Kimberly A. Hochstedler Webb
kah343@cornell.edu

RESEARCH STATEMENT

In my research, I aim to answer questions in public health and the social sciences by leveraging the complex mechanisms that underlie the data generation process. In observational studies, data may be imperfectly measured and using the observed, but potentially mismeasured, data can cause bias in statistical estimates. Importantly, this measurement problem does not dissipate with increasing sample size. Thus, even in the age of "big data", statistical methodology for handling data measured with error is still particularly important.

In designing novel statistical methodology for imperfect data, I develop tools to answer scientific questions and better understand the world around us. To further my goal of creating and applying useful statistical tools to address impactful problems, I seek out collaborations with social and public health scientists. I aim to share my methods with fellow statisticians and applied researchers by developing user-friendly software to accompany my methodological contributions.

Misclassification of binary variables

Much of my recent research is aimed at creating statistical methods for dealing with *differential misclassification* in binary variables. While existing work in this area largely focuses on correcting for bias caused by misclassification through validation studies, I instead consider the problem in cases where a gold standard measure is not available— making validation studies impossible.

In [1], I develop a statistical method for recovering unbiased parameter estimates in association studies with a misclassified binary outcome, without gold standard labels. I present a Markov Chain Monte Carlo (MCMC) algorithm and an Estimation-Maximization (EM) algorithm that leverage the data generation mechanism presented in Figure 1 to estimate both (1) the unbiased association between the predictor and true outcome of interest and (2) the rate at which the observed outcomes were misclassified. Misclassification models of this variety are known for pernicious identifiability problems; here, the challenge is a specific case of "label switching". To overcome this challenge, I introduce a novel "label switching correction" algorithm that relies only on the assumption that outcomes are correctly classified in at least 50% of the observations. I apply this method to study risk factors for myocardial infarction (MI) and show that the effects of risk factors for MI are stronger in a model that corrects for misdiagnosis of MI compared to a model that ignores potential misdiagnosis. This project demonstrates that developing methods for misclassified data can help us better understand risk factors for medical events. To share these methods with fellow statisticians, I wrote a user-friendly R software package called COMBO that is currently available on CRAN [2]. In 2023, this work was recognized with a Distinguished Student Paper Award from the Eastern North American Region International Biometric Society.

Recently, I extended the misclassification corrections in [1] to account for misclassified mediator variables. In [3], I develop an ordinary least squares correction algorithm for a misclassified binary mediator variable and a Normally distributed outcome. In addition, I provide an extension of the predictive value weighting approach to account for a misclassified mediator in a model with a binary, Normal, or count outcome. Both approaches update existing methodology to incorporate estimated misclassification rates, rather than assuming that the sensitivity and specificity of the mediator instrument are known. In [3], I also develop an EM algorithm to seamlessly account for misclassification in the mediating variable in a single step. I apply these methods to examine the mediating role of (potentially misclassified) gestational hypertension in the association between maternal age and risk of preterm birth.

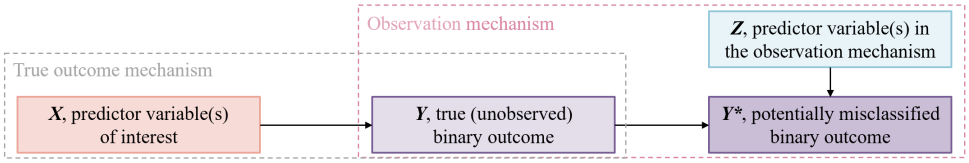

Figure 1: *Relationship between the predictors of interest, X, and true (unobserved) binary outcome, Y, in the true outcome mechanism. The observation mechanism models the association between predictor variables, Z, and the observed binary outcome, Y^*, given the latent true outcome.*

My previous work in developing methods for association studies with misclassified binary outcomes and with misclassified binary mediators, respectively, led me to consider cases where both the mediator and the outcome are suspected to be misclassified in a single system. In [4], I utilize a pseudolikelihood approach to model the misclassification in both the mediator and the response variable, within two separate EM algorithms. Importantly, this method does not require that the misclassification mechanisms are the same between the mediator and the outcome, allowing for realistic modeling in the context of public health applications.

Clinical trial design and methodology

Beyond observational studies, imperfect measurement can hinder effect estimation in other study designs, including clinical trials. In [5] I show that using my misclassification correction from [1] can recover unbiased average treatment effect estimates in standard clinical trial frameworks. In addition, I combine my misclassification correction with covariate adjustment methods to address misclassification in the outcome while simultaneously accounting for covariate imbalance in the trial and improving the precision of our estimators.

My methods for handling misclassification within a clinical trial framework are complemented by my research in clinical trial designs. Specifically, I have focused on designing clinical trials for specialized settings, including rare diseases. Because of the limited number of individuals affected by rare diseases, it is often difficult to accrue enough participants in trials for new treatments and thus these trials suffer from reduced power. In response to these challenges, I, along with my colleagues, developed a new small n, sequential, multiple assignment, randomized trial (snSMART) that is specifically intended for small samples [6]. In a typical SMART design, the goal is to develop dynamic treatment regimes that specify an initial treatment for patients, followed by subsequent treatments that are assigned based on the response to the initial treatment. In contrast, this snSMART design leverages the infrastructure of a traditional SMART design to efficiently share data across multiple trial stages via a Bayesian approach. This framework allows for more efficient estimation than a frequentist approach and enables estimation of a single superior treatment or dose of treatment in a small sample of patients.

Taken together, my work in [5] and [6] demonstrates my ability to incorporate realistic issues, like imperfect data and small sample sizes, into the design and analysis of clinical trials. This expertise will enable me to continue developing creative and practical trial designs in the future.

Collaborative Research

I have also played a leading role in impactful applied research. In [7], I developed methods to study a specialized decision-making structure within the Virginia pretrial system. My collaborators were primarily interested in risk factors associated with "pretrial failure", the event where defendants either fail to appear (FTA) to court or reoffend. Pretrial failure is not measured directly; instead the risk of pretrial failure is assessed prior to the trial via two imperfect proxies: (1) an automated assessment algorithm and then (2) a judge's decision, which is informed by the results of the assessment algorithm. I modeled these two stages of measurement using a two-stage misclassification model to analyze the risk factors for pretrial failure, while simultaneously

estimating misclassification rates for both the assessment algorithm and judges' decisions. In this work, I showed that the algorithm's recommendations and the court bail decisions may be differentially misclassified based on defendant race, leading black defendants to receive more punitive sentences, on average, than white defendants.

My other collaborative work has aimed to answer clinically valuable questions in cancer research [8, 9, 10, 11, 12, 13], with a focus on creating useful scientific tools for clinicians and statisticians. In [8], I, along with my colleagues, developed a suite of statistical models to predict the risk of severe side effects for patients undergoing radiation therapy for lung cancer. The clinical relevance of this project motivated me to create an open-source R Shiny web application that allows clinicians to predict their own patient's outcomes at various dosage levels using our models. I was also on a team that developed a statistical approach for quantifying the impact of a delay in cancer treatment vs. immediate care at the start of the COVID-19 pandemic. This work was developed into an R Shiny web application, allowing physicians to make real-time care decisions for their own patients [9]. Outside of cancer research, I was also on the development team for an R software package and R Shiny application for the analysis of the regression-discontinuity design [14]. This R package, entitled rddapp, is currently available on CRAN [15].

Future Work

Going forward, I will continue focusing on statistical methods for misclassified variables, while also devoting a significant proportion of my effort toward methods and applications for other complex and practical settings. A primary extension for my work on association studies with misclassified binary outcomes is to consider settings for nominal categorical outcomes with an arbitrary number of categories. Theoretically, my existing methods extend naturally to this setting, but I look forward to addressing the numerical issues and computational limitations associated with a greater number of outcome categories. In addition, I intend to explore the possibility of using constrained estimation approaches for misclassification models. A constrained estimation approach would allow me to estimate unbiased association parameters while directly applying the assumption that misclassification rates are less than 50%, rather than incorporating this assumption in an algorithm that is applied to existing estimates.

Another near-term project involves investigating response misclassification in SMART designs. Such designs rely on the assumption that patient response is measured without error. Recasting a SMART design as a system with misclassified patient outcomes allows tools from my previous work to be directly applied to new trial designs that are robust to mismeasurement.

An overarching goal of my research is not only to develop novel statistical methods, but also to develop tools that allow my methods to be used by fellow statisticians and applied researchers. As such, I intend to develop R packages and/or R Shiny web applications to accompany each of my methodological developments. These widely used tools not only share my methodological contributions with a broader audience, but their development also serves as an authentic and meaningful opportunity for undergraduate- and graduate-level student research projects.

In the longer term, my work shows continued promise in the era of "big data." Measurement error issues do not dissipate with increasing sample size, so methods for handling imperfect data will continue to be necessary and relevant, even as our datasets grow. In addition, the increasing use of algorithms for decision-making poses a challenge for society. We must understand inherent biases in algorithmic decision-making, and by viewing this bias as a "noisy labels" or misclassification problem, I can leverage my work to evaluate these algorithms now and in the future. My research goals revolve around developing methods that acknowledge the messy nature of data in the real world. I look forward to creating a research program that combines statistical methodology with collaborative research endeavors in public health and social sciences.

References

* First authors contributed equally, # Primary investigators contributed equally

[1] **Webb, Kimberly A. H.** and Wells, Martin T. "Statistical inference for association studies in the presence of binary outcome misclassification". In: *arXiv preprint arXiv:2303.10215* (2023).

[2] **Hochstedler, Kimberly A.** *COMBO: Correcting Misclassified Binary Outcomes in Association Studies*. R package version 1.0.0. 2023. URL: https://CRAN.R-project.org/package=COMBO.

[3] **Webb, Kimberly A. H.** and Wells, Martin T. "Effect estimation in the presence of a misclassified binary mediator". In: *In preparation* (2023).

[4] **Webb, Kimberly A. H.** and Wells, Martin T. "Regression methods for misclassified binary mediator and outcome variables". In: *In preparation* (2023).

[5] **Webb, Kimberly A. H.** "Improving precision and bias in estimating treatment effects in randomized trials with outcome misclassification". In: *In preparation* (2023).

[6] *Fang, Fang, ***Hochstedler, Kimberly A.**, Tamura, Roy, Braun, Tom M., and Kidwell, Kelley M. "Bayesian Methods to Compare Dose Levels to Placebo in a Small n Sequential Multiple Assignment Randomized Trial (snSMART)". In: *Statistics in Medicine* 40.4 (Nov. 2020), pp. 963–977. DOI: 10.1002/sim.8813.

[7] **Webb, Kimberly A. H.**, Riley, Sarah A., and Wells, Martin T. "An assessment of racial disparities in pretrial decision-making using misclassification models". In: *In preparation* (2023).

[8] *McFarlane, Matthew R., ***Hochstedler, Kimberly A.**, Laucis, Anna M., Sun, Yilun, Chowdhury, Aulina, Matuszak, Martha M., Hayman, James, Bergsma, Derek, Boike, Thomas, Kestin, Larry, Movsas, Benjamin, Grills, Inga, Dominello, Michael, Dess, Robert T., Schonewolf, Caitlin, Spratt, Daniel E., Pierce, Lori, Paximadis, Peter, #Jolly, Shruti, and #Schipper, Matthew. "Predictors of Pneumonitis After Conventionally Fractionated Radiotherapy for Locally Advanced Lung Cancer". In: *International Journal of Radiation Oncology*Biology*Physics* 111.5 (2021), pp. 1176–1185. ISSN: 0360-3016. DOI: https://doi.org/10.1016/j.ijrobp.2021.07.1691.

[9] Hartman, Holly E., Yilun, Sun, Devasia, Theresa, Chase, Elizabeth, Jairath, Neil, Dess, Robert T., Jackson, William C., Morris, Emily, Li, Pin, **Hochstedler, Kimberly A.**, Abbott, Madeline R., Kidwell, Kelley M., Walter, Vonn, Wang, Ming, Wang, Xi, Zaorsky, Nicholas, #Schipper, Matthew J., and #Spratt, Daniel E. "Integrated Survival Estimates from Cancer Treatment Delay during the COVID-19 Pandemic". In: *JAMA Oncology* 6 (Dec. 2020), pp. 1881–1889. DOI: 10.1001/jamaoncol.2020.5403.

[10] Herr, Daniel J., **Hochstedler, Kimberly A.**, Yin, Huiying, Dess, Robert T., Matuszak, Martha, Grubb, Margaret, Dominello, Michael, Movsas, Benjamin, Kestin, Larry L., Bergsma, Derek, Dragovic, Aleksandar F., Grills, Inga S., Hayman, James A., Paximadis, Peter, Schipper, Matthew, and Jolly, Shruti. "Effect of Education and Standardization of Cardiac Dose Constraints on Heart Dose in Patients With Lung Cancer Receiving Definitive Radiation Therapy Across a Statewide Consortium". In: *Practical Radiation Oncology* 12 (5 2022), e376–e381. ISSN: 1879–8500. DOI: https://doi.org/10.1016/j.prro.2022.01.002.

[11] Laucis, Anna M. B., **Hochstedler, Kimberly A.**, Schipper, Matthew J., Paximadis, Peter A., Boike, Thomas P., Bergsma, Derek P., Movsas, Benjamin, Kretzler, Annette, Spratt, Daniel E., Dess, Robert T., Mietzel, Melissa A., Dominello, Michael M., Matuszak, Martha M., Jagsi, Reshma, Hayman, James A., Pierce, Lori J., and Jolly, Shruti. "Racial Differences in Treatments and Toxicity in Patients With Non–Small-Cell Lung Cancer Treated With Thoracic Radiation Therapy". In: *JCO Oncology Practice* 18.6 (2022), e1034–e1044. DOI: 10.1200/OP.21.00224.

[12] Wilkie, Joel R, **Hochstedler, Kimberly A.**, Schipper, Matthew J, Matuszak, Martha M, Paximadis, Peter A, Dominello, Michael M, Grills, Inga S, Hayman, James A, Dess, Robert T, Pierce, Lori J, Spratt, Daniel E, Bergsma, Derek P, Boike, Thomas P, Movsas, Benjamin, and Jolly, Shruti. "Association between Adverse Events and Quality of Life in Patients Treated with Definitive Radiotherapy for Locally Advanced Lung Cancer in a Statewide Consortium". In: *International Journal of Radiation Oncology, Biology, Physics* 112 (4 2021), pp. 942–950. DOI: 10.1016/j.ijrobp.2021.11.024.

[13] Ramanathan, Siddharth, **Hochstedler, Kimberly A.**, Laucis, Anna M, Movsas, Benjamin, Stevens, Craig W, Kestin, Larry L, Dominello, Michael M, Grills, Inga S, Matuszak, Martha M, Hayman, James A, Paximadis, Peter A, Schipper, Matthew J, Jolly, Shruti, and Boike, Thomas P. "Predictors of Early Hospice or Death in Patients with Inoperable Lung Cancer Treated with Curative Intent". In: *In Preparation* (2022).

[14] Thommes, Felix, Liao, Wang, Papst, Irena, Jin, Ze, Zhang, Wenyu, and **Hochstedler, Kimberly A.** "`rddapp`: An `R` package for multi-assignment regression-discontinuity designs". In: (2022).

[15] Jin, Ze, Liao, Wang, Papst, Irena, Zhang, Wenyu, **Hochstedler, Kimberly A.**, and Thoemmes, Felix. *rddapp: Regression Discontinuity Design Application*. R package version 1.3. 2021.