson, S. (2014). Estimating overall exposure effects for zero-inflated regression models with application to dental caries. *Statistical Methods in Medical Research*, 23(3):257–278.

Austin, P. C. (2009). Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in Medicine*, 28(25):3083–3107.

Austin, P. C. and Stuart, E. A. (2015). Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in Medicine*, 34(28):3661–3679.

Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973.

Bangsberg, D. R., Hecht, F. M., Charlebois, E. D., Chesney, M., and Moss, A. (2001). Comparing objective measures of adherence to HIV antiretroviral therapy: electronic medication monitors and unannounced pill counts. *AIDS and Behavior*, 5(3):275–281.

Bodnar, L. M., Davidian, M., Siega-Riz, A. M., and Tsiatis, A. A. (2004). Marginal structural models for analyzing causal effects of time-dependent treatments: an application in perinatal epidemiology. *American Journal of Epidemiology*, 159(10):926–934.

Bohensky, M. A., Jolley, D., Sundararajan, V., Evans, S., Pilcher, D. V., Scott, I., and Brand, C. A. (2010). Data linkage: a powerful research tool with potential problems. *BMC Health Services Research*, 10(1):1–7.

Böhning, D., Dietz, E., Schlattmann, P., Mendonca, L., and Kirchner, U. (1999). The zero-inflated poisson model and the decayed, missing and filled teeth index in dental epidemiology. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 162(2):195–209.

Bolfarine, H. and Zacks, S. (1992). *Prediction Theory for Finite Populations*. Springer Science & Business Media.

Brumback, B. A., Hernán, M. A., Haneuse, S. J., and Robins, J. M. (2004). Sensitivity analyses for unmeasured confounding assuming a marginal structural model for repeated measures. *Statistics in Medicine*, 23(5):749–767.

Casella, G. and Berger, R. L. (2002). *Statistical inference*, volume 2. Duxbury Pacific Grove, CA.

Cassel, C. M., Särndal, C. E., and Wretman, J. H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, 63(3):615–620.

Center for Behavioral Health Statistics and Quality (2019). 2017 National Survey on Drug Use and Health Methodological Resource Book, Section 11: Person-Level Sampling Weight Calibration. Technical report, Substance Abuse and Mental Health Services Administration.

Chow, S.-C., Shao, J., Wang, H., and Lokhnygina, Y. (2017). *Sample Size Calculations in Clinical Research*. Chapman and Hall/CRC.

Cole, S. R. and Frangakis, C. E. (2009). The consistency statement in causal inference: a definition or an assumption? *Epidemiology*, 20(1):3–5.

Cole, S. R. and Hernán, M. Á. (2008). Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiology*, 168(6):656–664.

Cummings, T. H., Hardin, J. W., McLain, A. C., Hussey, J. R., Bennett, K. J., and Wingood, G. M. (2015). Modeling heaped count data. *The Stata Journal*, 15(2):457–479.

Daniel, J. (2012). *Sampling Essentials: Practical Guidelines for Making Sampling Choices*. Sage Publications.

Daniel, R. M., Cousens, S., De Stavola, B., Kenward, M. G., and Sterne, J. (2013). Methods for dealing with time-dependent confounding. *Statistics in Medicine*, 32(9):1584–1618.

Deville, J.-C. and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87(418):376–382.

Folsom, R. E. and Singh, A. C. (2000). The generalized exponential model for sampling weight calibration for extreme values, nonresponse, and poststratification. In *Proceedings of the American Statistical Association, Survey Research Methods Section*.

Ford, J. B., Roberts, C. L., and Taylor, L. K. (2006). Characteristics of unmatched maternal and baby records in linked birth records and hospital discharge data. *Paediatric and Perinatal Epidemiology*, 20(4):329–337.

Funk, M. J., Westreich, D., Wiesen, C., Stürmer, T., Brookhart, M. A., and Davidian, M. (2011). Doubly robust estimation of causal effects. *American Journal of Epidemiology*, 173(7):761–767.

Gabler, S., Häder, S., and Lahiri, P. (1999). A model based justification of Kish's formula for design effects for weighting and clustering. *Survey Methodology*, 25:105–106.

Garcia-Aymerich, J., Varraso, R., Danaei, G., Camargo, C. A. J., and Hernán, M. A. (2013). Incidence of adult-onset asthma after hypothetical interventions on body mass index and physical activity: An application of the parametric g-formula. *American Journal of Epidemiology*, 179(1):20–26.

Gibbard, A. and William, L. (1981). Counterfactuals and two kinds of expected utility. *Foundations and Applications of Decision Theory*, 1.

Hansen, M. H. (1987). Some history and reminiscences on survey sampling. *Statistical Science*, 2(2):180–190.

Harron, K., Dibben, C., Boyd, J., Hjern, A., Azimaee, M., Barreto, M. L., and Goldstein, H. (2017). Challenges in administrative data linkage for research. *Big Data & Society*, 4(2):1–12.

Harron, K., Wade, A., Gilbert, R., Muller-Pebody, B., and Goldstein, H. (2014). Evaluating bias due to data linkage error in electronic healthcare records. *BMC Medical Research Methodology*, 14(1):1–10.

Heitjan, D. F. (1989). Inference from grouped continuous data: a review. *Statistical Science*, pages 164–179.

Heitjan, D. F. and Rubin, D. B. (1990). Inference from coarse data via multiple imputation with application to age heaping. *Journal of the American Statistical Association*, 85(410):304–314.

Hernán, M. Á., Brumback, B., and Robins, J. M. (2000). Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology*, pages 561–570.

Hernán, M. Á. and Robins, J. M. (2020). *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960.

Judson, D. H., Parker, J. D., and Larsen, M. D. (2013). Adjusting sample weights for linkage-eligibility using SUDAAN. Technical report, National Center for Health Statistics.

Kalton, G., Brick, J., and Lê, T. (2005). Household sample surveys in developing and transition countries. Technical Report 96, F, United Nations: Statistics Division, Department of Economic and Social Affairs.

Kang, J. D. and Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 22(4):523–539.

Kiaer, A. (1897). The representative method of statistical surveys (1976 English translation of the original Norwegian). *Central Bureau of Statistics of Norway*.

Kish, L. (1965). *Survey Sampling*. John Wiley & Sons.

Kish, L. (1992). Weighting for unequal pi. *Journal of Official Statistics*, 8(2):183–200.

Klein, J. P. and Moeschberger, M. L. (2006). *Survival analysis: Techniques for Censored and Truncated data*. Springer Science & Business Media.

Klesges, R. C., Debon, M., and Ray, J. W. (1995). Are self-reports of smoking rate biased? Evidence from the Second National Health and Nutrition Examination Survey. *Journal of Clinical Epidemiology*, 48(10):1225–1233.

Knittel, A. K., Shook-Sa, B. E., Rudolph, J., Edmonds, A., Ramirez, C., Cohen, M., Adedimeji, A., Taylor, T., Michel, K. G., Milam, J., Cohen, J., Donohue, J., Foster, A., Fischl, M., Konkle-Parker, D., and Adimora, A. A. (2020). Incarceration and number of sexual partners after incarceration among vulnerable US women, 2007–2017. *American Journal of Public Health*, 110(S1):S100–S108.

Kong, A. (1992). A note on importance sampling using standardized weights. *University of Chicago, Dept. of Statistics, Tech. Rep*, 348:1–4.

Kong, A., Liu, J. S., and Wong, W. H. (1994). Sequential imputations and Bayesian missing data problems. *Journal of the American Statistical Association*, 89(425):278–288.

Kott, P. S. (2005). Randomization-assisted model-based survey sampling. *Journal of Statistical Planning and Inference*, 129(1-2):263–277.

Kott, P. S. (2006). Using calibration weighting to adjust for nonresponse and coverage errors. *Survey Methodology*, 32(2):133–142.

Kott, P. S. and Day, C. D. (2014). Developing calibration weights and standard-error estimates for a survey of drug-related emergency-department visits. *Journal of Official Statistics*, 30(3):521–532.

Kott, P. S. and Liao, D. (2012). Providing double protection for unit nonresponse with a nonlinear calibration-weighting routine. *Survey Research Methods*, 6(2):105–111.

Kott, P. S. and Liao, D. (2015). One step or two? Calibration weighting from a complete list frame with nonresponse. *Survey Methodology*, 41(1):165–181.

Lambert, D. (1992). Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1):1–14.

Lavrakas, P. J. (2008). *Encyclopedia of survey research methods*. Sage Publications.

Lawless, J. F. (1987). Negative binomial and mixed poisson regression. *Canadian Journal of Statistics*, 15(3):209–225.

Lee, B. K., Lessler, J., and Stuart, E. A. (2011). Weight trimming and propensity score weighting. *PLOS ONE*, 6(3):1–6.

Little, R., Lewitzky, S., Heeringa, S., Lepkowski, J., and Kessler, R. (1997). Assessment of weighting methodology for the national comorbidity survey. *American Journal of Epidemiology*, 146(5):439–449.

Little, R. J. (2004). To model or not to model? competing modes of inference for finite population sampling. *Journal of the American Statistical Association*, 99(466):546–556.

Little, R. J. and Rubin, D. B. (2019). *Statistical Analysis with Missing Data*. John Wiley & Sons.

Little, R. J. and Vartivarian, S. (2005). Does weighting for nonresponse increase the variance of survey means? *Survey Methodology*, 31(2):161–168.

Lohr, S. L. (2010). *Sampling: Design and Analysis*. Chapman and Hall/CRC.

Long, D. L., Preisser, J. S., Herring, A. H., and Golin, C. E. (2014). A marginalized zero-inflated poisson regression model with overall exposure effects. *Statistics in Medicine*, 33(29):5151–5165.

Lunceford, J. K. and Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine*, 23(19):2937–2960.

McCaffrey, D. F., Griffin, B. A., Almirall, D., Slaughter, M. E., Ramchand, R., and Burgette, L. F. (2013). A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Statistics in Medicine*, 32(19):3388–3414.

McCaffrey, D. F., Ridgeway, G., and Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods*, 9(4):403–425.

McCullagh, P. and Nelder., J. (1989). *Generalized Linear Models*. Chapman Hall.

Myers, J. A., Rassen, J. A., Gagne, J. J., Huybrechts, K. F., Schneeweiss, S., Rothman, K. J., Joffe, M. M., and Glynn, R. J. (2011). Effects of adjusting for instrumental variables on bias and precision of effect estimates. *American Journal of Epidemiology*, 174(11):1213–1222.

Neyman, J. (1923). On the application of probability theory to agricultural experiments: essay on principles, Section 9. *Translated in Statistical Science 1990*, 5:465–480.

North Carolina HIV/STD/Hepatitis Surveillance Unit (2017). 2016 North Carolina HIV/STD/Hepatitis surveillance report. Report, North Carolina Department of Health and Human Services, Division of Public Health, Communicable Disease Branch.

Pearl, J. (2010). On the consistency rule in causal inference: axiom, definition, assumption, or theorem? *Epidemiology*, 21(6):872–875.

Preisser, J. S., Das, K., Long, D. L., and Divaris, K. (2016). Marginalized zero-inflated negative binomial regression with application to dental caries. *Statistics in Medicine*, 35(10):1722–1735.

Qayad, M. G. and Zhang, H. (2009). Accuracy of public health data linkages. *Maternal and Child Health Journal*, 13(4):531–538.

Roberts, J. M. and Brewer, D. D. (2001). Measures and tests of heaping in discrete quantitative distributions. *Journal of Applied Statistics*, 28(7):887–896.

Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9-12):1393–1512.

Robins, J. (1998). Marginal structural models. In *1997 proceedings of the American Statistical Association, section on Bayesian statistical science*, pages 1–10.

Robins, J. M., Hernán, M. Á., and Brumback, B. (2000). Marginal structural models and causal inference in Epidemiology. *Epidemiology*, 11(5):550–560.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Rothman, K. J. (2012). *Epidemiology: an Introduction*. Oxford University Press.

Royall, R. M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, 57(2):377–387.

Royall, R. M. (1976). The linear least-squares prediction approach to two-stage sampling. *Journal of the American Statistical Association*, 71(355):657–664.

Royall, R. M. and Cumberland, W. G. (1978). Variance estimation in finite population sampling. *Journal of the American Statistical Association*, 73(362):351–358.

RTI International (2012). SUDAAN language manual, release 11.0. Manual, RTI International.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701.

Rubin, D. B. (1977). Assignment to treatment group on the basis of a covariate. *Journal of Educational Statistics*, 2(1):1–26.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, pages 34–58.

Rubin, D. B. (1980). Randomization analysis of experimental data: The Fisher randomization test comment. *Journal of the American Statistical Association*, 75(371):591–593.

Rubin, D. B. (1997). Estimating causal effects from large data sets using propensity scores. *Annals of internal medicine*, 127(8_Part_2):757–763.

Salganik, M. J. (2006). Variance estimation, design effects, and sample size calculations for respondent-driven sampling. *Journal of Urban Health*, 83(1):98.

Särndal, C. (2010). Models in survey sampling. In Carlson, N. and Villani, editors, *Official Statistics - Methodology and Applications in Honour of Daniel Thorburn*, pages 15–28. Official Statistics.

Särndal, C.-E., Swensson, B., and Wretman, J. (2003). *Model assisted survey sampling*. Springer Science & Business Media.

Särndal, C.-E., Thomsen, I., Hoem, J. M., Lindley, D., Barndorff-Nielsen, O., and Dalenius, T. (1978). Design-based and model-based inference in survey sampling [with discussion and reply]. *Scandinavian Journal of Statistics*, pages 27–52.

Sato, T. and Matsuyama, Y. (2003). Marginal structural models as a tool for standardization. *Epidemiology*, 14(6):680–686.

Saul, B. and Hudgens, M. (2020). The calculus of M-estimation in R with geex. *Journal of Statistical Software, Articles*, 92(2):1–15.

Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94(448):1096–1120.

Shook-Sa, B. E., Kott, P., Berzofsky, M., Couzens, G. L., Moore, A., Lee, P., Langton, L., and Planty, M. (2017). Maintaining precision in survey estimates while adjusting for conditional bias at the subnational level through calibration weighting. *Survey Research Methods*, 11:405–414.

Singh, A. and Folsom, R. (2000). Bias corrected estimating function approach for variance estimation adjusted for poststratification. *Proceedings of Survey Research and Methodology Section, American Statistical Association*.

Singh, K., Suchindran, C., and Singh, R. (1994). Smoothed breastfeeding durations and waiting time to conception. *Social Biology*, 41(3-4):229–239.

St. Sauver, J. L., Grossardt, B. R., Yawn, B. P., Melton III, L. J., and Rocca, W. A. (2011). Use of a medical records linkage system to enumerate a dynamic population over time: the Rochester epidemiology project. *American Journal of Epidemiology*, 173(9):1059–1068.

Stefanski, L. A. and Boos, D. D. (2002). The calculus of M-estimation. *The American Statistician*, 56(1):29–38.

Taubman, S. L., Robins, J. M., Mittleman, M. A., and Hernán, M. A. (2009). Intervening on risk factors for coronary heart disease: an application of the parametric g-formula. *International Journal of Epidemiology*, 38(6):1599–1611.

United Nations Statistical Division (2008). *Designing Household Survey Samples: Practical Guidelines*, volume 98. United Nations Publications.

U.S. Census Bureau Population Division (2018). Annual Estimates of the Resident Population: April 1, 2010 to July 1, 2017. Technical report, U.S. Census Bureau.

Valliant, R., Dever, J. A., and Kreuter, F. (2013). *Practical Tools for Designing and Weighting Survey Samples*. Springer.

Vansteelandt, S., Bekaert, M., and Claeskens, G. (2012). On model selection and model misspecification in causal inference. *Statistical Methods in Medical Research*, 21(1):7–30.

Vera Institute of Justice (2014). Incarceration trends dataset. Available at: https://github.com/vera-institute/incarceration_trends.

Verma, V., Scott, C., and O'Muircheartaigh, C. (1980). Sample designs and sampling errors for the world fertility survey. *Journal of the Royal Statistical Society: Series A (General)*, 143(4):431–463.

Waernbaum, I. (2012). Model misspecification and robustness in causal inference: comparing matching with doubly robust estimation. *Statistics in Medicine*, 31(15):1572–1581.

Wang, H. and Heitjan, D. F. (2008). Modeling heaping in self-reported cigarette counts. *Statistics in Medicine*, 27(19):3789–3804.

Weisberg, S. (2005). *Applied Linear Regression*, volume 528. John Wiley & Sons.

Wiederman, M. W. (1997). The truth must be in here somewhere: Examining the gender discrepancy in self-reported lifetime number of sex partners. *Journal of Sex Research*, 34(4):375–386.

Wohl, D. A., Golin, C., Rosen, D. L., May, J. M., and White, B. L. (2013). Detection of undiagnosed HIV among state prison entrants. *JAMA*, 310(20):2198–2199.

Wright, D. E. and Bray, I. (2003). A mixture model for rounded data. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 52(1):3–13.

Young, J. G., Cain, L. E., Robins, J. M., O'Reilly, E. J., and Hernán, M. A. (2011). Comparative effectiveness of dynamic treatment regimes: an application of the parametric g-formula. *Statistics in Biosciences*, 3(1):119.

Zhang, Z. and Sun, J. (2010). Interval censoring. *Statistical Methods in Medical Research*, 19(1):53–70.