



Breaking the MAR Paradigm: Estimation, Bounding, and Sensitivity When Data Are Missing Not at Random

Citation

Ocampo, Alex. 2020. Breaking the MAR Paradigm: Estimation, Bounding, and Sensitivity When Data Are Missing Not at Random. Doctoral dissertation, Harvard University, Graduate School of Arts & Sciences.

Link

<https://nrs.harvard.edu/URN-3:HUL.INSTREPOS:37365839>

Terms of use

This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material (LAA), as set forth at

<https://harvardwiki.atlassian.net/wiki/external/NGY5NDE4ZjgzNTc5NDQzMGIzZWZhMGFIOWI2M2EwYTg>

Accessibility

<https://accessibility.huit.harvard.edu/digital-accessibility-policy>

Share Your Story

The Harvard community has made this article openly available.
Please share how this access benefits you. [Submit a story](#)

Breaking the MAR Paradigm: Estimation, Bounding, and Sensitivity When Data are Missing Not at Random

a dissertation presented
by
Alex Ocampo
to
The Department of Biostatistics
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of
Biostatistics

Harvard University
Cambridge, Massachusetts
May 2020

©2020 – Alex Ocampo
All rights reserved.

Breaking the MAR Paradigm: Estimation, Bounding, and Sensitivity When Data are Missing Not at Random

Abstract

Statistical methods for unobserved, or missing data often rely on an assumption defined over 40 years ago; namely, that the data are missing at random (MAR). Simply put, MAR is when the probability of a missing value is based on other observed values, not the missing value itself. Armed with this assumption, statisticians can use methods that leverage the available data such as multiple imputation, likelihood approaches, and inverse probability weighting to obtain consistent and efficient estimates. If the MAR assumption is unrealistic, then only a few tools are available. The consequences of departures from MAR can be quantified by conducting a sensitivity analysis, however, there is no consensus on how to best carry out such an analysis, and many current approaches are tedious, technical, and time-consuming.

This dissertation aims to provide a path forward in certain situations when the MAR assumption does not hold. Chapter 1 provides conditions for identifying treatment effects when a continuous outcome is missing not at random. Identification is possible by reframing the estimand of a trimmed means estimator when the missing outcome comes from the "poor" tail of its treatment distribution. Chapter 2 proposes an efficient approach to estimate upper and lower bounds of parameters that account for the uncertainty in the missing data. The bounding approach utilizes the influence function of the statistical functional at hand to identify the best and worst possible imputations of an incomplete data set. Chapter 3 argues that a ratio type estimator may make the MAR assumption more plausible. The approach is used to estimate vaccination coverage rates in the 77 communes of Benin where only some communes have survey data from random samples but all have administrative data available. The chapter additionally demonstrates how departures from MAR can be explored intuitively in the Bayesian framework through the introduction of sensitivity parameters.

Contents

1	Identifying Treatment Effects using Trimmed Means when Data are Missing Not at Random	1
1.1	Introduction	2
1.2	Statistical Methods	4
1.2.1	The Trimmed Means Estimator	4
1.2.2	Combining Trimmed Means with Multiple Imputation	6
1.3	Properties of the Estimand	7
1.3.1	Equivalence to the Population Treatment Effect	7
1.3.2	Intercurrent Events	9
1.4	Numerical Studies	11
1.4.1	Simulation Objectives	11
1.4.2	Simulation Design	11
1.4.3	Simulation Results	13
1.5	Application to a Clinical Trial	22
1.6	Discussion	25
2	Using Influence Functions to Bound Regression Parameters for Incomplete Data	27
2.1	Introduction	28
2.2	Methods for Estimating Bounds	29
2.2.1	Notation	29
2.2.2	Positive Jackknife Approach	30
2.2.3	Influence Function Approach	31
2.3	Theoretical Properties	34
2.3.1	Asymptotic Equivalence of Positive Jackknife to Influence Function	34
2.3.2	Asymptotic Variance of the Bounds	35
2.4	Simulation Studies	36
2.4.1	Estimation of Upper and Lower Bounds	36
2.4.2	Computational Efficiency	38
2.5	Data Application	39
2.6	Conclusion	42
3	Extrapolating Hybrid Estimation to Districts with Missing Probability Survey Data	44
3.1	Introduction	45
3.2	Data Overview	47
3.2.1	The Probability Surveys	47
3.2.2	Administrative Data	48
3.3	Methodology	49
3.3.1	Estimating Denominators	49
3.3.2	Predicting Denominators	51
3.3.3	Bayesian approach	52
3.4	Application to Benin Data	54
3.4.1	Initial results	55
3.4.2	Bayesian Approach	55
3.4.3	Bayesian Sensitivity Analysis	56
3.5	Discussion	58

Appendix A	Proofs for Chapter 1	61
Appendix B	Supplementary Results for Chapter 3	64
References		71

List of figures

1.1	<i>m</i> -graphs for Scenarios a, b ,c, and d of the Numerical Study	12
2.1	Computation Times	38
2.2	Effect of Missing Data Rate on Bounds Interval	42
3.1	Communes with Probability Survey Data	48
3.2	Model Covariates vs. Estimated Denominator	54
3.3	Children’s Polio Vaccination Coverage Estimates (12-59 Months)	55
3.4	Children’s Polio Vaccination Coverage Bayesian Estimates (12-59 Months)	56
3.5	Sensitivity Analysis of Δ_3	57
3.6	Bayesian Sensitivity Analysis of Δ_3	58
B.1	Children’s Vitamin A Coverage Estimates (12-59 Months)	64
B.2	Children’s Polio Vaccination Coverage Estimates (6-11 Months)	65
B.3	Children’s Vitamin A Coverage Estimates (6-11 Months)	65
B.4	Children’s Vitamin A Coverage Bayesian Estimates (12-59 Months)	66
B.5	Children’s Polio Vaccination Coverage Bayesian Estimates (6-11 Months)	66
B.6	Children’s Vitamin A Coverage Bayesian Estimates (6-11 Months)	67

List of tables

1.1	Trimmed Means with MCAR	13
1.2	Multiple Imputation with MCAR	14
1.3	Trimmed Means with MAR	15
1.4	Multiple Imputation with MAR	16
1.5	Trimmed Means with MNAR	17
1.6	Multiple Imputation with MNAR	17
1.7	Trimmed Means with a Mixture of Missing Data Types	18
1.8	Multiple Imputation with a Mixture of Missing Data Types	19
1.9	Trimmed Means + MI with a Mixture of Missing Data Types	20
1.10	Comparison of Adaptive and Fixed α	21
1.11	Treatment Discontinuation Reasons	22
1.12	Clinical Trial Analysis Results	24
2.1	MCAR Simulation Results	37
2.2	MNAR Simulation Results	38
2.3	Variable Descriptions	39
2.4	Cox Model Hazard Ratio Bounds	41

Acknowledgments

I owe my deepest gratitude to my dissertation advisor, Marcello Pagano, for all his invaluable support and guidance throughout. I am extremely thankful for your support and freedom to pursue this challenging research topic. In addition, your keen attention to the details of the methodology as well as the articulation of the research greatly improved the work herein. Most of all, I'm grateful for your personal support and advice during difficult and challenging times during my PhD studies. I am not only a better statistician, but a better person because of your mentorship.

I would also like to sincerely thank my dissertation committee members, Bethany Hedt-Gauthier and Joseph Valadez. Bethany, your insights during committee meetings led to major improvements in all my research projects: the simulation in chapter 1, the data application in chapter 2, and the sensitivity analysis in chapter 3. Joe, I'm especially grateful for the stimulating project for my third paper and your guidance from across the pond.

Without going into specific names, as the list would be too long, I would also like to take this opportunity to thank all the students, staff, and faculty who I had the pleasure of working with in the Department of Biostatistics at Harvard. The environment in the department is incredibly enriching and challenged me to achieve new heights in statistics.

Lastly I would like to thank my family - Mom, Dad, and Kristen - for all the support throughout the years. You were my biggest champions, as well as my emotional and spiritual support during my graduate studies. I'm so lucky to have all of you in my life.

This dissertation is dedicated to my grandparents, Guadelupe and Eliseo Ocampo, both of whom passed away while I was pursuing my PhD. I know how much you wanted to see me finish, and I know you were here with me in spirit as I crossed the finish line.

Identifying Treatment Effects using Trimmed Means when Data are Missing Not at Random

Alex Ocampo¹, Heinz Schmidli², Peter Quarg², Francesca Callegari², Marcello Pagano¹

1 Department of Biostatistics, Harvard University

2 Novartis Pharma AG, Basel, Switzerland

Abstract

Patients often discontinue treatment in a clinical trial because their health condition is not improving. Consequently, the patients still in the study at the end of the trial have better health outcomes on average than the initial patient population would have had if every patient had completed the trial. If we only analyze the patients who complete the trial, then this missing data problem biases the estimator of a medication's efficacy because study outcomes are missing not at random (MNAR). One way to overcome this problem - the trimmed means approach for missing data - sets missing values as the worst observed outcome and then trims away a fraction of the distribution from each treatment arm before calculating differences in treatment efficacy (Permutt 2017, Pharmaceutical statistics 16.1: 20-28). In this paper we derive sufficient and necessary conditions for when this approach can identify the average population treatment effect. Numerical studies show the trimmed means approach's ability to effectively estimate treatment efficacy when data are MNAR and missingness is strongly associated with an unfavorable outcome, but trimmed means fail when data are missing at random (MAR). If the reasons for discontinuation in a clinical trial are known, analysts can improve estimates with a combination of multiple imputation (MI) and the trimmed means approach when the assumptions of each hold. We compare the methodology to existing approaches using data from a clinical trial for chronic pain. When the assumptions are justifiable, using trimmed means can help identify treatment effects notwithstanding MNAR data.

1.1 Introduction

Restricting statistical analysis to patients who complete a clinical trial can lead to biased results (Little and Rubin, 2019). Patients with unfavorable endpoints often discontinue the trial prematurely and thus the remaining patients no longer provide a truly representative sample of patients, even if the original sample did. This situation can be viewed as a missing data problem – the endpoints are not measured for the patients who have left the study and are thus labeled as missing. There is no consensus on how best to adjust a statistical analysis for missing data. This is in part because the reason for missing data impacts the choice of analytical methods to use. The best understood situation is when the data are missing at random (MAR) (Rubin, 1976). This means that one observes all the data necessary to explain the missingness in the data - i.e. the missing values themselves did not contribute to the fact that they are missing. This is not only an untestable assumption, but in many settings it is an unrealistic one. Its complementary situation, when data are missing not at random (MNAR), occurs in clinical trials. An example of this is trials of chronic pain, where subjects are more likely to leave the study if they experience little or no decrease in pain. Consequently, the outcomes of those who complete the study differs from those who do not, often even when accounting for observed information. In these situations, ignoring the violations of the assumptions underpinning popular missing data methods designed for MAR data, such as multiple imputation (Rubin, 2004) and inverse probability weighting (Robins et al., 1995), leads to biases that make the analyses inadequate and of little scientific value.

The paucity of data analytic methods for when the data are MNAR may contribute to the widespread use of inappropriate methods. The National Research Council (NRC) report on missing data suggests two general paths forward when data are MNAR: Selection Models and Pattern-Mixture Models (National Research Council, 2010). Both of these models are limited in that they rely on parameters that cannot be inferred from the observed data. Therefore, identification of treatment effects using these models is not possible. This limitation is emphasized by Little (2009) who highlights that all MNAR models are subject to a fundamental lack of identification. A shadow variable approach can identify causal effects under MNAR (Miao and Tchetgen Tchetgen, 2016); however, it relies on the presence of a surrogate outcome - a shadow variable - that is closely related to the missing outcome and unrelated to why the missing outcome is unobserved. One may not observe such a variable in the clinical trial setting. Jump to reference imputation (Carpenter et al., 2013)

is a popular approach to handling MNAR data where the missing outcomes of patients in the experimental treatment arm are imputed using observed outcomes from the reference arm. The logic of this approach makes intuitive sense; in the experimental arm MNAR dropouts have worse outcomes than those who complete the study. So to reflect this difference one can leverage the worse outcomes of reference/reference arm for imputations of experimental dropouts. The main drawback for the MNAR setting is that missing outcomes in the reference arm are imputed by reference completers which effectively assumes MAR. If reference arm dropouts are MNAR, then reference imputations may be optimistic and the treatment effect can be underestimated. Especially in placebo-controlled trials, one generally expects more lack of efficacy dropouts in the reference arm which could be problematic for jump to reference unless initial measurements leveraged for imputations reveal the poor outcome trajectory.

The “trimmed means approach” for missing data was first introduced by [Permutt and Li \(2017\)](#). The method is designed for settings where missing values can be assumed to be poor health outcomes that are the continuous endpoints of the study; precisely the example introduced above. We focus on this method in the remainder of the paper. The trimmed means approach is simple to implement in a comparative trial. The final statistic is the arithmetic mean after trimming the poor performing parts of the distribution in each arm of the study, after assigning a value to the missing outcomes that is worse than the worst observed outcome. Applying this approach only requires the ability to assign a rank to all outcomes and ranks missing outcomes at the tail end of the distribution prior to trimming. The trimmed means approach can be extended to include covariates such as in an ANCOVA and can be applied in mixed models for repeated measures (MMRM) ([Mallinckrodt et al., 2001](#)). The trimmed-means-approach was designed to estimate an estimand different than most standard analyses. Instead of estimating treatment difference in the whole study population, the approach estimates the treatment difference in a subset of the best performing patients since those performing poorly are trimmed out of the analysis. As a result, some of the data is lost, and that is the price one pays for obtaining an analysis that accommodates MNAR data. The trimmed-means-approach was evaluated under various missing data generating mechanisms ([Wang et al., 2018](#)), which reveals some of its limitations.

The rest of the paper concentrates on extending the utility of the trimmed-means-approach for missing data. Section 1.2 describes the trimmed means approach and extends its use to a combination with multiple imputation. Section 1.3 discusses the estimand and provides a proof for settings under which the trimmed means approach can identify the population treatment effect. Section 1.4 evaluates the finite sample prop-

erties of the approach in a numerical study under various missing data mechanisms. Section 1.5 provides recommendations on how to apply the approach in the context of a randomized clinical trial for a chronic pain medication. Section 1.6 concludes the article with a discussion.

1.2 Statistical Methods

1.2.1 The Trimmed Means Estimator

In 2017, Permutt and Li introduced the trimmed means estimator. They provide a thorough motivation of the approach as well as how to implement it in practice. [Mehrotra et al. \(2017\)](#) demonstrate an implementation of the approach in SAS. This approach does not rely on any parametric model or imputation of the missing values; it only depends on the ability to rank outcomes. The trimmed means approach for missing data utilizes one sided trimmed means. Consider a continuous outcome Y_i for observation i . The one-sided trimmed mean is an L-estimator defined as:

$$\hat{\mu}_T = \frac{1}{n_T} \sum_{i: Y_i > \hat{F}^{-1}(\alpha)} Y_i$$

The above is simply the average of the observations that fall above the quantile $\hat{F}^{-1}(\alpha)$ of the empirical distribution function of Y_i where α represents the proportion of trimmed outcomes. For example, $\alpha = 0.3$ represents trimming away the bottom 30% of the distribution. Here, $n_T = \sum_{i=1}^n \mathbf{1}(Y_i > \hat{F}^{-1}(\alpha))$ is the sample size after trimming. The expectation of the trimmed mean is the population trimmed mean:

$$\mu_T = E[Y|Y > F^{-1}(\alpha)] = \frac{1}{1 - \alpha} \int_{F^{-1}(\alpha)}^{\infty} y f(y) dy$$

Note that for this example above, poor outcomes correspond to low values of Y , so the lower part of the distribution is trimmed. This could be switched based on the clinical context. For example, in pain trials a decrease in pain intensity is a good outcome, so one would rather trim away the upper part of the distribution.

To implement this approach, first consider three observed variables: A a binary indicator for treatment, R a binary missing data indicator, and the clinical outcome Y which is a continuous value when $R = 1$ and is missing when $R = 0$. Operationally, the first step of the trimmed means approach is to remove the missing

data by ranking all missing outcomes as slightly worse than the poorest observed outcome in the trial. To do so one defines a new outcome denoted by U for each subject i :

$$U_i = \begin{cases} Y_i & \text{if } R_i = 1 \\ \min(\mathbf{Y}) - \epsilon & \text{if } R_i = 0 \end{cases} \quad \text{for } \epsilon > 0$$

Note the above corresponds to low values representing poor outcomes. If high values of the outcome reflected poor values then if $R = 0$ the missing outcomes would be set to $\max(\mathbf{Y}) + \epsilon$. After ranking, α proportion of each treatment arm is trimmed away from the end of the distribution of U associated with poor outcomes. The analyst has some flexibility in determining α , the proportion of data trimmed from each distribution, before calculating the treatment effect. This value can be fixed a priori with a value of α chosen that anticipates the amount of missing data. Alternatively, trimming can be adaptive and chosen to be the maximum between the proportions of missing data in each arm. Thus the values that the missing observations were set to are never actually used in the analysis, but do serve the important function of informing the quantiles at which each distribution is to be trimmed. After trimming, then calculate the mean of the remaining $100 \times (1 - \alpha)\%$ observations in each arm $\hat{\mu}_{T1}$ and $\hat{\mu}_{T0}$. The final estimate is obtained by taking the difference of these trimmed means between each arm: $\hat{\mu}_{T\Delta} = \hat{\mu}_{T1} - \hat{\mu}_{T0}$.

Inference can be conducted via a permutation test that conditions on the observed data and randomly permutes the treatment assignments. The resulting permutation distribution of treatment differences formed under the null hypothesis can be used to determine significant differences and confidence intervals. To reject the null hypothesis of no treatment difference, the point estimate should fall above the upper 2.5th percentile of the permutation distribution. If $\hat{\mu}_{T\Delta}$ is the treatment difference calculated after trimming, a 95% confidence interval can be constructed by adding the 97.5th and 2.5th percentiles of the permutation distribution to $\mu_{T\Delta}$. This can be generalized for any significance level γ such that $(\hat{\mu}_{T\Delta} - y_{\gamma/2}, \hat{\mu}_{T\Delta} + y_{1-\gamma/2})$ yields a $(1 - \gamma)\%$ confidence interval where y_γ is the γ percentile of the permutation distribution. Since these confidence intervals are constructed using the permutation distribution generated under the null hypothesis, the intervals will be conservative when the null hypothesis is false (Good, 2006).

1.2.2 Combining Trimmed Means with Multiple Imputation

In well-conducted clinical trials, the reason for study discontinuation is collected for each patient who drops out of the study. Treating certain types of dropout as poor outcomes and ranking them at the low end of the distribution, as the trimmed means approach does, would lead to biases. Knowing the reason for dropping out of a study, and using that information, should lead to more precise analyses. To that end, consider the expanded indicator:

$$R = \begin{cases} r_1 & \text{if } Y \text{ observed} \\ r_2 & \text{if MAR or MCAR} \\ r_3 & \text{if MNAR} \end{cases}$$

Assume the complete data of Y is partitioned into the observed and missing components as follows $Y = [Y_{r_1}, Y_{r_2}, Y_{r_3}]$ which denote the observed, missing at random, and missing not at random components of Y respectively. Here we propose imputing Y_{r_2} and trimming Y_{r_3} . We can perform multiple imputation of Y_{r_2} when the conditional distribution $f(Y_{r_2}|Y_{r_1} = y_{r_1}, A, \mathbf{X})$ is a valid imputation model given the MAR assumption. Here y_{r_1} denotes the observed outcomes, A is the treatment assignment, and $\mathbf{X}$ is a matrix of auxiliary covariates that may or may not be available and of use for imputations. Using this conditional distribution, one can draw m samples for the MAR and MCAR data $Y_{r_2}^{(1)}, Y_{r_2}^{(2)}, \dots, Y_{r_2}^{(m)}$ to derive a set of data that is now complete for Y_{r_2} where missing values Y_{r_3} remain. Let $\hat{\mu}_{T\Delta} = \hat{\mu}_{T\Delta}(Y_{r_1}, Y_{r_2}, \mathbf{1}(R = r_3))$ denote the trimmed means statistic given that complete data on Y_{r_2} were available. Note we do not need to observe Y_{r_3} since the trimmed means approach will trim these observations out of the analysis. Multiple imputation relies on the asymptotically normal distribution of $\hat{\mu}_{T\Delta}$, which applies to the trimmed mean (Stigler, 1973). Since data on Y_{r_2} are missing, the imputed data are utilized to calculate trimmed means estimates of the form $\hat{\mu}_{T\Delta}^{(\ell)} = \hat{\mu}_{T\Delta}(Y_{r_1}, Y_{r_2}^{(\ell)}, \mathbf{1}(R = r_3))$ for the m imputed datasets. Lastly, we Rubin's rules [3] to summarize the results of the trimmed means applied to each partially imputed dataset.

$$\begin{aligned} \bar{\mu}_{T\Delta} &= \frac{1}{m} \sum_{\ell=1}^m \hat{\mu}_{T\Delta}^{(\ell)} = \frac{1}{m} \sum_{\ell=1}^m \hat{\mu}_{T1}^{(\ell)} - \hat{\mu}_{T0}^{(\ell)} \\ Var(\bar{\mu}_{T\Delta}) &= \frac{1}{m} \sum_{\ell=1}^m (\hat{\sigma}^{(\ell)})^2 + \left(1 + \frac{1}{m}\right) \left(\frac{1}{m-1} \sum_{\ell=1}^m (\hat{\mu}_{T\Delta}^{(\ell)} - \bar{\mu}_{T\Delta})^2\right) \end{aligned}$$

where $\hat{\sigma}^{(\ell)}$ is the estimated standard error of the trimmed means estimate in the ℓ th imputed dataset. This combination approach is only valid if all unobserved values of Y_{r_3} fall below the quantile of the distributions that are trimmed, a condition discussed in the subsequent section.

1.3 Properties of the Estimand

1.3.1 Equivalence to the Population Treatment Effect

We focus on a randomized clinical trial, where the estimand of interest is the treatment difference in the population means of a clinical endpoint. Consider counterfactual outcomes Y_a where $a \in 0, 1$ indicates potential treatment assignments. In addition, $A \in 0, 1$ is the binary indicator for observed treatment in the trial. Denote U_a as the counterfactual version of the composite outcome defined in Section 1.2. The trimmed means estimand is most similar to a composite estimand, using the terminology of the ICH E9 addendum. The trimmed means estimand in counterfactual notation is $E[U_1|U_1 > F_{U_1}^{-1}(\alpha)] - E[U_0|U_0 > F_{U_0}^{-1}(\alpha)]$ where $F_{U_a}^{-1}(\alpha)$ represents the inverse cdf of the counterfactual distributions of U_a evaluated at α . This is a unique estimand defined for a sub-population of the trial, interpreted as the difference between treatments in endpoint means in the best $100 \times (1 - \alpha)\%$ of patients from each arm. This is not, however, the treatment effect among all randomized patients.

The advantage of the trimmed means approach's composite estimand is that it gives us a strategy for handling the missing data; however, the main drawback of using a composite estimand is that the clinical relevance can be unclear. It may be difficult and unfamiliar to interpret an estimated treatment effect in the best $100 \times (1 - \alpha)\%$ of patients as compared to an estimate of efficacy for all patients in a particular indication. The estimand for treatment efficacy in the population from which all randomized patients are drawn is the difference in counterfactual means $E[Y_1] - E[Y_0]$, however using the trimmed means approach only the difference in trimmed means of the composite outcomes U_1 and U_0 are estimated. Herein, the sufficient and necessary conditions under which the trimmed means estimand and the population estimand for treatment effect are identical are formalized.

Theorem 1. *If the outcomes among the treated and untreated are identically distributed relative to a shift and all unobserved values fall below the trim, then the treatment difference estimated by the trimmed means*

approach is equivalent to the treatment difference in the population, i.e.:

$$E[U_1|U_1 > F_{U_1}^{-1}(\alpha)] - E[U_0|U_0 > F_{U_0}^{-1}(\alpha)] = E[Y_1] - E[Y_0]$$

The proof is given in the appendix.

Theorem 1 uses the following two conditions in order to prove the equality of the trimmed means estimand and treatment difference in the whole population.

1. **Location family assumption.** The distribution of potential outcomes had the patient taken the experimental treatment $Y_1 \sim f_1(y)$ is in the same location family as the distribution of potential outcomes had the patient taken the reference treatment $Y_0 \sim f_0(y)$. Consider some constant Δ then:

$$f_0(y) = f_1(y + \Delta)$$

2. **Quantile assumption.** All missing values fall below the point at which the distributions are trimmed. Explicitly, the quantile assumption states:

$$Y_a|R_a = 0 < F_a^{-1}(\alpha)$$

The quantile assumption ensures that the composite outcome U_a is trimmed at the same value as Y_a for all percentiles above the maximum rate of missing data between the two arms, i.e. $F_{U_a}^{-1}(\alpha) = F_a^{-1}(\alpha) \forall \alpha : \alpha > Pr[R_a = 0]$. It also guarantees that the untrimmed distribution of U_a is identical to that of Y_a .

If Y_1 and Y_0 can be identified from the observed data, then the trimmed means approach can estimate the causal estimand of treatment effect in the population given the above two assumptions as shown:

$$\begin{aligned} E[Y_1] - E[Y_0] &= E[Y_1|Y_1 > F_1^{-1}(\alpha)] - E[Y_0|Y_0 > F_0^{-1}(\alpha)] \quad (1) \\ &= E[U_1|U_1 > F_{U_1}^{-1}(\alpha)] - E[U_0|U_0 > F_{U_0}^{-1}(\alpha)] \quad (2) \\ &= E[U_1|U_1 > F_{U_1}^{-1}(\alpha), A = 1] - E[U_0|U_0 > F_{U_0}^{-1}(\alpha), A = 0] \\ &= E[U|U > F_{U_1}^{-1}(\alpha), A = 1] - E[U_0|U_0 > F_{U_0}^{-1}(\alpha), A = 0] \end{aligned}$$

The location family assumption makes the average difference of the entire population of counterfactuals equivalent to the difference in the sub-population of counterfactuals that are not trimmed (1). The quantile

assumption makes the trimmed means of the counterfactuals equivalent to the trimmed means of the composite outcome (2). Then randomization and consistency allow us to identify the counterfactuals from the observed data (Cole and Frangakis, 2009). Note that using Theorem 1 only the difference in the counterfactual means can be recovered, albeit with a smaller sample size than if there were no MNAR data. One cannot accurately estimate the marginal means of Y_1 and Y_0 in the presence of MNAR data.

Theorem 2. *If the difference in untrimmed means between two counterfactual distributions is equivalent to the difference in one sided trimmed means for all percentiles,*

$$E[Y_1|Y_1 > F_1^{-1}(\alpha)] - E[Y_0|Y_0 > F_0^{-1}(\alpha)] = E[Y_1] - E[Y_0] = \Delta \quad \forall \alpha(0, 1)$$

then the counterfactual distributions are a location shift of one another

$$f_0(y) = f_1(y + \Delta)$$

Proof is given in the appendix.

Theorem 2 demonstrates that using the trimmed means approach to estimate the population treatment effect is only relevant for treatments with an additive effect. For all possible α , the difference in trimmed means and the population mean are equivalent if and only if the distributions being compared are a location shift of one another. There are conditions where the difference in trimmed means and population means are equivalent when the quantile assumption is not true (i.e. the MCAR case). Thus, theorem 2 reveals that the location family assumption is a sufficient and necessary condition, while the quantile assumption is only a sufficient condition for the equivalence of the estimands.

1.3.2 Intercurrent Events

Theorem 1 extends the utility of the trimmed means approach by proving under what assumptions one can estimate the estimand representing the treatment effect based on all randomized patients rather than a subset of the best performing patients. It is important to discuss how this result fits into the estimand framework outlined in the ICH E9 addendum. The trimmed means approach is a statistical analysis that specifies how

to deal with missing data, not necessarily a particular strategy to deal with intercurrent events (IE), which may often, but do not deterministically, lead to missing data. The IE strategy depends on how one handles data after observing an IE and is one of the four components in defining the estimand.

As discussed above, the trimmed means can be used as a composite approach, whereby IE that lead to missing data - and potentially others - are ranked poorly and trimmed out of the analysis. The resulting composite estimand can be thought of as a measure of treatment difference that is penalized by the amount of dropout in each arm. This penalty comes by placing the IE towards the poor end of the distribution irrespective of if the unobserved data are MAR or MNAR. In other words, they are ranked as the worst outcome even if their outcome would not have been poor had they continued in the trial. This type of estimand seems ideal for IE such as adverse events that outweigh the benefit of treatment or death. To estimate this composite estimand, the location family and quantile assumptions do not need to hold. Should these assumptions be realistic however, the opportunity arises to use the trimmed means approach to estimate two other types of estimands: 1) the hypothetical and 2) intention to treat (ITT) estimands.

The hypothetical estimand postulates what would have happened had the intercurrent event not occurred and the patient remained on treatment for the duration of the trial. Even if post IE data is collected, it is discarded and treated as missing data. For the ITT estimand the IE is irrelevant, and one is interested in data that occur after the patient discontinues treatment. If this post IE data is not missing, it is used as a valid endpoint in the analysis. Both the ITT and hypothetical estimands are estimable using the trimmed means approach if the quantile and location family assumptions of Theorem 1 hold. The difference between using the trimmed means approach for the hypothetical and ITT estimands is that the assumptions are made on different counterfactuals that are determined by whether or not treatment is continued post-IE. It seems more likely that the location family assumption in particular would hold for the hypothetical estimand, especially when the drug has an additive effect. The quantile assumption seems more likely for the ITT estimand, but is ultimately based on the process generating the missing data. In addition, if information is collected on the reasons for missing data, and if some of these missing data can be assumed to be MAR and others MNAR below the quantile then one could use a combination of multiple imputation and trimmed means to estimate these estimands.

1.4 Numerical Studies

1.4.1 Simulation Objectives

Numerical studies herein evaluate the finite sample properties of the trimmed means approach in estimating treatment efficacy under various missing data generating mechanisms. The simulation presented is motivated by the design of Wang et al. (2018). This earlier work is extended in a number of ways. Firstly, a comparison to MI under the various missing data generating mechanisms of the simulation is demonstrated. Additionally, when there exists a mixture of missing data types the combination approach is evaluated and compared to applying the trimmed means approach and MI globally. Furthermore, the relationship between bias and violation of the quantile assumption of Theorem 1 is considered under different MNAR scenarios. Lastly, the choice of α is explored. The comparison of trimmed means to MI as well as the combination approach would not be possible without Theorem 1 because it demonstrates that the approaches can estimate the same estimand. Data are imputed using the mice package in R, which leverages the same methodology used by PROC MI in SAS (Buuren and Groothuis-Oudshoorn, 2010).

1.4.2 Simulation Design

We design the study using four different ways to generate the missing data: (a) Missing Completely at Random (MCAR), (b) Missing at Random (MAR), (c) Missing Not at Random (MNAR), and (d) a mixture of all three types. Figure 1.1 displays the causal diagrams for these simulation designs using the *m*-graphs (Mohan et al., 2013).

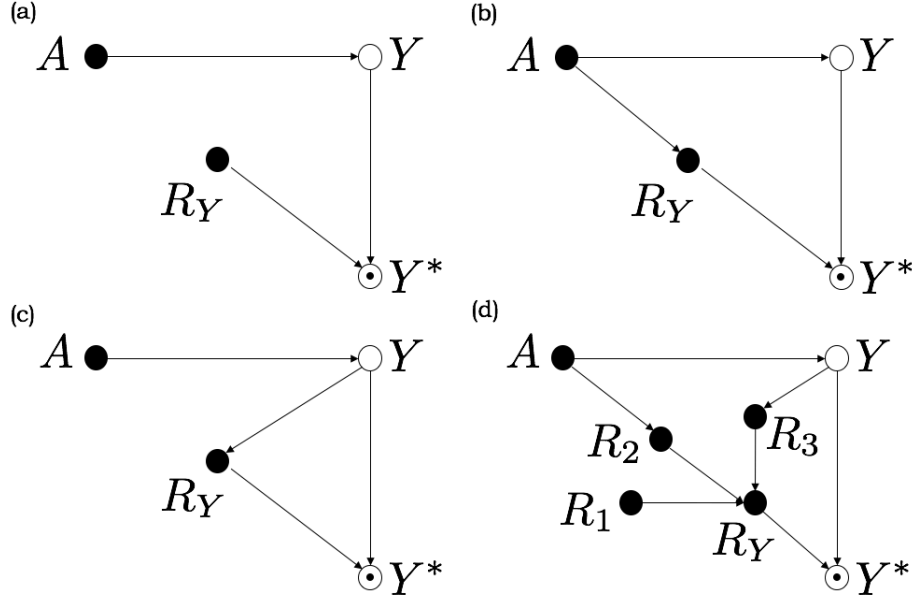


Figure 1.1 m -graphs for Scenarios a, b, c, and d of the Numerical Study

In these diagrams A represents a binary indicator for treatment, Y a continuous clinical endpoint of the study, and R_Y a binary indicator variable that is equal to 1 if Y is observed and 0 if Y is missing. Variable Y^* is the observed outcome, and its partially filled in node on the graph indicates that it has some missing values. Filled in nodes represent fully observed variables (i.e. A and R_Y). Nodes that are not filled in represent unobserved variables (i.e. Y). The arrows in these graphs make explicit the assumptions about which variables have a causal effect on missingness.

We use a study sample size of $N = 100$ ($n = 50$ per treatment arm) in each of the four scenarios. Each scenario was replicated $K = 5000$ times. The α parameter that determines which percentile to trim in the analysis is chosen adaptively, unless stated otherwise. The upper part of the distribution is trimmed, corresponding to lower values reflecting better outcomes. The underlying model for the continuous outcome remains the same in all simulations:

$$Y = \beta_0 + \beta_A A + \epsilon$$

Y is the continuous outcome variable and A is the binary variable representing experimental treatment if 1 and reference treatment if 0. Here, the error term is normally distributed $\epsilon \sim N(0, \sigma^2)$. The goal is to estimate β_A , the difference of the means between treatments. In all scenarios, the values of the parameters for the outcome model are $\beta_0 = -1, \beta_A = -1, \sigma = 1.5$. We chose $\sigma = 1.5$ to obtain a benchmark $\sim 90\%$

power in a one-sided t-test when there is no missing data.

The missing data in outcome Y were generated via the following logit model:

$$Pr(R_Y = 1|A, Y) = \text{logit}^{-1}(\alpha_0 + \alpha_A A + \alpha_Y Y)$$

Where R_Y is the binary variable indicating that Y has been observed if equal to 1. In this model, setting parameters $\alpha_A = \alpha_Y = 0$ corresponds to MCAR because the missing values are unrelated to treatment or outcome. Setting parameter $\alpha_Y = 0$ corresponds to MAR because the missing values are only dependent on the observed values and not the unobserved outcome. If $\alpha_Y \neq 0$ then the model represents an MNAR missing data generating mechanism.

1.4.3 Simulation Results

(a) MCAR

In the MCAR setting the α_0 parameter is set to values of 2.94, 2.20, 1.74, and 1.39 to induce missing data rates of 5, 10, 15, and 20 percent while keeping $\alpha_A = \alpha_Y = 0$. The missing data rates are the same on average in each arm since unobserved outcomes are completely random.

Table 1.1 Trimmed Means with MCAR

Missing Rate, %			Trimmed Means			
$A = 1$	$A = 0$	Exp	Ref	Diff (% bias)	Coverage	Power
5	5	-2.17	-1.17	-1.00 (0%)	0.96	0.86
10	10	-2.16	-1.16	-1.00 (0%)	0.96	0.83
15	15	-2.15	-1.16	-1.00 (0%)	0.96	0.80
20	20	-2.39	-1.39	-1.00 (0%)	0.96	0.70

Under a completely random missing data generating mechanism (MCAR), the trimmed means approach estimates the true treatment difference without bias and with appropriate coverage even as the proportion of

data missing varies (Table 1.1). As expected, power decreases as the amount of data not trimmed decreases. MI performed similarly in that bias and coverage were accurate (Table 1.2). However, as the amount of missing data increases, power does not deteriorate as quickly using MI as when using trimmed means. This is because the trimmed means approach performs inference on the subset of the observations post-trimming and thus uses a smaller effective sample size.

Table 1.2 Multiple Imputation with MCAR

Missing Rate, %		Multiple Imputation (MI)				
$A = 1$	$A = 0$	Exp	Ref	Diff (% bias)	Coverage	Power
5	5	-2.00	-1.00	-1.00 (0%)	0.94	0.89
10	10	-2.00	-1.00	-1.00 (0%)	0.94	0.87
15	15	-1.99	-1.00	-0.99 (1%)	0.94	0.84
20	20	-1.99	-1.01	-0.99 (1%)	0.94	0.82

(b) MAR

In the MAR setting, we first set the α_A parameter to values of -8.61, -8.27, -7.80, and -7.06 to induce missing data rates of 20, 15, 10, and 5 percent in the experimental arm while keeping $\alpha_Y = 0$ and $\alpha_0 = 10$ in order to maintain all outcomes observed in the reference arm. Next, we set $\alpha_Y = 0$ and $\alpha_A = 10$ in order to fully observe outcomes in the experimental arm while varying α_0 to values of 2.94, 2.20, 1.73, and 1.39 to induce missing data rates of 5, 10, 15, and 20 percent in the reference arm.

Table 1.3 Trimmed Means with MAR

Missing Rate, %		Trimmed Means				
$A = 1$	$A = 0$	Exp	Ref	Diff (% bias)	Coverage	Power
20	0	-1.99	-1.51	-0.48 (52%)	0.74	0.22
15	0	-2.00	-1.41	-0.59 (40%)	0.81	0.37
10	0	-2.00	-1.28	-0.72 (28%)	0.88	0.54
5	0	-2.00	-1.15	-0.84 (16%)	0.93	0.74
0	5	-2.16	-1.01	-1.15 (-15%)	0.95	0.95
0	10	-2.28	-1.00	-1.28 (-28%)	0.90	0.97
0	15	-2.40	-1.00	-1.40 (-20%)	0.85	0.98
0	20	-2.51	-1.00	-1.51 (-51%)	0.77	0.99

As expected, the trimmed means estimator is biased in all scenarios when the missing data is truly MAR (Table 1.3). The bias increases when the fraction of missing data increases. The direction of the bias is positive when the reference arm has more missing data and negative when the active arm has more missing data. This directionality of the bias has an impact on power, with more MAR data in the active arm leading to a drastic decrease in power and more MAR data in the reference arm causing unreasonably high power. The trimming is directional as all missing values are placed at the poor end of each respective treatment distribution when in reality under MAR they come from all areas of the distribution. MI obtains valid estimation in this setting as it was designed explicitly for situations where data are MAR (Table 1.4).

Table 1.4 Multiple Imputation with MAR

Missing Rate, %		Multiple Imputation (MI)				
$A = 1$	$A = 0$	Exp	Ref	Diff (% bias)	Coverage	Power
20	0	-1.99	-1.00	-0.99 (-1%)	0.94	0.86
15	0	-2.00	-1.00	-1.00 (0%)	0.94	0.88
10	0	-2.00	-1.00	-1.00 (0%)	0.95	0.89
5	0	-1.99	-1.00	-1.00 (0%)	0.95	0.91
0	5	-2.00	-1.00	-1.00 (0%)	0.94	0.90
0	10	-2.00	-1.00	-1.00 (0%)	0.95	0.90
0	15	-2.00	-1.01	-1.00 (0%)	0.95	0.88
0	20	-2.00	-1.01	-1.00 (0%)	0.94	0.86

(c) MNAR

In the MNAR setting, the α_Y parameter was set to values of -1, -2.5, -5, and -10 causing higher values of Y to be more likely to be missing while keeping $\alpha_0 = 2.85$ and $\alpha_A = 0$. Here, α_Y is negative because a decrease in Y reflects a better outcome. This setup induces missing data rates in the experimental vs reference arms of 2 vs 5, 3 vs 10, 5 vs 15, and 7 vs 20 respectively. The missing data are not simulated strictly as quantile but a general MNAR missing data mechanism.

Table 1.5 Trimmed Means with MNAR

Missing Rate, %		Trimmed Means				
$A = 1$	$A = 0$	Exp	Ref	Diff (% bias)	Coverage	Power
2	5	-2.14	-1.11	-1.04 (-4%)	0.96	0.90
3	10	-2.28	-1.26	-1.02 (-2%)	0.96	0.90
5	15	-2.41	-1.41	-1.00 (0%)	0.96	0.90
7	20	-2.51	-1.51	-1.00 (0%)	0.95	0.89

The trimmed means approach is fairly unbiased, obtains ideal coverage, and maintains its power in the MNAR setup as the amount of missing data increases (Table 1.5). While the marginal means in each arm are biased, the means in each arm increase at equal rates and keep the estimate of their difference unbiased. Multiple imputation increases bias, reduces coverage of the true effect, and loses power as the fraction of MNAR data increases (Table 1.6).

Table 1.6 Multiple Imputation with MNAR

Missing Rate, %		Multiple Imputation (MI)				
$A = 1$	$A = 0$	Exp	Ref	Diff (% bias)	Coverage	Power
2	5	-2.10	-1.10	-1.00 (0%)	0.95	0.92
3	10	-2.14	-1.26	-0.89 (11%)	0.93	0.88
5	15	-2.19	-1.41	-0.78 (22%)	0.87	0.81
7	20	-2.22	-1.52	-0.70 (30%)	0.79	0.74

(d) Mixture: MCAR, MAR, and MNAR

Table 1.7 Trimmed Means with a Mixture of Missing Data Types

Missing Rate, %					Trimmed Means				
Trt	R1	R2	R3	Overall	Exp	Ref	Diff (% bias)	Coverage	Power
$A = 1$	5	23	2	30	-2.05	-1.65	-0.40 (60%)	0.71	0.13
$A = 0$	5	0	5	10					
$A = 1$	5	17	3	25	-2.12	-1.55	-0.56 (44%)	0.80	0.30
$A = 0$	5	0	10	15					
$A = 1$	5	10	5	20	-2.24	-1.47	-0.77 (23%)	0.90	0.56
$A = 0$	5	0	15	20					
$A = 1$	5	3	7	15	-2.48	-1.54	-0.93 (7%)	0.94	0.78
$A = 0$	5	0	20	25					

Having a mixture of reasons for missing data reflects the information one would have in a closely monitored clinical trial. In many trials, data are missing for a combination of reasons such as lack of efficacy, intolerability, and administrative reasons. In order to generate such data the deletion strategies used in the previous three scenarios are combined. MNAR data (R3) were deleted first at rates of 2 vs 5, 3 vs 10, 5 vs 15, and 7 vs 20 in the experimental vs reference arms respectively. MAR data (R2) were then generated in the experimental group at rates of 23, 17, 10, and 3. MCAR data (R1) were generated at a rate of 5 percent in each arm. Overall, the missing data rates in the four mixture scenarios in the experimental vs reference arms are 10 vs 30, 15 vs 25, 20 vs 20, and 15 vs 25 respectively.

Table 1.8 Multiple Imputation with a Mixture of Missing Data Types

Missing Rate, %					Multiple Imputation (MI)				
Trt	R1	R2	R3	Overall	Exp	Ref	Diff (% bias)	Coverage	Power
$A = 1$	5	23	2	30	-2.04	-1.13	-0.91 (9%)	0.92	0.79
$A = 0$	5	0	5	10					
$A = 1$	5	17	3	25	-2.10	-1.31	-0.79 (21%)	0.87	0.73
$A = 0$	5	0	10	15					
$A = 1$	5	10	5	20	-2.14	-1.41	-0.73 (27%)	0.83	0.70
$A = 0$	5	0	15	20					
$A = 1$	5	3	7	15	-2.20	-1.54	-0.66 (34%)	0.75	0.66
$A = 0$	5	0	20	25					

In the mixture setting, the performance of trimmed means applied globally is directly related to the proportion of MAR missing data (Table 1.7). This fraction of MAR data decreases across the four scenarios and consequently bias, coverage, and power improve across the scenarios. Contrarily, MI applied globally performs well with a large fraction of MAR data and its performance weakens as the proportion of MNAR data increases (Table 1.8). The combination of trimmed means and MI improves bias, coverage, and power as compared to each method applied individually (Table 1.9). Bias is at most 3%, no matter the variation in the fraction of missing data due to MAR and MNAR. Coverage and Power are near the optimal 0.95 and 0.90.

Table 1.9 Trimmed Means + MI with a Mixture of Missing Data Types

Missing Rate, %					Trimmed Means + MI				
Trt	R1	R2	R3	Overall	Exp	Ref	Diff (% bias)	Coverage	Power
$A = 1$	5	23	2	30	-2.17	-1.14	-1.03 (-3%)	0.92	0.88
$A = 0$	5	0	5	10					
$A = 1$	5	17	3	25	-2.32	-1.31	-1.01 (-1%)	0.93	0.88
$A = 0$	5	0	10	15					
$A = 1$	5	10	5	20	-2.42	-1.42	-1.00 (-0%)	0.94	0.88
$A = 0$	5	0	15	20					
$A = 1$	5	3	7	15	-2.53	-1.54	-0.99 (1%)	0.94	0.89
$A = 0$	5	0	20	25					

(e) Choice of α and Quantile Assumption

The analyst chooses α – the proportion trimmed from each treatment arm – which can be set to any value above the maximum proportion of missing data between the two arms. For all previous simulations the value of α was chosen adaptively. Herein, the adaptive choice of α is compared to a fixed choice where $\alpha = 0.5$. In addition, to investigate the quantile assumption of Theorem 1, the percent of missing values that would have fallen below the trim point had they been observed is calculated for each scenario. The adaptive and fixed α approaches are evaluated under 10 different MNAR data generating mechanisms using the same logit model as before. The α_Y parameter was set to values of -0.5, -1, -1.5, -2, -2.5, -3, -4, -5, -7.5 and -10 while keeping $\alpha_0 = 2.85$ and $\alpha_A = 0$. Rates of missing data and results are shown in Table 1.10.

Table 1.10 Comparison of Adaptive and Fixed α

Missing Rate, %		Adaptive α				Fixed $\alpha = 0.5$			
$A = 1$	$A = 0$	Diff (% bias)	sMNAR	SE	MSE	Diff (% bias)	sMNAR	SE	MSE
3	4	-1.038 (-3.8%)	11.4%	0.321	0.104	-1.015 (1.5%)	74.6%	0.355	0.126
2	5	-1.038 (-3.8%)	28.2%	0.313	0.100	-1.012 (1.2%)	91.1%	0.352	0.124
2	7	-1.032 (3.2%)	45.9%	0.310	0.097	-1.009 (0.9%)	97.0%	0.351	0.123
3	8	-1.025 (2.5%)	60.6%	0.309	0.096	-1.008 (0.8%)	99.1%	0.351	0.123
3	10	-1.019 (1.9%)	70.3%	0.308	0.095	-1.008 (0.8%)	99.6%	0.351	0.123
3	11	-1.015 (1.5%)	76.4%	0.308	0.095	-1.007 (0.7%)	99.8%	0.350	0.123
4	14	-1.003 (0.3%)	83.8%	0.314	0.099	-0.999 (0.1%)	100.0%	0.356	0.127
5	15	-1.002 (0.2%)	87.9%	0.312	0.098	-1.003 (0.3%)	100.0%	0.355	0.126
6	18	-0.998 (0.2%)	93.4%	0.314	0.099	-1.003 (0.3%)	100.0%	0.355	0.126
7	20	-0.995 (0.5%)	95.7%	0.313	0.098	-1.001 (0.1%)	100.0%	0.351	0.123

Overall, both the fixed and adaptive α accurately estimate the true difference in treatment effects ($\beta_A = -1$). As α_Y moves further from 0, the percentage of missing values falling below the trim point (i.e. sMNAR) increases. As a consequence, bias decreases which is consistent with the theoretical result of Theorem 1. Unlike in imputation, bias when using trimmed means is not directly related to the fraction of missing data, but rather due to sMNAR. Thus, it is possible to observe lower bias despite larger amounts of missing data, as is demonstrated here. The fixed α approach consistently has a higher sMNAR than the adaptive approach. This explains why using the fixed α has lower bias in all scenarios: the underlying missing values have a higher likelihood of falling below the more extreme trimming quantile. Using the adaptive α has a smaller variance than the fixed α because the adaptive approach trims the least amount of data possible. The smaller variance of the adaptive approach translated to a smaller MSE than the fixed approach despite the fixed approach having less bias. This simulation highlights the bias vs variance tradeoff associated with increasing the percentage of data trimmed.

1.5 Application to a Clinical Trial

We applied the methodologies described above to data from a double-blind randomized clinical trial of two treatments (A and B) conducted in patients with neuropathic pain due to diabetic neuropathy. Seventy-one patients were randomized to treatment A and seventy to treatment B. The outcome of interest was change in pain severity from baseline to week 16, as assessed on a Visual Analog Scale (VAS). VAS is a well-studied instrument for recording pain where a score of 100 reflects the “worst pain possible” and a score of 0 reflects “no pain” (Becker et al., 1993). Pain scores were recorded in a digital diary daily by patients. At most, there were 16 weekly pain measurements for each patient, produced by averaging daily pain recordings during each week.

Study discontinuation in this clinical trial was common, as there were 53 (38%) patients who did not stay on trial for 16 weeks. Discontinuation differed among treatment arms, 33 (46%) in the treatment A arm and 20 (29%) in the treatment B arm. The reason for discontinuing the study was recorded for each dropout and categorized as Adverse Event (AE), Loss of Efficacy (LoE), or Administrative (Table 1.11). The rates of study discontinuation, the time at which they occurred, and the observed data before dropout were used to inform missing data assumptions. AE and LoE generally occurred during the first half of the study period while administrative dropout occurred uniformly throughout the trial. On average AE and LoE occurred after 6.77 and 6.83 weeks on trial, respectively, and administrative dropouts after 10.6 weeks. Based on this exploratory data analysis and our clinical knowledge, we assume that the dropouts classified as AE & LoE were MNAR and administrative dropouts were MCAR.

Table 1.11 Treatment Discontinuation Reasons

Dropout Type	Treatment A	Treatment B
Adverse Events	18	4
Loss of Efficacy	3	3
Administrative	12	14
Total	33	21

We applied 5 different methods to the trial data. First, the trimmed means approach was applied globally to all dropouts (i.e. assumes all dropouts MNAR). The fraction trimmed was chosen adaptively and thus corresponded to the amount of dropout in the treatment A arm (i.e. $\alpha = 0.46$). To test the location shift assumption, we performed a Kolmogorov-Smirnov test between the distribution of treatment A shifted by the treatment effect compared to the observed distribution of treatment B. The test failed to reject that the untrimmed outcome distributions were a location shift of one another ($D=0.0946$, $p=0.9849$). We also applied multiple imputation to all dropouts in an ANOVA model despite the MAR assumption being unlikely for many dropouts. Next, we applied the approach that combines trimmed means and multiple imputation. To do this, we trimmed AE and LoE (MNAR) and imputed administrative dropout data (MCAR). As a consequence, the fraction trimmed was reduced to $\alpha = 0.30$ in each of the imputed datasets. Lastly, we applied two more approaches for historical reference, a complete case analysis of all patients completing the trial (i.e. assumes all dropouts MCAR) and a Last Observation Carried Forward (LOCF) analysis.

Table 1.12 contains the results of each of these approaches. The trimmed means applied to all dropouts showed the largest treatment difference of -14.48 points lower on the pain VAS; however, the trimming inflated the standard error to 7.61. Similarly, the combination of trimmed means and MI had the second largest effect size -12.67. Contrarily, the combination approach trimmed less data which resulted in a less inflated standard error. These two methods, which involve trimming, resulted in a larger effect size than the other methods because they allow for the fact that the worse-performing treatment had a higher dropout rate. The other approaches presented are not appropriate given our missing data assumptions, but were included for illustrative purposes as a reference, especially the relative standard errors.

Table 1.12 Clinical Trial Analysis Results

Method	Pain Difference	SE	95% CI	<i>p</i> -value
Trimmed Means	-14.48	7.61	[-29.38, 0.43]	0.055
Trimmed Means + MI	-12.67	6.21	[-24.83, -0.49]	0.041
Complete Case Analysis	-3.74	5.39	[-14.45, 6.97]	0.497
Multiple Imputation	-2.71	4.68	[-11.88, 6.45]	0.537
LOCF	-1.76	4.20	[-10.06, 6.54]	0.675

Focusing on the SE column, we see that LOCF has the smallest standard error, 4.20, as expected since the method does not admit to missing data; it replaces all missing data with the last available data point, in time, for each patient based on a single imputation. Some object to this method being unrealistic in this situation. The next smallest entry, 4.68, Multiple Imputation, treats all missing data as missing at random, which may not be plausible for all discontinuation reasons in this trial. Once again, the method creates data whenever they are missing, except not once, as in LOCF, but multiple times in order to better reflect the uncertainty induced by the missing data. The complete case analysis has the next smallest value, 5.39. This was only included for completeness. In general, the shortcomings of this method are well known. The next smallest is the Trimmed Means + MI, 6.21. This approach achieves an unbiased comparison by maintaining an “equal percentage” of the data for those who prematurely leave the study for cause. This method also permits distinguishing between observations that are truly missing at random and other observations for which the missing at random assumption is not plausible (i.e. missing not at random). The last one, the Trimmed Mean method has the largest standard error, 7.61. This is easily explained because this is the method that discards the most data. This is excessive in that the fraction of the missing data that are missing at random are best handled as missing at random and thus amenable to multiple imputation.

1.6 Discussion

Our work extends the utility of the trimmed means approach for missing data in two key ways: 1) It determines sufficient conditions for which the trimmed means approach can identify the estimand for the population treatment effect; and 2) It demonstrates that when different types of missing data are present and can be distinguished, one could combine the trimmed means approach with multiple imputation to improve estimation.

The trimmed means approach was originally designed to estimate a unique estimand: the treatment difference in the best $(100 \times \alpha)$ % of patients of each arm. The work herein allows us to view the trimmed means approach in a different way, not a method estimating a unique estimand, but a method that targets the usual estimand of a clinical trial where accuracy depends on how well the assumptions are satisfied.

Missing data inferences are not possible without assumptions. The quantile assumption, similar to the MAR assumption, is untestable. It is a conservative assumption that assumes every missing value falls below the trimming quantile. It would be rare for this assumption to hold perfectly; however, when applied to dropouts reporting loss of efficacy, where poor outcomes are the primary cause of dropout, the assumption may hold for enough of the missing data to justify adopting the trimmed means approach. Also the numerical studies show that the trimmed means estimator can still perform well under MNAR scenarios that are not explicitly consistent with the quantile assumption. Thus, the assumption may be robust to slight deviations. One interesting paradox is that while trimming more data leads to a loss in efficiency, theoretically it allows the quantile assumption to become more plausible since missing values are then more likely to be trimmed. The simulation comparing a fixed to adaptive choice of α demonstrates this. This bias/variance trade-off should be considered when choosing the value of α . The location shift assumption may be more realistic, is testable among the observed values, and is often assumed in many statistical methods. In practice, the untrimmed fractions of the distributions for each treatment arm of the study should be compared using a Kolmogorov-Smirnov test to assess the validity of this assumption. Of course should neither the location family nor the quantile assumptions be plausible, the trimmed means approach can still be useful in estimating the original composite estimand in the sub-population of the trial for which it was originally developed.

This research highlights the importance for administrators and physicians conducting clinical trials to document the reasons for dropout. If close collaboration between statisticians and clinicians can inform

which dropouts are MCAR, MAR, or MNAR then analysts may have a combination of data that can be imputable and other data that should be trimmed away using the trimmed means approach. This combination approach can protect analysts from penalizing themselves using trimmed means globally for all missing data but also respects the assumption that patients may drop out of the trial due to poor health outcomes. If a fraction of dropouts are missing because of factors unrelated to their unobserved outcomes (MAR), the bias and loss of power using trimmed means for all dropouts can be drastic. Choosing which missing data to treat as MAR or MNAR will vary from trial to trial, and will also be dependent on the particular estimand of interest. Clinical input is crucial for the mixture approach to be effective. One limitation of the combination approach is that MNAR dropouts must precede MAR dropouts; otherwise the complete cases leveraged for imputations may have a different outcome distribution than the MAR outcomes.

The clinical trial example highlights the importance of combining trimmed means and multiple imputation. MCAR data is by definition evenly distributed between each arm of the study; therefore, it does not cause bias when applying trimmed means. However, there is considerable power loss as shown in MCAR situation of our simulations. In the clinical trial analysis, applying trimmed means to all dropouts produced a larger estimate of the difference in treatment effects, but because 46% of observations were missing in one arm of the study the standard errors were inflated, making it harder to reject the null hypothesis of no difference between treatments. The combination approach, however, preserved a similar estimate of the treatment comparison and did not inflate standard errors as drastically. The combination approach leverages a larger effective sample size than applying trimmed means alone.

The trimmed means approach is a creative solution to estimating treatment effects in a clinical trial when missing data can safely be assumed to be due to poor outcomes. As is the case in any missing data analysis, especially those with MNAR data, no analytical method replaces a good sensitivity analysis to determine the plausible range of what could have happened. While no method can fully or confidently rectify the issues caused by missing data, a combination of multiple imputation and/or trimmed means could be useful when the assumptions of both methods are satisfied.

Acknowledgements

We thank Jaffer Zaidi and Zack McCaw for useful discussions regarding the proofs of Theorems 1 and 2, respectively.

2

Using Influence Functions to Bound Regression Parameters for Incomplete Data

Alex Ocampo

Department of Biostatistics, Harvard University

Abstract

In lieu of making untestable assumptions about missing data, analysts can bound parameters that take the incomplete observations into account. Despite their cautious nature, bounds can inform the presence, direction, and magnitude of parameters. Herein a novel approach is proposed to estimate these bounds which leverages influence functions. Influence functions can reveal which potential values of a particular missing observation would most drastically change the estimate; after imputing these observations one recalculates the estimate to obtain lower and upper bounds. If the influence function of an estimator is bounded, then sharp lower and upper bounds can be identified. The asymptotic variance for the estimates of these bounds follow directly from the approach due to the close relationship between influence functions and variance estimation. The approach is computationally efficient and speed is dependent on the number of missing values and their event space. The bounding procedure is applied to assess the effect of the drug indinavir in a time to event analysis using the Cox proportional hazards model. This strategy for estimating bounds is a tool that can quickly characterize the impact missing data are having on analyses.

2.1 Introduction

When confronted with missing data statisticians typically make untestable assumptions about the mechanism that generates missing data, most notably missing at random (MAR) (Rubin, 1976). This allows underlying parameters to be identifiable. Then energy is focused on obtaining efficiency given the assumption. Because these assumptions are untestable, and sometimes unrealistic, it is then recommended to perform a sensitivity analysis to see how estimates change as one deviates from the assumptions made about the missing data. However, sensitivity analyses are often neglected. One review found that 65% of clinical trials failed to report a sensitivity analysis (Horton and Kleinman, 2007). In addition, there is no consensus view on how to best conduct a sensitivity analysis. The scientific community needs more easily implementable tools that quantify the impact that missing data could be having on research conclusions. An alternative unambiguous approach to a sensitivity analysis is to find upper and lower bounds of what could have happened under the worst and best case realizations of the missing values. Abstaining from untestable assumptions about the nature of the missing data, upper and lower bounds can be estimated which could inform the presence, direction, and magnitude of population parameters.

Bounds on population parameters were independently written about by Robins (1989) and Manski (1990) in the context of comparing counterfactual means between randomized treatment groups. Such a problem is analogous to bounding with missing data since no single counterfactual is completely observed for all observations. Similar bounds for treatment effects for studies with imperfect compliance have been derived (Balke and Pearl, 1997). Horowitz and Manski (2000) derived bounds for nonparametric analysis when treatments are fully observed but covariate and outcome data may be partially missing. In finite samples these bounds can be estimated from the observed data and confidence intervals for the bounds constructed empirically, or more simply using the bootstrap. Horowitz et al. (2003) formalized the problem as a mathematical programming problem with linear constraints and showed that their previous work on bounds is a simplification of this problem where a simple analytical solution exists. In some situations, estimating these bounds is a fractional linear program that can be solved rather efficiently (Zaffalon, 2002). For many situations, however, a simple solution does not exist and computationally intensive algorithms must be employed to estimate the bounds. Various algorithms proposed included a genetic algorithm, simulated annealing, branch-and-bound methods, and a simulation approach where missing values are drawn from the uniform distribution of all

possible values. The genetic algorithm solved the programming problem the quickest, however barriers in efficiency, defining convergence, and non-generality of its implementation may deter analysts from using the approach to estimate bounds. More efficient, interpretable, and easily implementable tools would encourage analysts to compute bounds in practice.

In this paper a fast and accurate approach to estimating bounds for parameters in many statistical models is proposed. This is accomplished by leveraging the influence function, which measures the contamination of a distribution by one observation, to determine how to fill in incomplete observations in order to take the largest step towards the upper and lower bounds. Section 2.2 describes the influence function and how to use it in the estimation of bounds. Section 2.3 describes some theoretical properties of the approaches. Section 2.4 explores the finite sample properties of the algorithm in a simulation study that compares some approaches in their accuracy and speed. Section 2.5 provides an application of the influence bounding approach to the treatment affect parameter for the antiretroviral drug indinavir in a Cox proportional hazards model. Section 2.6 concludes the article with discussion of recommendations and future outlook.

2.2 Methods for Estimating Bounds

The goal is to identify bounds, estimating the lowest and highest possible values of our statistic given the observed data. By considering what the missing values could have been had they been observed and imputing these data to form a complete dataset, one can estimate the lowest and highest possible realizations of a statistic.

2.2.1 Notation

Let $z_i = (y_i, \mathbf{x}_i)$, $i = 1, \dots, n$ be a sample of n i.i.d. random vectors, where y_i is the outcome of interest and $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})^T$ is a p -dimensional vector of covariates. Assume that $\beta = (\beta_1, \dots, \beta_p)^T \in \mathbb{R}^p$ is a parameter of interest. Denote $p^* \in \{1, \dots, p\}$ as the index of the parameter of primary interest for which bounds are desired. Let n_{cc} be the number of complete cases for which all variables are observed and denote n_{mis} the number of vectors with at least one missing value. Note that $n = n_{cc} + n_{mis}$. Consider indexing the complete observations by $i \in \{1, \dots, n_{cc}\}$ and the incomplete observations by $j \in \{(n_{cc} + 1), \dots, n\}$. In addition, let $\mathcal{D}_{obs}$ denote the observed but incomplete dataset, which can be partitioned into $\mathcal{D}_{obs} =$

$[\mathcal{D}_{cc}, \mathcal{D}_{mis}]$ for the complete and incomplete components of the data.

2.2.2 Positive Jackknife Approach

First, consider estimating bounds using the positive jackknife, which is only different than the usual jackknife (Quenouille, 1949) in that one considers adding an additional observation rather than removing an observed one. Consider a random sample $(z_1, \dots, z_n)$ of i.i.d random vectors defined above. Consider the estimator $T_n(z_1, \dots, z_n) = \hat{\beta}$ for parameter β . The positive jackknife for a new observation z_{n+1} is defined by

$$PJ(z_{n+1}) = (n + 1) (T_{n+1}(z_1, \dots, z_n, z_{n+1}) - T_n(z_1, \dots, z_n))$$

This provides a measure of the change in the statistic if observation z_{n+1} is added to the sample. To estimate the missing data bounds using the positive jackknife one could proceed as follows:

1. Use the complete cases $\mathcal{D}_{cc} = \{z_1, \dots, z_{n_{cc}}\}$ to compute $T_{n_{cc}}(\mathcal{D}_{cc}) = \hat{\beta}$
2. Take the first incomplete observation j and enumerate all possible realizations z_j of the missing values conditional on its observed values
3. For each enumeration calculate the positive jackknife $PJ(z_j)$
4. Impute the permutation with the most negative change in $PJ(z_j)$ relative to p^* for the lower bound and most positive change for the upper bound
5. Repeat steps 2-4 for all n_{mis} observations with incomplete data to impute $\mathcal{D}_{min}$ and $\mathcal{D}_{max}$
6. Then estimate the upper and lower bounds by $T(\mathcal{D}_{max}) = \bar{\beta}$ and $T(\mathcal{D}_{min}) = \underline{\beta}$

The above considers all possible imputations for a missing observation and measures how the estimator changes. Note that the above approach is only feasible if it is possible to enumerate all possible realizations of the missing data. Obviously this is only possible if the potential missing outcomes form a finite set, which is the case when all variables in the analysis are categorical. Now that the intuition of the problem has been motivated using the positive jackknife, one can now dive into how to derive and then use the influence function directly to estimate bounds. There is close connection between the positive jackknife and influence function which will be discussed in detail in Section 3.

2.2.3 Influence Function Approach

The positive jackknife accomplishes our goal, but it may be computationally expensive to recalculate a statistic T_n for each potential value of the incomplete observations. Also, the approach is not possible with missing continuous covariates since there are an infinite number of imputations to consider. Deriving the influence function for an estimator can overcome these challenges. Influence functions come from the robustness literature and characterize the effect that one observation has on an estimate (Hampel, 1974). They also play an important role in characterizing the asymptotic efficiency of an estimator (Tsiatis, 2007).

To derive an influence function it is useful to express estimators as statistical functionals. Statistics T_n estimate population parameters of a particular distribution F_X which gives rise to the X_i . These population parameters can be explicitly defined using functionals. A few simple examples of statistical functionals are the mean $T(F) = \int x dF(x)$, the variance $T(F) = \int (x - \mu)^2 dF(x)$, and the quantiles $T(F) = F^{-1}(p)$. One can also define parameters estimated by regressions as functionals with respect to the joint distribution of the outcome y and covariates x , i.e. F_{yx} . For example, in linear regression,

$$T(F_{yx}) = \left(\int x' x F_{yx} \right)^{-1} \int x' y dF_{yx}$$

as well as logistic regression,

$$T(F_{yx}) = \left(\int x' x F_{yx} \right)^{-1} \int \{x' \text{logit}(P[Y = 1|X = x])\} dF_{yx}$$

These population quantities are being expressed as functionals $T(F)$ which take as input the distribution F and return the parameter of interest. A statistical functional that estimates parameter β is denoted by $T(F) = \beta$ which has a corresponding estimator $T(\hat{F}_n) = \hat{\beta}$ where $\hat{F}_n$ is the empirical distribution estimated from sample of size n . Using functional notation our estimator from Section 2.2.2 can be rewritten $T_n(z_1, \dots, z_n) = T(\hat{F}_n) = \hat{\beta}$.

The influence function measures the influence that one observation x_j has on a statistical functional $T(F)$. The formal definition considers what would happen if the distribution F is contaminated an infinitesimal amount by one observation x_j . One can write this contaminant distribution as $(1 - \epsilon)F + \epsilon\delta_x$ where $\delta_{x_j} =$

$\mathbf{1}(X < x_j)$ is the Heaviside step function. The influence function is defined as:

$$\varphi_{T,F}(x_j) = \lim_{\epsilon \downarrow 0} \frac{T((1-\epsilon)F + \epsilon\delta_{x_j}) - T(F)}{\epsilon} = \left. \frac{\partial}{\partial \epsilon} T((1-\epsilon)F + \epsilon\delta_x) \right|_{\epsilon=0}$$

The influence function measures the rate at which the functional T changes when F is contaminated by a small probability ϵ of obtaining an observation x , and thus is a measure of the influence of such a contamination. It is formally the Gâteaux derivative of functional T at F in the direction of δ_x . From this definition one can derive the influence function directly for wide range of estimators. The j -th influence function for any maximum likelihood estimator of $T(F) = \beta$ with score equation $S(Z, \beta)$ and information matrix $I(\beta) = E[S(Z, \beta)S^T(Z, \beta)]$ is:

$$\varphi_{\beta,F}(z_j) = \{I(\beta)\}^{-1} S(z_j, \beta)$$

Another example is for a M-Estimator $\hat{\theta}$ which is the solution to an equation of the form $\sum_i \eta(Z_i, \hat{\theta}) = 0$. Then the influence function of estimator $\hat{\theta}$ is:

$$\varphi_{\theta,F}(z_j) = - \left[E_F \left\{ \frac{\partial \eta(Z, \theta)}{\partial \theta} \right\} \right]^{-1} \eta(z_j, \theta)$$

Many the most common statistical methods fit into the maximum likelihood or M-estimation framework. This includes tools such as the generalized linear models, estimators that solve GEE, and estimators with differentiable objective functions which all have well defined influence functions. In fact, the influence function uniquely characterizes any asymptotically linear estimator i.e. any estimator of the form $\sqrt{n}(\hat{\beta} - \beta) = n^{-1/2} \sum_{i=1}^n \varphi(X_i) + o_p(1)$ (Tsiatis, 2007). For all these estimators we can use their influence functions to estimate missing data bounds. Before estimating the bounds, note that the distribution F used to define the influence function is generally unknown. Thus we use the plug in estimator using the empirical distribution $\hat{F}_n$ known as the empirical influence function:

$$\varphi_{T,\hat{F}_n}(x) = \lim_{\epsilon \downarrow 0} \frac{T((1-\epsilon)\hat{F}_n + \epsilon\delta_x) - T(\hat{F}_n)}{\epsilon}$$

This now allows us to consider the different influence of a new observation after using the complete case data to estimate $\hat{F}$. Now using the influence function we can estimate the missing data bounds of parameter $T(F) = \beta$ as follows:

Algorithm 1: Influence Function Bounding Algorithm

Input : $\mathcal{D}_{obs}, T(F) = \beta, p^*$

Output : Finite Sample Upper Bound ($\bar{\beta}$) and Lower Bound ($\underline{\beta}$)

Initialize : $\mathcal{D}_{min} = \mathcal{D}_{max} = \mathcal{D}_{cc}$

begin

Use the complete case data $\mathcal{D}_{cc}$ to estimate $T(\hat{F}_{cc}) = \hat{\beta}$

for $j \in \{(n_{cc} + 1), \dots, n\}$ **do**

$\bar{z}_j = \underset{z_j}{\operatorname{argmax}} \varphi_{\hat{\beta}, \hat{F}_{cc}}(z_j)_{[p^*]}$

$\underline{z}_j = \underset{z_j}{\operatorname{argmin}} \varphi_{\hat{\beta}, \hat{F}_{cc}}(z_j)_{[p^*]}$

Append $\underline{z}_j$ to $\mathcal{D}_{min}$

Append $\bar{z}_j$ to $\mathcal{D}_{max}$

Use $\mathcal{D}_{min}$ and $\mathcal{D}_{max}$ to estimate $T(\hat{F}_{min}) = \underline{\beta}$ and $T(\hat{F}_{max}) = \bar{\beta}$

return $\bar{\beta}, \underline{\beta}$

While the procedures using the positive jackknife and the influence approach employ the same strategy, there are some advantages to using the derived influence function of an estimator. It has the potential to be much faster as there is no need to recalculate the statistic/functional many times. The relative computational speeds of these approaches is explored in our simulation studies in Section 3. The influence function also allows us to move past the situation where all data are categorical and include continuous covariates. Using the derived influence function one does not have to enumerate all possibilities of the missing data, instead one needs to maximize and minimize $\varphi(z)$. If an influence function of an estimator is bounded, then we can always identify $\bar{z}_i$ and $\underline{z}_i$. If it is unbounded then often by making some assumptions about the range of the continuous variables we can bound the influence function and then proceed with the approach. This same strategy is employed in robust regression (Rousseeuw and Leroy, 2005) and Huber's M-estimator (Huber, 1964).

Using influence functions helps us formalize the impact each incomplete observation could be having on the effect of interest, and doing so allows us to formulate an efficient strategy for estimating bounds of statistical parameters that take into account uncertainty due to missing data.

2.3 Theoretical Properties

2.3.1 Asymptotic Equivalence of Positive Jackknife to Influence Function

The jackknife, a leave one out variance estimation approach, is an asymptotic approximation to the influence function (Lehmann, 2004). In addition, it is straightforward to show that the positive jackknife also approximates the influence function. As a consequence, one can show that the two algorithms shown above are asymptotically equivalent. Start by rewriting the positive jackknife in functional notation

$$\begin{aligned} PJ(x_j) &= (n+1) (T_{n+1}(x_1, \dots, x_n, x_j) - T_n(x_1, \dots, x_n)) \\ &= (n+1) \left(T(\hat{F}_{(+j)}) - T(\hat{F}_n) \right) \\ &= \frac{\left(T(\hat{F}_{(+j)}) - T(\hat{F}_n) \right)}{\epsilon_n} \end{aligned}$$

where $\epsilon_n = \frac{1}{n+1}$ so that $\epsilon_n \rightarrow 0$ as $n \rightarrow \infty$. Next, decompose $\hat{F}_{(+j)}$ into two respective parts

$$\hat{F}_{(+j)} = \frac{1}{n+1} \left\{ \delta_{x_j} + \sum_{i=1}^n \delta_{x_i} \right\} = \frac{1}{n+1} \left\{ \delta_{x_j} + n\hat{F}_n \right\} = \frac{1}{n+1} \delta_{x_j} + \frac{n}{n+1} \hat{F}_n = \epsilon_n \delta_{x_j} + (1 - \epsilon_n) \hat{F}_n$$

Plugging the above into the functional form of the positive jackknife the influence function appears

$$\begin{aligned} PJ(x_j) &= \frac{\left(T(\hat{F}_{(+j)}) - T(\hat{F}_n) \right)}{\epsilon_n} \\ &= \frac{\left(T(\epsilon_n \delta_{x_j} + (1 - \epsilon_n) \hat{F}_n) - T(\hat{F}_n) \right)}{\epsilon_n} \\ &= \varphi_{T, \hat{F}_n}(x_j) \text{ as } n \rightarrow \infty \end{aligned}$$

Our approaches defined in Section 2 are asymptotically identical. Deciding between which of the two approaches to use is a choice of the implementer; both approaches should be concordant in reasonably sized samples. Should the analyst not desire to derive the closed form solution of the influence function, then the positive jackknife is simple to implement and only requires taking the difference of the statistic with and without adding new observation x_j into the sample. For most estimators the influence function is available in

standard literature, for more complicated estimators the influence function could be more difficult to derive. The algorithms may differ in their computational efficiency depending on the functional of interest.

2.3.2 Asymptotic Variance of the Bounds

Confidence intervals can be constructed for the estimates of the bounds generated by our algorithms. Since most previously published algorithms for estimating bounds in these settings are often complicated, instead of explicitly deriving asymptotic variances, a bootstrap of the procedure is recommended. Over many iterations a bootstrap provides a valid standard errors of these bounding estimates (Efron and Tibshirani, 1994). When procedures for estimating bounds take a long time to estimate, then it becomes computationally unrealistic to repeat these procedures 100's of times for a bootstrap. Luckily, not only is the influence function approach more feasible to bootstrap, it also fits naturally into estimating variances. This is because influence functions characterize the asymptotic properties of an estimator. To see how, first note that asymptotically linear estimators have the following representation:

$$\sqrt{n}(\hat{\beta} - \beta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \varphi_{T,F}(x_i) + o_p(1)$$

With $\varphi_{T,F}(x_i)$ such that $E[\varphi_{T,F}(x_i)] = 0$ and $E[\varphi\varphi^T]$ exists, is finite, and non-singular. The shorthand φ , assumes that the influence function uses the distribution F that generates the data indexed by the true parameter β . From the representation above the Central Limit Theorem applies,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \varphi_{T,F}(x_i) \xrightarrow{\mathcal{D}} N(0, E[\varphi\varphi^T])$$

and by Slutsky's theorem,

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{\mathcal{D}} N(0, E[\varphi\varphi^T])$$

Or equivalently, $\hat{\beta} \sim AN(\beta, E[\varphi\varphi^T]/n)$. Knowing this limiting distribution of our statistic $\hat{\beta}$ allows one to estimate the asymptotic variance with the empirical estimate (Boos and Stefanski, 2013):

$$\hat{E}[\varphi\varphi^T] = \frac{1}{n} \sum_{i=1}^n \left\{ \varphi_{\hat{F}}(x_i) \varphi_{\hat{F}}^T(x_i) \right\}$$

Where $\varphi_{\hat{F}}(x_i)$ indicates that the influence function has been estimated with our observed statistic $\hat{\beta}$. Since $Var(\hat{\beta}) = E[\varphi\varphi^T]/n$ then our estimate of the variance is $\hat{Var}(\hat{\beta}) = \hat{E}[\varphi\varphi^T]/n$. This reveals the convenience of our influence function approach. This influence estimate of the asymptotic variance relies on a sum of the i -th influence functions for each observation x_i .

Now let's consider our incomplete data setting. The goal is to find the asymptotic variance for $\underline{\beta}$ and $\bar{\beta}$, the lower and upper bounds of β identified by the algorithm. Since the influence contribution of each imputed value - e.g. $\underline{x}_j$ and $\bar{x}_j$ - is calculated as apart each iteration of the algorithm, estimating the asymptotic variance of the bounds adds no computational burden. One can see this by decomposing this empirical influence variance estimate in its observed and imputed components,

$$\hat{Var}(\underline{\beta}) = \frac{1}{n} \left\{ \frac{1}{n_{cc}} \sum_{i=1}^{n_{cc}} \left\{ \varphi_{\hat{F}}(x_i) \varphi_{\hat{F}}^T(x_i) \right\} + \frac{1}{n_{mis}} \sum_{j=1}^{n_{mis}} \left\{ \varphi_{\hat{F}}(\underline{x}_j) \varphi_{\hat{F}}^T(\underline{x}_j) \right\} \right\}$$

Here n_{cc} indicates the number of complete case observations and n_{mis} the number of incomplete observations. Above, the first summand is the variance computed from the complete case estimator in the first step of the influence algorithm. The individual j components of the second summand are identified during each iteration of the influence algorithm and can be stored. Thus, as the bounds are estimated, their variances are concurrently estimated. For the analogous estimated variance of the upper bound - i.e. $\hat{Var}(\bar{\beta})$ - simply replace $\underline{x}_j$ with $\bar{x}_j$.

2.4 Simulation Studies

2.4.1 Estimation of Upper and Lower Bounds

The numerical study herein aims to evaluate the finite sample properties for estimation of the lower and upper bounds of a sample statistic given all possible realizations of the missing data. First, two uncorrelated Bernoulli random variables X_2 , and X_3 are generated with success probabilities of 0.25, and 0.75 respectively. Next, covariate X_1 given X_2 are generated under the model $P(X_1 = 1|X_2) = \text{logit}^{-1}(-0.25 + X_2)$. Lastly, given these covariates, the outcome is generated using the following logit model:

$$P(Y = 1|\mathbf{X}) = \text{logit}^{-1}(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3)$$

Where β are sampled from a multivariate normal distribution on each simulation iteration. Next the missing data are created by using both a MCAR and a MNAR missing data generating mechanism. Under MCAR, for the outcome, Y , 5% of the values were deleted randomly, 10% from X_1 , 5% from X_2 , and 5% from X_3 . This results in at least 80% of observations that are complete cases, and at most 20% of observations with at least one missing value. Under MNAR, observations are deleted based on the unobserved variables according to the following scheme: $P(R_{Yi} = 0|Y_i) = 0.08 \times \mathbf{1}(Y_i = 1)$, $P(R_{X1i} = 0|X_{1i}) = 0.16 \times \mathbf{1}(X_{1i} = 1)$, $P(R_{X2i} = 0|X_{2i}) = 0.16 \times \mathbf{1}(X_{2i} = 1)$, and $P(R_{X3i} = 0|X_{3i}) = 0.04 \times \mathbf{1}(X_{3i} = 1)$.

Three approaches are compared to the true upper and lower bounds of each simulated dataset. The true lower and upper bounds of the statistic are obtained by fitting the logit model on all possible permutations of the missing data. This is possible in this particular simulation, as there are roughly 2^{20} possible permutations, which is large but realistic to compute. The positive jackknife and influence function approaches are applied as described in Section 2. Lastly, we applied a simulation approach described in [Horowitz et al. \(2003\)](#). Random samples of sizes $K = 1000, 10000$, and 100000 are drawn from the uniform distribution over the set of logically possible values and imputed for the missing values. Results in terms of bias and computational efficiency of the various approaches are in Tables 2.1 and 2.2.

Table 2.1 MCAR Simulation Results

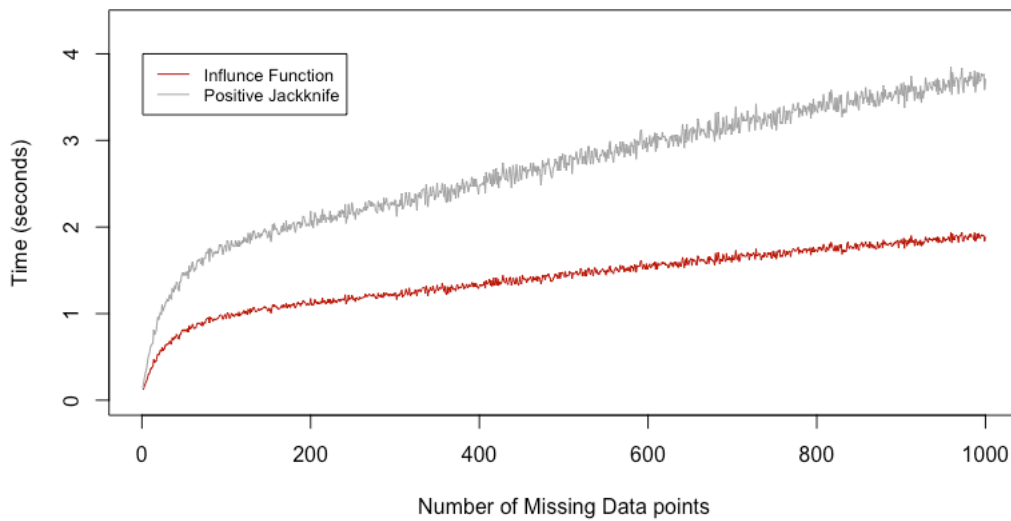
Method	Bias Lower Bound	Bias Upper Bound	Time
Influence Function	0.002	-0.0008	0.03 sec
Positive Jackknife	0.002	-0.0008	0.06 sec
Random Permutations ($K = 100,000$)	0.013	-0.011	3.36 min
All Possible Permutations	0	0	35.15 min

Table 2.2 MNAR Simulation Results

Method	Bias Lower Bound	Bias Upper Bound	Time
Influence Function	0.005	-0.0000007	0.03 sec
Positive Jackknife	0.004	0	0.06 sec
Random Permutations ($K = 100,000$)	0.017	-0.016	2.73 min
All Possible Permutations	0.004	0	36.06 min

2.4.2 Computational Efficiency

The computational efficiency of the positive jackknife to the influence function approach were compared using the simulation design above but with a larger sample size of $N = 10,000$. The number of missing values was varied from 1 to 1,000. For each of these 1,000 scenarios we simulated 10 datasets and averaged the computation time across each. The results are shown in Figure 2.1.

**Figure 2.1** Computation Times

2.5 Data Application

The bounding approach was applied to a real world dataset which was collected to assess the effect of the drug indinavir in a dataset from $N = 1151$ patients living with HIV (Hosmer et al., 2008). The outcome of interest was time to either AIDS progression or death. For a fraction of participants, the outcome was censored and consequently a time to event framework was adopted. Five explanatory variables were considered, three were categorical and two were continuous. The exposure of interest, which is to be bounded, was taking an antiretroviral (ARV) therapy combination that included indinivir vs a combination therapy without indinavir. A detailed description of the variable codings can be found in Table 2.3.

Table 2.3 Variable Descriptions

Variable	Description	Codes/Values
time	Time to AIDS or death	Months
censor	Event Indicator	1 = AIDS defining diagnosis or death 0 = censored
trt	Treatment Indicator	1 = Combination ARV with indinavir 0 = Combination ARV without indinavir
cd4	Baseline CD4 Count	Cells/ml
karnof	Karnofsky Performance Scale	100 = Normal 90 = Some symptoms 80 = More Symptoms normal activity with effort 70 = Normal Activity not possible
age	Age at Enrollment	Years
female	Female Sex	0 = Male 1 = Female

Given the time to event nature of the dataset, the following Cox proportional hazards model was postulated:

$$\lambda(t|X_i) = \lambda_0(t)\exp(\beta_1\text{trt}_i + \beta_2\text{cd4}_i + \beta_3\text{age}_i + \beta_4\text{female}_i + \beta_5\text{karnof}_i)$$

Where the primary parameter of interest is $\exp(\beta_1)$ for the hazard ratio comparing the risk of AIDS progression/death between the treatment regimen with indinavir vs without indinavir while controlling for baseline cd4, Karnofsky performance scale, age, and sex. Given the model above, the influence bounding approach can be used to estimate upper and lower bounds for the hazard ratio that account for the missing data.

Data were deleted to create missing values, as the original dataset contained no missing observations. This approach was chosen over a dataset with a particular amount of missing data because it allowed the assessment of the effect that varying percentages of missing data have on the bounds. Missing data were induced in each covariate using a MNAR deletion mechanism. This mechanism took covariates in the bottom quartile of their observed distribution and deleted 20% of them which resulted in 5% missing data per covariate. As alluded to above, this percentage was varied in later analyses.

To implement the influence function bounds, the influence function for the Cox proportional hazards model was used. Before writing down the influence function explicitly, as derived by Reid and Crépeau (1985), first recall that the Cox proportional hazards model maximizes the partial likelihood which is equivalent to solving the estimating equation

$$\sum_{i=1}^n \delta_i \left\{ x_i - \left\{ \frac{\sum_{\mathcal{R}_i} x_j \exp(x_j^T \beta)}{\sum_{\mathcal{R}_i} \exp(x_j^T \beta)} \right\} \right\} = 0$$

Where δ_i is the censoring indicator, x_i is a vector of covariates, and $\mathcal{R}_i$ denotes the risk set for observation i . Which leads to a point estimate for $\hat{\beta}$. To find the asymptotic variance for $\hat{\beta}$ one uses the negative inverse of the second derivative of the partial likelihood

$$H(\beta) = \frac{1}{n} \sum_{i=1}^n \delta_i \left[\left\{ \frac{\sum_{\mathcal{R}_i} x_j x_j^T \exp(x_j^T \beta)}{\sum_{\mathcal{R}_i} \exp(x_j^T \beta)} \right\} - \left\{ \frac{\sum_{\mathcal{R}_i} x_j \exp(x_j^T \beta)}{\sum_{\mathcal{R}_i} \exp(x_j^T \beta)} \right\} \left\{ \frac{\sum_{\mathcal{R}_i} x_j \exp(x_j^T \beta)}{\sum_{\mathcal{R}_i} \exp(x_j^T \beta)} \right\}^T \right]$$

The influence function evaluated at observation $z_i = (t_i, x_i, \delta_i)$ is then,

$$\varphi_{\beta}(z_i) = H^{-1}(\beta) \delta_i \left\{ x_i - \left\{ \frac{\sum_{\mathcal{R}_i} x_j \exp(x_j^T \beta)}{\sum_{\mathcal{R}_i} \exp(x_j^T \beta)} \right\} \right\} + H^{-1}(\beta) C_i(\beta)$$

The first term of the influence for the Cox model resembles the standard influence function of a M-estimator. Only uncensored observations contribute to this component. The second component reflects the influence that observation i has on the risk set, where

$$C_i(\beta) = \exp(x_i^T \beta) \left(\sum_{t_j \leq t_i} \delta_j \left[\sum_{\mathcal{R}_j} z_k \exp(z_k^T \beta) / \left\{ \sum_{\mathcal{R}_j} z_k \exp(z_k^T \beta) \right\}^2 \right] - z_i \sum_{t_j \leq t_i} \delta_j \{ 1 / \sum_{\mathcal{R}_j} \exp(z_k^T \beta) \} \right)$$

Using this influence function, the influence function bounding approach was applied as described in Section 2.2.3. As required by the approach, continuous variables must be restricted to plausible ranges. CD4 counts on the interval $[0, 320]$ and ages on the interval $[15, 75]$ were considered. These ranges were consistent with ranges of the observed data. The results are presented in Table 2.4. The hazard ratio estimated from the complete cases was 0.54 suggesting that indinivir decreases time to AIDS progression our death among the complete cases. This was slightly higher than the original analysis with no missing data (HR = 0.51, 95% CI $[0.34, 0.79]$). The best and worst possible imputations identified by the influence function approach reveal that the possible observed hazard ratios could have ranged between 0.38 and 0.64 . One can interpret the 95% confidence interval of the bounds as follows: Under the extreme case scenarios of the missing data, we are 95% confident that the the hazard ratio comparing the risk of AIDS progression/death between the treatment regimen with indinavir vs without indinavir is between 0.25 and 0.98.

Table 2.4 Cox Model Hazard Ratio Bounds

Covariate	Hazard Ratio	(Lower Bound, Upper Bound)	95% Confidence Interval
Indinavir	0.54	(0.38, 0.64)	[0.25, 0.98]

In addition to the analysis above with an overall missing data rate of 5% for each covariate, other missing data rates ranging from as low as 1% to as high as 10% were considered in order to understand the impact this would have on the width of our bounds. As anticipated, the width of the interval for the bounds increased with the amount of missing data. For these cases, the estimate of the upper bound for the hazard ratio was always well below 1, although the confidence interval for the upper bound was above 1 from 6% missing data and thereafter. Plots of the complete case estimated hazard ratios (solid dots), bounds interval (solid

line), and 95% confidence interval for the bounding interval (dotted lines) are presented in Figure 2.2.

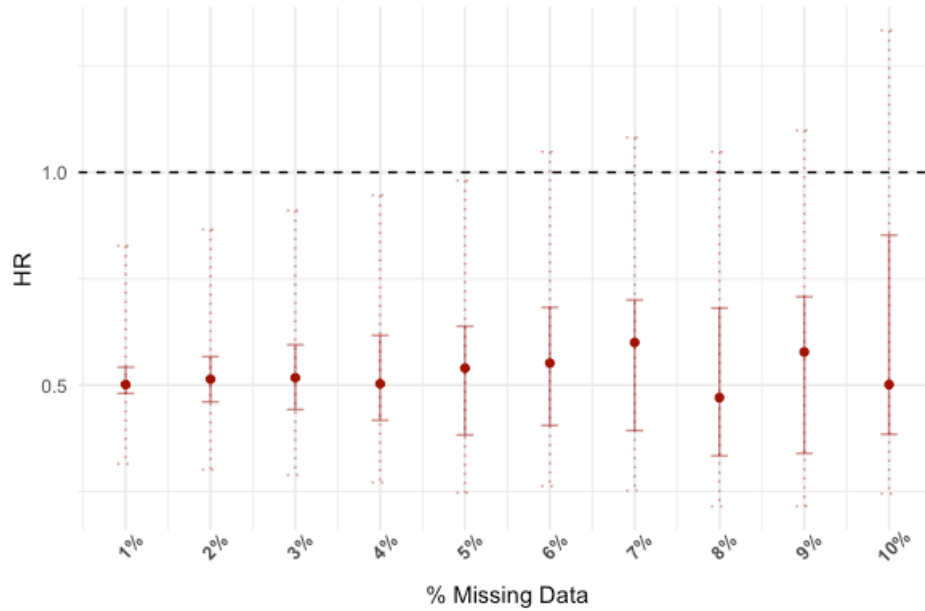


Figure 2.2 Effect of Missing Data Rate on Bounds Interval

2.6 Conclusion

Leveraging influence functions can provide a fast and accurate method for estimating bounds that account for potential possibilities of the missing values. Estimating these bounds makes no assumptions about the missing data generating mechanism (i.e. MCAR, MAR, MNAR etc). Untestable assumptions such as MAR allow estimation but may be unrealistic in certain contexts. For MNAR models even stronger assumptions must be made for identification of parameters such as the presence of a shadow variable (Miao and Tchetgen Tchetgen, 2016) or that the missing values occur below some quantile (Ocampo et al., 2019). By leveraging influence functions, one can use this approach for estimating bounds that consider the most extreme possibilities of the missing data. In addition, using the influence function is efficient. This paper also demonstrates that for more complicated estimators where an influence function is difficult to derive or does not exist, then the positive jackknife approach can be implemented which is asymptotically equivalent to the influence function approach.

This approach intuitively adapts itself for the incorporation of continuous covariates. Influence functions can be functions of continuous covariates, so instead of having to enumerate all possible realizations of

a continuous variable (which is infinite), one can simply maximize the influence function with respect to these variables, which often takes place at the extremes in linear models. If not at the extremes, modern optimization techniques such as simulated annealing could be used to maximize the influence function over a range of plausible values. As a consequence, one does have to restrict the range of the continuous variable. Other bounding strategies require that all data be categorical and coarsen continuous variables to a fine grid for computational purposes (Horowitz et al., 2003) (Zaffalon, 2002).

I recommend that any analyst who confronts missing data begin by first implementing these bounds. This influence bounding approach will take only a matter of seconds for most reasonably sized datasets. If the bounds are tight, then the results are robust to the missing data and one may use this fact to justify the results from a complete case analysis. This is advantageous for scientists without the statistical skills to perform a technical sensitivity analysis. If the bounds are wide, then the analyst may have no choice but to make untestable assumptions about missing data. If bounds were reported on all studies with missing data, it would increase transparency among scientists regarding uncertainties due to missing data.

Acknowledgements

Sincere thanks to Kaitlyn Cook and Zack McCaw for fruitful discussions and comments.

3

Extrapolating Hybrid Estimation to Districts with Missing Probability Survey Data

Alex Ocampo¹, Joseph Valadez², Bethany Hedt-Gauthier^{1,3}, Marcello Pagano¹

¹ Department of Biostatistics, Harvard University

² Department of International Public Health, Liverpool School of Tropical Medicine

³ Department of Global Health and Social Medicine, Harvard Medical School

Abstract

Administrative health data have inherent biases when used to estimate district level prevalences. Statistical inferences using such data can be improved upon when the data are combined with random probability survey data to form hybrid estimators. Sometimes, probability survey data are only available for some, but not all districts. We present methods for extending estimations to all districts. We accomplish this hybrid extrapolation by estimating denominators using the survey data and then building a statistical model using population estimates from different sources as covariates to estimate denominators in adjacent districts. By dividing administrative numerators by the model-estimated denominators we obtain extrapolated hybrid prevalence estimates. Framing the problem in the Bayesian paradigm guarantees estimated prevalence rates that fall within the appropriate ranges and allows a convenient way to incorporate a sensitivity analysis. Using the methodology, we obtain estimates of polio vaccination rates in all 77 communes of Benin. Probability survey data available on 19 communes are leveraged to formulate estimates for the 58 communes with only administrative data from a National Immunization Day. Polio vaccination coverage estimates fall within ranges consistent with those from the probability surveys (mean: 0.87, sd: 0.09). Extrapolated hybrid estimation can now be used for national assessments of primary health coverage and reduce expenses of national household surveys.

3.1 Introduction

Disparities and unequal access to resources exist across all aspects of the global healthcare landscape, and access to reliable health care data is no exception. The lack of reliable health data in low- and middle-income countries is well documented (Ndabarora et al., 2014) (WHO, 2003). Now more than ever, the global community recognizes the essential need for quality data from all countries, not just for those with the resources to aggregate and validate them. One promising trend is the continued adoption of electronic Health Management Information Systems (HMIS) in developing countries (Lippeveld et al., 2000). As a result, administrative data aggregated as a part of routine governmental reports, public health outreach initiatives, and healthcare visits have improved their accuracy (Chaudhry et al., 2006) (Nisingizwe et al., 2014). This advance can lead to more dependable estimates of public health indicators which are crucial for evidence-based public health decision making (Haux et al., 2007) (AbouZahr and Boerma, 2005). However, methods for validating their accuracy are still needed.

Despite recent improvements, administrative data still have deficiencies. Most notable is that they are often convenience samples, and thus generalizations to the general population are questionable without adjustments from alternative data sources (Hedt and Pagano, 2011). Within this biased sample, there are additional structural, behavioral, and technical issues with administrative data in under-resourced countries. Lack of clear policy for health information systems makes designing, aligning information, and integrating HMIS difficult (WHO, 2003)(Lima et al., 2009) (Smith et al., 2008). Absence of collaboration and feedback between information technology experts and end users can lead to HMIS that fail to fulfill their intended protocols (Odhiambo-Otieno, 2005). Pressure for good results tied to resource allocation can lead to false reporting (Mutemwa, 2006). Lack of resources, management, training, and skills can be a hindrance at the data entry level (WHO, 2003) (Mutemwa, 2006).

To combat these limitations, the WHO released a set of guidelines for developing countries in the hopes of improving the reliability of administrative data (WHO, 2003). The guidelines specify a series of checks that can be implemented to assess data quality and flag inconsistencies in administrative data. Without such data quality checks, basing national health planning on administrative data could be misguided (Nisingizwe et al., 2014).

A viable strategy to approximate an accurate measure of health indicators is to utilize all available data.

Optimal statistical methods can synthesize knowledge from different sources, taking advantage of their disparate strengths. One approach to form more accurate estimators is to combine administrative data with random surveys and leverage both of their statistical properties associated with randomness. Probability surveys are representative of the population because they employ randomization to sample from either the whole population or within well-defined strata (Kish, 1995). Hedt and Pagano (2011) demonstrated that the operating characteristics of such estimators reduced bias and improved efficiency. Their work sets the stage for how incorporating less expensive random samples into estimates can correct biases inherent to using administrative data alone.

Similarly, Jeffery et al. (2018) presented an approach called hybrid prevalence estimation which combines large-scale administrative data from a Child Health Day with Lot Quality Assurance Sampling (LQAS) survey data. LQAS has been used widely throughout the developing world to assess disease prevalence, immunization coverage, and behavioral health indicators (Robertson and Valadez, 2006) (Pagano and Valadez, 2010) (Valadez, 1991). The statistics from both administrative and LQAS survey data are combined using weights estimated using the relative efficiency of the statistics to formulate one final hybrid estimate and standard error. To demonstrate the utility of their approach, they estimated vitamin A supplementation (VAS) and polio vaccination rates in children from Benin. We utilize the same data set to demonstrate an extension of their approach to situations where the survey data is not available at all locations.

While hybrid prevalence estimation provides suitable estimates in districts in which random samples have been conducted, it is not always feasible to perform probability surveys in all administrative areas. We thus have a missing data problem. There is to date no hybrid estimator solution for areas with administrative data, but without a corresponding random sample with which to the anneal administrative data. Herein, we propose a solution that leverages regions with both probability surveys and administrative data to posit a model that can extrapolate estimates to an entire catchment area without conducting additional probability surveys.

This paper is organized as follows. Section 3.2 gives an overview of the available data. Section 3.3 illustrates methods for constructing estimates for all geographic areas including framing the problem from the Bayesian perspective. Section 3.4 uses the methods to estimate polio vaccination rates and vitamin A supplementation in Benin. Section 3.5 concludes the article with a brief discussion.

3.2 Data Overview

To illustrate our innovative approach for estimating intervention coverage rates for each area in a given country, we consider data on polio vaccinations and vitamin A supplementation given to children less than five years of age in Benin during a Child Health Day (JJ et al., 2017). The goal is to estimate coverage rates, i.e., the percentage of children receiving the intervention, in each of the 77 communes which constitute the geographical administrative units in Benin. Although these data have been described in detail elsewhere (Jeffery et al., 2018), we give a brief overview in this section to illustrate the fundamental issues and the need for new methodology. We leverage two types of data available for communes in Benin: (1) small sample probability surveys estimating the percentage of children receiving the intervention and (2) large administrative data counting the number of children receiving a polio vaccination and VAS during the Child Health Day (analogous to a National Immunization Day). While administrative data are available for each of the 77 communes in Benin, probability surveys are only available for 19 communes. Having widespread administrative data and incomplete probability survey data has been seen elsewhere, notably Madagascar (Cisse M, 2016). This paper postulates the question: Using the above two sources of data, is it possible to improve upon estimates based on only the administrative data in areas without probability surveys?

3.2.1 The Probability Surveys

Probability surveys were conducted in 19 of Benin's 77 communes between Nov 16-20, 2015. Selected communes were divided into administrative units called supervision areas (SA). Each commune had five SAs, in each of which 19 children were surveyed, for a total of 95 children surveyed per SA. The only exception was the commune of Ouinhi, which had four SA, resulting in only 76 children surveyed. The total sample size was 1,786 respondents. The sampling scheme was designed in the context of a LQAS and has been published previously in detail (Robertson and Valadez, 2006) (Turner et al., 1996) (Davis and Valadez, 2014). Structured questionnaires were used to measure multiple health indicators, including both polio vaccination and VAS in two age strata, children 6-11 months and 12-59 months of age. The final commune-level coverage rate is not a simple average of the SA, but a weighted average with the weights proportional to estimated population size.

Benin aggregates communes into departments. The 19 sampled communes came from three areas of

Benin: Nine communes from the departments of Alibori and Atacora in the Northeast, all nine communes of the central department of Zou, and one commune near the coast and capital. The locations of these communes and probability survey estimates are mapped in Figure 3.1. Since this set of communes was not chosen by random chance those communes with missing probability surveys cannot be classified as missing completely at random (Rubin, 1976).

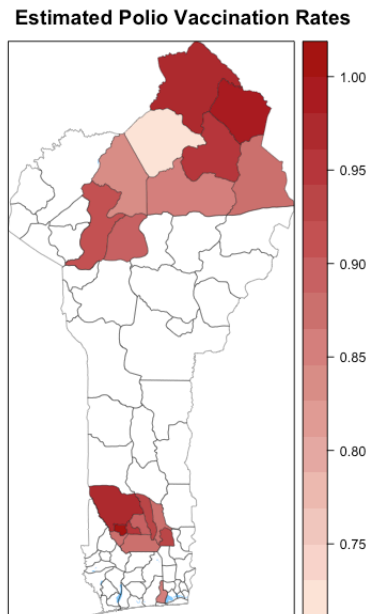


Figure 3.1 Communes with Probability Survey Data

3.2.2 Administrative Data

The Ministry of Health (MOH) in Benin collects data on the administration of VAS and polio vaccination to every eligible child during a Child Health Day (JJ et al., 2017) which recurs every six months. We focus on data from the 2015 initiative. Numbers of children served are recorded on tally sheets for each of Benin's 77 communes. Tally sheets differ from typical HMIS data, which are aggregated from patient registration books at specific health facilities. These tallies could be useful, should the resulting number of children receiving services during these campaigns be indicative of VAS and polio vaccination rates in a given commune. In order to estimate the proportion of children receiving services one must divide this administrative numerator data by the denominator: the number of children in the population. However, since the last census in Benin was conducted in 2013 this denominator can only be estimated. These denominator estimates either use

the numerator from six months ago (Jeffery et al., 2018) or extrapolate old census data adjusting for the population growth rate (United Nations Development, 2017). These approaches are sometimes questionable, especially when they lead to coverage rates that are larger than 100%, which was the case in Jeffery 2018. This motivates our subsequent approach for estimating denominators that leverages the probability survey data.

3.3 Methodology

3.3.1 Estimating Denominators

Simple Estimation of Denominators and their Variances. We can use the probability survey data to estimate denominators using the relationship:

$$p = \frac{n}{d}$$

where for each commune, p is the probability of polio vaccination, n is the number of children vaccinated, and d is the number of children in that administrative region. When estimating, for n , we use the number of children vaccinated from the administrative data, available in all communes; however, we have no estimates of d due to the lack of 2015 census data. Because in some communes we have a reliable estimate of p from the probability surveys, we can estimate d for these communes by:

$$\hat{d} = \frac{n}{\hat{p}}$$

Here, $\hat{p}$ is the probability survey estimate, which in our example in Benin, are available for 19 communes. We can calculate $\hat{d}$ for these 19 communes, and subsequently use them in conjunction with the n from the administrative data.

Since we know the variance of $\hat{p}$ from the probability surveys, we can obtain estimates of the variances of $\hat{d}$ by the delta method. First, note that the LQAS estimates are asymptotically normal as the sample size $N \rightarrow \infty$ by the central limit theorem,

$$\sqrt{N}(\hat{p} - p_0) \rightarrow N(0, \sigma_p^2)$$

where σ_p^2 is the variance. We obtain an estimate of this variance $\hat{\sigma}_p^2 = \hat{p}(1 - \hat{p})$ from the probability surveys. As shown above, $\hat{d}$ is a simple transformation of $\hat{p}$ where the first derivative exists and is non-zero. Define this transformation as $g(\hat{p}) = \frac{n}{\hat{p}}$. Then by the delta method,

$$\sqrt{N}(g(\hat{p}) - g(p_0)) \rightarrow N(0, \sigma_p^2 [g'(p)]^2)$$

where $[g'(p)]^2 = \left[\frac{\partial}{\partial p} \frac{n}{p} \right]^2 = \frac{n^2}{p^4}$. Therefore, we can estimate the variance of these denominators $\hat{\sigma}_d^2 = \hat{\sigma}_p^2 \frac{n^2}{\hat{p}^4}$.

Estimating Denominators with Multiple Health Indicators. Above we have considered the case of one health indicator $\hat{p}$ being estimated by a probability survey associated with one numerator n for that same indicator provided by administrative data. It may be plausible multiple indicators are measured by both probability survey and administrative data. For example, in Benin we present data on two health indicators, polio vaccination and VAS coverage. While the coverage rates for these indicators do differ across communes, the denominator we are estimating in a commune, e.g. the number of children age 12-59 months, is the same for both indicators. Therefore, we can combine data across health indicators to improve estimates of the denominators. Consider the following estimated denominators:

$$\hat{d}_1 = \frac{n_1}{\hat{p}_1}$$

$$\hat{d}_2 = \frac{n_2}{\hat{p}_2}$$

where the 1 subscript represents polio vaccination and the 2 subscript represents VAS coverage. We can average these estimated denominators to get a global estimate of the denominators:

$$\hat{d} = \frac{\hat{d}_1 + \hat{d}_2}{2}$$

Alternatively, we can calculate an inverse variance weighted denominator estimate. Our original probability

survey estimates $\hat{p}_1$ and $\hat{p}_2$ come with estimated variances. When these estimates are combined with numerators to form the estimates of the denominators $\hat{d}_1$ and $\hat{d}_2$, then using the delta method above we can obtain valid variance estimates for the denominators. We would expect an estimated denominator with a smaller variance to be a more reliable estimate. In that case, giving it a larger weight when averaging the denominators estimated by multiple health indicators from the probability surveys makes sense. For data with $k \in 1, \dots, K$ indicators, consider the weights $w_k = \frac{1}{\hat{\sigma}_{\hat{d}_k}^2}$. Then the inverse variance weighted denominator estimate is:

$$\hat{d} = \frac{\sum_{k=1}^K w_k \hat{d}_k}{\sum_{k=1}^K w_k}$$

Not only do we provide greater weight to more reliable denominator estimates, but in addition the inverse-variance weighted average has the least variance among all weighted averages of unbiased estimators. While the equations above demonstrate estimating denominators for the case of Benin which has probability survey data for 2 health indicators, this approach can be extended to include more estimates. In the next section, we leverage these denominator estimates and their estimated variances to estimate denominators for communes without probability surveys.

3.3.2 Predicting Denominators

The above strategy for estimating denominators — the number of children in each commune — allows us to construct an unbiased estimate of these denominators in the communes for which we have probability surveys. Our goal now is to use this information to extrapolate denominator estimates for the communes without probability surveys. Then by combining these estimated denominators with the true numerators, we can get an estimate of the intervention coverage rates in each commune. Those familiar with Laplace's famous work using ratio estimators to estimate the population of France in 1802 (Cochran, 1978) will note similarities. Indeed, our approach makes a similar assumption, that the ratio of the administrative numerator to polio vaccination coverage is similar between communes with and without probability surveys (Sen, 1993). However, we take the additional steps of estimating denominators as above and considering population size covariates known for all communes.

We estimate denominators in the 58 communes that do not have probability surveys by first establishing

a regression relationship within the 19 communes that do have probability surveys, and then use this model predictively in the 58 communes. Covariate data and the modeling assumption can be as large and complex as one wishes, we keep our description intentionally general for this section and demonstrate a detailed application using the Benin data in the next section. Consider fitting the following regression model:

$$\hat{d}_j = \mathbf{X}_j \beta + \epsilon_j \quad (3.1)$$

where $\hat{d}_j$ are the estimated denominators for the $j = 1, \dots, 19$ communes with probability surveys, $\mathbf{X}_j$ is a design matrix containing covariate information, and ϵ_j is the residual term where $E[\epsilon_j] = 0$ and $Var[\epsilon_j] < \infty$. This yields estimates, $\hat{\beta}$. Since the $\mathbf{X}_j$ area available for all communes, we can then apply the formula,

$$\hat{d}_j = \mathbf{X}_j \hat{\beta} \quad (3.2)$$

for $j = 20, \dots, 77$ the communes without probability surveys, because $\mathbf{X}_j$ is available on all communes. Variance estimates can be obtained in the standard way. Then by combining these predicted denominators with the known numerators we can get an estimates of the $\hat{p}_j$ for communes $j = 20, \dots, 77$.

3.3.3 Bayesian approach

One drawback of the approach above is that the linear model postulated for the denominator above does not guarantee that the estimated denominators be larger than the known numerators. As a consequence, it is possible to obtain estimates of $\hat{p}_j > 1$; which, we know would be incorrect. If this is the case, we could use the ad hoc estimate, $\hat{p}_j = 1$, which has no other justification than that it does not exceed one. By taking a Bayesian approach we can rectify this problem and benefit in other ways as well. Framing the problem in the Bayesian approach offers three main advantages:

1. Using probability distributions explicitly admits to uncertainty about $\hat{p}_j$ for the communes without hard data
2. Estimates for $\hat{p}_j$ are guaranteed to be less than 1 by truncating the posterior distribution
3. A sensitivity analysis can be easily incorporated into the framework

The proposed Bayesian approach picks up where the previous approach ends its inference on the denominators. The next step would be to postulate that the denominators d_j for $j = 20, \dots, 77$, are random variables

centered at the respective predicted estimates with variance consistent with the out of sample prediction interval ($\sigma_{\hat{d}_j}^2$). With the d_j independent, we can take draws from our posterior predictive distribution using:

$$d_j \sim N(\hat{d}_j, \sigma_{\hat{d}_j}^2)$$

This distribution might quantify our uncertainty about the estimated denominator thereby satisfying advantage (1) of the Bayesian approach listed above. Dividing the distribution by the numerator leads to the distribution on the coverage probability p_j as well.

We can refine our model by truncating the probability density of d_j to ensure that estimates of p_j are less than 1. If $p^*(d_j)$ is the density of constructed above for the denominator in commune j then we truncate by the numerator as follows:

$$p(d_j) = \frac{p^*(d_j)\mathbf{1}(d_j > n_j)}{\int_{n_j}^{\infty} p^*(\theta)d\theta}$$

This ensures the proper support for the density function, and thus guarantees an estimate of p_j less than 1. We call this 'truncation by reality' since we know for a fact the denominator cannot be less than the number of children vaccinated in a particular commune.

The last listed advantage, (3), of the formulating this problem in the Bayesian framework, allows us to easily conduct a sensitivity analysis by shifting the posterior predictive distributions in a systemic way. Suppose we have the probability density of $p(d_j)$. This estimate may be too high or too low, but by introducing sensitivity parameters (Daniels and Hogan, 2008) we can shift our Bayesian distributions and see how our inferences might change. Consider the sensitivity parameters Δ_{1j} for the mean and Δ_{2j} for the variance in each commune j . Then we can easily alter the posterior predictive density of the coverage as follows:

$$d_j \sim N(\hat{d}_j + \Delta_{1j}, \Delta_{2j}\sigma_{\hat{d}_j}^2)$$

In addition, sensitivity to the numerator in our specific context can also be incorporated by introducing a third sensitivity parameter Δ_{3j} . We can vary our choices of these sensitivity parameters, perform the truncation using $n_j \times \Delta_{3j}$, and conduct inferences as above. Such a sensitivity analysis is important especially when the model is based on data that may not be representative of the country as a whole. Adjusting Δ_{3j} can prove useful in considering bias in our administrative numerator.

3.4 Application to Benin Data

Consider estimating polio vaccination and VAS coverage among children at the commune level in Benin in 2015. There are two populations of interest, defined by age group: 6-11 months and 12-59 months each having independent probability samples and tally sheets. We wish to leverage the survey data from 19 communes to the 58 communes for which we only have administrative data. While analyses were conducted for both indicators in both populations, for brevity we present here the results from polio vaccinations for the 12-59 month old children and include the remainder of the results in the Appendix.

For covariates we utilize population estimates for each commune from the 2013 census (INSAE, 2013) as well as previously published estimates of the denominator (Jeffery et al., 2018). Both of these variables are strongly associated with our estimated denominators $\hat{d} = \frac{n}{\hat{p}}$ (Figure 3.2).

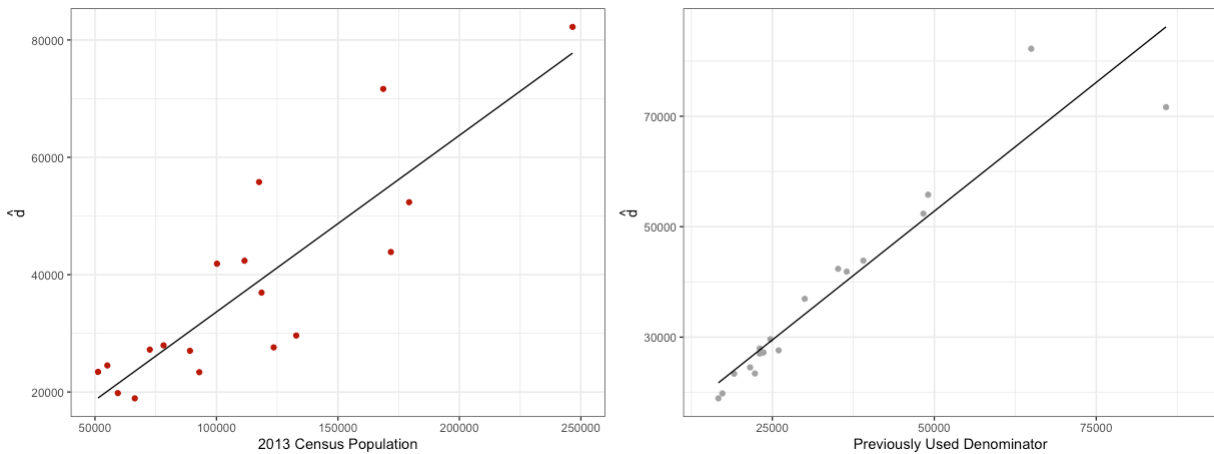


Figure 3.2 Model Covariates vs. Estimated Denominator

Using these covariates we fit the regression model as defined in (3.1) using the inverse variance weighted estimate of the denominator as our outcome variable. We fit this model on the 19 communes that had probability surveys for which we have an unbiased estimate of the denominator. This model had an adjusted $R^2 = 0.9452$ and variance of the error term was estimated to be $\sigma = 4479$. Using the covariates from the communes without probability surveys this model allows us to estimate denominators for these communes. Since covariates are available for all communes, we can use this model to extrapolate estimates to communes without probability surveys for which hybrid estimation is not possible.

3.4.1 Initial results

The estimated polio vaccination rates and corresponding confidence interval for children ages 12-59 months in each commune are shown in (Figure 3.3). There is an obvious issue with these results: The confidence intervals often extend above 1. In fact, a few of the point estimates are even above 1, which is of course not consistent with reality. In our model, 6 out of 77 communes have point estimates are above 1. This is an improvement compared to previously published attempts to estimate these denominators which led to 58 of 77 communes with coverage estimates above 1. Correcting for this violation of having estimates and confidence intervals above 1 motivates taking a Bayesian approach.

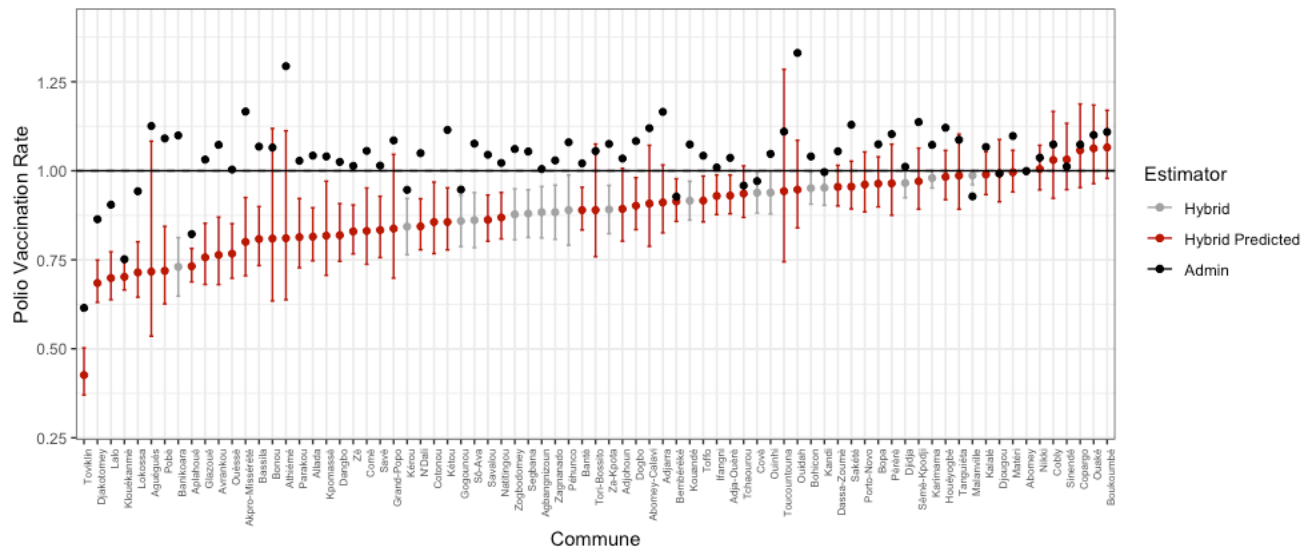


Figure 3.3 Children's Polio Vaccination Coverage Estimates (12-59 Months)

Six communes have confidence intervals noticeably larger than the other 71 communes. Such small communes fall outside the experimental region defined by the communes within which we had probability surveys, and thus have large predictive confidence intervals.

3.4.2 Bayesian Approach

The posterior medians and resulting 95% credible intervals as a result of the Bayesian formulation for the same polio vaccination rates above are shown in (Figure 3.4). This figure confirms that it is impossible to obtain either point estimates or intervals above 1 using this approach. In addition, estimates are temporally consistent with those from the non-Bayesian approach.

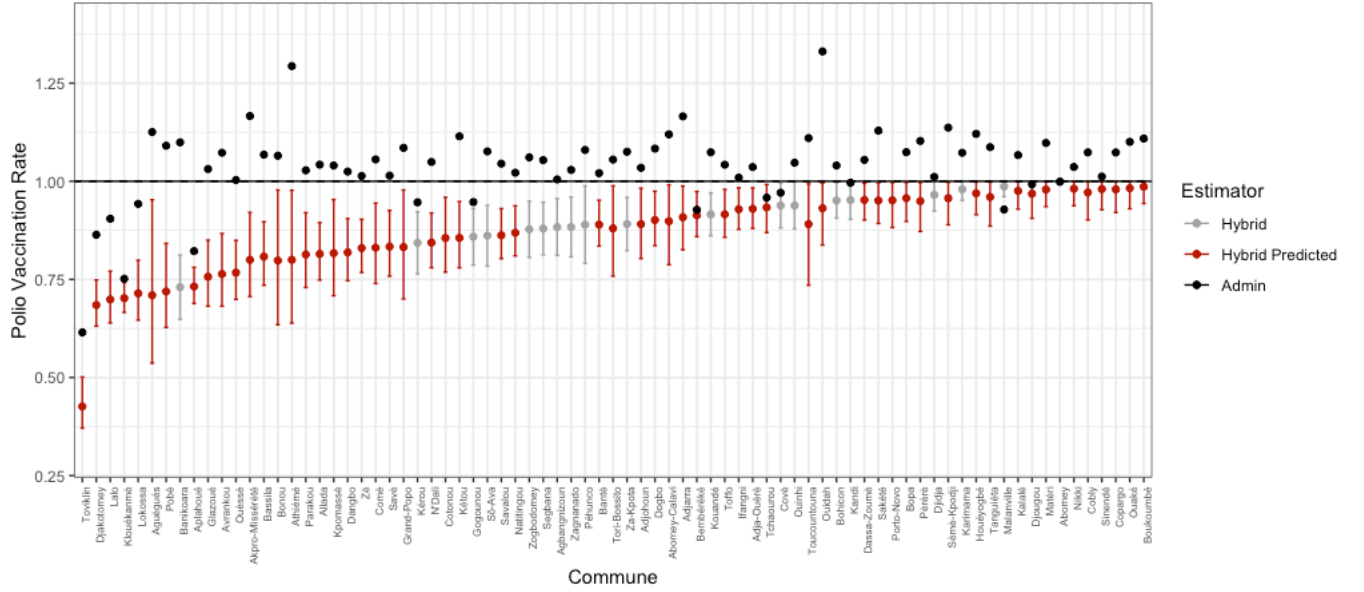


Figure 3.4 Children's Polio Vaccination Coverage Bayesian Estimates (12-59 Months)

3.4.3 Bayesian Sensitivity Analysis

In Section 3.3.3, we describe how we can incorporate a sensitivity analysis into our approach by introducing the sensitivity parameters Δ_1 , Δ_2 and Δ_3 . These sensitivity parameters are each J dimensional, where J is the number of administrative regions for which we are extrapolating results. We therefore write Δ_{kj} to denote the k th sensitivity parameter in commune j , although in subsequent sensitivity analysis, we will consider a constant value for each entry in Δ_k across all communes.

We now describe and present results from a sensitivity analysis of the administrative numerator in the communes without probability survey data. This corresponds to tuning the sensitivity parameter Δ_3 . The approach taken for Δ_3 is analogous to that of investigating the sensitivity of the estimated denominators (Δ_1) and the variance of their estimation (Δ_2). Note that when $\Delta_1 = 0$, $\Delta_2 = 1$, and $\Delta_3 = 1$ this indicates the communes used to fit the regression model are conditionally exchangeable to those on which predictions are made given the covariate data. This exchangeability is comparable the missing at random (MAR) assumption.

Recall that when $\Delta_3 \neq 1$ we are assuming that our administrative numerator is biased and considering deviations from the observed data. The sensitivity parameter Δ_3 has a multiplicative effect on our observed administrative numerator, i.e. $n_j \times \Delta_{3j}$. By using the numerator adjusted by Δ_3 we can proceed with the

posterior truncation and transformation in the same way as described above.

We considered scenarios where the observed administrative numerator n_j was 5% and 10% biased in either direction (i.e. -10%,-5%,5%,10%). This corresponds to choosing values of the sensitivity parameter in the set $\Delta_3 \in \{0.9, 0.95, 1.05, 1.1\}$. The resulting estimates and confidence intervals are shown in Figure 3.5:

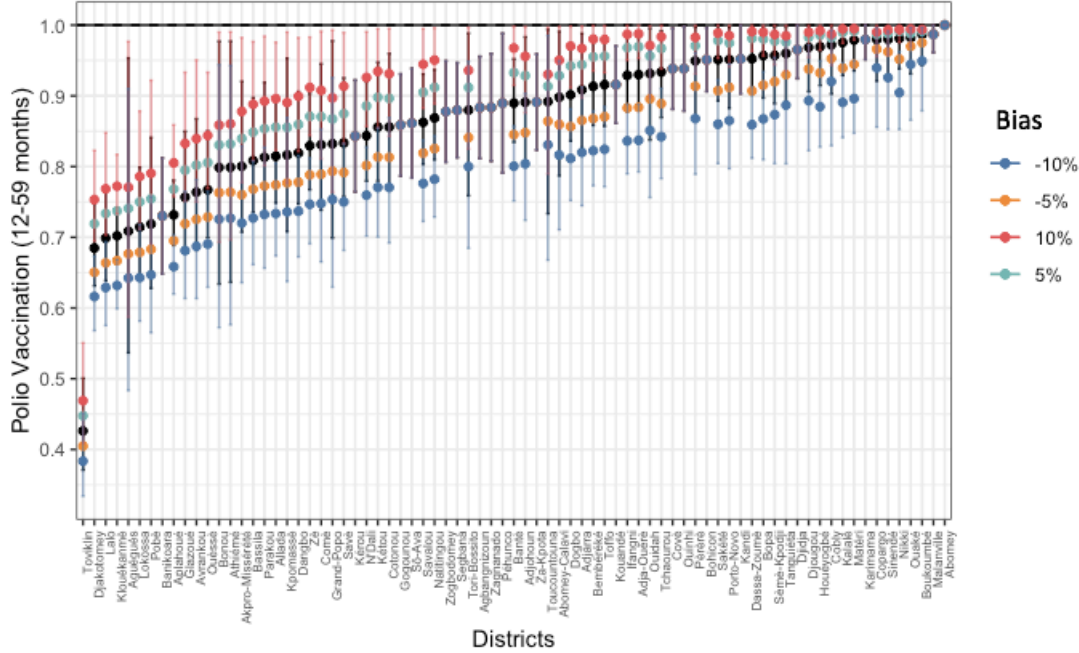


Figure 3.5 Sensitivity Analysis of Δ_3

We see the original estimates in black and the sensitivity estimates for the varying choices of our sensitivity parameter labeled by the figure legend. Lowering the administrative numerators by 10% lowers the coverage estimates by 6%, on average. Upping the administrative numerators by 10%, increases the coverage estimates by 4%, on average. The upward bias has a smaller effect near the margins because our probabilities are bounded above by 1. We find inferences to be fairly robust despite 10% bias in the administrative numerator.

Choosing specific values for Δ_3 is the frequentist approach for a sensitivity analysis. Alternatively, because we are in the Bayesian paradigm, we can consider priors for our sensitivity parameters. In fact, the frequentist sensitivity analysis corresponds to summarizing conditional posteriors, where conditioning is done on fixed values of Δ_3 as shown above.

Consider using the prior $\Delta_{3j} \sim N(1, 0.1)$ to quantify our belief about the bias in our administrative numerator. Because we have adopted a Bayesian framework we then simply incorporate the added variability/bias of this prior into our posterior predictive distribution. The results are shown in Figure 3.6:

Figure 3.6 Bayesian Sensitivity Analysis of Δ_3

3.5 Discussion

Our methodology adds to a body of tools that aim to leverage both administrative and probability survey data. Administrative data tend to be larger, indeed, they purport to be exhaustive, but the information may be

biased. Estimates published using administrative data alone resulted in 75% of the coverage rates for children aged 12-59 months to be above 1. The advantage of probability survey data is that when sampling is done appropriately, estimates are unbiased, but, for economic reasons, sample sizes may be small and limited to certain regions. For instance, the probability survey data used here only contained estimates for 19 out of 77 communes in Benin, and roughly 100 households were sampled in each commune. Our methodology aims to take the best from both sources of data.

Our approach and analysis are subject to limitations. While our regression model fits the data well, we only utilized two covariates and assumed a simple linear form. Analyses could benefit from more complex models that take into account a larger and more diverse set of covariates. In addition, some of the confidence intervals of our estimates are wide for certain communes. These communes had covariate values that fell outside of the ranges on which the model was fit. Specifically, these were communes with very small populations, having reported less than 60,000 inhabitants in the 2013 census. These communes were outside the interquartile range of population sizes included in the probability sample. The median population size for communes in our model fit was 100,197 (IQR: 75,383 – 128,180). This problem could potentially be avoided by ensuring a greater range of communes with large and small population sizes in subsequent probability samples. This can be accomplished by over sampling communes with those characteristics. Lastly, a great deal of confidence is placed in the administrative numerators in forming initial denominator estimates for which our model was fit. The administrative data may be biased, which highlights the importance of a sensitivity analysis.

In general, formulating this problem from the Bayesian perspective seems appropriate. We are uncertain what the coverage rates for polio vaccination and VAS are in communes without probability survey data. The Bayesian approach makes explicit our uncertainty by presenting a probability distribution that we believe characterizes our belief for coverage rates for a particular commune in question. It also enables a simple and flexible approach to a sensitivity analysis. Most importantly, it forces our coverage estimates below 1.

This approach motivates researchers to further consider the merits of combining administrative data with results from random probability survey data. Our work demonstrates that not only could this have implications for geographic regions with both sources of data available, but adjacent regions can benefit from annealing data sources as well. Although more research is needed, extrapolating hybrid estimation is a new tool which can now be used for more complete assessment of Child Health Days, and other campaigns such as Mass Drug Administration and distribution of insecticide impregnated bed-nets ([Maroto-Camino et al.](#),

2019). The next step is to apply it to a wider range of indicators measured with HMIS.

Acknowledgements

We thank Caroline Jeffrey for her assistance in getting acquainted with the data from Benin.

A

Proofs for Chapter 1

Proof of Theorem 1.

Consider counterfactuals Y_1 and Y_0 with absolutely continuous distribution functions $f_1(y)$ and $f_0(y)$ respectively both defined over a common domain $(-\infty, \infty)$. Here, R_a is a binary indicator of Y_a being observed or missing. Here, binary treatment $a \in \{0, 1\}$ determines which of the two counterfactuals is observed and is intervened on through randomization. Lastly consider the transformation of counterfactual Y_a such that:

$$U_a = \begin{cases} Y_a & \text{if } R_a = 1 \\ \min(\mathbf{Y}_a | R_a = 1) - \epsilon & \text{if } R_a = 0 \end{cases} \text{ for } \epsilon > 0$$

There exist two sufficient conditions in order to prove the equality of the trimmed means estimand and treatment difference in the whole population. The first condition – the location family assumption - is that the distribution of potential outcomes from the experimental group $Y_1 \sim f_1(y)$ is in the same location family as the distribution of potential outcomes from the reference group $Y_0 \sim f_0(y)$. Consider some constant Δ then $f_0(y) = f_1(y + \Delta)$. If two distributions are a location shift of one another then $E[Y_1] = E[Y_0] + \Delta$ because the mean is the location parameter of a distribution. Also, note that as a consequence all quantiles of these distributions are a location shift of one another i.e. $F_0^{-1}(\alpha) + \Delta = F_1^{-1}(\alpha)$. The second condition – quantile assumption – is that all missing values fall below the point at which the distributions are trimmed, i.e. $Y_a | R_a = 0 < F_a^{-1}(\alpha)$. The quantile assumption ensures that the composite outcome U_a is trimmed at the same value as Y_a for all percentiles above the maximum rate of missing data between the two arms, i.e. $F_{U_a}^{-1}(\alpha) = F_a^{-1}(\alpha) \forall \alpha : \alpha > Pr[R_a = 0]$. Leveraging both assumptions we can demonstrate the equality of the two estimands:

$$\begin{aligned}
E[U_1|U_1 > F_{U_1}^{-1}(\alpha)] - E[U_0|U_0 > F_{U_0}^{-1}(\alpha)] &= E[Y_1|Y_1 > F_1^{-1}(\alpha)] - E[Y_0|Y_0 > F_0^{-1}(\alpha)] \\
&= \frac{1}{1-\alpha} \int_{F_1^{-1}(\alpha)}^{\infty} y f_1(y) dy - \frac{1}{1-\alpha} \int_{F_0^{-1}(\alpha)}^{\infty} y f_0(y) dy \\
&= \frac{1}{1-\alpha} \left[\int_{F_1^{-1}(\alpha)}^{\infty} y f_1(y) dy - \int_{F_0^{-1}(\alpha)}^{\infty} y f_0(y) dy \right] \\
&= \frac{1}{1-\alpha} \left[\int_{F_1^{-1}(\alpha)}^{\infty} y f_1(y) dy - \int_{F_0^{-1}(\alpha)}^{\infty} y f_1(y + \Delta) dy \right] \\
&= \frac{1}{1-\alpha} \left[\int_{F_1^{-1}(\alpha)}^{\infty} y f_1(y) dy - \int_{F_1^{-1}(\alpha)}^{\infty} (x - \Delta) f_1(x) dx \right] \quad x = y + \Delta \\
&= \frac{1}{1-\alpha} \left[\int_{F_1^{-1}(\alpha)}^{\infty} y f_1(y) dy - \int_{F_1^{-1}(\alpha)}^{\infty} x f_1(x) dx + \Delta \int_{F_1^{-1}(\alpha)}^{\infty} f_1(x) dx \right] \\
&= \frac{1}{1-\alpha} \left[\Delta \int_{F_1^{-1}(\alpha)}^{\infty} f_1(x) dx \right] \\
&= \frac{1}{1-\alpha} [\Delta [1 - \alpha]] \\
&= \Delta \\
&= E[Y_0] - E[Y_0] + \Delta \\
&= E[Y_1] - E[Y_0]
\end{aligned}$$

Which completes the proof.

Proof of Theorem 2.

Consider two distributions $F_1(\cdot)$ and $F_0(\cdot)$ which are absolutely continuous distribution functions defined over a common domain $(-\infty, \infty)$. Also, both distributions have an expectation and that expectation is finite. Assume that the differences between the α -trimmed means are the same constant $\Delta \in \mathbb{R}$ for all $\alpha \in [0, 1]$ that is:

$$E[Y_1|Y_1 > F_1^{-1}(\alpha)] - E[Y_0|Y_0 > F_0^{-1}(\alpha)] = \Delta$$

$$\begin{aligned}
E[Y_0|Y_0 > F_0^{-1}(\alpha)] &= E[Y_1|Y_1 > F_1^{-1}(\alpha)] - \Delta \\
\frac{1}{1-\alpha} \int_{F_0^{-1}(\alpha)}^{\infty} y f_0(y) dy &= \frac{1}{1-\alpha} \int_{F_1^{-1}(\alpha)}^{\infty} y f_1(y) dy - \Delta \\
\frac{1}{1-\alpha} \int_{F_0^{-1}(\alpha)}^{\infty} y f_0(y) dy &= \frac{1}{1-\alpha} \int_{F_1^{-1}(\alpha)}^{\infty} y f_1(y) dy - \frac{(1-\alpha)}{(1-\alpha)} \Delta \\
\frac{1}{1-\alpha} \int_{F_0^{-1}(\alpha)}^{\infty} y f_0(y) dy &= \frac{1}{1-\alpha} \int_{F_1^{-1}(\alpha)}^{\infty} y f_1(y) dy - \frac{1}{(1-\alpha)} \Delta \int_{F_1^{-1}(\alpha)}^{\infty} f_1(y) dy \\
\frac{1}{1-\alpha} \int_{F_0^{-1}(\alpha)}^{\infty} y f_0(y) dy &= \frac{1}{1-\alpha} \left[\int_{F_1^{-1}(\alpha)}^{\infty} y f_1(y) dy - \int_{F_1^{-1}(\alpha)}^{\infty} \Delta f_1(y) dy \right] \\
\frac{1}{1-\alpha} \int_{F_0^{-1}(\alpha)}^{\infty} y f_0(y) dy &= \frac{1}{1-\alpha} \left[\int_{F_1^{-1}(\alpha)}^{\infty} (y - \Delta) f_1(y) dy \right] \\
\frac{1}{1-\alpha} \int_{F_0^{-1}(\alpha)}^{\infty} y f_0(y) dy &= \frac{1}{1-\alpha} \left[\int_{F_1^{-1}(\alpha) - \Delta}^{\infty} x f_1(x + \Delta) dx \right] \quad x = y - \Delta, \quad dx = dy \\
\int_{F_0^{-1}(\alpha)}^{\infty} y f_0(y) dy &= \int_{F_1^{-1}(\alpha) - \Delta}^{\infty} x f_1(x + \Delta) dx
\end{aligned}$$

At this stage perform the substitution of $y = F_0^{-1}(\beta)$, $d\beta = f_0(y)dy$ for the integral on the left side of the equation and $x = F_1^{-1}(\beta) - \Delta$, $d\beta = f(x + \Delta)dx$ for the integral on the right side of the equation such that:

$$\begin{aligned}
\int_{\alpha}^1 F_0^{-1}(\beta) d\beta &= \int_{\alpha}^1 [F_1^{-1}(\beta) - \Delta] d\beta \\
\frac{d}{d\alpha} \int_{\alpha}^1 F_0^{-1}(\beta) d\beta &= \frac{d}{d\alpha} \int_{\alpha}^1 [F_1^{-1}(\beta) - \Delta] d\beta \\
-F_0^{-1}(\alpha) &= -[F_1^{-1}(\alpha) - \Delta], \quad \alpha \in (0, 1) \\
F_0^{-1}(\alpha) &= F_1^{-1}(\alpha) - \Delta
\end{aligned}$$

From here it follows that $F_1 - \Delta = F_0$ and it is proven that F_1 is a location shift of F_0 almost everywhere on the domain i.e. $f_0(y) = f_1(y + \Delta)$.

B

Supplementary Results for Chapter 3

Both administrative data and LQAS surveys are available for Vitamin A supplementation as well as for Polio. In addition, coverage estimates are generally stratified by children aged 6-11 months as well as those 12-59 months. We showcase the additional results here. The first set of Figures B.1-B.3 present the results of the frequentist analysis.

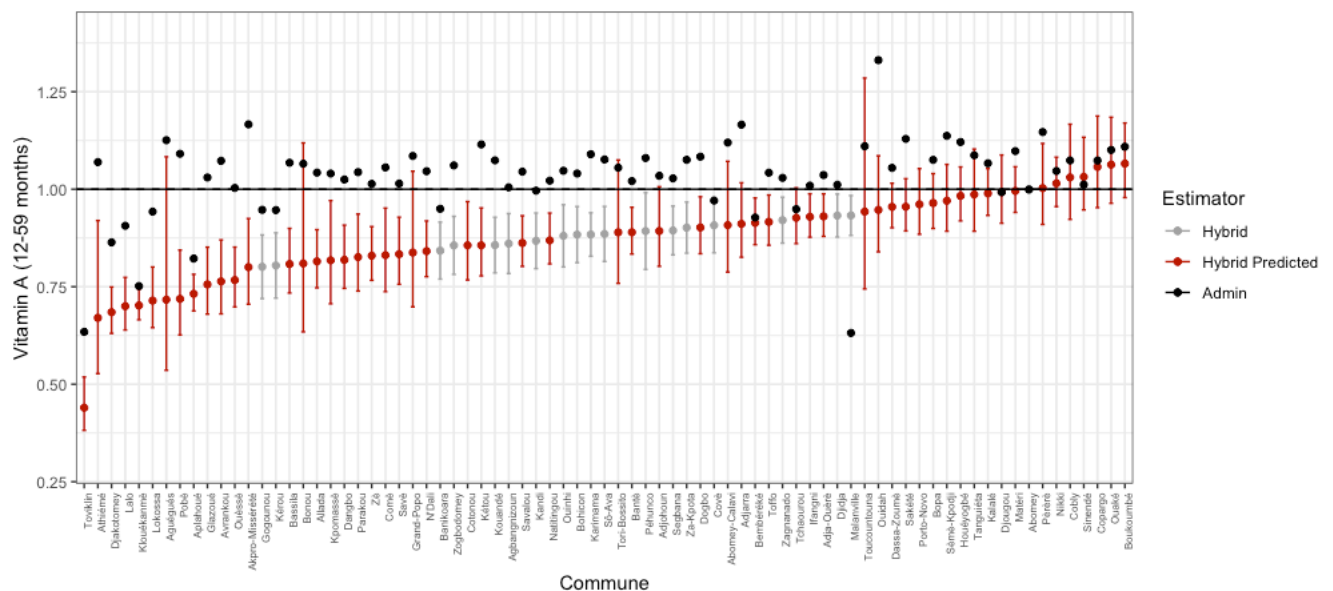


Figure B.1 Children's Vitamin A Coverage Estimates (12-59 Months)

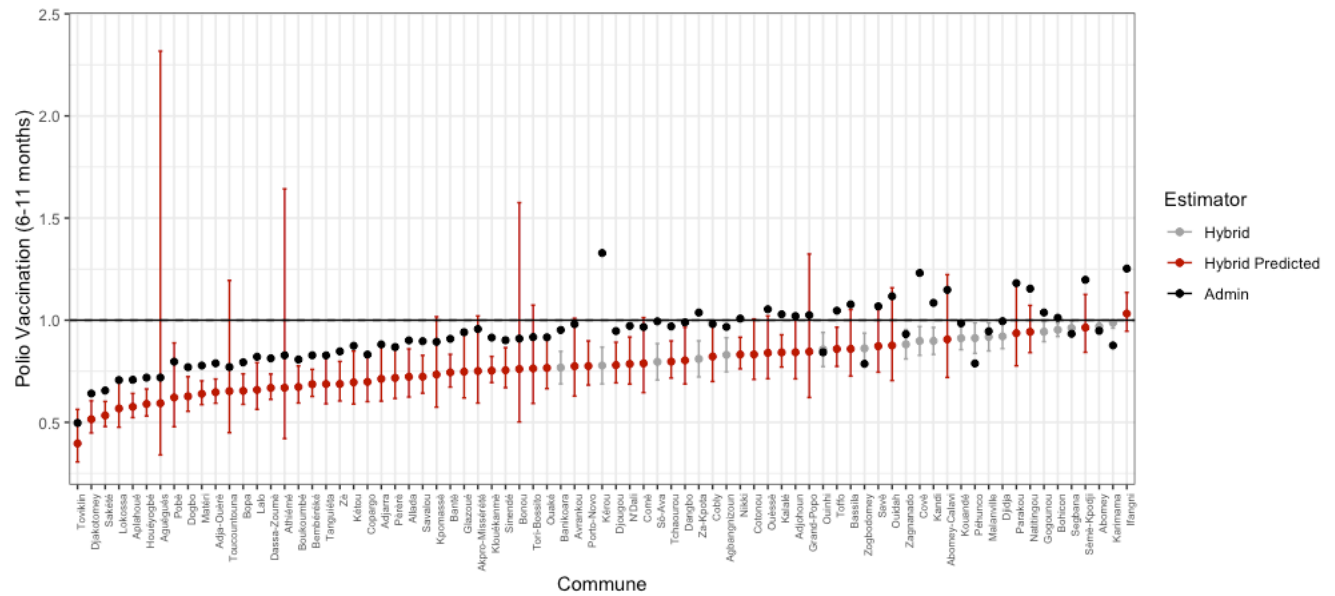


Figure B.2 Children's Polio Vaccination Coverage Estimates (6-11 Months)

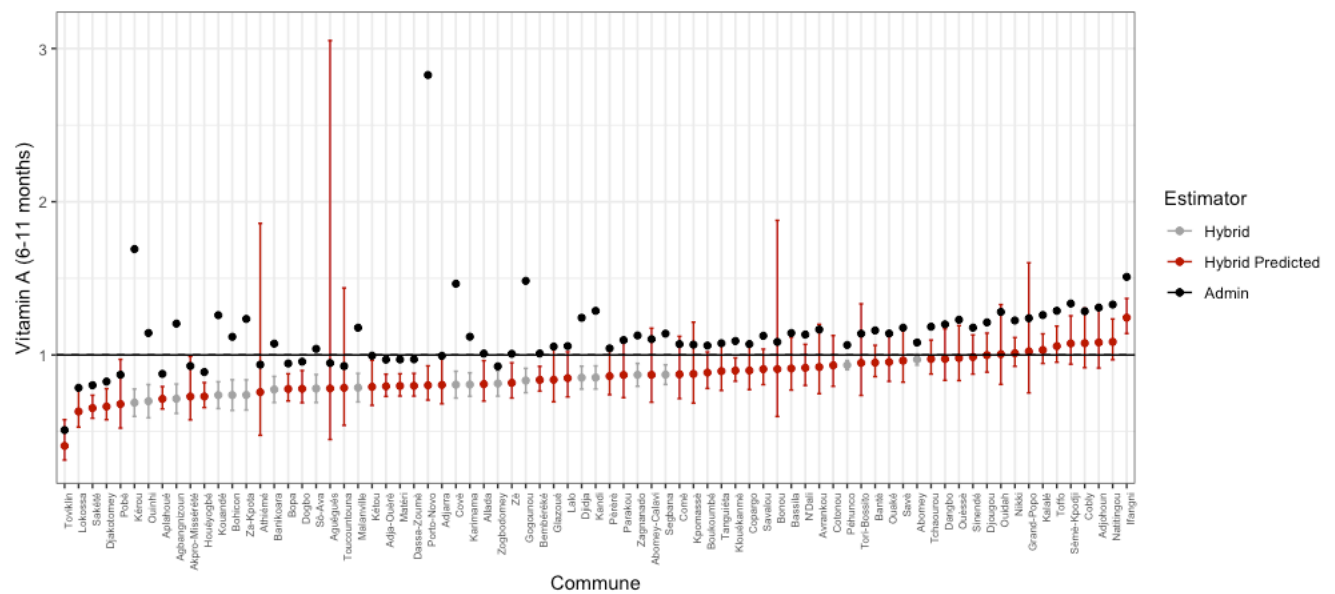


Figure B.3 Children's Vitamin A Coverage Estimates (6-11 Months)

Figures B.4-B.6 present the resulting Bayesian estimates for the same above indicators. We note the Bayesian adjustment forces estimates and intervals to fall in the appropriate ranges (i.e. below 1).

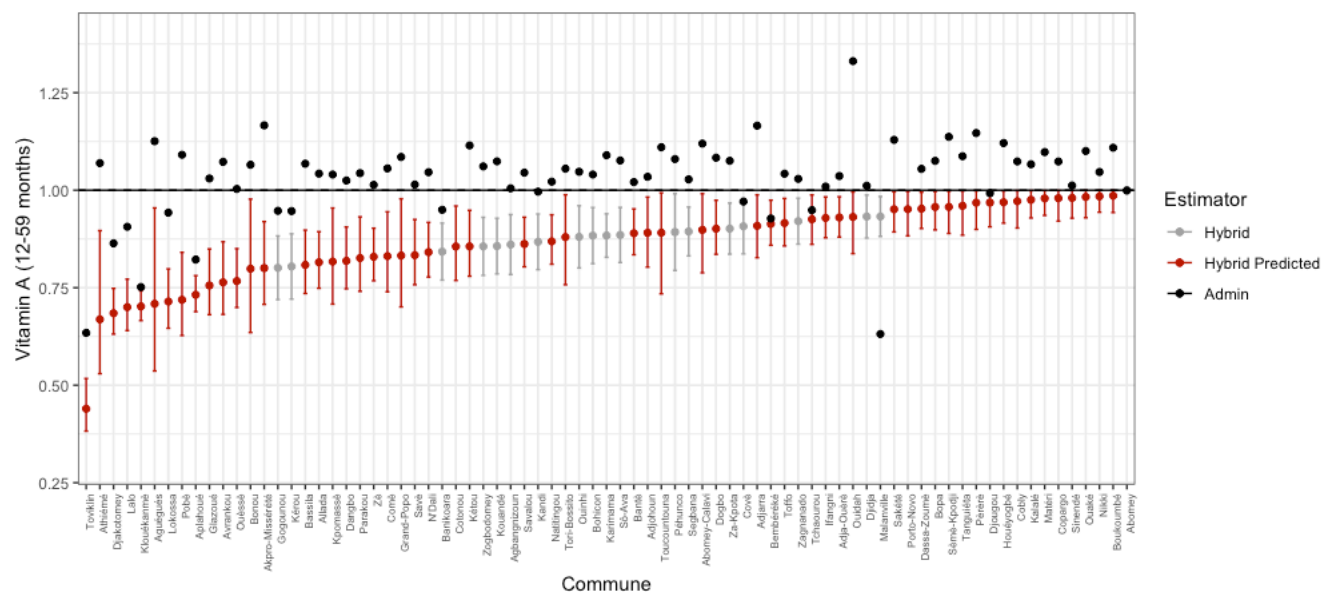


Figure B.4 Children's Vitamin A Coverage Bayesian Estimates (12-59 Months)

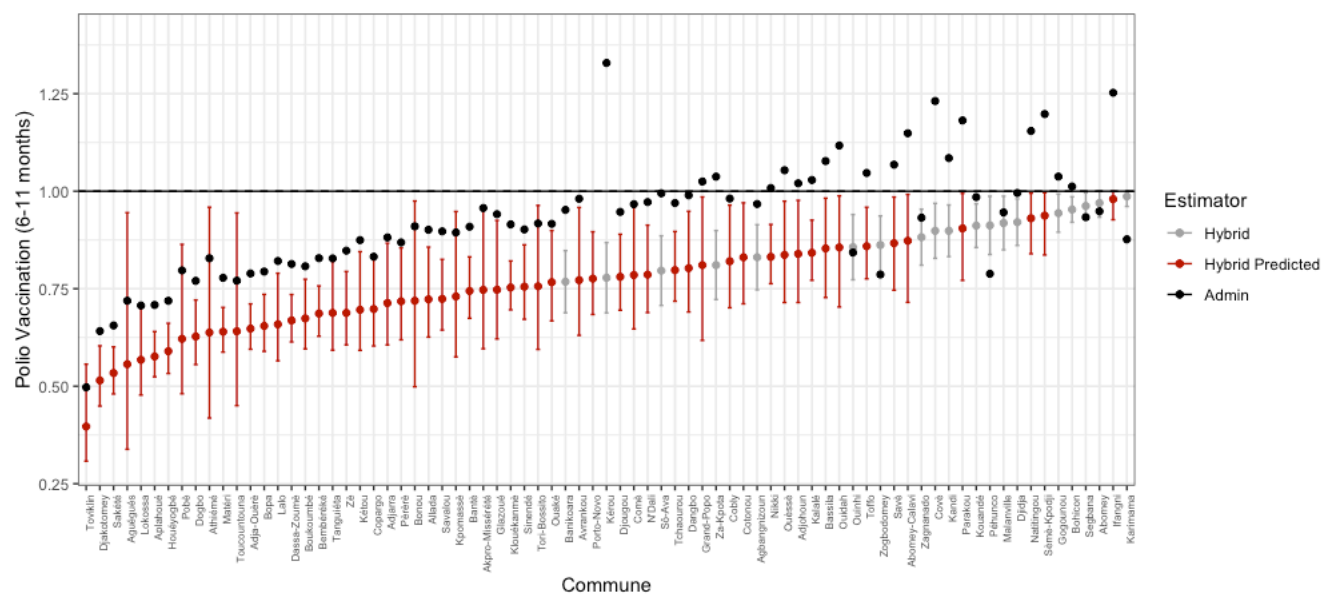


Figure B.5 Children's Polio Vaccination Coverage Bayesian Estimates (6-11 Months)

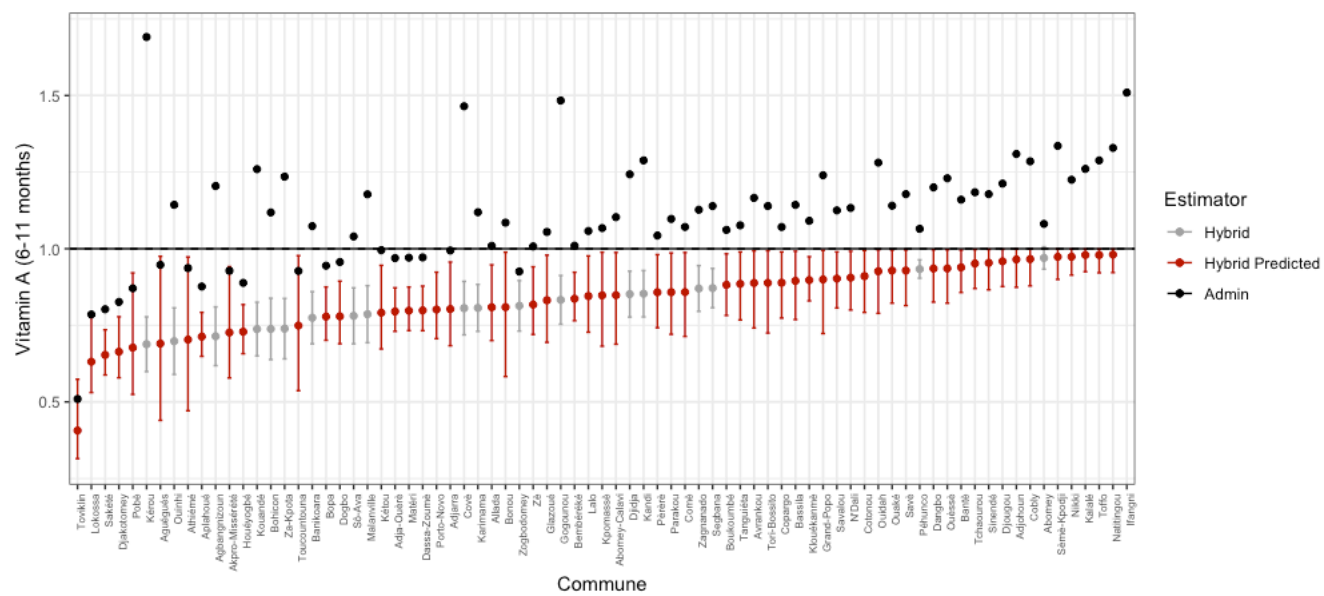


Figure B.6 Children's Vitamin A Coverage Bayesian Estimates (6-11 Months)

References

- C. AbouZahr and T. Boerma. Health information systems: the foundations of public health. *Bulletin of the World Health Organization*, 83:578–583, 2005.
- A. Balke and J. Pearl. Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, 92(439):1171–1176, 1997.
- D. E. Becker, C. D. Yingling, and G. Fein. Identification of pain, intensity and p300 components in the pain evoked potential. *Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section*, 88(4):290–301, 1993.
- D. Boos and L. Stefanski. *Essential Statistical Inference*. Springer, 2013.
- S. v. Buuren and K. Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, pages 1–68, 2010.
- J. R. Carpenter, J. H. Roger, and M. G. Kenward. Analysis of longitudinal trials with protocol deviation: a framework for relevant, accessible assumptions, and inference via multiple imputation. *Journal of biopharmaceutical statistics*, 23(6):1352–1371, 2013.
- B. Chaudhry, J. Wang, S. Wu, M. Maglione, W. Mojica, E. Roth, S. C. Morton, and P. G. Shekelle. Systematic review: Impact of health information technology on quality, efficiency, and costs of medical care. *Annals of Internal Medicine*, 144(10):742–752, 2006.
- J. C. Cisse M, Vargas W. Evaluation de la validité des données administratives de la semaine de la santé de la mère et de l’enfant d’octobre 2015 dans trois districts de madagascar. *United Nations International Children’s Emergency Fund, Antananarivo, Madagascar*, 2016.
- W. G. Cochran. *Laplace’s ratio estimator*, pages 3–10. Elsevier, 1978.
- S. R. Cole and C. E. Frangakis. The consistency statement in causal inference: a definition or an assumption? *Epidemiology*, 20(1):3–5, 2009.
- M. J. Daniels and J. W. Hogan. *Missing data in longitudinal studies: Strategies for Bayesian modeling and sensitivity analysis*. CRC Press, 2008.
- R. H. Davis and J. J. Valadez. Improving the collection of knowledge, attitude and practice data with community surveys: a comparison of two second-stage sampling methods. *Health policy and planning*, 29(8):1054–1060, 2014.
- B. Efron and R. J. Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.
- P. I. Good. *Permutation, parametric, and bootstrap tests of hypotheses*. Springer Science & Business Media, 2006.
- F. R. Hampel. The influence curve and its role in robust estimation. *Journal of the american statistical association*, 69(346):383–393, 1974.
- R. Haux, P. Knaup, and F. Leiner. On educating about medical data management. *Methods of information in medicine*, 46(01):74–79, 2007.

- B. L. Hedt and M. Pagano. Health indicators: eliminating bias from convenience sampling estimators. *Statistics in medicine*, 30(5):560–568, 2011.
- J. L. Horowitz and C. F. Manski. Nonparametric analysis of randomized experiments with missing covariate and outcome data. *Journal of the American statistical Association*, 95(449):77–84, 2000.
- J. L. Horowitz, C. F. Manski, M. Ponomareva, and J. Stoye. Computation of bounds on population parameters when the data are incomplete. *Reliable Computing*, 9(6):419–440, 2003.
- N. J. Horton and K. P. Kleinman. Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *The American Statistician*, 61(1):79–90, 2007.
- D. Hosmer, S. Lemeshow, and S. May. Applied survival analysis. hoboken, 2008.
- P. J. Huber. Robust estimation of a location parameter. *Annals of Mathematical Statistics*, 35:73–101, 1964.
- INSAE. Recensement général de la population et de l’habitation. *Disponível em*, 2013.
- C. Jeffery, M. Pagano, J. Hemingway, and J. J. Valadez. Hybrid prevalence estimation: Method to improve intervention coverage estimations. *Proceedings of the National Academy of Sciences*, 115(51):13063–13068, 2018.
- V. JJ, J. C, B. T, and V. N. Monitoring vitamin a supplementa- tion and polio vaccination during child health days: What are the options? *United Nations International Children’s Emergency Fund*, 2017.
- L. Kish. Wiley, Berlin, 1995.
- E. L. Lehmann. *Elements of large-sample theory*. Springer Science & Business Media, 2004.
- C. Lima, J. Schramm, and C. M. Coeli. Review of data quality dimensions and applied methods in the evaluation of health information systems. *Cadernos de saude publica*, 25(10):2095–2109, 2009.
- T. Lippeveld, R. Sauerborn, C. Bodart, and WHO. Design and implementation of health information systems. 2000.
- R. J. Little. Comments on: Missing data methods in longitudinal studies: a review. *Test*, 18(1):47–50, 2009.
- R. J. Little and D. B. Rubin. *Statistical analysis with missing data*, volume 793. John Wiley & Sons, 2019.
- C. H. Mallinckrodt, W. S. Clark, and S. R. David. Accounting for dropout bias using mixed-effects models. *Journal of biopharmaceutical statistics*, 11(1-2):9–21, 2001.
- C. F. Manski. Nonparametric bounds on treatment effects. *The American Economic Review*, 80(2):319–323, 1990.
- C. Maroto-Camino, P. Hernandez-Pastor, N. Awaca, L. Safari, J. Hemingway, M. Massangaie, D. Whitson, C. Jeffery, and J. J. Valadez. Improved assessment of mass drug administration and health district management performance to eliminate lymphatic filariasis. *PLoS neglected tropical diseases*, 13(7):e0007337, 2019.
- D. V. Mehrotra, F. Liu, and T. Permutt. Missing data in clinical trials: control-based mean imputation and sensitivity analysis. *Pharmaceutical statistics*, 16(5):378–392, 2017.

- W. Miao and E. J. Tchetgen Tchetgen. On varieties of doubly robust estimators under missingness not at random with a shadow variable. *Biometrika*, 103(2):475–482, 2016.
- K. Mohan, J. Pearl, and J. Tian. Graphical models for inference with missing data. In *Advances in neural information processing systems*, pages 1277–1285, 2013.
- R. I. Mutemwa. Hmis and decision-making in zambia: re-thinking information solutions for district health management in decentralized health systems. *Health Policy and Planning*, 21(1):40–52, 2006.
- National Research Council. *The prevention and treatment of missing data in clinical trials*. National Academies Press, 2010.
- E. Ndabarora, J. A. Chipps, and L. Uys. Systematic review of health data quality management and best practices at community and district levels in lmic. *Information Development*, 30(2):103–120, 2014.
- M. P. Nisingizwe, H. S. Iyer, M. Gashayija, L. R. Hirschhorn, C. Amoroso, R. Wilson, E. Rubyutsa, E. Gaju, P. Basinga, A. Muhire, et al. Toward utilization of data for program management and evaluation: quality assessment of five years of health management information system data in rwanda. *Global health action*, 7(1):25829, 2014.
- A. Ocampo, H. Schmidli, P. Quarg, F. Callegari, and M. Pagano. Identifying treatment effects using trimmed means when data are missing not at random. *arXiv preprint arXiv:1908.01044*, 2019.
- G. W. Odhiambo-Otieno. Evaluation of existing district health management information systems: a case study of the district health systems in kenya. *International journal of medical informatics*, 74(9):733–744, 2005.
- M. Pagano and J. J. Valadez. Commentary: Understanding practical lot quality assurance sampling. *International journal of epidemiology*, 39(1):69–71, 2010.
- T. Permutt and F. Li. Trimmed means for symptom trials with dropouts. *Pharmaceutical statistics*, 16(1):20–28, 2017.
- M. H. Quenouille. Problems in plane sampling. *The Annals of Mathematical Statistics*, 20(3):355–375, 1949.
- N. Reid and H. Crépeau. Influence functions for proportional hazards regression. *Biometrika*, 72(1):1–9, 1985.
- S. E. Robertson and J. J. Valadez. Global review of health care surveys using lot quality assurance sampling (lqas), 1984-2004. *Social science & medicine*, 63(6):1648–1660, 2006.
- J. M. Robins. The analysis of randomized and non-randomized aids treatment trials using a new approach to causal inference in longitudinal studies. *Health service research methodology: a focus on AIDS*, pages 113–159, 1989.
- J. M. Robins, A. Rotnitzky, and L. P. Zhao. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the american statistical association*, 90(429):106–121, 1995.
- P. J. Rousseeuw and A. M. Leroy. *Robust regression and outlier detection*, volume 589. John wiley & sons, 2005.

- D. B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.
- D. B. Rubin. *Multiple imputation for nonresponse in surveys*, volume 81. John Wiley & Sons, 2004.
- A. Sen. Some early developments in ratio estimation. *Biometrical journal*, 35(1):1–13, 1993.
- M. Smith, S. Madon, A. Anifalaje, M. Lazarro, Malecela, and E. Michael. Integrated health information systems in tanzania: experience and challenges. *The Electronic journal of information systems in developing countries*, 33(1):1–21, 2008.
- S. M. Stigler. The asymptotic distribution of the trimmed mean. *The Annals of Statistics*, pages 472–477, 1973.
- A. Tsiatis. *Semiparametric theory and missing data*. Springer Science & Business Media, 2007.
- A. G. Turner, R. J. Magnani, and M. Shuaib. A not quite as quick but much cleaner alternative to the expanded programme on immunization cluster survey design. *International journal of epidemiology*, 25(1):198–203, 1996.
- United Nations Development. A propos de madagascar. *United Nations Development Program, Antananarivo, Madagascar*, 2017.
- J. J. Valadez. *Assessing child survival programs in developing countries*. Dept. of Population and International Health, Harvard School of Public Health, 1991.
- M. D. Wang, J. Liu, G. Molenberghs, and C. Mallinckrodt. An evaluation of the trimmed mean approach in clinical trials with dropout. *Pharmaceutical statistics*, 17(3):278–289, 2018.
- WHO. Improving data quality: a guide for developing countries, 2003.
- M. Zaffalon. Exact credal treatment of missing data. *Journal of Statistical Planning and Inference*, 105(1): 105–122, 2002.