2022

Statistical Approaches To Reducing Bias And Improving Variance Estimation In The Presence Of Covariate And Outcome Measurement Error

Lillian Boe
University of Pennsylvania

Statistical Approaches To Reducing Bias And Improving Variance Estimation In The Presence Of Covariate And Outcome Measurement Error

Abstract

Large epidemiologic studies with self-reported or routinely collected electronic health records (EHR) data are frequently being used as cost-effective ways to conduct clinical research, but these types of data are often prone to measurement error. While large epidemiologic studies play a crucial role in understanding the relationship between risk factors and health outcomes, such as disease incidence, these relationships cannot be properly understood unless methods are developed that reduce the bias caused by errors in both exposure variables and time-to-event outcome variables. Furthermore, variance estimates for outcome model regression parameters can be quite large in the presence of complex error-prone exposures and outcomes, yet strategies to improve variance estimation have been given little attention in the measurement error literature. Throughout this dissertation, we address these gaps in the literature by developing methodology that focuses on (1) reducing the bias that occurs from both error-prone exposures and outcomes in large epidemiologic cohort studies with periodic follow-up, (2) improving statistical efficiency by leveraging error-prone, auxiliary data alongside validated outcome data, and (3) considering alternative, better-behaved variance estimation strategies that may be used when techniques for adjusting for measurement error are applied. In Chapter 2, we present a method that combines an approach for addressing errors in event classification variables with regression calibration, a popular technique for addressing exposure error. This method reduces the bias induced by measurement errors in a discrete time-to-event setting. We apply our method to data from the Women's Health Initiative (WHI) study to evaluate the association between dietary energy and protein and incident diabetes. Chapter 3 develops an approach for incorporating error-prone, auxiliary data into the analysis of an interval-censored time-to-event outcome. Here, the key goal is to improve statistical efficiency in the estimation of exposure-disease associations. We extend our methodology to handle data from a complex survey design and to be used in conjunction with regression calibration. Using this approach, we assess the association between energy and protein and the risk of diabetes in our motivating study, the Hispanic Community Health Study/Study of Latinos (HCHS/SOL). In Chapter 4, we propose a sandwich variance estimator as an approach for accounting for the uncertainty added by using an estimated exposure when regression calibration is applied to adjust for covariate error. This variance approach broadly applies to other two-stage regression settings. We outline a procedure for easily computing the sandwich in standard software and assess its properties through a numerical study and through illustrative data examples from the WHI and HCHS/SOL studies. Our results show that this method may have advantages over commonly applied, resampling-based variance estimation approaches.

Degree Type

Dissertation

Degree Name

Doctor of Philosophy (PhD)

Graduate Group

Epidemiology & Biostatistics

First Advisor

Pamela A. Shaw

Keywords

measurement error, misclassification, proportional hazards, regression calibration, survival analysis

STATISTICAL APPROACHES TO REDUCING BIAS AND IMPROVING VARIANCE ESTIMATION IN THE PRESENCE OF COVARIATE AND OUTCOME MEASUREMENT ERROR

Lillian Boe

A DISSERTATION

in

Epidemiology and Biostatistics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2022

Supervisor of Dissertation

Pamela A. Shaw, Adjunct Associate Professor of Biostatistics

Graduate Group Chairperson

Nandita Mitra, Professor of Biostatistics

Dissertation Committee

Mary E. Putt, Professor of Biostatistics
Sharon X. Xie, Professor of Biostatistics
Shivan J. Mehta, Assistant Professor of Medicine

STATISTICAL APPROACHES TO REDUCING BIAS AND IMPROVING VARIANCE ESTIMATION IN THE PRESENCE OF COVARIATE AND OUTCOME MEASUREMENT ERROR

This dissertation is dedicated to my parents, Christine Washburn Boe Naher and Todd Boe. Mom and Dad, I wouldn't be who I am without you, and I certainly wouldn't have made it through graduate school without you. Thank you for giving me the world, and I hope I have made you proud. I love you.

ACKNOWLEDGEMENT

I would like to express my sincere gratitude and appreciation to my dissertation advisor, Dr. Pamela Shaw, who has provided excellent guidance, mentorship, and friendship throughout the last five years. In addition to being a patient and enthusiastic mentor, Pam has also sparked my passion for working in the measurement error space. With Pam's help, I have been able to grow my research skills and become a better statistician. I will always be thankful that I had the opportunity to work with Pam and I couldn't imagine a better dissertation advisor for my time in graduate school.

Thank you so much to my terrific dissertation committee members, Dr. Mary Putt, Dr. Sharon Xie, and Dr. Shivan Mehta, for the comprehensive advice throughout the years and countless hours of time they have dedicated to serving on my committee. I was fortunate to complete my first lab rotation at Penn under Mary's supervision, which planted the seeds for a fruitful and rewarding relationship throughout the rest of my time at Penn. Mary has been a terrific mentor and has always pointed me in the right direction as chair of my dissertation committee. Sharon, who I was lucky to have as a teacher for advanced survival analysis, has also been an incredibly important member of my committee. Sharon's expertise in survival analysis and measurement error methods have made her both a valuable mentor and an integral voice on the committee. As an Assistant Professor of Medicine, Shivan has been tremendously helpful in getting me to think about my research through a clinical lens. I am very grateful for Shivan's scientific advice and knowledge, which have always been crucial in helping me think about the big picture goals of my research efforts.

I will forever be thankful for the wonderful faculty in the GGEB, who were excellent teachers both in and out of the classroom. It was a joy to get to know them and have the opportunity to learn from so many of them over the years. I want to say a special thank you to Dr. Yimei Li and Dr. Kelly Getz, who supervised my M.S. thesis and continued to serve as excellent mentors for my entire time in graduate school. I also want to express my sincere

thanks to the chair of the GGEB, Dr. Nandita Mitra, for being a friendly face who always checked in with me throughout my graduate school journey.

I would also like to thank my collaborators who provided valuable contributions and input to the work in this dissertation. Dr. Lesley Tinker's expertise with the Women's Health Initiative data enabled me to maximize the quality of my data analysis for the work completed in Chapter 2. This work could not have been completed without the input of Dr. Thomas Lumley, who provided incredibly valuable contributions to the work in Chapter 3 and Chapter 4 in this dissertation. Thomas has helped fill in many gaps in my knowledge of survey sampling and has also generally provided excellent statistical advice throughout my years in graduate school. Thank you also to Dr. Daniela Sotres-Alvarez, who provided extremely valuable guidance in all data analyses involving the Hispanic Community Health Study/Study of Latinos.

I also want to acknowledge Dr. Bryan Shepherd and all members of the Penn/Vanderbilt/Auckland research group. Our biweekly research calls always served as a helpful space for sharing research ideas and getting feedback on my dissertation work. Thank you also to the members of STRATOS TG4, especially Dr. Larry Freedman, Dr. Doug Midthune, Dr. Victor Kipnis, and Dr. Paul Gustafson, who I had the chance to work with during my time in graduate school and who provided me with helpful advice as an early career statistician.

I am also incredibly grateful for all of the wonderful students I had the chance to meet during my time as a graduate student in the GGEB. I know I'll reflect back on my time in graduate school with fondness for the culture of community and support that we had as students. Thank you to so many of you for your help in tackling challenging homework assignments, for your many pieces of advice, and most importantly, for your friendship. I was especially lucky to enter graduate school with the most amazing and supportive cohort, who quickly became lifelong friends.

Thank you also to all investigators and members of the Publications Committees from the

Women's Health Initiative and the Hispanic Community Health Study/Study of Latinos for allowing me to use their data throughout this dissertation.

My entire graduate school experience would not have been possible without the love and support of my family and friends. To my mom and dad, Christine and Todd, thank you for literally everything. Your compassion, kindness, love, and generosity have been vital to my wellbeing during my time in school. Thank you for each taking me in on separate occasions during the pandemic and for being my support system when graduate school became remote. I'll always treasure the opportunity I had to live with each of you as an adult and for the love you showed me as I worked toward my PhD. And thank you to their spouses, my stepparents, Rick and Lori, who are our wonderful, more-recent additions to the family and have never failed to show me endless love and kindness.

Thank you also to my brother, Andrew, who has made life more fun these last five years. I can't thank you enough for always being one of my biggest fans, near or far, and for just being the most awesome little brother I could ever ask for.

And finally, to my partner, Ian, who has been there for me from the day I decided to apply to graduate school six years ago up until the moment I wrote the very last sentence of my dissertation: thank you. I can't thank you enough for supporting my dreams, even though it meant living in different cities for several years. Thank you for sticking by my side (literally or figuratively) all these years. I truly could not have done this without you.

I'm also very lucky to have a large extended family and many dear friends that always provided a very strong culture of support throughout my graduate school experience. Thank you all. I am forever grateful.

ABSTRACT

STATISTICAL APPROACHES TO REDUCING BIAS AND IMPROVING VARIANCE ESTIMATION IN THE PRESENCE OF COVARIATE AND OUTCOME MEASUREMENT ERROR

Lillian Boe

Pamela A. Shaw

Large epidemiologic studies with self-reported or routinely collected electronic health records (EHR) data are frequently being used as cost-effective ways to conduct clinical research, but these types of data are often prone to measurement error. While large epidemiologic studies play a crucial role in understanding the relationship between risk factors and health outcomes, such as disease incidence, these relationships cannot be properly understood unless methods are developed that reduce the bias caused by errors in both exposure variables and time-to-event outcome variables. Furthermore, variance estimates for outcome model regression parameters can be quite large in the presence of complex error-prone exposures and outcomes, yet strategies to improve variance estimation have been given little attention in the measurement error literature. Throughout this dissertation, we address these gaps in the literature by developing methodology that focuses on (1) reducing the bias that occurs from both error-prone exposures and outcomes in large epidemiologic cohort studies with periodic follow-up, (2) improving statistical efficiency by leveraging error-prone, auxiliary data alongside validated outcome data, and (3) considering alternative, better-behaved variance estimation strategies that may be used when techniques for adjusting for measurement error are applied. In Chapter 2, we present a method that combines an approach for addressing errors in event classification variables with regression calibration, a popular technique for addressing exposure error. This method reduces the bias induced by measurement errors in a discrete time-to-event setting. We apply our method to data from the Women's Health Initiative (WHI) study to evaluate the association between dietary energy and protein and

incident diabetes. Chapter 3 develops an approach for incorporating error-prone, auxiliary data into the analysis of an interval-censored time-to-event outcome. Here, the key goal is to improve statistical efficiency in the estimation of exposure-disease associations. We extend our methodology to handle data from a complex survey design and to be used in conjunction with regression calibration. Using this approach, we assess the association between energy and protein and the risk of diabetes in our motivating study, the Hispanic Community Health Study/Study of Latinos (HCHS/SOL). In Chapter 4, we propose a sandwich variance estimator as an approach for accounting for the uncertainty added by using an estimated exposure when regression calibration is applied to adjust for covariate error. This variance approach broadly applies to other two-stage regression settings. We outline a procedure for easily computing the sandwich in standard software and assess its properties through a numerical study and through illustrative data examples from the WHI and HCHS/SOL studies. Our results show that this method may have advantages over commonly applied, resampling-based variance estimation approaches.

TABLE OF CONTENTS

ACKNOWLEDGEMENT iv

ABSTRACT vii

LIST OF TABLES xi

LIST OF ILLUSTRATIONS xxi

CHAPTER 1 : Introduction 1

CHAPTER 2 : An Approximate Quasi-Likelihood Approach for Error-Prone Failure Time Outcomes and Exposures 5

2.1 Abstract 5

2.2 Introduction 5

2.3 Methods 8

2.4 Numerical Study 16

2.5 Women's Health Initiative (WHI) Example 22

2.6 Discussion 26

CHAPTER 3 : An Augmented Likelihood Approach for the Discrete Proportional Hazards Model Using Auxiliary and Validated Outcome Data – with Application to the HCHS/SOL Study 36

3.1 Abstract 36

3.2 Introduction 36

3.3 Methods 39

3.4 Numerical Study 45

3.5 Hispanic Community Health Study/Study of Latinos (HCHS/SOL) Data Example 51

3.6 Discussion 54

CHAPTER 4 : Practical considerations for sandwich variance estimation in two-stage regression settings 64

4.1 Abstract 64

4.2 Introduction 64

4.3 Motivating Data Examples 66

4.4 Methods 67

4.5 Simulation Study 74

4.6 Reanalysis of WHI and HCHS/SOL Data 75

4.7 Discussion 77

CHAPTER 5 : Discussion 86

APPENDICES 91

BIBLIOGRAPHY 151

LIST OF TABLES

TABLE 2.1 The mean percent (%) biases, average standard errors (ASE), empirical standard errors (ESE) and coverage probabilities (CP) are given for 1000 simulated data sets for the proposed method and naive method with $\beta_{X1} = \log(1.5)$, $\beta_{Z1} = \log(0.7)$, and $\beta_{Z2} = \log(1.3)$; e is normally distributed with mean zero. 31

TABLE 2.2 The mean percent (%) biases, average standard errors (ASE), empirical standard errors (ESE) and coverage probabilities (CP) are given for 1000 simulated data sets for the proposed method and naive method with $\beta_{X1} = \log(3)$, $\beta_{Z1} = \log(0.7)$, and $\beta_{Z2} = \log(1.3)$; e is normally distributed with mean zero. 32

TABLE 2.3 The mean percent (%) biases, average standard errors (ASE), empirical standard errors (ESE) and coverage probabilities (CP) are given for 1000 simulated data sets for the proposed method and naive method with $\beta_{X1} = \log(1.5)$, $\beta_{Z1} = \log(0.7)$, and $\beta_{Z2} = \log(1.3)$; e is distributed as either a t with 4 df or as $.4\mathcal{N}(0,1) + .6\mathcal{N}(2,1.5)$. . . 33

TABLE 2.4 The mean percent (%) biases, average standard errors (ASE), empirical standard errors (ESE) and coverage probabilities (CP) are given for 1000 simulated data sets for the proposed method and naive method, when both allow for strata-specific baseline hazards. We assume four equally-sized strata. Let $\beta_{X1} = \log(1.5)$, $\beta_{Z1} = \log(0.7)$, and $\beta_{Z2} = \log(1.3)$; e is normally distributed with mean zero. 34

TABLE 2.5 Type I error results for $\beta_{X1} = 0$ are given for 1000 simulated data sets for the proposed method. Let $\beta_{X1} = \log(1.5)$, $\beta_{Z1} = \log(0.7)$, and $\beta_{Z2} = \log(1.3)$; e is normally distributed with mean zero. 35

TABLE 2.6 Hazard Ratio (HR) and 95% confidence interval (CI) estimates of incident diabetes for a 20% increase in consumption of energy (kcal/d), protein (g/d), and protein density (% energy from protein/d) based on the naive method ignoring error in the outcome and covariate, the regression calibration method that corrects for covariate error only, and the proposed method. Here, sensitivity = 0.61, specificity = 0.995, and negative predictive value = 0.96. 35

TABLE 3.1 Simulation results are shown for exponential failure times assuming the Cox proportional hazards model with $X \sim Gamma(0.2, 1)$ and $\beta = \log(1.5)$. The median percent (%) bias, median standard errors (ASE), empirical median absolute deviation (MAD) and coverage probabilities (CP) are given for 1000 simulated data sets for the proposed method and the standard interval-censored approach that does not incorporate auxiliary data. Here, $Se = 0.80$ and $Sp = 0.90$ for the auxiliary data. 58

TABLE 3.2 Simulation results are shown for exponential failure times assuming the Cox proportional hazards model with $X \sim Normal(0.2, 1)$ and $\beta = \log(1.5)$.The median percent (%) bias, median standard errors (ASE), empirical median absolute deviation (MAD) and coverage probabilities (CP) are given for 1000 simulated data sets for the proposed method and the standard interval-censored approach that does not incorporate auxiliary data. Here, $Se = 0.80$ and $Sp = 0.90$ for the auxiliary data. 59

TABLE 3.3 Simulation results are shown for data simulated to be from a complex survey with exponential failure times assuming the Cox proportional hazards model with $X \sim Gamma(\text{shape}_s + \omega_{gs}, \text{scale}_s + \rho_{gs})$ for an individual in block group g and stratum s and $\beta = \log(1.5)$. The median percent (%) bias, median standard errors (ASE), median absolute deviation (MAD) and coverage probabilities (CP) are given for 1000 simulated data sets for the weighted proposed estimator and the weighted interval-censored approach that does not incorporate auxiliary data when both use a sandwich variance estimator to address within-cluster correlation. Here, $Se = 0.80$ and $Sp = 0.90$ for the auxiliary data. 60

TABLE 3.4 Simulation results are shown for data simulated to have a similar structure to the complex survey design of HCHS/SOL, assuming exponential failure times and the Cox proportional hazards model with $\beta = \log(1.5)$. The median percent (%) bias, median standard errors (ASE), median absolute deviation (MAD) and coverage probabilities (CP) are given for 1000 simulated data sets for the proposed estimator and the interval-censored approach that does not incorporate auxiliary data when both apply regression calibration to address covariate error. Variance estimation is performed using the resampling-based multiple imputation procedure of Baldoni et al. (2021). 61

TABLE 3.5 Type I error results for $\beta = 0$ are given for 1000 simulated data sets for the proposed method when data are simulated using exponential failure times and assuming the Cox proportional hazards model with $X \sim Gamma(0.2, 1)$. Here, $Se = 0.80$ and $Sp = 0.90$ for the auxiliary data. 62

TABLE 3.6 HCHS/SOL Data Analysis on a random subset ($N = 8,200$) of study participants using baseline sensitivity ($Se = 0.61$) and specificity ($Sp = 0.98$) values. Hazard Ratio (HR) and 95% confidence interval (CI) estimates of incident diabetes for a 20% increase in consumption of energy (kcal/d), protein (g/d), and protein density (% energy from protein/d) based on the proposed estimator and the interval-censored approach that does not incorporate auxiliary data. 63

TABLE 4.1 The median percent (%) bias, median standard errors (ASE), empirical median absolute deviation (MAD) and coverage probabilities (CP) for 1000 simulated data sets from a simple random sample for a logistic regression stage 2 model fit to true exposure, naive exposure, and calibrated exposure with naive (model-based) standard errors, sandwich standard errors, and bootstrap standard errors. We vary the correlation between X and Z, the sample size (N), and the measurement error variance (σ^2). Sample size of the calibration subset is $n = 450$. 82

TABLE 4.2 The median percent (%) bias, median standard errors (ASE), empirical median absolute deviation (MAD) and coverage probabilities (CP) for 1000 simulated data sets from a simple random sample for a logistic regression stage 2 model fit to true exposure, naive exposure, and calibrated exposure with naive (model-based) standard errors, sandwich standard errors, and MI standard errors. We vary the correlation between X and Z, the sample size (N), and the measurement error variance (σ^2). Sample size of the calibration subset is $n = 450$. 83

TABLE 4.3 WHI data analysis (N=77,805) results from the Cox Proportional Hazards model for incident diabetes with dietary exposures of energy (kcal/d), protein (g/d), and protein density (% energy from protein). Results are shown for each stage 2 model fit to the calibrated exposure with naive (model-based) standard errors, sandwich standard errors, and bootstrap standard errors. 84
TABLE 4.4 HCHS/SOL data analysis ($N = 8,176$) results from the linear regression of baseline systolic blood pressure and the logistic regression of hypertension status each on log-transformed intake of potassium. Results are shown for each stage 2 model fit to the calibrated exposure with naive (model-based) standard errors, sandwich standard errors, and standard errors from the multiple imputation (MI) approach of Baldoni et al. (2021). 85
TABLE A.1 The mean percent (%) biases, average standard errors (ASE), empirical standard errors (ESE) and coverage probabilities (CP) are given for 1000 simulated data sets for the proposed method, the naive method, a method that corrects for covariate error only, and a method that corrects for outcome error only, with $\beta_{X1} = \log(1.5)$, $\beta_{Z1} = \log(0.7)$, and $\beta_{Z2} = \log(1.3)$; e is normally distributed with mean zero; Sensitivity (Se)=0.80; Specificity (Sp)=0.90. 108
TABLE A.2 The mean percent (%) biases, average standard errors (ASE), empirical standard errors (ESE) and coverage probabilities (CP) are given for 1000 simulated data sets for the proposed method and naive method with $\beta_{X1} = \log(1.5)$, $\beta_{Z1} = \log(0.7)$, and $\beta_{Z2} = \log(1.3)$; e is normally distributed with mean zero. The censoring rate is fixed at 0.90. Here, we vary sensitivity, specificity, and negative predictive value. 109

TABLE A.3 The mean percent (%) biases, average standard errors (ASE), empirical standard errors (ESE) and coverage probabilities (CP) are given for 1000 simulated data sets for the proposed method and naive method with $\beta_{X1} = \log(1.5)$, $\beta_{Z1} = \log(0.7)$, and $\beta_{Z2} = \log(1.3)$; e is normally distributed with mean zero. The censoring rate is fixed at 0.90. Here, we vary sensitivity, specificity, and probability of missingness at each visit. 110

TABLE A.4 The mean percent (%) biases, average standard errors (ASE), empirical standard errors (ESE) and coverage probabilities (CP) are given for 1000 simulated data sets for the proposed method, naive method, method that corrects for covariate error only, and method that corrects for outcome error only for a simulated dataset with similar features to the Women's Health Initiative (WHI) data. Here, Sensitivity (Se)=0.61, Specificity (Sp)=0.995, Negative Predictive Value (η) = 0.96, $\beta_{X1} = \log(1.5)$, $\beta_{Z1} = \log(0.7)$, $\beta_{Z2} = \log(1.3)$, e is normally distributed with mean zero, and the censoring rate for the error-prone indicator is fixed at 0.95. 111

TABLE A.5 Hazard Ratio (HR) and 95% confidence interval (CI) estimates of incident diabetes for a 20% increase in consumption of energy (kcal/d), protein (g/d), and protein density (% energy from protein/d) based on the naive method ignoring error in the outcome and covariate, the method corrected for error in the covariate only, and the proposed method. Here, sensitivity = 0.61, specificity = 0.995, and negative predictive value = 1. 112

TABLE A.6 Sensitivity Analysis varying sensitivity and specificity of diabetes self-reports across WHI DM-C and WHI OS participants. We consider separate models for dietary energy, protein, and protein density. Each model is adjusted for potential confounders, including BMI, and is stratified on age (10-year categories) and DM or OS cohort membership. We show HR estimates of incident diabetes for a 20% increase in consumption of energy (kcal/d), protein (g/d), and protein density (% energy from protein/d). 112
TABLE A.7 Sensitivity Analysis for different stratification strategies using a modeling approach similar to that of Tinker et al. Tinker et al. (2011) We examine hazard ratio (HR) and 95% confidence interval (CI) estimates of incident diabetes for a 20% increase in consumption of energy (kcal/d), protein (g/d), and protein density (% energy from protein/d) based on discrete proportional hazards analyses and continuous Cox proportional hazards models that correct for error in the covariate (X) only. 113
TABLE A.8 For model used by Tinker et al., Tinker et al. (2011), we examine the sensitivity of results to choices of how BMI is treated in analyses. We present hazard ratio (HR) and 95% confidence interval (CI) estimates of incident diabetes for a 20% increase in consumption of energy (kcal/d), protein (g/d), and protein density (% energy from protein/d) based on the discrete proportional hazards model with a post-hoc correction for covariate error, the continuous Cox model with a post-hoc correction for covariate error, and the continuous Cox model with the non-post-hoc traditional regression calibration correction for covariate error. 114

TABLE B.1 Simulation results are shown for exponential failure times assuming the Cox proportional hazards model with $X \sim Gamma(0.2, 1)$ and $\beta = \log(1.5)$ for (1) the grouped time survival approach that uses the true outcome data from all periodic visits and (2) the standard interval-censored approach that does not incorporate auxiliary data. The median percent (%) bias, median standard errors (ASE), empirical median absolute deviation (MAD) and coverage probabilities (CP) are given for 1000 simulated data sets. 132

TABLE B.2 Simulation results are shown for exponential failure times assuming the Cox proportional hazards model with $X \sim Gamma(0.2, 1)$, $\beta = \log(1.5)$, and values of $Se = 0.90$ and $Sp = 0.80$ for the auxiliary data. The median percent (%) bias, median standard errors (ASE), empirical median absolute deviation (MAD) and coverage probabilities (CP) are given for 1000 simulated data sets for the proposed method and the standard interval-censored approach that does not incorporate auxiliary data. 133

TABLE B.3 Simulation results are shown for exponential failure times assuming the Cox proportional hazards model with $X \sim Gamma(0.2, 1)$ and $\beta = \log(3)$. The median percent (%) bias, median standard errors (ASE), empirical median absolute deviation (MAD) and coverage probabilities (CP) are given for 1000 simulated data sets for the proposed method and the standard interval-censored approach that does not incorporate auxiliary data. Here, $Se = 0.80$ and $Sp = 0.90$ for the auxiliary data. 134

TABLE B.4 Simulation results are shown for data simulated to be from a complex survey with exponential failure times assuming the Cox proportional hazards model with $X \sim Normal(\text{shape}_s + \omega_{gs}, \text{scale}_s + \rho_{gs})$ for an individual in block group g and stratum s and $\beta = \log(1.5)$. The median percent (%) bias, median standard errors (ASE), median absolute deviation (MAD) and coverage probabilities (CP) are given for 1000 simulated data sets for the weighted proposed estimator and the weighted interval-censored approach that does not incorporate auxiliary data when both use a sandwich variance estimator to address within-cluster correlation. Here, $Se = 0.80$ and $Sp = 0.90$ for the auxiliary data. 135

TABLE B.5 Sensitivity analysis using HCHS/SOL data on a subset of study participants with visit 2 sensitivity ($Se = 0.77$) and specificity ($Sp = 0.92$) values. Hazard Ratio (HR) and 95% confidence interval (CI) estimates of incident diabetes for a 20% increase in consumption of energy (kcal/d), protein (g/d), and protein density (% energy from protein/d) based on the proposed estimator and the interval-censored approach that does not incorporate auxiliary data. 136

TABLE C.1 Simulation results are shown for the Cox proportional hazards regression model (event rate = 0.38) for data simulated to be from a simple random sample. The median percent (%) bias, median standard errors (ASE), empirical median absolute deviation (MAD) and coverage probabilities (CP) are given for 1000 simulated data sets for the outcome model fit to true exposure, naive exposure, calibrated exposure with naive (model-based) standard errors, calibrated exposure with standard errors from the sandwich approach, and calibrated exposure with standard errors from the bootstrap approach ($B = 500$ bootstrap samples). We vary the correlation between the error-prone and precisely-measured covariates (0.3 or 0.7), the sample size (N), and the measurement error variance (σ^2). Sample size of the calibration subset is $n = 450$. 149

TABLE C.2 Simulation results are shown for logistic regression (event rate = 0.38) for the outcome model fit to true exposure, naive exposure, calibrated exposure with naive (model-based) standard errors, calibrated exposure with standard errors from the sandwich approach, and calibrated exposure with standard errors from the bootstrap approach, with bootstrap confidence intervals constructed in 3 ways for $B = 1000$ bootstrap samples. The median percent (%) bias, median standard errors (ASE), empirical median absolute deviation (MAD) and coverage probabilities (CP) are given for 1000 simulated data sets with $N = 1000$ each. We vary the correlation between the error-prone and precisely-measured covariates (0.3 or 0.7) and the measurement error variance (σ^2). Sample size of the calibration subset is $n = 450$. 150

LIST OF ILLUSTRATIONS

FIGURE 4.1 Code for obtaining the sandwich matrix using functions from the survey package for a simple random sample 80

FIGURE 4.2 Code for obtaining the sandwich matrix using functions from the survey package for a complex survey design 81

FIGURE A.1 Estimated nonparametric maximum likelihood estimators (NPMLEs) of the survival distribution for the error-prone outcomes compared to true outcomes for the simulation study, fit using the R package 'interval.'(Fay and Shaw, 2010) Panel A corresponds to censoring rate = 0.90 (baseline hazard = 0.012) with observation times $(2, 5, 7, 8)$. Panel B corresponds to censoring rate = 0.55 (baseline hazard = 0.094) with observation times $(1, 3, 4, 6)$. Vertical lines represent observation times. Simulated from data with $\beta_{X1} = \log(1.5)$, $\beta_{Z1} = \log(0.7)$, $\beta_{Z2} = \log(1.3)$, $e \sim \mathcal{N}(0, 1.31)$, sensitivity = 0.90, and specificity = 0.80. 107

CHAPTER 1

Introduction

There has recently been increasing interest in using data from large epidemiologic studies with routinely collected electronic health records (EHR) data as cost-effective ways to conduct clinical research. Large, prospective cohort studies are essential for identifying risk factors for chronic diseases (e.g. diabetes, cardiovascular disease, and pulmonary disease) in at-risk populations, but oftentimes have methodological complexities that need to be addressed in the statistical analysis stage. In particular, large cohort studies with routinely collected or self-reported data have the potential for measurement error. Measurement error has been shown to impact studies focused on nutritional epidemiology, physical activity, air pollution, HIV, and several other areas relevant to public health (Keogh et al., 2020; Shepherd and Shaw, 2020).

Measurement errors in both covariate and outcome variables pose a considerable challenge to the reliable estimation of exposure-disease associations in cohort studies. The impact of covariate measurement error has been well-studied, and consequently, methods have been developed to reduce the bias caused by errors in covariates (Carroll et al., 2006; Yi, 2017). Specifically, regression calibration is the most widely-used approach for addressing exposure errors, likely because it is relatively intuitive to understand and straightforward to implement. This method involves building a "calibration" model for the expected value of the true exposure given the error-prone variable and other observed, error-free covariates (Prentice, 1982). Regression calibration is an exact fix in linear models, but only an approximate correction for non-linear models. While outcome errors in time-to-event settings have been less studied compared to covariate errors, a few methods have been developed to address errors in binary outcome variables in discrete time-to-event settings (Meier et al., 2003; Magaret, 2008; Gu et al., 2015). Although correcting for measurement error is needed to make valid statistical inference, there are few methods available that simultaneously correct errors in both covariates and outcomes (Shaw et al., 2018). One of the key goals of this dissertation is

addressing the errors that may occur in both covariates and outcomes in large, prospective cohort studies.

Measurement error in covariates or outcome variables can apprexiably increase the uncertainty in study estimates. Thus, in cohort studies where there is substantial exposure measurement error, variance estimates can be quite large. Any ways to improve the variance estimation of outcome model regression parameters have the potential to be very impactful. One strategy for reducing the estimated variance in these settings is to leverage all available outcome data. In some large cohort studies, error-prone outcome variables are available alongside gold or reference standard outcome variables. In these settings, the error-prone data can be thought of as “auxiliary" data that may be incorporated into the analyses in order to improve statistical efficiency. Some authors have previously considered using auxiliary data to improve statistical efficiency in time-to-event settings (Pepe, 1992; Zee et al., 2015; Fleming et al., 1994). However, none of these authors have considered the setting common to epidemiologic cohort studies in which the auxiliary, error-prone outcome is obtained through periodic follow-up and thus may be observed more frequently (potentially both before and after) the corresponding gold standard outcome variable. Additionally, no existing methods that leverage error-prone auxiliary outcome data have considered the setting where exposure variables may be recorded with error as well.

When methodology to address measurement error is applied, additional steps are required in the variance estimation stage. For example, when regression calibration is used to correct for covariate error, the standard errors of the outcome model regression parameters need to be further adjusted for the extra uncertainty added by the calibration model step. Very little attention has been given to the comparison of competing approaches for variance estimation that may be used to account for this extra uncertainty, such as the sandwich variance estimator obtained by stacking the estimating equations from the calibration and outcome models (Boos and Stefanski, 2013). In fact, many investigators often rely on resampling-based approaches like the popular bootstrap, despite the fact that these methods tend to be

computationally intensive and may also result in problematic confidence intervals (Efron, 1979, 1987). Alternative variance estimation strategies like the sandwich, which may have better performance in certain scenarios, thus warrant consideration when applying methods like regression calibration to adjust for measurement error. In this dissertation, we also focus on improved variance estimation in the presence of measurement error, accomplished by (1) leveraging error-prone outcome data to improve statistical efficiency and (2) considering alternative, less popular strategies for variance estimation when techniques for adjusting for covariate error are applied.

In Chapter 2, we introduce a practical approach to correcting errors in covariates and a discrete time-to-event outcome that results in nearly unbiased estimates of the regression parameters of interest in typical applied settings. This method uses a regression calibration-type approach to correct biases in the outcome model regression parameters when continuous covariates are recorded with error. Our method can accommodate errors in a discrete failure time outcome variable when the sensitivity and specificity of the error-prone outcome are known. We apply this method to data from the Women's Health Initiative (WHI) study to assess the association between dietary energy, protein, and protein density (percentage of energy from protein) and incident diabetes when both the exposure and outcome variables are recorded by self-report and hence subject to substantial error.

In Chapter 3, we develop an augmented likelihood approach that incorporates error-prone, auxiliary data into the analysis of a gold-standard time-to-event outcome. The key goal of this approach is to improve statistical efficiency in the estimation of exposure-disease associations. To properly leverage the auxiliary data, we incorporate known values of sensitivity and specificity into our analysis to correct for the bias induced by errors in our event classification variable. This method is extended to accommodate data from a complex survey design so that it can be applied in our motivating study, the Hispanic Community Health Study/Study of Latinos (HCHS/SOL). The proposed method is also used in combination with regression calibration to additionally adjust for exposure error in order to assess the

association between energy and protein and the risk of diabetes in the HCHS/SOL cohort.

In Chapter 4, we introduce a practical approach for computing the sandwich variance estimator in two-stage regression model settings, such as regression calibration, to correct for covariate error. Regression calibration can be viewed as a two-stage model setting where in stage 1, a calibration model is fit to a subset to estimate the expected value of the unobserved true exposure on all study participants and in stage 2, the outcome model is fit to this estimated exposure and other confounders of interest. The sandwich variance estimator is one strategy for variance estimation that accounts for the uncertainty added by the calibration model step and the estimated exposure. We propose the sandwich estimator as an alternative to the popular resampling variance estimators used in practice, i.e. the bootstrap for the case where the data are a simple random sample setting and the multiple imputation procedure of Baldoni et al. (2021) for the case where the data are from a complex survey sampling design. A key reason the sandwich variance estimator has not been frequently implemented is the lack of available software to readily compute this estimate. We provide code illustrating how to easily compute the sandwich variance by taking advantage of existing functions from R software and we have developed functions in R, available on GitHub, which directly compute estimates of the sandwich variance for two stage regression model settings for a wide variety of regression models. We illustrate the performance of the sandwich alongside its competing variance estimators using a simulation study and by applying these approaches to data from the WHI and HCHS/SOL studies. In Chapter 5, we conclude by summarizing our research and highlighting how our work contributes to the measurement error literature. Additionally, we discuss a few important areas of future research that have been motivated by this dissertation.

CHAPTER 2

An Approximate Quasi-Likelihood Approach for Error-Prone Failure Time Outcomes and Exposures

2.1. Abstract

Measurement error arises commonly in clinical research settings that rely on data from electronic health records or large observational cohorts. In particular, self-reported outcomes are typical in cohort studies for chronic diseases such as diabetes in order to avoid the burden of expensive diagnostic tests. Dietary intake, which is also commonly collected by self-report and subject to measurement error, is a major factor linked to diabetes and other chronic diseases. These errors can bias exposure-disease associations that ultimately can mislead clinical decision-making. We have extended an existing semiparametric likelihood-based method for handling error-prone, discrete failure time outcomes to also address covariate error. We conduct an extensive numerical study to compare the proposed method to the naive approach that ignores measurement error in terms of bias and efficiency in the estimation of the regression parameter of interest. In all settings considered, the proposed method showed minimal bias and maintained coverage probability, thus outperforming the naive analysis which showed extreme bias and low coverage. This method is applied to data from the Women's Health Initiative to assess the association between energy and protein intake and the risk of incident diabetes mellitus. Our results show that correcting for errors in both the self-reported outcome and dietary exposures leads to considerably different hazard ratio estimates than those from analyses that ignore measurement error, which demonstrates the importance of correcting for both outcome and covariate error.

2.2. Introduction

Chronic diseases are often recorded primarily by self-reported diagnosis in large observational cohort studies. For example, in comparison to reference (gold) standard measures for detecting diabetes, such as fasting glucose and hemoglobin A1c (HbA1c), self-reported dia-

betes status is inexpensive and easily attainable. However, not all people who are diagnosed with diabetes or other conditions will self-report that they have the disease. Reasons for failing to report having a chronic condition include failure to be diagnosed, lack of understanding about the disease, and a belief that the disease has gone away if it is being properly treated (Centers for Disease Control and Prevention, 2017; Shah and Manuel, 2008). Conversely, a positive disease status is occasionally reported when the disease is not actually present (Ning et al., 2016; Schneider et al., 2012). Dietary intake, which is also commonly recorded by self-report, is thought to play a crucial role in determining the risk of chronic diseases such as diabetes and cardiovascular disease. In nutritional epidemiology, estimates of diet-disease associations can be distorted due to measurement error in both self-reported dietary exposures and disease outcomes. A new analytic approach is needed to properly relate error-prone exposures with error-prone disease outcomes of interest. In this paper, we have extended an existing semiparametric model for handling failure time outcomes assessed through interval-censored, error-prone measures to also address measurement error in the exposure variable.

There is ample literature available on methods for adjusting analyses with error-prone exposures in the case of time-to-event outcomes (Carroll et al., 2006). In existing epidemiological analyses, regression calibration is one of the most popular methods for addressing covariate measurement error (Shaw et al., 2018). This method relies on building a calibration model that relates the expected value of the unobserved true exposure to the observed data. Prentice (1982) introduced the method for time-to-event outcomes. Rosner et al. considered it for logistic regression, where a single or multiple covariates were error-prone Rosner et al. (1989, 1990). In non-linear models, such as Cox and logistic regression, regression calibration is considered a quasi-likelihood approach as it is generally only an approximate correction,(Buonaccorsi, 2010) but it has been observed to do well for modest β and low event rates. (Prentice, 1982; Carroll et al., 2006) The popularity of this approach likely has to do with the intuitive appeal of the method and the ease of implementation. The method proposed in this manuscript uses regression calibration in order to develop an estimator that

will correct for both covariate and outcome error.

Compared to methods for addressing covariate error, there has been notably less investigation into methods that correct for errors that occur in the time-to-event outcomes themselves. In epidemiologic cohort studies, the time-to-event of interest is often ascertained through periodic follow-up, thus resulting in data captured in fixed intervals. Thus, methods that address errors in the event indicator at each interval are of particular interest. Balasubramanian and Lagakos (2003) developed estimation methods for the distribution of the time-to-event that consider various periods of exposure and diagnostic tests with different levels of accuracy. Meier et al. (2003) presented an adjusted proportional hazards model for estimating hazard ratios in discrete time survival analysis when the outcome is measured with error. Magaret (2008) considered methods that adjusted the proportional hazards model to incorporate data from validation subsets for the case where the sensitivity and the specificity of the diagnostic tests are unknown. All of this existing work assumes that the covariates included in the time-to-event analyses are error-free, which is often untrue with clinical data.

This manuscript specifically builds on the work of Gu et al., Gu et al. (2015) which introduced a semiparametric likelihood-based approach for estimating the association of covariates with an error-prone discrete failure time outcome. Motivated by an example from the Women's Health Initiative (WHI), we extend this method to incorporate a regression calibration fix that additionally adjusts for covariate measurement error and also allows for strata-specific baseline hazards. Our method can be applied to a study cohort that has collected follow-up data on an error-prone disease status variable at two or more distinct visit times and has information available at baseline on specific covariates of interest. In the presence of covariate measurement error, the proposed method can be considered when there is data that informs the measurement error model. We must assume that (1) information is available regarding the sensitivity and specificity of the outcome measure (2) a second measure of the error-prone covariate(s) is available on at least a subset.

Section 2.3 introduces the theoretical development of the method by providing notation, constructing the likelihood function and discussing the proposed adjustment method that corrects for outcome and covariate error. Next, we examine the numerical performance of the proposed method with a simulation study in Section 2.4. In Section 2.5, we apply the proposed method to evaluate the association between dietary energy, protein, and protein density intake and incident diabetes in a subset of women enrolled in the WHI. Finally, we highlight the important findings of this work and discuss potential extensions in Section 2.6.

2.3. Methods

2.3.1. Notation and Time-to-Event Model

Let T_i be the unobserved time-to-event of interest for subjects $i = 1, ..., N$. Consider a study with periodic follow-up where each subject may have a slightly different visit schedule or missed visits. Define $\tau_1, ..., \tau_J$ as the distinct possible visit times among all N subjects. Denote $\tau_0 = 0$ and $\tau_{J+1} = \infty$. We assume that the time to the event of interest is continuous, but follow-up occurs at discrete visit times. The follow-up time period can then be divided into $J + 1$ disjoint intervals, listed as follows: $[\tau_0, \tau_1), [\tau_1, \tau_2), ...[\tau_J, \tau_{J+1})$. Assume that all subjects in the study are event-free at time τ_0. Later, we will relax this assumption. Let n_i be the number of visits for the i^{th} subject, which we assume is random. In our motivating data example, each subject self-reports his or her disease status at each visit time, potentially with error, up until the first positive. Our method can also be applied to the more general setting for error-prone outcomes in which follow-up continues beyond the first positive. Define $\mathbf{Y_i}$ and $\mathbf{t_i}$ as the random $1 \times n_i$ vector of error-prone outcomes and corresponding vector of visit times for subject i. Specifically, define Y_{ij} as 1 if the j^{th} error-prone outcome for i^{th} subject is positive, and 0 otherwise. Then, the joint probability of the observed data for the i^{th} subject is:

$$f(\mathbf{Y_i}, \mathbf{t_i}, n_i) = \sum_{j=1}^{J+1} \theta_j \Pr(\mathbf{Y_i}, \mathbf{t_i}, n_i | \tau_{j-1} < T_i \leqslant \tau_j), \tag{2.1}$$

where $\theta_j = \Pr(\tau_{j-1} < T_i \leqslant \tau_j)$.

We make the additional assumption that conditioned on the true event time T_i, the n_i error-prone outcomes Y_{ij} are independent, i.e. $\Pr(\mathbf{Y_i}|T_i, \mathbf{t_i}) = \prod_{l=1}^{n_i} \Pr(Y_{il}|T_i, t_{il})$. Thus, other observed error-prone outcomes do not provide additional information about a specific error-prone outcome beyond what is already given by the true time of event. Following the notation and logic of Balasubramanian and Lagakos, Balasubramanian and Lagakos (2003) it can be shown that for the case of a prespecified visit schedule, the likelihood becomes:

$$f(\mathbf{Y_i}, \mathbf{t_i}, n_i) = \sum_{j=1}^{J+1} \theta_j \left[\prod_{l=1}^{n_i} \Pr(Y_{il}|\tau_{j-1} < T_i \leqslant \tau_j, t_l) \right] = \sum_{j=1}^{J+1} \theta_j C_{ij}, \tag{2.2}$$

where $C_{ij} = \prod_{l=1}^{n_i} \Pr(Y_{il}|\tau_{j-1} < T_i \leqslant \tau_j, t_l)$. In Appendix A.2, we show how equation (2.2) becomes the expression for a subject's likelihood contribution.

For ease of presentation, we calculate C_{ij} for the case of no missed visits, but the formula can be easily adapted to accommodate missed visits by summing up the θ_j for the $(\tau_{j-1}, \tau_j]$ that define each subject's observational interval. We assume constant and known sensitivity (Se) and specificity (Sp); namely, $Se = \Pr(Y_{il} = 1|\tau_{j-1} < T_i \leqslant \tau_j, t_l \geqslant \tau_j)$ and $Sp = \Pr(Y_{il} = 0|\tau_{j-1} < T_i \leqslant \tau_j, t_l \leqslant \tau_{j-1})$. Then, the C_{ij} terms take the following form:

$$\begin{aligned}
C_{i1} &= Se^{\sum_{j=1}^{n_i} Y_{ij}} (1 - Se)^{\sum_{j=1}^{n_i} (1 - Y_{ij})}, \\
C_{i2} &= Sp^{(1-Y_{i1})} (1 - Sp)^{Y_{i1}} Se^{\sum_{j=2}^{n_i} Y_{ij}} (1 - Se)^{\sum_{j=2}^{n_i} (1 - Y_{ij})}, \\
&\ldots \\
C_{i(J+1)} &= Sp^{\sum_{j=1}^{n_i} (1 - Y_{ij})} (1 - Sp)^{\sum_{j=1}^{n_i} Y_{ij}}.
\end{aligned}$$

Now suppose we have the proportional hazards model, $S(t) = S_0(t)^{\exp(x^T \beta_X + z^T \beta_Z)}$. We assume that one or more covariates are recorded with error. Define X_i^* as a $p-$dimensional vector of covariates of interest that may be observed with error, and X_i a corresponding $p-$dimensional vector of unobserved true exposure variables. We describe the error structure of the observed error-prone covariate X^* in Section 2.3.2. Let Z_i be a $q-$dimensional vector of precisely measured covariates (i.e. error-free) that may be correlated with X_i. Define $\beta = (\beta_X, \beta_Z)^T$. The likelihood can be rewritten in terms of the baseline survival probabilities

$\mathbf{S} = (S_1, S_2, ..., S_{J+1})^T$ defined by the random variable T_0 with survival function $S_0(t)$, where $S_j = \Pr(T_0 > \tau_{j-1})$. One then has $1 = S_1 > S_2 > ... > S_{J+1} > 0$ and $S_j = \sum_{h=j}^{J+1} \theta_h$. We can define a linear $(J+1) \times (J+1)$ transformation matrix M such that $\theta = M\mathbf{S}$. Finally, define the $(N) \times (J+1)$ matrix $D = CM$, where $C_{N\times(J+1)}$ consists of the C_{ij} elements defined above. Following Gu et al., Gu et al. (2015) the log-likelihood can be rewritten as:

$$l(\mathbf{S}, \beta) = \sum_{i=1}^{N} \log \left(\sum_{j=1}^{J+1} D_{ij} S_j^{\exp(x_i^T \beta_X + z_i^T \beta_Z)} \right). \tag{2.3}$$

The D_{ij} components of the log-likelihood consist of elements of the matrix D defined above and are functions of the observed data, (X_i, Z_i, Y_i, t_i), as well as Se and Sp. One can apply the usual maximum likelihood approach to solve for the unknown parameters β_X, β_Z, S_2,...,S_{J+1}. The covariance matrix can be found by inverting the Hessian matrix. Note that the model above introduced by Gu et al. (2015) is considered semiparametric because we do not make any assumptions about the form of the baseline survival probabilities, S_j, for $j = 1, ..., J+1$.

2.3.2. Proposed Method for Outcome and Covariate Error

We now extend the above method that corrects for outcome error in the discrete proportional hazards model to also adjust for covariate error by adopting a regression calibration type approach. In this section, we describe the regression calibration approach for covariate measurement error, present our proposed method to adjust for covariate and outcome error, extend our method to accommodate a baseline hazard that varies across strata, and extend the method to handle false negatives that are mistakenly included in the analysis.

Regression calibration for covariate error

Regression calibration is an approach to correcting biases in regression parameters when exposure variables are recorded with error, in which a calibration equation for the unobserved exposure X is estimated. Namely, one builds a model for $E(X|X^*, Z)$, where X^* is the error-prone observation or surrogate for X while Z are the other precisely observed covariates in

the outcome model (2.3). Regression calibration may be used when X^* follows the classical measurement error model or when X^* has both systematic and classical random error. These error settings will be explained in further detail in the subsequent section. Rosner et al. (1989) introduced a post-hoc calibration fix in the logistic regression setting when there is measurement error in a single covariate of interest and Rosner et al. (1990) extended the method to handle multiple error-prone covariates in logistic regression. In each of these approaches, the calibration equation is used to correct the naive parameter estimates that are obtained from first fitting the outcome regression that ignores the measurement error. An asymptotic formula for the variance that incorporates the uncertainty of the calibration equation is derived using the Delta method. We will employ a similar post-hoc calibration fix-up for the estimator that first corrects for outcome measurement error. We further justify why this post-hoc correction approach is expected to work well in our discrete-time proportional hazards setting at the end of this section.

Proposed approach for outcome and covariate error

Recall that X_i is a $p-$dimensional vector of true, unobserved covariates, while Z_i is a $q-$dimensional vector of observed, precisely measured covariates possibly correlated with X_i. Instead of observing X_i, we assume an error-prone X_i^* is observed, where X_i^* is assumed to be linearly related with X_i and possibly other covariates Z_i. This error model has been commonly applied in many settings, including nutritional epidemiology (Carroll et al., 2006; Keogh et al., 2020). The regression calibration model then takes the following form:

$$X_i = \delta_{(0)} + \delta_{(1)} X_i^* + \delta_{(2)} Z_i + U_i, \tag{2.4}$$

where U_i is a random, mean 0 error term, which is independent of X_i^* and Z_i. Equation (2.4) directly implies that our observed, error-prone variable X_i^* follows the linear measurement error model, i.e. $X_i^* = \alpha_{(0)} + \alpha_{(1)} X_i + \alpha_{(2)} Z_i + e_i$, where the random error e_i is independent of X_i and Z_i (Keogh et al., 2020). Note that we also assume non-differential error, i.e. the distribution of T conditional on (X, X^*, Z) is equal to the distribution of T conditional on

(X, Z). The model parameters in equation (2.4) are identifiable if we have a calibration subset available in which we observe the error-prone measure X_i^*, as well as a measure X_i^{**} that is unbiased for the true X_i and follows the classical measurement error model:

$$X_i^{**} = X_i + \epsilon_i, \tag{2.5}$$

where ϵ_i is random, mean 0 error that is independent of X_i. X^{**} is often referred to as an imperfect reference or alloyed gold standard Shaw et al. (2020); Spiegelman et al. (1997). Note that ϵ_i are assumed to be independent of all variables in the outcome model (2.3). Observing the exact true exposure X_i in the ancillary data is a special case of observing X_i^{**} where the error variance is 0, and the subset is typically called a validation subset. A special case of the linear measurement error model occurs when $\alpha_{(0)} = \alpha_{(2)} = 0$ and $\alpha_{(1)} = 1$, and thus the observed error-prone measurement X_i^* has classical measurement error. In this scenario, we can estimate the parameters of the calibration model by assuming that we observe replicates of X^*. Ancillary data of this type is typically referred to as a reliability subset.

When a calibration or validation subset is available, one can adopt a regression calibration type approach to further correct the regression coefficients for error in the exposure variable. In the case of a calibration subset, we regress X_i^{**} on the error-prone exposure, X_i^*, and other covariates of interest Z_i to fit the model:

$$X_i^{**} = \delta_{(0)} + \delta_{(1)} X_i^* + \delta_{(2)} Z_i + V_i, \tag{2.6}$$

where V_i, is random, mean 0 error. Note the model in equation (2.6) differs from that in equation (2.4) only in that the error term V_i incorporates the extra variability introduced by the error term in X_i^{**}. Estimates of the coefficients from fitting this linear regression can then be used to correct the β coefficients from the time-to-event model. Following the

approach of Rosner et al., Rosner et al. (1990) the corrected β can be found by solving:

$$\hat{\beta} = \hat{\beta}^* \hat{\Delta}^{-1}, \tag{2.7}$$

where $\hat{\beta}^*$ is the partially "naive" regression coefficient obtained from the time-to-event model ignoring the error in X^*, and $\hat{\Delta}$, the estimated multivariate correction factor, is defined as:

$$\hat{\Delta} = \begin{bmatrix} \hat{\delta}_{(1)p\times p} & \hat{\delta}_{(2)p\times q} \\ 0_{q\times p} & I_{q\times q} \end{bmatrix}. \tag{2.8}$$

The variance-covariance matrix Σ for $\hat{\beta}$ is calculated using the multivariate delta method. Assume that $\hat{\beta}^*$ and $\hat{\Delta}$ are independent, which holds if the calibration subset is an independent group of individuals from the main study (i.e. the main study data and the calibration subset are either independent data sets or are mutually exclusive subsets of the same set of data) and approximately holds if the number of subjects in the calibration subset, n_c, is a small percentage of the main study sample size, N (Rosner et al., 1989). Once we make this connection, we see that we can apply the same formulas as Rosner et al. (1990) and therefore the $(j_1, j_2)^{th}$ element of $\hat{\Sigma}$ for $\hat{\beta}$ is

$$\hat{\Sigma}_\beta(j_1, j_2) \cong \left(\hat{A}'\hat{\Sigma}_{\beta*}\hat{A}\right)_{j_1,j_2} + \hat{\beta}^*\hat{\Sigma}_{A,j_1,j_2}\hat{\beta}^{*\prime}, \tag{2.9}$$

where $\hat{A} = \hat{\Delta}^{-1}$, $\hat{\Sigma}_{\beta*}$ is the corresponding estimated variance-covariance matrix, and $\hat{\Sigma}_{A,j_1,j_2}$ is described below. Note that $\hat{\Sigma}_{\beta*}$ can be estimated from the model introduced above that only adjusts for outcome error. The matrix $\hat{\Sigma}_\beta(j_1, j_2)$ is essentially a sum of two pieces: the first can be viewed as the contribution of the uncertainty in estimating β^* and the second is a contribution of the uncertainty in the calibration coefficients. Following Rosner et al., Rosner et al. (1990) the $(i_1, i_2)^{th}$ element of $\hat{\Sigma}_{A,j_1,j_2}$, for $i_1, i_2, j_1, j_2 = 1, ..., w, (w = p + q)$ is

$$\hat{\Sigma}_{A,j_1,j_2} \cong \sum_{r=1}^{w}\sum_{s=1}^{w}\sum_{t=1}^{w}\sum_{u=1}^{w} \hat{A}_{i_1 r}\hat{A}_{sj_1}\hat{A}_{i_2 t}\hat{A}_{uj_2}Cov(\hat{\Delta}_{rs}, \hat{\Delta}_{tu}). \tag{2.10}$$

In the simple linear regression case, the post-hoc correction presented in equation (2.7) reduces to the following familiar form: $\hat{\beta} = \frac{\hat{\beta}^*}{\hat{\delta}}$, where $\hat{\beta}^*$ is the estimate for β obtained from the "naive" regression using X_i^* that ignores the error in the covariate of interest, and $\hat{\delta}$ is the estimate of the attenuation coefficient from the simple linear regression correction. Similarly, the variance estimator for this correction is easily calculated using the univariate delta method as $var(\hat{\beta}) = \frac{1}{\hat{\delta}^2} var(\hat{\beta}^*) + \frac{\hat{\beta}^{*2}}{\hat{\delta}^4} var(\hat{\delta})$.

Rosner et al. (1990) justified this proposed correction for logistic regression for small β. One can use a Taylor series approximation to show when this method can be expected to work similarly for the Cox proportional hazards model. Specifically, Green and Symons (1983) used a linear Taylor series expansion to illustrate the approximate mathematical equivalence between the logistic regression model and the Cox proportional hazards model when the event of interest is rare, the follow-up time is short, and the baseline hazard in the Cox model is constant. The post-hoc regression parameter correction developed for logistic regression is expected to do similarly well for the Cox proportional hazards model for settings that uphold these assumptions. We explore this further with a numerical study. In Appendix A.3, we establish the asymptotic properties of our estimator. Computational details and R code for implementing the proposed method are presented in Appendix A.1. The R code used to implement all simulations is available on GitHub at https://github.com/lboe23/Outcome-Error-RC.

Strata-specific baseline hazards

For a continuous failure time outcome, the proportional hazards model takes the familiar form $S(t) = S_0(t)^{\exp(x^T\beta_X + z^T\beta_Z)}$. Under this assumption, the baseline survival function $S_0(t)$ and baseline hazard function $\lambda_0(t)$ are shared by all subjects in the data. Oftentimes, however, this assumption is invalid and we expect baseline survival to differ across groups defined by one or more covariates. To address the issue of non-proportional hazards, we let the survival function for a subject from stratum k be $S_k(t) = S_{0k}(t)^{\exp(x^T\beta_X + z^T\beta_Z)}$, $k = 1, ..., K$, where $S_{0k}(t)$ is the baseline survival for all individuals in stratum k.

In a discrete proportional hazards model that incorporates stratification, we allow strata-specific versions of the baseline survival function introduced in Section 2.3.1, such that $\mathbf{S}_k = (S_{1k}, S_{2k}, ..., S_{(J+1)k})^T$. We can accordingly modify the log-likelihood function from equation (2.3) to allow for stratification on one or more predictors. As in the continuous time setting, the stratified log-likelihood for all N subjects is a simple sum of the log-likelihood for each stratum. Now, in our discrete failure time setting, the log-likelihood function for the N_k subjects in stratum k is given by:

$$l_k(\mathbf{S}_k, \beta) = \sum_{i=1}^{N_k} \log \left(\sum_{j=1}^{J+1} D_{ij} S_{jk}^{\exp(x_i^T \beta_X + z_i^T \beta_Z)} \right). \tag{2.11}$$

Correspondingly, the log-likelihood for all N subjects is calculated as follows:

$$l(\mathbf{S}_k, \beta) = \sum_{k=1}^{K} \left[\sum_{i=1}^{N_k} \log \left(\sum_{j=1}^{J+1} D_{ij} S_{jk}^{\exp(x_i^T \beta_X + z_i^T \beta_Z)} \right) \right]. \tag{2.12}$$

Using this likelihood, we can solve for the unknown parameters β_X, β_Z, S_{2k},...,$S_{(J+1)k}$, $k = 1, ..., K$ and compute the estimated covariance matrix as described in Section 2.3.1. Although the baseline survival functions are different for each stratum, the coefficients β_X and β_Z are assumed to be uniform across all strata. Note, in the setting without misclassification in the event indicator, strata should be chosen such that each stratum contains subjects with the event of interest, as a stratum with no events does not contribute any information to the analysis (Harrell Jr, 2015). However, with a sensitivity less than 1, events and non-events of a stratum both contribute to the likelihood (2.12). Under this model, we can apply the same post-hoc fix introduced in Section 2.3.2 to also correct the estimated coefficients for exposure error.

Adjusting for false negatives at baseline

The proposed method can be modified to handle the case in which individuals with a baseline false negative test are erroneously included into the analysis. This simple extension of the method applies to scenarios in which subjects are only included in the study if they report

being event-free at baseline. This extension is motivated by the analysis approach of Tinker et al., Tinker et al. (2011) which excluded anyone with a positive self-report at baseline. To allow for a non-zero probability of a baseline false negative test, we will now assume $S_1 < 1$.

Let R_i and E_i be the observed error-prone event status at baseline and the unobserved true event status at baseline, respectively. Consider all subjects in the study that have a negative error-prone outcome at baseline, i.e. $R_i = 0$, and are therefore included in the analysis population. Define η as the negative predictive value, or the probability that a subject with a negative error-prone outcome is truly disease-free, i.e. $\eta = \Pr(E_i = 0|R_i = 0)$, which we assume is constant across all N subjects. Further assume all subjects with a negative error-prone outcome who are truly disease-free constitute a random sample of all subjects who are truly disease-free at baseline, so that $\Pr(\mathbf{Y_i}, \mathbf{t_i}, n_i|E_i = 0, R_i = 0) = \Pr(\mathbf{Y_i}, \mathbf{t_i}, n_i|E_i = 0)$. Then, the likelihood function for subject i can be expressed as follows:

$$\begin{aligned} f(\mathbf{Y_i}, \mathbf{t_i}, n_i) &= \Pr(\mathbf{Y_i}, \mathbf{t_i}, n_i|R_i = 0) \\ &= \eta \Pr(\mathbf{Y_i}, \mathbf{t_i}, n_i|E_i = 0, R_i = 0) + (1 - \eta) \Pr(\mathbf{Y_i}, \mathbf{t_i}, n_i|E_i = 1, R_i = 0) \\ &= \eta \sum_{j=1}^{J+1} D_{ij} S_j^{\exp(x_i^T \beta_X + z_i^T \beta_Z)} + (1 - \eta) D_{i1} S_1^{\exp(x_i^T \beta_X + z_i^T \beta_Z)}. \end{aligned}$$

Thus, the log-likelihood for all N subjects is

$$l(\mathbf{S}, \beta) = \sum_{i=1}^{N} \log \left(D_{i1} S_1^{\exp(x_i^T \beta_X + z_i^T \beta_Z)} + \eta \sum_{j>1}^{J+1} D_{ij} S_j^{\exp(x_i^T \beta_X + z_i^T \beta_Z)} \right). \tag{2.13}$$

2.4. Numerical Study

We examine the numerical performance of our proposed estimator using a simulation study. We compare our estimator to the results from the "true" model, in which a discrete proportional hazards model is fit with the true (error-free) event time and covariate values, and the "naive" model, which fits the same model with the error-prone outcome and covariate. In all simulations, we assume a single error-prone covariate of interest. We assume that there

are two precisely measured covariates, which are moderately correlated with the error-prone variable. Our results show how our estimator performs under different levels of outcome sensitivity and specificity, error variance in the covariate, sample size, and censoring rates. We present percent biases, average standard errors (ASE), empirical standard errors (ESE), and 95% coverage probabilities (CP) across these various settings. Mean percent bias is calculated as follows: $\frac{\hat{\beta}-\beta}{\beta} \times 100$, where β is the target regression parameter of interest. The ASE is defined as the mean of the estimated standard errors from the model, while the ESE is the empirical standard deviation of the estimated coefficients across simulations. Additionally, we present type I error results for $\beta_{X1} = 0$ and $\alpha = 0.05$, where β_{X1} is the regression parameter corresponding to the error-prone covariate.

2.4.1. Simulation Setup

We present results from 1000 simulations run in R version 3.5.2 (R Core Team, 2018). The three covariates, X_1, Z_1, and Z_2 were generated from a multivariate normal distribution, all with mean 0 and a covariance matrix with all diagonal elements equal to 1 and all off-diagonal elements equal to 0.3. We generated our error-prone covariate X_1^* using the linear measurement error model, $X_1^* = \alpha_0 + \alpha_1 X_1 + \alpha_2 Z_1 + \alpha_3 Z_2 + e$, with $\alpha_0 = 1$, $\alpha_1 = 0.8$, $\alpha_2 = 0.3$, and $\alpha_3 = 0.5$. We assumed $e \sim N(0, \sigma^2)$ and considered σ^2 values of 0.59 and 1.72, which correspond to estimated $\delta_{(1)}$ values of approximately 0.60 and 0.30, respectively.

Later, we assess how our method performs when error is not normally distributed, but instead $e \sim .4\mathcal{N}(0,1) + .6\mathcal{N}(2,1.5)$ and e distributed as a t with 4 degrees of freedom (df). For all simulations, there are $N = 1000$ subjects in the main study data. We assume our calibration subset is a random sample of $n_C = 500$ subjects from the main study. The measure approximating X_1 in the calibration subset, X_1^{**}, is generated to follow the classical measurement error model from equation (2.5), where $\epsilon \sim \mathcal{N}(0, 0.06)$.

We considered typical settings for which regression calibration has been observed to perform well, including a moderate β_{X1} and a higher censoring rate (Shaw and Prentice, 2012). The true log hazard ratios were selected to be $\beta_{X1} = \log(1.5)$, $\beta_{Z1} = \log(0.7)$, and $\beta_{Z2} = \log(1.3)$.

Later, we set $\beta_{X1} = \log(3)$ to assess how the method performs under a more extreme regression coefficient corresponding to the error-prone covariate. The true time-to-event was generated from a continuous time exponential distribution. To mimic the settings of real data, we considered a follow-up schedule with four possible visit times. To obtain an average true censoring rate (CR) of approximately 0.90, we set the visit times to be $\{2, 5, 7, 8\}$ with baseline hazard rates of 0.012 and 0.008 for $\beta_{X1} = \log(1.5)$ and $\beta_{X1} = \log(3)$, respectively. Fixing the visit times at $\{1, 3, 4, 6\}$ and baseline hazard rates at 0.094 and 0.076 for $\beta_{X1} = \log(1.5)$ and $\beta_{X1} = \log(3)$, respectively, leads to an average true CR of approximately 0.55. Note that the visit times are not required to be equally spaced. Figure A.1 in the Appendix depicts the estimated nonparametric maximum likelihood estimators of the survival distribution for the true and error-prone outcomes under the two CRs for $\beta_{X1} = \log(1.5)$ for a single simulated data set.

To assess how our method performs when the baseline hazard varies across strata, we simulate four approximately equal sized strata. For test times at $\{2, 5, 7, 8\}$, we let the four baseline hazard rates be 0.008, 0.010, 0.011, and 0.019, which resulted in an overall censoring rate of approximately 90%. Similarly, to obtain an overall censoring rate of approximately 55%, the baseline hazard rates for each stratum were fixed at 0.090, 0.080, 0.075, and 0.131 for visit times at $\{1, 3, 4, 6\}$.

To capture the interval in which each simulated event occurred, we created an indicator for whether or not the current visit time was greater than the actual event time itself. This indicator variable was "corrupted" using sensitivity and specificity values in order to create the error-prone vector of outcomes, $\mathbf{Y_i}$. To mimic a diagnostic test with different levels of accuracy, we considered the case where sensitivity = 0.90 while specificity = 0.80, and sensitivity = 0.80 while specificity = 0.90. Later, we assess the performance of the proposed method when a baseline negative predictive value (η) less than 1 is incorporated into the analyses to adjust for erroneously included false negative participants. We vary η between 0.98 and 0.90. To simulate this scenario, we set the true time-to-event equal to 0 for a fixed

proportion of subjects, η, included in the data. This represents an event time prior to the start of the study. Additionally, we show that the proposed method can handle different visit structures by allowing each visit to be subject to a constant, independent probability of missingness, which mimics the Missing Completely at Random (MCAR) setting for missing data. To simulate this, we create a binary variable indicating whether the jth visit is missing for each subject using a fixed probability P_{Miss} of either 0.10 or 0.40. We further assess the method under parameters that mimic the structure of the WHI data example, with $N = 65,000$, $n_C = 500$, $Se = 0.61$ and $Sp = 0.995$, $\eta = 0.96$, and a censoring rate of 95% for the error-prone discrete failure time. To simulate self-reported outcomes in the WHI data, we stopped visit times for each subject after the first positive error-prone outcome. Regression coefficients for the discrete time Cox proportional hazards model and the true data can be estimated by fitting a generalized linear model assuming the binomial outcome and complementary log-log link (Hashimoto et al., 2011).

2.4.2. Simulation Results

Tables 2.1-2.4 present estimates of mean percent bias, ASE, ESE, and 95% CP across the various settings described above. For Table 2.1, we consider the case where $\beta_{X1} = \log(1.5)$. Overall, we see that the proposed method improves over naive analyses in bias and in the nominal coverage of 95% confidence intervals. In fact, under various different settings, the percent bias of our parameters of interest never exceeds 5%. Additionally, we maintain nominal coverage for a 95% confidence interval. Furthermore, our ASEs closely resemble the ESEs, demonstrating that our standard error estimates also performed well. In contrast, for the analyses that ignore measurement error, estimates of β_{X1}, β_{Z1}, and β_{Z2} have bias as high as -96.33% and attain very little coverage. Table A.1 in the Appendix further shows results for the method that corrects for covariate error only and the method that corrects for outcome error only under these same simulation settings. Regression parameters for the method correcting for only covariate error have absolute mean percent biases ranging from 47.05 to 82.31, while the method correcting solely outcome error has bias ranging from 5.840 to 70.28. Unsurprisingly, the proposed method greatly improves over all three alternative

approaches that ignore measurement error to some degree.

In Table 2.2, we set $\beta_{X1} = \log(3)$. The method still performs reasonably well when the censoring rate is high (CR = 0.90), as absolute percent bias stays below 12% and nominal coverage is maintained. However, when the censoring rate decreases to 0.55, we begin to see an increase in bias and a steep decrease in coverage, particularly for β_{X1}. This is unsurprising, as regression calibration is known to break down with a larger β coefficient and a higher event rate (Shaw and Prentice, 2012). We observe that even in the most challenging scenarios for the proposed method, i.e. a more extreme β_{X1}, less censoring, and more covariate measurement error, the percent attenuation bias (coverage) was 17% (77%) compared to 91%(0%) for the naive analysis.

In Table 2.3, we examine the relative performance of our proposed method when the error in X^* no longer follows a normal distribution. Here, we let the error in X^* follow either a t distribution with 4 df, or a mixture of two normals, as described in the simulation setup. On average, we observe $\delta_{(1)} = 0.27$ when the error in X^* follows the t distribution and $\delta_{(1)} = 0.21$ when the error follows the mixture distribution, which reflects substantial error in our simulated covariate of interest in all scenarios. Since the applied regression calibration method assumes a first order approximation to estimate $E(X|X^*, Z)$, we expect the proposed method to perform best when the error in X^* is normally distributed. Thus, it is unsurprising that the mean percent bias for the proposed method is a bit higher for β_{X1} under these settings, particularly when the error follows a t distribution. Nonetheless, absolute percent bias stays under 4% in all scenarios. Most intervals still come very close to achieving the nominal level of 95% CP. Our proposed approach still outperforms the naive method, which again shows severe bias of up to -97.65% and poor coverage.

Table 2.4 shows the performance of the proposed method alongside the naive method in terms of mean percent bias, ASE, ESE, and 95% CP when both approaches allow for stratification. In this table, we revert to letting the error in X follow a normal distribution and set $\beta_{X1} = \log(1.5)$. We assume that there are four equally-sized strata. Similarly to what we observed

in Table 2.1, we see that the method performs well in terms of bias and coverage. Absolute bias for β_{X1}, corresponding to the error-prone covariate, ranges from 0.310% to 1.893% and is therefore quite low in all scenarios. The standard error estimator works well, as indicated by the attainment of nominal coverage. Again, we see extremely high bias for the naive approach, ranging from -68.31% to -96.68% for β_{X1}.

Type I error results for the coefficient corresponding to the error-prone covariate are presented in Table 2.5. Type I error values ranged from 0.039 to 0.058 across different values of Se, Sp, $\delta_{(1)}$, and CR. With 1000 simulations, a 95% confidence interval based on the true error rate $\alpha = 0.05$ is $(0.036, 0.064)$. All calculated error rates in Table 2.5 are within simulation error of the truth, indicating that type I error is preserved in the proposed method for all settings.

Table A.2 of the Appendix demonstrates the performance of the proposed method, now including adjustment for an imperfect baseline negative predictive value. Under different levels of covariate error and changes to the sensitivity and specificity, the bias of our parameters remains under 6% and nominal coverage for a 95% confidence interval is maintained, illustrating that the method performs well. We observe that the performance of the proposed method surpasses that of the naive method, which shows excessive bias in the parameters of interest, ranging from -79.01% to -97.33%.

In Table A.3 of the Appendix, we show that the proposed method can accommodate missed visits. Our approach performs well in all scenarios, maintaining an absolute mean percent bias of under 4.353% when we let each visit to be subject to either 10% or 40% missingness. When there are missed visits, the proposed method outperforms the naive method, which shows extreme mean percent bias of up to -96.85%.

Finally, we present results for the simulations that mimic the structure of the WHI data in Table A.4 of the Appendix. We see that the proposed method works well under measurement error settings similar to that of the WHI, maintaining an absolute percent bias of under 0.8%

for all scenarios. Again, the proposed method outperforms the naive method, in which we see absolute percent bias as high as 89.53% for the regression parameter of interest and 0% coverage probability for many scenarios. Similarly, the methods that correct for covariate error only and outcome error only both show extreme bias and inadequate coverage under these settings.

2.5. Women's Health Initiative (WHI) Example

2.5.1. WHI Study

The Women's Health Initiative is a collection of studies launched in 1993 that together investigated the major causes of morbidity and mortality in US post-menopausal women (The Women's Health Initiative Study Group, 1998). We seek to examine the association between energy, protein and protein density (percentage of energy from protein) intakes with the risk of diabetes when all three exposures as well as diabetes status are self-reported and subject to error (Neuhouser et al., 2008; Gu et al., 2015). We analyze data on post-menopausal women aged 50-79 who participated in either the comparison arm of the Dietary Modification trial (DM-C) or the Observational Study (OS) and who had an average follow-up of approximately 9 years (Ritenbaugh et al., 2003; Langer et al., 2003). Neither women from the DM-C nor the OS received study interventions. The WHI also included the nutritional biomarker study which collected objective recovery biomarkers for energy and protein intake, thought to have only classical measurement error, on a subset of participants ($n_C = 544$). These biomarkers were previously used to develop calibration equations for the self-reported intakes of energy, protein and protein density (Neuhouser et al., 2008). Using these calibration equations, Tinker et al. (2011) reported incident diabetes hazard ratios in this cohort for energy, protein, and protein density that were corrected for the error in self-reported dietary exposures. Self-reported diabetes in the WHI has been reported to be subject to error (Margolis et al., 2008). We apply our proposed method to correct for error in both the exposure and the diabetes failure time outcome. Our goal was to answer a similar research question as Tinker et al., Tinker et al. (2011) only to use our method

that additionally adjusts for error in the diabetes outcome. We adopted the same exclusion criteria as Tinker et al. (2011) in order to arrive at our final analytic data set of 65,358 participants. In short, these criteria attempt to align the characteristics of DM-C and OS cohorts and exclude those with missing data or who reported diabetes at baseline. Baseline was defined as the time of the first self-reported dietary assessment post-enrollment, year 1 for the DM-C and year 3 for the OS. Further details are provided in Appendix A.4.

We started with the previously developed calibration equations for dietary energy, protein, and protein density from Neuhouser et al., Neuhouser et al. (2008) which we call our "base" calibrations. Body mass index (BMI), age, race-ethnicity, income, and physical activity were included in the energy calibration model; BMI, age, race-ethnicity, income, and education for protein; and BMI, age, and smoking status for protein density. To avoid bias, regression calibration requires the calibration model to include the same covariates as the outcome model (Rosner et al., 1990; Kipnis et al., 2009). We only considered the form of regression calibration in which the variables in the calibration and outcome models are exactly aligned. Thus, we extended each base calibration to include all predictors from our outcome model. Specifically, education, hypertension, and alcohol use were added to all calibrations. For each of the three nutrients, the calibration equation was fit by regressing the biomarker value (X^{**}) on the corresponding self-reported value and participant characteristics, as described above.

In the WHI, prevalent diabetes was recorded via a self-reported questionnaire at baseline. We consider data from 8 years of annual follow-up visits in our analyses. Only the censored event-time was recorded in continuous time in our analytical dataset. Thus, we discretized the available data by dividing the follow-up time into 9 possible intervals. Then, for all 65,358 women in our analytic cohort, we considered the time at which the first occurrence of self-reported diabetes or censoring time was recorded and assumed that the occurrence of the censored self-reported outcome happened in the annual interval that the event time fell into. We note that in other settings our method could accommodate an increase in the

number of time intervals if follow-up occurred more frequently than once a year (e.g. a bi-annual visit structure).

Self-reported diabetes in the WHI was previously reported to have a sensitivity of 0.61, specificity of 0.995, and a baseline negative predictive value of 0.96 (Gu et al., 2015). We incorporated these values into our analyses. We also considered a sensitivity analysis in which we examined the results for a negative predictive value of 1 and explored cohort-specific values of sensitivity and specificity. All diabetes risk models were adjusted by standard risk factors, also included in the calibration equations. Additionally, we stratified our discrete proportional hazards models on age in 10-year categories and DM-C or OS membership to better approximate previous analyses. Because BMI may be only a mediator for energy intake or may possibly also be an independent risk factor, it is not clear whether adjusting for BMI in our diabetes risk model is appropriate due to the challenge of overcontrolled or undercontrolled models, as discussed in Tinker et al. Tinker et al. (2011) Thus, we ran each outcome model with and without BMI.

To fit the naive model, we used the binomial generalized linear model with the complementary log-log link. To fit the model corrected for covariate error only, we used this same approach, then adopted the post-hoc matrix correction and corresponding variance adjustment described in the body of this paper. We applied our proposed approach to correct for error in both the self-reported diabetes outcome and dietary exposures. In all models, we used log values of dietary energy, protein, and protein density. We present hazard ratios (HR) and 95% confidence intervals (CI) associated with a 20% increase in consumption.

2.5.2. Results

Incident diabetes was reported in 3053 (4.7%) of the 65,358 participants of analytic cohort. Table 2.6 shows the results for the three different analysis approaches. In the BMI-adjusted analysis, the HR (95% CI) for a 20% increase in energy intake was 0.822 (0.512, 1.318) for the proposed approach compared to 1.041 (0.758, 1.492) for the covariate-error adjusted method and 1.002 (0.986, 1.018) for the naive approach. Note, however, that the incident diabetes

is not significantly associated with increasing energy in any of these three models. Without BMI in the outcome model, the proposed method estimated a HR of 1.189 (0.836, 1.692) for a 20% increase in energy intake, compared to 1.421 (1.043, 1.938) for the covariate-error adjusted method and 1.024 (1.008, 1.040) for the naive method. In this case, adjusting for error in the self-reported outcome led to qualitatively different results in that the HR was about 20% smaller and no longer significant.

When we apply the proposed method, a 20% increase in protein intake is associated with a 1.077 (0.978, 1.186) HR, compared to a HR of 1.121 (1.036, 1.213) for the covariate-error adjusted method and 1.024 (1.010, 1.039) for the naive approach. When we do not adjust for BMI, all three approaches result in HRs that are significantly associated with an increase in protein consumption. For protein density, whether or not we adjust for BMI, all three approaches show that a 20% increase in intake is positively associated with risk of diabetes. When we adjust for BMI, the HR estimated by the proposed method, 1.266 (1.115, 1.436), is fairly similar to the HR estimated by the method that adjusts for covariate error only, 1.243 (1.125, 1.374), and somewhat higher than the HR estimated by the naive method, 1.100 (1.064, 1.137). We note some of our HRs differ from the results reported by Tinker et al. Tinker et al. (2011) We believe this is due to a few discrepancies in the analytical dataset and model and is discussed further in Appendix A.4.

In Table A.5 in the Appendix, we present a WHI data analysis results table that ignores the issue of an imperfect baseline self-report and assumes the negative predictive value is 1. For energy and protein density, assuming baseline self-reports are perfect does not qualitatively change our results. However, for protein, the HR (95% confidence interval) estimated by the proposed method is 1.077 (0.978, 1.186) when the negative predictive value is set to 0.96, but changes to 1.107 (1.025, 1.195) when the negative predictive value is set to 1. Here, we see that because our estimate is so close to a boundary, incorporating the uncertainty at baseline into our analyses does slightly change our results.

Since we analyzed data on participants from two different cohorts, the WHI DM-C trial and

the WHI OS, we investigated how cohort-specific sensitivity and specificity might impact our HR estimates. We used a weighted-average approach to select sensitivity and specificity values for the DM-C and OS trials such that the overall values worked out to be 0.61 and 0.995, respectively. One might hypothesize that the clinical trial (WHI DM-C) recorded data with higher accuracy than the larger observational study (OS), though in our analysis we also consider the possibility that sensitivity and specificity are higher for the observational study. Table A.6 in the Appendix presents the results of this analysis. We observe that implementing slightly variable cohort-specific sensitivity and specificity values was not enough to qualitatively impact our conclusions regarding the significance of the association between an increase in intake of dietary energy, protein, or protein density with the risk of diabetes.

2.6. Discussion

In settings such as large epidemiological studies, where outcomes or complex exposures are often collected by self-report, both the exposure and outcome of interest can be subject to measurement error. This was observed in our data example from the WHI, but has also been observed in other cohorts where data were reliant on routinely collected electronic health records data (Shepherd and Yu, 2011; Oh et al., 2019). This paper presents a method to accommodate errors in continuous covariates and a discrete failure time outcome variable when sensitivity and specificity of the error-prone outcome are known; when error rates are unknown, our method can be used as a sensitivity analysis using hypothesized values. The proposed method can be applied when, for a subset, there is either a gold standard measure of the exposure or a second measure with independent, unbiased (classical) measurement error available. For the WHI, the calibration subset containing the variable with classical measurement error was sampled after baseline with the assumption that the measurement error model did not change over time.

We studied the relative performance of the proposed method under various settings of sensitivity, specificity, error variance of the exposure, and censoring rate, including those where

ignoring the measurement error led to extreme bias in the regression parameters of interest. In all settings studied, our method led to nearly unbiased estimates of the regression parameters, maintaining bias of less than approximately 19% for non-zero regression parameters and generally much less bias when the underlying log-hazard parameter β was of moderate size (e.g., log(1.5)). Furthermore, our variance estimator performed favorably, as evidenced by the coverage probability and ASEs that closely resembled ESEs. Our variance estimator assumes approximate independence of $\hat{\beta}^*$ and $\hat{\delta}$. While we have not verified independence of these components for all settings, even in our settings where the calibration subset was 50% of the cohort, we observed no appreciable correlation between these estimates (data not shown). If there is concern that this approximate independence does not hold, one could instead consider a bootstrap approach for variance estimation. For our simulations where $\beta_{X1} = 0$, we observed that type I error rates were preserved. Our adjustment for covariate error relied on a regression calibration type adjustment. As expected from previous literature, this method performs best when the regression parameter corresponding to the error-prone covariate is of modest size, the error in the covariate is normally distributed, and the censoring rate is high (i.e. the event of interest is rare). Our method in particular shows more appreciable bias when the regression parameter is of large size, e.g. $\beta_{X1} = \log(3)$, especially for a lower censoring rate. This method proved to be fairly robust to changes in the distribution of the error in X studied; for more extreme deviations from normality, this may no longer be true. Our method also performs favorably after stratifying on one or more covariates. Lastly, the proposed method works well under simulation parameters that mimic the structure of the WHI data. In all scenarios explored, the proposed method substantially outperformed the naive method, which repeatedly showed severe bias and minimal coverage. For settings different from those studied, one might consider conducting additional numerical studies.

The method introduced in this paper is applied to data from 65,358 post-menopausal women enrolled in the WHI to assess the association between energy, protein, and protein density intake and the risk of incident diabetes, adjusting for error in self-reported exposures and

outcome. Hazard ratios obtained for all exposures were considerably different than those from the naive analyses ignoring the error in both diabetes status and dietary intake and those that only adjusted for error in dietary intake. In some cases, our proposed method led to qualitatively different conclusions in that the parameter of interest was no longer statistically significant. For the case of non-differential outcome error, this stems largely from the increased uncertainty in the results coming from the uncertain outcomes. These conclusions demonstrate the importance of adjusting for errors in both outcomes and covariates.

Our proposed method offers a practical approach to estimating the association between a covariate and a discrete time-to-event outcome, when both are recorded with error. A limitation of our approach stems from the curse of dimensionality that can accompany discrete data in settings where the visit times are irregular, which can cause the number of parameters to grow with the number of subjects in the data. It is impractical to assume that in a real data setting, all subjects' visit times fall on the same schedule in the study (e.g. exactly annually). Thus, we must make a compromise depending on how many parameters the data can stably support. Ultimately, the data should help inform a reasonable decision regarding the number of intervals to consider for analyses of this type. Sensitivity analyses can be also be conducted to examine whether the number or choice of discrete time intervals affected study estimates. In many cohort studies with long-term follow-up like the WHI, there is a specified visit schedule in the study protocol. If all subjects adhere to this schedule with little variation, this naturally leads to the discrete-time framework with a common set of possible visit times across all individuals. Frequently in these studies, including the WHI, the observed visit schedule varies across subjects. To apply the proposed method in our WHI example, we made some simplifying assumptions. Since our analytical data set included only the amount of time that elapsed between enrollment and the first occurrence of self-reported diabetes or censoring time recorded on a continuous timescale, we rounded the censored event-time to the nearest annual visit date and assumed the outcome or censoring event occurred sometime between that visit and the prior annual visit. If data are available on the timing of all visits, the likelihood could be adapted to allow for longer intervals

between visits for some individuals (i.e. missed visits).

We note that for the case of self-reported data, we assume that each subject is followed up until the first positive, as it is not expected that a new diagnosis would be subsequently recorded. This assumption corresponds to the applied setting in which self-reported disease incidence stops after the first positive report. However, the model by Gu et al. (2015) and thus the proposed approach do allow for a more flexible framework and can accommodate repeated testing. As an example, this approach can be applied to a data set containing repeat blood test results, such as those used in monitoring for cancer relapse.

A potential limitation of our work is the reliance of the proposed method on the assumption that given the true disease status at each visit, the error-prone outcomes are independent. In the WHI data, we assume the self-reported outcomes are far enough apart that there are a number of random processes affecting a subject's knowledge and interpretation of the outcomes questionnaire that make this independence assumption reasonable; however, this assumption may not always be realistic, particularly for the case of self-reported data. We note that our method is applicable more generally to settings where the error-prone outcome of interest is not self-reported, but derived say from an objective biomarker for which this assumption may be more reasonable. In future work, one might consider a similar framework to the one proposed which relaxes this assumption by positing a more complex error model for the outcome of interest, such as one with sensitivity and specificity potentially dependent on covariates or previous responses.

The increasing reliance of clinical research on self-administered questionnaires or administrative databases in epidemiological studies has led to more attention being given to methods to correct for measurement error. Gu et al. (2015) conducted a sensitivity analysis to show how changes in sensitivity, specificity, and negative predictive shifted the estimated hazard ratio of statin use on the risk of incident diabetes in data from the WHI. The results showed that the estimated hazard ratio is highly sensitive to changes in specificity and modestly sensitive to changes in sensitivity and negative predictive value. This analysis helps illus-

trate the importance of having accurate values of sensitivity and specificity in the proposed method. Our sensitivity analysis showed that while varying sensitivity and specificity by cohort did not qualitatively change the results in our particular example, the hazard ratio estimates are much more vulnerable to changes in specificity when the event of interest is as rare as it is in the WHI data (diabetes incidence = 4.7%). Thus, we emphasize the importance of employing correct values of sensitivity and specificity, especially when they might vary by some demographic factor or group membership.

This paper explored the incorporation of the negative predictive value into the analyses to handle misclassification at baseline. Evidence suggests that some women in the WHI who provided a negative self-report of diabetes at baseline were actually diabetic. A question of interest is whether mistakenly excluding women who were false positives can induce bias. It has been previously reported that when all potential confounders are adjusted for in the outcome model and the missing at random (MAR) assumption is satisfied, missing data should not cause bias (Groenwold et al., 2011). Furthermore, given that positive predictive value is assumed to be quite high in the motivating data example, we did not explore the issue further in this paper. This exclusion criteria-related matter may be more relevant in other cohorts, particularly if the reason for exclusion is related to some unobserved characteristic.

A worthwhile extension of this work might consider incorporating covariate-specific or even subject-specific sensitivity and specificity, particularly when these values are no longer assumed to be known constants and need to be estimated along with the outcome model parameters. Such an extension would require a validation or calibration subset to also contain information on the measurement error structure of the self-reported outcome. When the outcome is rare, such a cohort can be difficult to construct prospectively as validation subsets are generally of fairly modest size due to cost. Efficient choices of a validation sampling design and development of analysis methods that provide consistent estimates of the target parameter are two important areas of future research.

Table 2.1: The mean percent (%) biases, average standard errors (ASE), empirical standard errors (ESE) and coverage probabilities (CP) are given for 1000 simulated data sets for the proposed method and naive method with $\beta_{X1} = \log(1.5)$, $\beta_{Z1} = \log(0.7)$, and $\beta_{Z2} = \log(1.3)$; e is normally distributed with mean zero.

Se[1]$= 0.80, Sp$[2]$= 0.90$			Proposed				Naive			
$\hat{\delta}_{(1)}$[3]	CR[4]	β	% Bias	ASE	ESE	CP	% Bias	ASE	ESE	CP
0.60	0.90	β_{X1}	1.616	0.200	0.204	0.950	−88.03	0.046	0.046	0.000
		β_{Z1}	−1.094	0.143	0.142	0.945	−79.22	0.057	0.058	0.002
		β_{Z2}	−3.731	0.143	0.143	0.945	−84.07	0.057	0.054	0.021
	0.55	β_{X1}	−1.231	0.093	0.094	0.949	−68.11	0.038	0.038	0.000
		β_{Z1}	−1.055	0.067	0.066	0.958	−43.46	0.047	0.046	0.079
		β_{Z2}	−3.018	0.066	0.065	0.957	−53.48	0.046	0.045	0.133
0.30	0.90	β_{X1}	1.840	0.283	0.286	0.954	−93.88	0.033	0.033	0.000
		β_{Z1}	−1.233	0.151	0.151	0.947	−82.46	0.054	0.055	0.001
		β_{Z2}	−4.212	0.151	0.150	0.945	−79.74	0.054	0.052	0.025
	0.55	β_{X1}	−2.246	0.131	0.133	0.940	−84.02	0.027	0.027	0.000
		β_{Z1}	−1.967	0.071	0.069	0.951	−52.48	0.045	0.044	0.008
		β_{Z2}	−3.899	0.070	0.068	0.956	−42.08	0.045	0.044	0.306
$Se = 0.90, Sp = 0.80$			Proposed				Naive			
$\hat{\delta}_{(1)}$	CR	β	% Bias	ASE	ESE	CP	% Bias	ASE	ESE	CP
0.60	0.90	β_{X1}	0.391	0.210	0.209	0.957	−93.08	0.037	0.037	0.000
		β_{Z1}	−3.692	0.150	0.153	0.942	−91.96	0.046	0.045	0.001
	0.55	β_{X1}	−1.246	0.094	0.093	0.960	−77.95	0.034	0.035	0.000
		β_{Z1}	−1.188	0.068	0.067	0.951	−61.05	0.042	0.042	0.001
		β_{Z2}	−3.502	0.067	0.066	0.953	−68.46	0.042	0.042	0.014
0.30	0.90	β_{X1}	0.665	0.296	0.291	0.967	−96.33	0.026	0.026	0.000
		β_{Z1}	−0.963	0.158	0.160	0.951	−90.21	0.044	0.044	0.000
		β_{Z2}	−4.214	0.158	0.160	0.947	−89.56	0.044	0.043	0.001
	0.55	β_{X1}	−2.034	0.133	0.130	0.964	−88.87	0.024	0.024	0.000
		β_{Z1}	−1.994	0.072	0.070	0.950	−67.22	0.040	0.040	0.000
		β_{Z2}	−4.420	0.071	0.069	0.959	−60.63	0.040	0.040	0.029
$Se = 1, Sp = 1$			Truth							
	CR	β	% Bias	ASE	ESE	CP				
	0.90	β_{X1}	0.163	0.108	0.109	0.944				
		β_{Z1}	0.205	0.107	0.107	0.953				
		β_{Z2}	−0.586	0.107	0.109	0.949				
	0.55	β_{X1}	0.639	0.052	0.052	0.948				
		β_{Z1}	0.345	0.052	0.051	0.949				
		β_{Z2}	−0.383	0.052	0.052	0.952				

[1] Se = Sensitivity [2] Sp = Specificity [3] $\hat{\delta}_{(1)}$ = Estimate of attenuation coefficient
[4] CR = True censoring rate

Table 2.2: The mean percent (%) biases, average standard errors (ASE), empirical standard errors (ESE) and coverage probabilities (CP) are given for 1000 simulated data sets for the proposed method and naive method with $\beta_{X1} = \log(3)$, $\beta_{Z1} = \log(0.7)$, and $\beta_{Z2} = \log(1.3)$; e is normally distributed with mean zero.

Se[1]$= 0.80, Sp$[2]$= 0.90$			Proposed				Naive			
$\hat{\delta}_{(1)}$[3]	CR[4]	β	% Bias	ASE	ESE	CP	% Bias	ASE	ESE	CP
0.60	0.90	β_{X1}	−3.442	0.211	0.213	0.946	−88.61	0.047	0.048	0.000
		β_{Z1}	−6.773	0.146	0.145	0.941	−78.03	0.057	0.059	0.002
		β_{Z2}	−9.280	0.145	0.142	0.948	−88.07	0.057	0.054	0.012
	0.55	β_{X1}	−12.71	0.111	0.101	0.752	−72.45	0.040	0.038	0.000
		β_{Z1}	−12.57	0.075	0.068	0.916	−45.77	0.047	0.047	0.078
		β_{Z2}	−14.42	0.075	0.066	0.952	−67.90	0.047	0.045	0.032
0.30	0.90	β_{X1}	−4.532	0.296	0.295	0.951	−94.26	0.033	0.033	0.000
		β_{Z1}	−8.063	0.156	0.152	0.944	−86.52	0.054	0.056	0.000
		β_{Z2}	−11.64	0.155	0.149	0.951	−77.03	0.054	0.052	0.025
	0.55	β_{X1}	−16.88	0.154	0.137	0.766	−86.56	0.028	0.027	0.000
		β_{Z1}	−16.75	0.080	0.071	0.899	−67.26	0.045	0.045	0.000
		β_{Z2}	−18.69	0.080	0.069	0.956	−42.46	0.045	0.044	0.299
$Se = 0.90, Sp = 0.80$			Proposed				Naive			
$\hat{\delta}_{(1)}$	CR	β	% Bias	ASE	ESE	CP	% Bias	ASE	ESE	CP
0.60	0.90	β_{X1}	−3.581	0.220	0.221	0.945	−93.60	0.038	0.039	0.000
		β_{Z1}	−6.164	0.152	0.153	0.936	−87.76	0.046	0.047	0.000
		β_{Z2}	−8.663	0.151	0.150	0.954	−94.13	0.046	0.045	0.000
	0.55	β_{X1}	−12.65	0.112	0.103	0.764	−80.88	0.035	0.035	0.000
		β_{Z1}	−12.64	0.076	0.071	0.915	−62.26	0.042	0.043	0.001
		β_{Z2}	−14.65	0.076	0.068	0.939	−78.22	0.042	0.042	0.001
0.30	0.90	β_{X1}	−4.585	0.309	0.298	0.963	−96.73	0.027	0.027	0.000
		β_{Z1}	−7.171	0.163	0.159	0.947	−92.46	0.044	0.045	0.000
		β_{Z2}	−11.06	0.162	0.157	0.954	−88.01	0.044	0.043	0.000
	0.55	β_{X1}	−16.67	0.156	0.138	0.772	−90.62	0.025	0.024	0.000
		β_{Z1}	−16.65	0.082	0.073	0.901	−77.08	0.040	0.041	0.000
		β_{Z2}	−18.86	0.081	0.070	0.943	−60.56	0.040	0.040	0.025
$Se = 1, Sp = 1$			Truth							
	CR	β	% Bias	ASE	ESE	CP				
	0.90	β_{X1}	0.565	0.115	0.116	0.951				
		β_{Z1}	−0.222	0.108	0.108	0.949				
		β_{Z2}	−0.347	0.108	0.110	0.948				
	0.55	β_{X1}	0.605	0.063	0.064	0.944				
		β_{Z1}	0.264	0.054	0.054	0.952				
		β_{Z2}	−0.162	0.054	0.052	0.955				

[1] Se = Sensitivity [2] Sp = Specificity [3] $\hat{\delta}_{(1)}$ = Estimate of attenuation coefficient
[4] CR = True censoring rate

Table 2.3: The mean percent (%) biases, average standard errors (ASE), empirical standard errors (ESE) and coverage probabilities (CP) are given for 1000 simulated data sets for the proposed method and naive method with $\beta_{X1} = \log(1.5)$, $\beta_{Z1} = \log(0.7)$, and $\beta_{Z2} = \log(1.3)$; e is distributed as either a t with 4 df or as $.4\mathcal{N}(0,1) + .6\mathcal{N}(2,1.5)$.

Se[1]$= 0.80, Sp$[2]$= 0.90$			Proposed				Naive			
e [3]	CR[4]	β	% Bias	ASE	ESE	CP	% Bias	ASE	ESE	CP
t[5]	0.90	β_{X1}	−2.238	0.291	0.300	0.953	−94.50	0.031	0.031	0.000
		β_{Z1}	0.622	0.152	0.157	0.951	−81.80	0.054	0.054	0.000
		β_{Z2}	2.529	0.152	0.148	0.957	−76.70	0.054	0.053	0.033
	0.55	β_{X1}	−0.646	0.140	0.153	0.940	−85.54	0.025	0.026	0.000
		β_{Z1}	−2.362	0.072	0.072	0.950	−53.37	0.045	0.044	0.013
		β_{Z2}	−3.303	0.071	0.072	0.950	−39.35	0.044	0.044	0.335
mix[6]	0.90	β_{X1}	0.394	0.335	0.330	0.955	−95.64	0.027	0.028	0.000
		β_{Z1}	−1.015	0.158	0.158	0.953	−83.29	0.054	0.055	0.000
		β_{Z2}	−0.800	0.156	0.152	0.962	−75.25	0.054	0.056	0.055
	0.55	β_{X1}	−1.081	0.156	0.151	0.958	−88.74	0.022	0.022	0.000
		β_{Z1}	−2.415	0.074	0.070	0.958	−55.28	0.044	0.045	0.010
		β_{Z2}	−2.083	0.073	0.070	0.964	−36.40	0.044	0.045	0.419
$Se = 0.90, Sp = 0.80$			Proposed				Naive			
$\hat{\delta}_{(1)}$	CR	β	% Bias	ASE	ESE	CP	% Bias	ASE	ESE	CP
t	0.90	β_{X1}	−3.792	0.305	0.316	0.942	−96.91	0.025	0.025	0.000
		β_{Z1}	1.848	0.160	0.165	0.948	−89.40	0.044	0.044	0.000
		β_{Z2}	3.386	0.159	0.158	0.959	−86.97	0.044	0.044	0.000
	0.55	β_{X1}	−1.119	0.141	0.159	0.933	−90.02	0.023	0.024	0.000
		β_{Z1}	−1.666	0.073	0.073	0.940	−67.86	0.040	0.040	0.000
		β_{Z2}	−3.048	0.072	0.074	0.944	−58.35	0.040	0.040	0.031
mix	0.90	β_{X1}	−0.975	0.350	0.346	0.952	−97.65	0.022	0.024	0.000
		β_{Z1}	−1.354	0.166	0.162	0.961	−90.94	0.043	0.044	0.000
		β_{Z2}	0.585	0.164	0.160	0.955	−86.57	0.043	0.045	0.000
	0.55	β_{X1}	−1.904	0.159	0.155	0.955	−92.26	0.020	0.021	0.000
		β_{Z1}	−2.590	0.075	0.072	0.954	−69.29	0.040	0.040	0.000
		β_{Z2}	−1.350	0.074	0.069	0.967	−56.67	0.040	0.039	0.036
$Se = 1, Sp = 1$			Truth							
	CR	β	% Bias	ASE	ESE	CP				
	0.90	β_{X1}	0.002	0.108	0.106	0.959				
		β_{Z1}	0.034	0.108	0.109	0.951				
		β_{Z2}	1.032	0.107	0.106	0.961				
	0.55	β_{X1}	0.395	0.053	0.052	0.952				
		β_{Z1}	−0.462	0.052	0.052	0.948				
		β_{Z2}	−0.300	0.052	0.050	0.954				

[1] Se = Sensitivity [2] Sp = Specificity
[3] e refers to the distribution of the error [4] CR = True censoring rate [5] t with 4 df
[6] Mixture of two normals, i.e. $.4\mathcal{N}(0,1) + .6\mathcal{N}(2,1.5)$

Table 2.4: The mean percent (%) biases, average standard errors (ASE), empirical standard errors (ESE) and coverage probabilities (CP) are given for 1000 simulated data sets for the proposed method and naive method, when both allow for strata-specific baseline hazards. We assume four equally-sized strata. Let $\beta_{X1} = \log(1.5)$, $\beta_{Z1} = \log(0.7)$, and $\beta_{Z2} = \log(1.3)$; e is normally distributed with mean zero.

Se[1]$= 0.80, Sp$[2]$= 0.90$			Proposed				Naive			
$\hat{\delta}_{(1)}$[3]	CR[4]	β	% Bias	ASE	ESE	CP	% Bias	ASE	ESE	CP
0.60	0.90	β_{X1}	1.893	0.202	0.199	0.961	−88.44	0.047	0.046	0.000
		β_{Z1}	3.249	0.145	0.148	0.954	−78.36	0.057	0.058	0.001
		β_{Z2}	−0.263	0.144	0.151	0.946	−81.41	0.057	0.058	0.044
	0.55	β_{X1}	−0.489	0.094	0.089	0.965	−68.31	0.038	0.038	0.000
		β_{Z1}	0.001	0.068	0.066	0.960	−42.99	0.047	0.047	0.095
		β_{Z2}	−0.885	0.067	0.066	0.958	−52.09	0.047	0.048	0.172
0.30	0.90	β_{X1}	1.036	0.286	0.280	0.959	−94.20	0.033	0.033	0.000
		β_{Z1}	2.777	0.153	0.154	0.956	−81.55	0.055	0.056	0.000
		β_{Z2}	−0.353	0.152	0.159	0.944	−77.15	0.055	0.056	0.046
	0.55	β_{X1}	−1.095	0.133	0.126	0.962	−84.08	0.027	0.027	0.000
		β_{Z1}	−0.866	0.071	0.070	0.964	−51.93	0.045	0.044	0.015
		β_{Z2}	−1.897	0.071	0.069	0.960	−40.78	0.045	0.047	0.337
$Se = 0.90, Sp = 0.80$			Proposed				Naive			
$\hat{\delta}_{(1)}$	CR	β	% Bias	ASE	ESE	CP	% Bias	ASE	ESE	CP
0.60	0.90	β_{X1}	0.986	0.214	0.217	0.949	−93.42	0.037	0.038	0.000
		β_{Z1}	3.516	0.153	0.159	0.948	−87.94	0.046	0.048	0.000
		β_{Z2}	0.025	0.151	0.162	0.945	−89.97	0.046	0.047	0.002
	0.55	β_{X1}	−0.488	0.096	0.092	0.958	−78.24	0.034	0.034	0.000
		β_{Z1}	−0.100	0.069	0.068	0.961	−60.97	0.042	0.043	0.000
		β_{Z2}	−0.953	0.068	0.067	0.957	−67.67	0.042	0.042	0.020
0.30	0.90	β_{X1}	−0.310	0.301	0.303	0.951	−96.68	0.027	0.027	0.000
		β_{Z1}	2.982	0.161	0.167	0.952	−89.72	0.044	0.046	0.000
		β_{Z2}	0.278	0.160	0.170	0.941	−87.53	0.044	0.045	0.002
	0.55	β_{X1}	−1.167	0.135	0.130	0.955	−89.05	0.024	0.024	0.000
		β_{Z1}	−0.943	0.073	0.072	0.962	−67.08	0.040	0.041	0.000
		β_{Z2}	−1.921	0.072	0.071	0.958	−59.89	0.040	0.041	0.039
$Se = 1, Sp = 1$			Truth							
	CR	β	% Bias	ASE	ESE	CP				
	0.90	β_{X1}	1.652	0.108	0.106	0.955				
		β_{Z1}	2.270	0.108	0.110	0.949				
		β_{Z2}	0.080	0.108	0.112	0.949				
	0.55	β_{X1}	1.252	0.053	0.052	0.961				
		β_{Z1}	1.153	0.053	0.052	0.961				
		β_{Z2}	0.223	0.052	0.053	0.937				

[1] Se = Sensitivity [2] Sp = Specificity [3] $\hat{\delta}_{(1)}$ = Estimate of attenuation coefficient
[4] CR = True censoring rate

Table 2.5: Type I error results for $\beta_{X1} = 0$ are given for 1000 simulated data sets for the proposed method. Let $\beta_{X1} = \log(1.5)$, $\beta_{Z1} = \log(0.7)$, and $\beta_{Z2} = \log(1.3)$; e is normally distributed with mean zero.

Se[1]	Sp[2]	$\hat{\delta}_{(1)}$[3]	CR[4]	Type I Error
0.80	0.90	0.30	0.55	0.048
			0.90	0.042
		0.60	0.55	0.058
			0.90	0.042
0.90	0.80	0.30	0.55	0.043
			0.90	0.039
		0.60	0.55	0.049
			0.90	0.044

[1] Se = Sensitivity [2] Sp = Specificity
[3] $\hat{\delta}_{(1)}$ = Estimate of attenuation coefficient [4] CR = True censoring rate

Table 2.6: Hazard Ratio (HR) and 95% confidence interval (CI) estimates of incident diabetes for a 20% increase in consumption of energy (kcal/d), protein (g/d), and protein density (% energy from protein/d) based on the naive method ignoring error in the outcome and covariate, the regression calibration method that corrects for covariate error only, and the proposed method. Here, sensitivity = 0.61, specificity = 0.995, and negative predictive value = 0.96.

		HR (95% CI)	
Model[1]	Method	Adjusted for BMI[2]	Not Adjusted for BMI
Energy (kcal/d)	Naive	1.002 (0.986, 1.018)	1.024 (1.008, 1.040)
	Regression Calibration	1.041 (0.758, 1.429)	1.421 (1.043, 1.938)
	Proposed	0.822 (0.512, 1.318)	1.189 (0.836, 1.692)
Protein (g/d)	Naive	1.024 (1.010, 1.039)	1.051 (1.035, 1.066)
	Regression Calibration	1.121 (1.036, 1.213)	1.231 (1.130, 1.342)
	Proposed	1.077 (0.978, 1.186)	1.241 (1.114, 1.384)
Protein Density	Naive	1.100 (1.064, 1.137)	1.128 (1.091, 1.167)
	Regression Calibration	1.243 (1.125, 1.374)	1.325 (1.181, 1.486)
	Proposed	1.266 (1.115, 1.436)	1.327 (1.183, 1.490)

[1] Each model is adjusted for potential confounders and is stratified on age (10-year categories) and Dietary Modification trial (DM) or Observational Study (OS) cohort membership. [2] BMI = Body Mass Index (kg/m^2)

CHAPTER 3

An Augmented Likelihood Approach for the Discrete Proportional Hazards Model Using Auxiliary and Validated Outcome Data – with Application to the HCHS/SOL Study

3.1. Abstract

In large epidemiologic studies, it is typical for an inexpensive, non-invasive procedure to be used to record disease status during regular follow-up visits, with less frequent assessment by a gold standard test. Inexpensive outcome measures like self-reported disease status are practical to obtain, but can be error-prone. Association analysis reliant on error-prone outcomes may lead to biased results; however, restricting analyses to only data from the less frequently observed error-free outcome could be inefficient. We have developed an augmented likelihood that incorporates data from both error-prone outcomes and a gold standard assessment. We conduct a numerical study to show how we can improve statistical efficiency by using the proposed method over standard approaches for interval-censored survival data that do not leverage auxiliary data. We extend this method for the complex survey design setting so that it can be applied in our motivating data example. Our method is applied to data from the Hispanic Community Health Study/Study of Latinos to assess the association between energy and protein intake and the risk of incident diabetes. In our application, we demonstrate how our method can be used in combination with regression calibration to additionally address the covariate measurement error in self-reported diet.

3.2. Introduction

In large epidemiologic or clinical studies with periodic follow-up, it is often impractical to obtain a gold standard or reference standard test on all subjects at each visit time throughout the study. Instead, an inexpensive measures is typically used to assess the outcome of interest at each follow-up visit , and the reference standard diagnostic test is obtained less frequently, if at all. Compared to some reference standard diagnostic tests that may involve invasive

or otherwise impractical biomarkers, self-reported disease status is inexpensive, noninvasive, and relatively easy to obtain in large cohorts. However, self-reported disease status is often prone to measurement error. For example, some studies have shown that the sensitivity and specificity of self-reported diabetes are imperfect compared to the reference instruments of fasting glucose and hemoglobin A1c (HbA1c).(Gu et al., 2015; Margolis et al., 2008)

There has been considerable interest in methods that use surrogate or auxiliary data to improve the efficiency of inference for time-to-event analyses. In this context, surrogate endpoints are defined as outcomes that are intended to replace the true, or gold standard, outcome of interest, while auxiliary data refers to variables that are used to improve the efficiency of the analysis of the gold standard endpoint.(Conlon et al., 2015) Pepe (1992)(Pepe, 1992) introduced an estimated likelihood method for general data structures in which surrogate outcomes are available on all subjects and true outcomes are available on a subset. Magaret (2008)(Magaret, 2008) extended this work to the setting of the discrete proportional hazards model. Zee et al. (2015)(Zee et al., 2015) proposed a similar semiparametric estimated likelihood approach for parameter estimation that allows for real-time validation and does not require true and surrogate censoring times to be equal when the surrogate outcome is censored. Fleming et al. (1994) (Fleming et al., 1994) presented an augmented likelihood approach that incorporates auxiliary information into the proportional hazards model for cases when true endpoints are available on all study subjects. In their method, the likelihood can be augmented for subjects using an auxiliary (surrogate) outcome whose true endpoints are censored prior to their auxiliary endpoints.

Several methods have been developed to correct errors in binary outcome variables for discrete time-to-event settings when gold or reference standard outcome data are not available. For these approaches, estimated values of sensitivity and specificity are incorporated into the analysis to correct for the bias induced by errors in the event classification variable. Specifically, Meier et al. (2003) (Meier et al., 2003) introduced an adjusted proportional hazards model for estimating hazard ratios in the presence of discrete failure time data subject

to misclassification. Gu et al. (2015) developed a likelihood-based method that models the association of a covariate with a discrete time-to-event outcome recorded by error-prone self-reports or imperfect diagnostic tests, assuming the proportional hazards model. Boe et al. (2021) extended this work by incorporating regression calibration to additionally adjust for covariate measurement error for cases in which one or more exposure variables of interest are also recorded with error. Each of these methods addressed the misclassification by incorporating externally estimated sensitivity and specificity into the estimation.

In this paper, we develop an augmented likelihood approach that incorporates error-prone auxiliary data into the analysis of an interval-censored, gold standard assessment of a time-to-event outcome. Our method is distinct from prior work in that we consider the setting where subjects have both frequent follow-up with an auxiliary outcome and infrequent follow-up with a gold standard evaluation. Our method may be applied when auxiliary outcome data, observed through periodically collected self-reports or diagnostic tests, are available either before or after the gold standard is scheduled to be observed. This work is motivated by the Hispanic Community Health Study/Study of Latinos (HCHS/SOL), a prospective longitudinal cohort with (1) a reference standard biomarker-defined diabetes status variable, using fasting glucose and/or hemoglobin A1c (HbA1c), available at baseline and once more after 4-10 years, and (2) self-reported diabetes status recorded annually, up to 4 years beyond the reference test.

We begin the next section by introducing notation and presenting the theoretical development of our augmented likelihood function. We also extend our method to handle data from a complex survey design and develop a sandwich variance estimator. In section 3.4, we provide an extensive numerical study to demonstrate how we can improve statistical efficiency by using the proposed method instead of standard approaches for interval-censored survival data that do not leverage the auxiliary data. Section 3.5 introduces the HCHS/SOL study and illustrates the results of applying the proposed approach to this data set to assess the association between dietary energy, protein, and protein density intake and incident diabetes.

For this analysis, we additionally address the covariate measurement error. We conclude by providing a discussion of our findings and potential extensions of this work in Section 3.6.

3.3. Methods

3.3.1. Notation and Time-to-Event Model

Define T_i as the unobserved, continuous event time of interest for subjects $i = 1, ..., N$. We assume the setting of a prospective cohort study where the participants follow-up occurs at regular visit intervals (e.g. annually) and all subjects are known to be disease-free at baseline, time τ_0. Let $0 = \tau_0 < \tau_1 < \ldots \tau_J$ be the possible visit times among the N subjects and $\tau_{J+1} = \infty$. Thus, the possible follow-up can be broken into $J + 1$ disjoint intervals as follows: $[\tau_0, \tau_1), [\tau_1, \tau_2), ...[\tau_J, \tau_{J+1})$. Let n_i be the number of visits for the i^{th} subject, which we assume is random. We consider the setting where each subject reports a potentially error-prone disease status at each visit until the first positive self-report or censoring time. Let $\mathbf{Y_i^*}$ be the vector of error-prone binary outcomes that indicates whether the ith subject self-reported the event at time j, and $\mathbf{T_i^*}$ be the corresponding vector of visit times. More specifically, we define Y_{ij}^* as the binary indicator that the j^{th} self-report for the i^{th} subject is positive. Motivated by the design of the HCHS/SOL study, we assume that a gold standard assessment of the disease is also obtained, but only once post-baseline. Namely, we assume at time τ_{V_i}, we observe Δ_i, a scalar binary indicator for each subject's true disease status recorded by a gold standard diagnostic test, where $V_i \in \{1, 2, ..., J\}$. Note that in some studies, the time of gold standard assessment is fixed at $\tau_{V_i} = \tau_J$ for all subjects, but we allow $\tau_{V_i} \leqslant \tau_J$, suggesting that follow-up by self-report may continue after the gold standard outcome is reported. Finally, we assume that as a result of loss to follow-up, Δ_i may be missing on a subset of study subjects and define M_i as the binary variable indicating whether Δ_i is missing. We assume this outcome is missing completely at random, though missing at random patterns could also be readily incorporated with application of a standard inverse probability weighting approach. We can now write the joint probability of the observed data for the i^{th} subject as $P(\mathbf{Y_i^*}, \mathbf{T_i^*}, \Delta_i, T_i) = P(\mathbf{Y_i^*}, \mathbf{T_i^*} | \Delta_i, T_i) P(\Delta_i, T_i) =$

$\sum_{j=1}^{J+1} P(\mathbf{Y_i^*}, \mathbf{T_i^*} | \Delta_i, \tau_{j-1} < T_i \leqslant \tau_j) P(\Delta_i, \tau_{j-1} < T_i \leqslant \tau_j)$.

Following previous work to address misclassified outcomes in the discrete proportional hazards model, Boe et al. (2021); Gu et al. (2015); Balasubramanian and Lagakos (2003) we assume the n_i error-prone outcomes Y_{ij}^* are conditionally independent given the true disease status and event time T_i, such that $P(\mathbf{Y_i^*} | T_i, \mathbf{T_i^*}, \Delta_i) = \prod_{l=1}^{n_i} P(Y_{il}^* | T_i, T_{il}^*, \Delta_i)$. We can re-express the joint probability of observed data for the ith subject as follows:

$$P(\mathbf{Y_i^*}, \mathbf{T_i^*}, \Delta_i, T_i) \quad = \quad \sum_{j=1}^{J+1} C_{ij} P(\Delta_i, \tau_{j-1} < T_i \leqslant \tau_j), \tag{3.1}$$

where $C_{ij} = \left[\prod_{l=1}^{n_i} P(Y_{il}^* | \tau_{j-1} < T_i \leqslant \tau_j, T_{il}^*, \Delta_i)\right]$. We first assume that sensitivity (Se) and specificity (Sp) are known constants and have the following definitions: $Se = \Pr(Y_{il}^* = 1 | \tau_{j-1} < T_i \leqslant \tau_j, T_l^* \geqslant \tau_j)$ and $Sp = \Pr(Y_{il}^* = 0 | \tau_{j-1} < T_i \leqslant \tau_j, T_l^* \leqslant \tau_{j-1})$. Then, the C_{ij} are simply functions of the sensitivity and specificity. See section B.2 of the Appendix for details.

We will now derive the likelihood contribution for subjects with observed Δ_i (i.e., $M_i = 0$). For these subjects, we can rewrite the likelihood in equation 3.1 as follows:

$$P(\mathbf{Y_i^*}, \mathbf{T_i^*}, \Delta_i, T_i) \quad = \quad \sum_{j=1}^{J+1} C_{ij} P(\tau_{j-1} < T_i \leqslant \tau_j | \Delta_i) P(\Delta_i). \tag{3.2}$$

Define $\theta_j = \Pr(\tau_{j-1} < T_i \leqslant \tau_j)$. If at time τ_{V_i}, subject i is identified as a validated positive, then we have $P(\Delta_i = 1) = P(T_i \leqslant \tau_{V_i}) = \sum_{l=1}^{V_i} \theta_l$ and:

$$P(\tau_{j-1} < T_i \leqslant \tau_j | \Delta_i = 1) \quad = \quad \begin{cases} \frac{\theta_j}{\sum_{l=1}^{V_i} \theta_l} & \text{for } 1 \leqslant j \leqslant V_i \\ 0 & \text{for } V_i < j \leqslant J+1. \end{cases}$$

If subject i is identified to be a validated negative at time τ_{V_i}, then $P(\Delta_i = 0) = P(T_i >$

$\tau_{V_i}) = \sum_{l=V_i+1}^{J+1} \theta_l$ and:

$$P(\tau_{j-1} < T_i \leqslant \tau_j | \Delta_i = 0) \quad = \quad \begin{cases} 0 & \text{for } 1 \leqslant j \leqslant V_i \\ \frac{\theta_j}{\sum_{l=V_i+1}^{J+1} \theta_l} & \text{for } V_i < j \leqslant J+1. \end{cases}$$

Next, we derive the likelihood for a subject who is lost to follow-up and is missing Δ_i (i.e. $M_i = 1$). In this scenario, the joint probability of observed data for the ith subject is $P(\mathbf{Y_i^*}, \mathbf{T_i^*}, \Delta_i, T_i) = \sum_{j=1}^{J+1} \left[\prod_{l=1}^{n_i} P(Y_{il}^* | \tau_{j-1} < T_i \leqslant \tau_j, T_{il}^*) \right] P(\tau_{j-1} < T_i \leqslant \tau_j)$, and thus:

$$P(\mathbf{Y_i^*}, \mathbf{T_i^*}, \Delta_i, T_i) \quad = \quad \sum_{j=1}^{J+1} C_{ij} \theta_j. \tag{3.3}$$

Define X_i as the p-dimensional vector of time-invariant covariates. We assume that X is related with the outcome through a Cox proportional hazards model, $S(t) = S_0(t)^{\exp(x'\beta)}$. We use this model to re-express the joint probability from equations 3.2 and 3.3 and write the likelihood in terms of the baseline survival probabilities, $\mathbf{S} = (S_1, S_2, ..., S_{J+1})'$, where $S_j = \Pr(T_0 > \tau_{j-1})$ and T_0 is a random variable that has survival function $S_0(t)$. Thus $1 = S_1 > S_2 > ... > S_{J+1} > 0$ and $S_j = \sum_{h=j}^{J+1} \theta_h$. It is convenient to define R as the linear $(J+1) \times (J+1)$ transformation matrix such that $\theta = R\mathbf{S}$ and to define the $N \times (J+1)$ matrix C that consists of the C_{ij} terms defined above. Finally, we define the matrix D as $D = CR$. Then the log-likelihood can then be expressed as:

$$\begin{aligned} l(S, \beta) = \sum_{i=1}^{N} l_i(S, \beta) \quad = \quad & \sum_{i=1}^{N} \Bigg[(1 - M_i) \Delta_i \log \left(\sum_{j=1}^{V_i} D_{ij} (S_j)^{\exp(x_i'\beta)} \right) + \\ & (1 - M_i)(1 - \Delta_i) \log \left(\sum_{j=V_i+1}^{J+1} D_{ij} (S_j)^{\exp(x_i'\beta)} \right) + \\ & M_i \log \left(\sum_{j=1}^{J+1} D_{ij} (S_j)^{\exp(x_i'\beta)} \right) \Bigg]. \end{aligned} \tag{3.4}$$

We can solve for the unknown vector of parameters ψ using standard maximum likelihood estimation. Define the score function $U_i(\psi) = \frac{\partial l_i(S,\beta)}{\partial \psi}$, where $\mathbf{S} = (S_1, S_2, ..., S_{J+1})'$ and ψ is the $(p + J + 1) \times 1$ parameter vector $[\beta, \mathbf{S}]$. Let $\hat{\psi}$ denote the solution to the equations $\sum_{i=1}^{N} U_i(\psi) = 0$. The covariance matrix can be found by inverting the Hessian matrix.

3.3.2. Survey Design and Probability Sampling Weights

In this section, we extend our proposed method that uses both auxiliary and gold standard outcomes to accommodate data from a complex survey sampling design, such as HCHS/SOL, that may includes cluster-based probability sampling. We develop a weighted analogue of our log-likelihood function from equation 3.4. Later, we outline how one might use a sandwich variance estimator to address within-cluster correlation and stratification.

Define π_i as the probability that subject i will be included in a sample, which we assume is known from the survey design. Subjects are sampled with probability π_i from a population of size N_{POP}, resulting in a sample of size N. Design-based inference makes the assumption that a subject sampled with a probability π_i represents $1/\pi_i$ subjects in the total population. (Lumley, 2011) Thus, $1/\pi_i$ becomes the sampling weight reflecting unequal probability of selection into the sample, which will be included in the weighted log-likelihood and score functions. The weighted log-likelihood equation becomes $l_\pi(S, \beta) = \sum_{i=1}^{N} \frac{1}{\pi_i} l_i(S, \beta) = \sum_{i=1}^{N} \breve{l}_i(S, \beta)$. We can then use standard maximum likelihood theory to solve the corresponding weighted estimating equation $\sum_{i=1}^{N} \breve{U}_i(\psi) = \sum_{i=1}^{N} \frac{1}{\pi_i} U_i(\psi) = 0$ for our vector of unknown parameters, ψ. To compute the variance for our estimator that addresses within-cluster correlation and stratification, we consider the implicit differentiation method proposed by Binder (1983).(Binder, 1983) Using a Taylor series linearization, the sandwich estimator for the asymptotic variance of $\hat{\psi}$ can be calculated as $\hat{\text{var}}[\hat{\psi}] \approx \left(\sum_{i=1}^{N} \frac{\partial \breve{U}_i(\hat{\psi})}{\partial \psi}\right)^{-1} \hat{\text{cov}}\left[\sum_{i=1}^{N} \breve{U}_i(\hat{\psi})\right] \left(\sum_{i=1}^{N} \frac{\partial \breve{U}_i(\hat{\psi})}{\partial \psi}\right)^{-1}$. Regularity conditions required for the consistency of $\hat{\text{var}}[\hat{\psi}]$ are stated in Binder (1983).(Binder, 1983) This variance estimate can easily be computed in R by applying vcov() to the svytotal() function from the survey package and providing the estimator's influence function as well as the survey

design. (Lumley, 2011)

3.3.3. Regression Calibration to Adjust for Covariate Measurement Error

Regression calibration is a popular analysis method for correcting bias in regression parameters when exposure variables are prone to error. (Prentice, 1982; Shaw et al., 2018) We will now outline how to use regression calibration with our proposed estimator in the setting of a complex sampling design.

Assume (X, Z) is a $(p+q)$-dimensional covariate in the outcome model of interest, where X_i is a p-dimensional vector that cannot be observed without error and Z_i is a q-dimensional vector of observed, error-free covariates. Assume instead of X_i, we observe X_i^*, the corresponding error-prone p-dimensional vector. To implement regression calibration, we build a calibration model for $\hat{X} = E(X|X^*, Z)$ and substitute this predicted value for the unknown, unobserved true exposure X in our outcome model. (Prentice, 1982; Keogh et al., 2020)

Measurement Error Model

We assume that the error-prone X_i^*, is linearly related with the target exposure X_i and other error-free covariates Z_i:

$$X_i = \delta_{(0)} + \delta_{(1)} X_i^* + \delta_{(2)} Z_i + \zeta_i, \tag{3.5}$$

where ζ_i is a random error term that has mean zero and variance $\sigma^2_{\zeta_i}$ and is independent of X_i^* and Z_i. Equation (3.5) is referred to as the *calibration model*. For ease of presentation, we assume $p = 1$. It follows that the observed, error-prone exposure X_i^* conforms to the linear measurement error model: $X_i^* = \alpha_{(0)} + \alpha_{(1)} X_i + \alpha_{(2)} Z_i + e_i$, where the random error e_i is independent of X_i and Z_i and has mean zero and variance $\sigma^2_{e_i}$. (Keogh et al., 2020) This error model has been commonly applied to model the error in the self-reported dietary intake exposures observed in our motivating example from the HCHS/SOL. (Keogh and White, 2014) Regression parameters in our calibration model are identifiable if, in a subset, we observe either the true exposure, X_i, or a second error-prone observation X_i^{**} with classical

measurement error, i.e., where $X_i^{**} = X_i + \epsilon_i$, where ϵ_i is random error that is independent of all variables, with mean 0 and variance $\sigma^2_{\epsilon_i}$. In many settings, it is more common to observe X_i^{**} in the ancillary data, which we call a calibration subset. We will assume a subset is available in which we observe X_i^{**}. Note that observing the true exposure X_i is a variation of observing X_i^{**} in which the measurement error variance $\sigma^2_{\epsilon_i}$ is equal to 0, and such a subset is referred to as a validation subset. In some applied settings, the error-prone measure X_i^* in the main data may only have classical measurement error, a special scenario where $\alpha_{(0)} = \alpha_{(2)} = 0$ and $\alpha_{(1)} = 1$ in the linear measurement error model. In this case, a replicate measure in the ancillary data (typically called a reliability subset) will ensure that the parameters in the calibration model are identifiable.

With the assumed calibration subset, we can regress X_i^{**} on the error-prone exposure, X_i^*, and other covariates of interest Z_i to fit the model $X_i^{**} = \delta_{(0)} + \delta_{(1)}X_i^* + \delta_{(2)}Z_i + W_i$, where W_i is random, mean 0 error with variance $\sigma^2_{W_i} = \sigma^2_{\zeta_i} + \sigma^2_{\epsilon_i}$. The error term W_i in this model now incorporates extra variability introduced by the error in X_i^{**}.

Applying Regression Calibration to the Outcome Model

Assuming that the measurement error models described above hold, we can use the predicted values from our calibration model to substitute the first moment $\hat{X}_i = E(X_i|X_i^*, Z_i)$ in place of X_i in our outcome model. Regression calibration is exact in linear models; however, this approach is only an approximate method with some bias in non-linear outcome models. (Carroll et al., 2006) Regression calibration has been observed to perform well in various settings, including when the regression parameter corresponding to the error-prone covariate is of modest size and when the event under study is rare. (Prentice, 1982; Buonaccorsi, 2010) Additionally, regression calibration has been been shown to work well under these same settings when also correcting for errors in time-to-event outcomes. (Boe et al., 2021)

As we described in Section 3.3.2, variance estimation for data from a complex survey design often requires extra steps to address within-cluster correlation. When regression calibration is applied, variance estimates from the outcome model need to be adjusted further to account

for the extra uncertainty added by the calibration model step. We adopt the variance estimation approach proposed by Baldoni et al. (2021)(Baldoni et al., 2021), in which the expected value of the latent true exposure is multiply imputed for all individuals by repeatedly sampling the calibration model coefficients required to estimate $\hat{X}_i$. New calibration coefficients can be sampled using either (1) their estimated asymptotic parametric distribution or (2) bootstrap resampling. At each step of the imputation, the outcome model is re-fit using the newly calibrated values. Using this approach, the final estimate of the variance of the jth regression coefficient $\hat{\beta}_j$ can be computed as $\hat{V}_j^* = \frac{1}{M}\sum_{m=1}^{M}\hat{V}_j^{(m)} + \frac{1}{M-1}\sum_{m=1}^{M}\left(\hat{\beta}_j^{(m)} - \bar{\hat{\beta}}_j\right)^2$, where $\bar{\hat{\beta}}_j = \frac{1}{M}\sum_{m=1}^{M}\hat{\beta}_j^{(m)}$ and $\hat{\beta}_j^{(m)}$ and $\hat{V}_j^{(m)}$ represent the estimated jth regression coefficient and its estimated variance, respectively, using the m-th completed data set with $m = 1, \ldots, M$. Further details on variance estimation can be found in Baldoni et al. (2021)(Baldoni et al., 2021) and code for implementing them is available on GitHub at https://github.com/plbaldoni/HCHSsim.

3.3.4. Asymptotic Theory

We assume the regularity conditions of Foutz (1977) (Foutz, 1977) and apply the techniques of Boos and Stefanski (2013) (Boos and Stefanski, 2013) for verifying asymptotic normality of standard maximum likelihood estimators to establish the asymptotic properties of the proposed estimator. In Section B.3 of the Appendix, we outline regularity conditions for the following three settings: (1) the proposed method estimator is applied to data from a simple random sample from the population; (2) the proposed method estimator is extended to accommodate data from a complex survey design; and (3) the proposed method estimator is extended to incorporate regression calibration in the presence of complex survey data.

3.4. Numerical Study

We now present a simulation study conducted to assess the numerical performance of the proposed method compared to the standard discrete proportional hazards model approach for the gold standard time-to-event outcome. Regression coefficients for the standard approach are obtained by fitting a generalized linear model with a binary response and complemen-

tary log-log link. (Hashimoto et al., 2011) We explore various settings to show when the proposed estimator improves over the standard interval-censored approach in terms of statistical efficiency. In particular, we vary the probability that the gold standard indicator Δ_i is missing for some subjects, the censoring rate (CR) of the latent true event time at the end of study (i.e. if Δ_i had been observed for all subjects), and the sample size, N. Additionally, we vary the missingness rate of our auxiliary outcome variable and consider different values for our true regression parameter of interest, β, different distributions of our simulated event times, and different values of sensitivity and specificity of the auxiliary data.

3.4.1. Simulation Setup

We first consider a set of simulations assuming a simple random sample. We simulate a single covariate of interest from either a gamma distribution with shape and scale parameters of 0.2 and 1, respectively (denoted Gamma$(0.2, 1)$) or a normal distribution with mean and variance parameters 0.2 and 1 (denoted Normal$(0.2, 1)$). We assume the proportional hazards model. We fix the true log hazard ratio at $\beta = \log(1.5)$ to represent a regression coefficient of moderate size. Later, we set $\beta = \log(3)$ to see how increasing the magnitude of our regression coefficient changes our efficiency gains. Additionally, we conduct simulations with $\beta = 0$ to check type I error rates, where $\alpha = 0.05$. All simulations were run in R version 4.1.0. (R Core Team, 2018)

True event times were generated from a continuous time exponential distribution. We simulated a follow-up schedule with four fixed visit times at which we collect the auxiliary outcome variables. We assume that at year four, a gold standard outcome variable is also recorded. To obtain average censoring rates (CR) for the latent true event of 0.9, 0.7, and 0.5, we considered baseline λ_b parameters of 0.023, 0.08, and 0.17, respectively, and simulated our event times using parameter $\lambda = \lambda_b \exp(x_i'\beta)$. We discretize the continuous event times by binary event indicators for each visit time, then use sensitivity and specificity values to "corrupt" this variable, resulting in the vector of error-prone auxiliary outcomes, $\mathbf{Y_i^*}$. We varied the accuracy of our auxiliary data by considering scenarios where sensitivity = 0.90

and specificity = 0.80, as well as sensitivity = 0.80 and specificity = 0.90. To simulate scenarios in which the gold standard outcome Δ_i is not observed for some subjects ($M_i = 1$), we vary the missingness rate (MR) of Δ_i at 0, 0.2 and 0.4. To simulate this missingness, we generated N variables U_i from a Uniform(0,1) distribution and then let Δ_i be missing for each subject if $U_i < MR$. We vary the sample size between $N = 1000$ and $N = 10,000$ subjects. When $MR = 0.0$, these sample sizes are exact for the proposed approach and the no auxiliary data approach. When $MR > 0.0$, $N = 1000$ and $N = 10,000$ represent the sample sizes for the proposed approach, but the true sample sizes for the standard (no auxiliary data) approach are smaller due to missingness in the gold standard indicator Δ. For all settings, we conducted 1000 simulation iterations.

We then performed a set of simulations with similar settings, except we sought to examine the performance of the proposed method with data having the structure of a complex survey design. Code for this set of simulations was developed and described by Baldoni et al. (2021)(Baldoni et al., 2021) and is available on GitHub at https://github.com/plbaldoni/HCHSsim. Briefly, this simulation pipeline creates a superpopulation of nearly 200,000 individuals in 89,777 households, across 376 block groups, and 4 geographic strata and then for each simulation iteration drew survey samples from it using a stratified three-stage sampling scheme. The resulting simulated data sets include sampling weights, stratification variables, and cluster indicators. To simulate our gamma covariate for this set of simulations, we considered different shape and scale parameters for the four strata: $\text{shape}_1 = 0.25$, $\text{scale}_1 = 1.25$; $\text{shape}_2 = 0.15$, $\text{scale}_2 = 0.75$; $\text{shape}_3 = 0.30$, $\text{scale}_3 = 1.50$; $\text{shape}_4 = 0.10$, $\text{scale}_4 = 0.50$. For each block group g within a certain stratum s, we created additional covariate differences by simulating variables ω_{gs} from a Uniform$(-0.15 * \text{shape}_s, 0.15 * \text{shape}_s)$ and ρ_{gs} Uniform$(-0.15 * \text{scale}_s, 0.15 * \text{scale}_s)$ distribution for $s = 1 \ldots, 4$. Then, the covariate for an individual in block group g and stratum s was simulated from a Gamma$(\text{shape}_s + \omega_{gs}, \text{scale}_s + \rho_{gs})$ distribution. To illustrate the performance of our method under the complex survey design with a normally distributed covariate, we also considered variables $X_i \sim$ Normal$(\text{shape}_s + \omega_{gs}, \text{scale}_s + \rho_{gs})$. All other settings, including setting $\beta = \log(1.5)$ and

the generation of the event times and the missingness in the gold standard, were kept the same between the random sample and complex survey for this set of simulations. Due to the randomness introduced by the complex survey sampling setting, we cannot fix the total number of individuals selected for a simulated sample, but we aimed for sample sizes of approximately $N = 1000$ and $N = 10,000$ as in prior tables.

We conducted one additional simulation that aimed to mimic the HCHS/SOL study, which included error-prone covariates. We aimed for an average sample of approximately $12,987$ in order to approximate the number of HCHS/SOL cohort subjects without baseline diabetes. We assumed eight fixed visit times at which the auxiliary outcome was recorded, with a simulated gold standard occurring at year four. Missingness in the gold standard indicator at year four was set at $MR = 0.29$, the censoring rate was fixed at roughly $CR = 90\%$, and the auxiliary data missingness rate was approximately 0.20. We simulated 3 covariates of interest: X, Z_1, and Z_2 to represent dietary intake, age, and body mass index (BMI), respectively. These covariates were simulated following the data generation structure of Baldoni et al. (2021)(Baldoni et al., 2021), where each subject's sex (male, female) and Hispanic/Latino background (Dominican, Puerto Rican, and other) were first simulated from a multinomial distribution. Next, self-reported dietary intake, age, and BMI were simulated for each combination of sex and Hispanic background following a multivariate normal distribution, with means and covariance matrices estimated from the HCHS/SOL Bronx field center data. We set $\beta_1 = \log(1.5)$, $\beta_2 = \log(0.7)$, $\beta_3 = \log(1.3)$. To simulate an error-prone covariate X^*, we use the linear measurement error model, $X^* = \alpha_{(0)} + \alpha_{(1)}X + \alpha_{(2)}Z_1 + \alpha_{(3)}Z_2 + e$, where $\alpha_{(0)} = 0.05$, $\alpha_{(1)} = 0.50$, $\alpha_{(2)} = 0.003$, and $\alpha_{(3)} = 0.0009$. We assumed $e \sim N(0, \sigma_e^2)$ and used a σ_e^2 value of 0.389. To represent the biomarker subset, we take a random sample of 450 participants on which we observe a measure with classical error, simulated as $X^{**} = X + \epsilon$, where $\epsilon \sim N(0, \sigma_\epsilon^2)$ and $\sigma_\epsilon^2 = 0.019$. These values of $\alpha_{(0)}, \alpha_{(1)}, \alpha_{(2)}, \alpha_{(3)}$, σ_e^2 and σ_ϵ^2 were chosen based on parameters fit for the self-reported and recovery biomarker measurements for protein density in the HCHS/SOL data. (Mossavar-Rahmani et al., 2015)

For all simulation settings we conducted 1000 simulation iterations and report median percent (%) biases, median standard errors (ASE), empirical median absolute deviation (MAD), 95% coverage probabilities (CP), and median relative efficiencies (RE), calculated as the median of the ratio of the estimated variance of the proposed method estimator to the estimated variance of the standard approach estimator. R code used to run our simulations can be found on GitHub at https://github.com/lboe23/AugmentedLikelihood.

3.4.2. Simulation Results

In Tables 3.1-3.5, we present results for the proposed method compared to the standard interval-censored approach without auxiliary data. Table 3.1 shows results for the simple random sample with the regression parameter of interest $\beta = \log(1.5)$ and a gamma distributed covariate. The proposed method performs well, maintaining an absolute median percent bias of under 2% for all settings and achieving nominal coverage for a 95% confidence interval. We also see that our variance estimator is working properly, as our ASE values closely approximate the MAD values. We note that substantial efficiency gains (1.2-69.9%) result from incorporating auxiliary data into the analysis. Our method shows larger efficiency gains when the missingness rate, MR, for the gold-standard indicator Δ is higher and when the censoring rate of the latent true event time at the end of study CR is lower. Table B.1 in the Appendix shows a benchmark for comparing the relative efficiency gains from the proposed method to the relative efficiency gains achieved if the gold standard were available at all four visit times. We can directly compare the relative efficiency improvements from the final column of Table S1 to those in the final column of Table 3.1 to see that for these particular settings, our method retains nearly 90% of the the ideal relative efficiency.

In Table B.2 from the Appendix, we change the sensitivity and specificity values for the auxiliary outcome and let $Se = 0.90$ while $Sp = 0.80$. We see that our method still performs well with these alternate values for Se and Sp in terms of mean percent bias, standard error estimation, and coverage probability. When $MR = 0.0$, relative efficiencies are similar between Table 3.1 ($Se = 0.80$, $Sp = 0.90$) and Table S2 ($Se = 0.90$, $Sp = 0.80$). For

example, when $CR = 0.50$ and $N = 10,000$, we have an efficiency gain of 1.186 in Table 3.1 and an efficiency gain of 1.178 in table Table S2. However, when $MR > 0$, we notice more substantial efficiency gains for Table 3.1, where sensitivity is lower and specificity is higher, e.g. 1.677 vs. 1.549 for $MR = 0.4$, $CR = 0.50$ and $N = 10,000$.

Table 3.2 shows the results when the covariate of interest follows a normal distribution. Relative efficiencies in this table range from 0.1% to 39.8%, indicating that efficiency gains are not as high for a normally distributed covariate. We also assess the gains in relative efficiency for the proposed method over the standard interval-censored approach for $\beta = \log(3)$ in Table B.3 in the Appendix. Increasing the magnitude of our regression coefficient leads to much larger increases in relative efficiency, ranging from 15.5% to 117%.

Table 3.3 presents results for data simulated from a complex survey. In all scenarios, the weighted proposed estimator has minimal finite sample bias. The sandwich variance estimator performs unfavorably in some settings for both the proposed and standard method, with coverage as low as 89.9%, particularly when the sample size is small ($N = 1000$) or the CR is high. We note, problematic finite sample performance of the sandwich variance has been observed in other settings where the number of observed events is modest and/or the covariate is from a skewed distribution (Carroll et al., 1998). For all settings, relative efficiency gains are observed to be quite high for the proposed method, ranging from 0.9% to 60.9%. In Table B.4 from the Appendix, we show results for data simulated from a complex survey design using a normally distributed covariate. With a symmetrical covariate, the sandwich variance estimator performs better, achieving empirical MADs that more closely resemble the ASEs and obtaining coverage closer to the nominal 95% level. However, as we observed for the random sample case, relative efficiency gains are not as large (1%-45.5%) using a normally distributed covariate.

We present results for the simulation that mimic the data structure and complex survey design of the HCHS/SOL study in Table 3.4. Median percent bias is -0.859% for the proposed estimator and we see that applying the multiple imputation-based variance correction

approach leads to well-behaved sandwich standard errors. We estimate a relative efficiency gain of 44.2%, suggesting that our approach can lead to substantial variance reductions under the data structure and measurement error settings similar to that of the HCHS/-SOL cohort. Finally, we assess type I error results in Table 3.5. Type I error rates ranged from 0.033 to 0.065 for different values of MR, CR, and N, indicating that type I error is preserved in the proposed method for all observed settings.

3.5. Hispanic Community Health Study/Study of Latinos (HCHS/SOL) Data Example

3.5.1. HCHS/SOL Study Description

The HCHS/SOL is an ongoing multicenter community-based cohort study of 16,415 self-identified Hispanics/Latino adults aged 18-74 years recruited from randomly selected households at 4 locations in the United States (Chicago, Illinois; Miami, Florida; Bronx, New York; San Diego, California). Households were selected using a stratified 2-stage area probability sample design. The sampling methods, design, and cohort selection for HCHS/SOL have been described previously. (Sorlie et al., 2010; LaVange et al., 2010) The study was designed to identify risk factors for chronic diseases including diabetes and to quantify morbidity and all-cause mortality. Prevalent diabetes was recorded using a biomarker-defined reference standard at the baseline, in-person clinical examination visit (2008-2011). The study design was such that all participants were scheduled to be assessed for incident diabetes using (1) a biomarker-defined reference standard at a second clinic visit (visit 2) 4-10 years after baseline, and (2) annual telephone follow-up assessments recorded by self-report. Participants have up to eight annual telephone follow-up calls. We found that most ($> 97\%$) participants' follow-up call dates rounded to exactly one year from the date of their prior call, so we used the assigned annual follow-up times to define the boundaries of the follow-up intervals. Follow-up time was divided into 9 possible intervals. To define the observation time for the reference standard at visit 2, we rounded the time between baseline and the second clinic visit to the nearest year. Visits that occurred after year 8 (1.51% of all visits)

were rounded down in order to preserve the visit schedule with 9 intervals. For the interval-censored, no auxiliary data approach, we assumed that visit 2 occurred at the same time for all participants that had the reference standard available. Note we made this simplifying assumption due to the lack of available software to handle the complex survey design for the interval-censored proportional hazards model. We used this as a comparative analysis that did not use auxiliary data.

We applied the proposed method to assess the association between energy, protein and protein density (percentage of energy from protein) dietary intakes and the risk of diabetes in HCHS/SOL using both the self-reported diabetes outcome (auxiliary data) and the reference standard. The dietary exposure variables were recorded using an error-prone, self-reported 24-hour recall instrument that is believed to follow to the linear measurement error model. A subset of 485 HCHS/SOL participants were enrolled in the Study of Latinos: Nutrition and Physical Activity Assessment Study (SOLNAS). (Mossavar-Rahmani et al., 2015) The SOLNAS subset included the collection of objective recovery biomarkers that conform to the classical measurement error model and therefore can be used to develop calibration equations for the self-reported dietary intake variables.

This work was motivated by more detailed, ongoing research looking to understand the relationship between several dietary factors and risk of chronic diseases, including diabetes and cardiovascular disease, in the HCHS/SOL cohort. The proposed method is applied to a random subset of $8,200$ eligible participants, which is half of the original HCHS/SOL cohort ($N = 16,415$). Eligibility included being diabetes-free at baseline and having complete covariate data. Details on eligibility and the selection of our random subset are provided in Section B.4 of the Appendix. Our calibration models for dietary energy, protein, and protein density included age, body mass index (BMI), sex, Hispanic/Latino background, language preference, income, and smoking status. We fit the calibration equation by regressing the biomarker value (X^{**}) on the corresponding self-reported measure and other covariates. We compared self-reported diabetes and the reference standard at baseline to determine that

self-reported diabetes in HCHS/SOL has a sensitivity of 0.61 and a specificity of 0.98. We also conduct a sensitivity analysis in which we use a sensitivity of 0.77 and a specificity of 0.92, which are the measures of agreement computed using self-reported diabetes and the reference standard diabetes measure at visit 2.

All analyses accounted for the HCHS/SOL complex survey design. To fit the model for the interval-censored reference standard diabetes measure from visit 2, we used the svyglm() function from the survey package in R. (Lumley, 2011) To apply our proposed approach, we maximized the weighted log-likelihood that included HCHS/SOL sampling weights, and obtained design-based standard errors using the approach outlined in Section 3.3.2. The models for both approaches are fit M times, once for each of the newly predicted intake values $\hat{X}_i^{(m)}$ from multiple imputation. The final variance estimate is computed using the approach described in Section 3.3.3. We chose $M = 25$ imputations for our analysis. In both models, we used biomarker calibrated values of dietary energy, protein, and protein density on the log scale. Both risk models were also adjusted by the standard risk factors included in the calibration equations. We present hazard ratios (HR) and 95% confidence intervals (CI) associated with a 20% increase in consumption.

3.5.2. Results

Of the 8,200 randomly selected participants, 5,922 (72.2%) had the reference standard diabetes status variable available at visit 2. Of participants who had visit 2 data, 5 (0.1%) participants returned to the clinic four years post-baseline, 1490 (25.2%) returned after five years, 3294 (55.6%) returned after six years, 739 (12.5%) returned after 7 years, and 394 (6.7%) returned after 8 years. Using the reference standard, 623 (10.5%) of the participants with visit 2 data had incident diabetes.

Table 3.6 shows results from applying the proposed method and the standard, no auxiliary data method to the HCHS/SOL data. The HR (95% CI) for a 20% increase in energy intake was 1.20 (0.47, 3.11) for the proposed approach compared to 1.20 (0.41, 3.82) for the no auxiliary data method. For energy, we observe a relative efficiency gain of 27%

by using the proposed method. While the estimated standard error for the no auxiliary data approach is larger compared to that of the proposed method, incident diabetes is not significantly associated with energy intake in either approach. For protein, the HR (95% CI) for a 20% increase in intake using the proposed method is estimated to be 1.30 (0.82, 2.06). Comparatively, we estimate an HR (95% CI) of 1.37 (0.74, 2.51) using the no auxiliary data approach, and estimate a corresponding relative efficiency gain of 74% using the proposed method. When the proposed method is applied, the HR for a 20% increase in protein density is estimated to be 1.01 (1.00, 1.02), compared to a HR of 1.01 (1.00, 1.03) for the no auxiliary data method. Our estimated relative efficiency gain using the proposed method over the standard approach is 63% when looking at protein density. We note that this large efficiency gain was from relatively small absolute changes on the log-hazard scale.

In Table B.5 of the Appendix, we present results from a sensitivity analysis that applies the proposed method using sensitivity and specificity values estimated at visit 2 ($Se = 0.77, Sp = 0.92$). For this investigation, we use the same subset of $8,200$ HCHS/SOL participants as in the primary analysis. We observe that changing the sensitivity and specificity values does not qualitatively change our results for any of the dietary intakes under study.

3.6. Discussion

In large cohort studies like HCHS/SOL, gold or reference standard outcome variables may be less readily available than error-prone auxiliary outcomes. We have introduced a method that leverages all available data by incorporating error-prone auxiliary variables into the analysis of an interval-censored outcome. We developed methods for both a simple random sample and complex survey design for the case of time-independent covariates. Our results suggest that making use of auxiliary outcome data may often lead to a considerable improvement in the efficiency of parameter estimates, particularly when the gold standard outcome is missing for a subset of study participants. We illustrate the practical use of our approach in a complex survey design by applying the proposed method to the HCHS/SOL study to assess the association between energy, protein, and protein density intake and the

risk of incident diabetes, while adjusting for error in the self-reported exposure. In HCHS/-SOL, the reference standard diabetes outcome variable was not practical to obtain annually, while self-reported diabetes status was easily attainable. This data example served as a compelling setting for which our method could contribute, reducing the estimated variance by up to 74%. In settings with substantial measurement error, where variance estimates can be quite large, relative efficiency improvements are extremely important and may inform cost reductions for future studies.

In the HCHS/SOL study, we observe a special case of interval-censored data in which the reference standard outcome is only observed at one time point. This type of data is often called current status data, or case I interval-censored data. (Zhang and Sun, 2010) In our data example, the current status data arise due to the study design, as the reference standard outcome was scheduled to be recorded only once at a predetermined time point post-baseline. However, under the discrete time framework in which there is a common set of assessment times for all individuals, our method could be easily adapted to accommodate a reference standard status variable recorded at multiple time points. For the continuous time setting, future work is needed to consider how our estimation methods for interval-censored data could be extended. Several approaches have been applied for the analysis of continuous time interval-censored data, many of which have been shown to be computationally complex. (Zeng et al., 2016; Zhang et al., 2010; Lindsey and Ryan, 1998) These methods, however, have not yet been adapted to handle error-prone and validated outcomes. A further extension would be to consider approaches able to handle time-varying covariates.

The application of the proposed method required defining a set of common, discrete visit times across participants to avoid the curse of dimensionality. We used the assigned annual visit times to define the boundaries of the visit intervals, thus ignoring that the annual visit may not occur on the participant's exact anniversary date. We deemed this appropriate because the observed visit times were generally quite close to the anniversary times. In other settings, where the fluctuations in visit times are more extreme, one might consider

dividing time into smaller intervals. For this approach, the choice of intervals will require us to consider to what the extent the data can support estimating the increased number of nuisance parameters from a finer grid. With discrete data, we must often make a pragmatic compromise that balances the bias induced from rounding event times and the problems that may arise from a large number of parameters. Extending our methods in a way that does not restrict the number of possible visit times and allows for more parameters to be stably estimated need further investigation.

One potential limitation of our analysis of the HCHS/SOL data was the assumption of constant sensitivity and specificity across visit times, as there was some apparent disagreement between these measures of accuracy at baseline compared to visit 2. We hypothesize that this difference in agreement is primarily a result of a larger lag time since the previous gold standard test at baseline compared to follow-up visits, but could also result from missing data in the reference measure at visit 2 that may impact the sensitivity and specificity values. We conducted a sensitivity analysis to explore how using visit 2 rather than baseline values of sensitivity and specificity may impact the results of our HCHS/SOL data analysis. In this example, incorporating slightly different measures of accuracy of the self-reported auxiliary outcome data did not substantially impact our results. However, we note that this may not always be the case, especially for more extreme changes in sensitivity and specificity. For many real data settings, it may be unreasonable to assume that the sensitivity and specificity of error-prone outcomes are time-invariant. Future methods might explore the possibility of incorporating time-varying values of sensitivity and specificity. A second potential limitation was our assumption that the gold standard outcome was missing completely at random. Using our proposed method for the complex survey design, we anticipate an extension could be readily developed to handle the missing at random case with the use of inverse probability weighting.

In our numerical study, we noticed that the sandwich variance estimator had some coverage issues in smaller sample settings using both the proposed method and the standard no

auxiliary data approach. While the sandwich variance estimator performed better with a normally distributed covariate, we noticed some numerical challenges when the covariate of interest had a long-tailed distribution (e.g. the gamma distribution). The numerical limitations of the sandwich variance estimator for complex survey data in non-linear models have been discussed previously. Bias in the sandwich estimator may be encountered with smaller sample sizes and rare outcomes, particularly for a covariate with a heavy-tailed distribution, since in these settings, the variability of regression parameters is underestimated. (Carroll et al., 1998; Rogers and Stoner, 2015) Despite these limitations, the sandwich estimator may be reasonable, as coverage remained above 89%, got closer to 95% in large samples, and it is very practical to implement.

There are several methods for variance estimation that may be considered when applying regression calibration. Further steps are typically required to incorporate the extra uncertainty added by the calibration model and make valid inference on the parameter of interest. In practice, the bootstrap estimator is often used due to its simple implementation. For cases in which the data are from a complex survey design like HCHS/SOL, additional considerations are needed to account for aspects of the sampling design and the standard bootstrap variance estimator may be less straightforward to apply. While the multiple imputation approach of Baldoni et al. (2021)(Baldoni et al., 2021) can be applied in these scenarios, we observed some instances of over-coverage of the 95% confidence intervals using these variance estimators (data not shown). This issue is discussed by Baldoni et al. (2021)(Baldoni et al., 2021) and is believed to be attributed to instability introduced by multicollinearity in the simulated data. Future work may consider alternative variance estimation strategies in the presence of regression calibration and the complex survey setting, such as a sandwich variance estimator obtained by stacking the calibration and outcome model estimating equations.

Table 3.1: Simulation results are shown for exponential failure times assuming the Cox proportional hazards model with $X \sim Gamma(0.2, 1)$ and $\beta = \log(1.5)$. The median percent (%) bias, median standard errors (ASE), empirical median absolute deviation (MAD) and coverage probabilities (CP) are given for 1000 simulated data sets for the proposed method and the standard interval-censored approach that does not incorporate auxiliary data. Here, $Se = 0.80$ and $Sp = 0.90$ for the auxiliary data.

			Proposed				No Auxiliary Data				
MR[1]	CR[2]	N[3]	% Bias	ASE	MAD	CP	% Bias	ASE	MAD	CP	RE[4]
0.0	0.9	1000	-0.958	0.159	0.150	0.956	-1.402	0.160	0.155	0.951	1.012
		10,000	1.351	0.048	0.050	0.947	1.279	0.048	0.051	0.951	1.010
	0.7	1000	0.824	0.103	0.100	0.947	0.614	0.107	0.106	0.950	1.053
		10,000	0.543	0.032	0.032	0.944	0.398	0.033	0.034	0.947	1.070
	0.5	1000	1.923	0.091	0.088	0.943	2.020	0.099	0.102	0.947	1.182
		10,000	0.521	0.028	0.029	0.946	0.382	0.031	0.034	0.951	1.186
0.2	0.9	1000	-1.071	0.172	0.170	0.957	-0.378	0.181	0.183	0.951	1.072
		10,000	1.199	0.052	0.050	0.958	0.769	0.054	0.055	0.952	1.087
	0.7	1000	1.333	0.109	0.106	0.953	0.377	0.120	0.116	0.954	1.184
		10,000	0.713	0.034	0.035	0.942	0.332	0.037	0.038	0.946	1.206
	0.5	1000	1.798	0.095	0.095	0.945	2.084	0.111	0.116	0.947	1.363
		10,000	0.534	0.029	0.030	0.945	0.247	0.034	0.036	0.952	1.370
0.4	0.9	1000	0.256	0.189	0.188	0.959	1.178	0.213	0.222	0.959	1.195
		10,000	1.444	0.056	0.057	0.951	2.122	0.062	0.064	0.960	1.221
	0.7	1000	1.228	0.115	0.111	0.946	1.616	0.140	0.138	0.958	1.419
		10,000	0.403	0.036	0.036	0.942	0.758	0.043	0.044	0.946	1.428
	0.5	1000	1.732	0.099	0.097	0.943	3.186	0.130	0.136	0.952	1.699
		10,000	0.350	0.031	0.030	0.946	0.122	0.040	0.043	0.945	1.677

[1] MR = Average probability that the gold standard indicator Δ is missing at year 4

[2] CR = Average censoring rate for the latent true event time at the end of study

[3] N = Sample size for proposed approach; if $MR > 0.0$, sample size for no auxiliary data approach is smaller because of missingness in gold standard indicator Δ.

[4] RE = median relative efficiency, calculated as the median of the ratio of the estimated variance of the standard, no auxiliary data approach estimator to the estimated variance of the proposed method estimator , e.g. $\frac{Var(\hat{\beta}_{Standard})}{Var(\hat{\beta}_{Proposed})}$

Table 3.2: Simulation results are shown for exponential failure times assuming the Cox proportional hazards model with $X \sim Normal(0.2, 1)$ and $\beta = \log(1.5)$.The median percent (%) bias, median standard errors (ASE), empirical median absolute deviation (MAD) and coverage probabilities (CP) are given for 1000 simulated data sets for the proposed method and the standard interval-censored approach that does not incorporate auxiliary data. Here, $Se = 0.80$ and $Sp = 0.90$ for the auxiliary data.

			Proposed				No Auxiliary Data				
MR[1]	CR[2]	N[3]	% Bias	ASE	MAD	CP	% Bias	ASE	MAD	CP	RE[4]
0.0	0.9	1000	-0.730	0.100	0.102	0.945	-0.887	0.100	0.102	0.945	1.002
		10,000	-0.199	0.032	0.032	0.952	-0.278	0.032	0.032	0.949	1.001
	0.7	1000	-0.545	0.059	0.055	0.951	-0.689	0.059	0.055	0.951	1.013
		10,000	0.019	0.018	0.018	0.950	0.064	0.019	0.019	0.949	1.014
	0.5	1000	0.157	0.046	0.044	0.953	0.166	0.047	0.049	0.948	1.056
		10,000	-0.194	0.014	0.015	0.948	-0.203	0.015	0.014	0.953	1.057
0.2	0.9	1000	-0.855	0.110	0.109	0.943	-0.940	0.112	0.110	0.944	1.044
		10,000	-0.060	0.035	0.036	0.951	-0.031	0.035	0.037	0.950	1.043
	0.7	1000	-0.676	0.063	0.058	0.954	-0.470	0.066	0.059	0.953	1.103
		10,000	-0.020	0.020	0.019	0.953	-0.058	0.021	0.021	0.948	1.103
	0.5	1000	0.072	0.049	0.050	0.954	0.583	0.053	0.054	0.939	1.184
		10,000	-0.197	0.015	0.015	0.949	-0.206	0.017	0.016	0.946	1.184
0.4	0.9	1000	-1.050	0.123	0.122	0.943	0.264	0.130	0.129	0.947	1.116
		10,000	-0.043	0.039	0.040	0.953	0.009	0.041	0.041	0.944	1.113
	0.7	1000	-0.470	0.068	0.066	0.956	-0.385	0.076	0.074	0.949	1.253
		10,000	-0.112	0.021	0.020	0.955	-0.248	0.024	0.023	0.960	1.252
	0.5	1000	0.124	0.052	0.051	0.949	0.723	0.061	0.065	0.937	1.398
		10,000	-0.258	0.016	0.016	0.948	-0.221	0.019	0.019	0.948	1.396

[1] MR = Average probability that the gold standard indicator Δ is missing at year 4

[2] CR = Average censoring rate for the latent true event time at the end of study

[3] N = Sample size for proposed approach; if $MR > 0.0$, sample size for no auxiliary data approach is smaller because of missingness in gold standard indicator Δ.

[4] RE = median relative efficiency, calculated as the median of the ratio of the estimated variance of the standard, no auxiliary data approach estimator to the estimated variance of the proposed method estimator , e.g. $\frac{Var(\hat{\beta}_{Standard})}{Var(\hat{\beta}_{Proposed})}$

Table 3.3: Simulation results are shown for data simulated to be from a complex survey with exponential failure times assuming the Cox proportional hazards model with $X \sim Gamma(\text{shape}_s + \omega_{gs}, \text{scale}_s + \rho_{gs})$ for an individual in block group g and stratum s and $\beta = \log(1.5)$. The median percent (%) bias, median standard errors (ASE), median absolute deviation (MAD) and coverage probabilities (CP) are given for 1000 simulated data sets for the weighted proposed estimator and the weighted interval-censored approach that does not incorporate auxiliary data when both use a sandwich variance estimator to address within-cluster correlation. Here, $Se = 0.80$ and $Sp = 0.90$ for the auxiliary data.

			Proposed				No Auxiliary Data				
MR[1]	CR[2]	N[3]	% Bias	ASE	MAD	CP	% Bias	ASE	MAD	CP	RE[4]
0.0	0.9	1000	3.528	0.137	0.155	0.903	2.591	0.140	0.161	0.901	1.009
		10,000	2.643	0.044	0.045	0.923	1.970	0.044	0.044	0.937	1.029
	0.7	1000	4.621	0.098	0.109	0.920	5.902	0.102	0.115	0.910	1.067
		10,000	1.862	0.033	0.032	0.928	1.707	0.034	0.035	0.927	1.093
	0.5	1000	4.714	0.092	0.100	0.927	4.735	0.099	0.107	0.917	1.167
		10,000	1.198	0.031	0.033	0.945	0.846	0.034	0.035	0.930	1.177
0.2	0.9	1000	3.048	0.146	0.165	0.902	0.135	0.153	0.183	0.912	1.053
		10,000	3.010	0.047	0.047	0.925	2.093	0.049	0.050	0.932	1.110
	0.7	1000	3.695	0.103	0.113	0.922	5.268	0.113	0.132	0.903	1.225
		10,000	2.438	0.034	0.035	0.926	1.823	0.038	0.038	0.919	1.218
	0.5	1000	3.180	0.097	0.103	0.923	3.578	0.110	0.117	0.931	1.308
		10,000	1.243	0.033	0.034	0.938	1.120	0.037	0.038	0.920	1.354
0.4	0.9	1000	4.535	0.158	0.175	0.899	0.796	0.180	0.206	0.916	1.169
		10,000	2.697	0.050	0.050	0.931	1.847	0.057	0.058	0.930	1.265
	0.7	1000	4.035	0.107	0.120	0.918	5.968	0.130	0.147	0.919	1.447
		10,000	2.805	0.036	0.037	0.925	2.259	0.043	0.047	0.924	1.455
	0.5	1000	3.168	0.099	0.111	0.932	4.136	0.126	0.149	0.927	1.573
		10,000	0.945	0.034	0.035	0.935	1.106	0.043	0.043	0.924	1.609

[1] MR = Average probability that the gold standard indicator Δ is missing at year 4

[2] CR = Average censoring rate for the latent true event time at the end of study

[3] (N) = Average sample size for proposed approach; if $MR > 0.0$, sample size for no auxiliary data approach is smaller because of missingness in gold standard indicator Δ.

[4] RE = median relative efficiency, calculated as the median of the ratio of the estimated variance of the standard, no auxiliary data approach estimator to the estimated variance of the proposed method estimator , e.g. $\frac{Var(\hat{\beta}_{Standard})}{Var(\hat{\beta}_{Proposed})}$

Table 3.4: Simulation results are shown for data simulated to have a similar structure to the complex survey design of HCHS/SOL, assuming exponential failure times and the Cox proportional hazards model with $\beta = \log(1.5)$. The median percent (%) bias, median standard errors (ASE), median absolute deviation (MAD) and coverage probabilities (CP) are given for 1000 simulated data sets for the proposed estimator and the interval-censored approach that does not incorporate auxiliary data when both apply regression calibration to address covariate error. Variance estimation is performed using the resampling-based multiple imputation procedure of Baldoni et al. (2021).

Proposed				No Auxiliary Data				
% Bias	ASE	MAD	CP	% Bias	ASE	MAD	CP	RE[1]
-0.859	0.203	0.189	0.949	-1.470	0.244	0.237	0.950	1.442

[1] RE = median relative efficiency, calculated as the median of the ratio of the estimated variance of the standard, no auxiliary data approach estimator to the estimated variance of the proposed method estimator , e.g. $\frac{Var(\hat{\beta}_{Standard})}{Var(\hat{\beta}_{Proposed})}$

Table 3.5: Type I error results for $\beta = 0$ are given for 1000 simulated data sets for the proposed method when data are simulated using exponential failure times and assuming the Cox proportional hazards model with $X \sim Gamma(0.2, 1)$. Here, $Se = 0.80$ and $Sp = 0.90$ for the auxiliary data.

		Type I Error Rate		
CR[1]	N[2]	MR[3]$= 0.0$	$MR = 0.2$	$MR = 0.4$
0.9	1000	0.045	0.033	0.049
	10,000	0.056	0.065	0.054
0.7	1000	0.043	0.049	0.061
	10,000	0.047	0.045	0.048
0.5	1000	0.050	0.049	0.057
	10,000	0.051	0.056	0.051

[1] CR = Average censoring rate for the latent true event time at the end of study [2] N = Sample size for proposed approach; if $MR > 0.0$, sample size for no auxiliary data approach is smaller because of missingness in gold standard indicator Δ.
[3] MR = Average probability that the gold standard indicator Δ is missing at year 4

Table 3.6: HCHS/SOL Data Analysis on a random subset ($N = 8,200$) of study participants using baseline sensitivity ($Se = 0.61$) and specificity ($Sp = 0.98$) values. Hazard Ratio (HR) and 95% confidence interval (CI) estimates of incident diabetes for a 20% increase in consumption of energy (kcal/d), protein (g/d), and protein density (% energy from protein/d) based on the proposed estimator and the interval-censored approach that does not incorporate auxiliary data.

	HR (95% CI)		
Model[1]	Proposed	No Auxiliary Data	RE[2]
Energy (kcal/d)	1.20 (0.47, 3.11)	1.20 (0.41, 3.82)	1.27
Protein (g/d)	1.30 (0.82, 2.06)	1.37 (0.74, 2.51)	1.74
Protein Density	1.01 (1.00, 1.02)	1.01 (1.00, 1.03)	1.63

[1] Each model is adjusted for potential confounders including age, body mass index (BMI), sex, Hispanic/Latino background, language preference, education, income, and smoking status.
[2] RE = relative efficiency, calculated as the ratio of the estimated variance of the standard, no auxiliary data approach estimator to the estimated variance of the proposed method estimator , e.g. $\frac{Var(\hat{\beta}_{Standard})}{Var(\hat{\beta}_{Proposed})}$

CHAPTER 4

Practical considerations for sandwich variance estimation in two-stage regression settings

4.1. Abstract

We present a practical approach for computing the sandwich variance estimator in two-stage regression model settings. As a motivating example for two-stage regression, we consider regression calibration, a popular approach for addressing covariate measurement error. The sandwich variance approach has been rarely applied in regression calibration, despite that it requires less computation time than popular resampling approaches for variance estimation, specifically the bootstrap. This is likely due to requiring specialized statistical coding. In practice, a simple bootstrap approach with Wald confidence intervals is often applied, but this approach can yield confidence intervals that do not achieve the nominal coverage level. We first outline the steps needed to compute the sandwich variance estimator. We then introduce a method of computation in R that we have developed for sandwich variance estimation, which leverages existing R functions and can be applied in the case of a simple random sample or complex survey design. We use a simulation study to compare the performance of the sandwich to a resampling variance approach for both data settings. Finally, we further compare these two variance estimation approaches for two data examples, the Women's Health Initiative (WHI) and Hispanic Community Health Study/Study of Latinos (HCHS/SOL).

4.2. Introduction

Two stage regression models arise in several settings in epidemiology and statistics, including those requiring correction for covariate measurement error, mediation analysis, and the use of instrumental variables in causal inference (Keogh et al., 2020; Baron and Kenny, 1986; Baiocchi et al., 2014). Typically, plug-in estimates for nuisance parameters are obtained in a stage 1 regression model and, in stage 2, an outcome model reliant on these plug-in estimates

is fit. As a motivating example, we consider regression calibration, a popular analysis approach for correcting biases in regression coefficients induced by covariate measurement error (Shaw et al., 2018; Prentice, 1982). When regression calibration is applied, additional steps are required for variance estimation for the outcome model coefficients to account for the uncertainty in the estimated, error-adjusted exposure. Usual standard errors obtained from the outcome model are generally too small, resulting in confidence intervals that are too narrow.

There are a few approaches for standard error estimation in two-stage regression. We consider settings where the data are either from a simple random sample (SRS) from the population or a complex survey sampling design. For the SRS case, two common approaches for standard error estimation rely on either a bootstrap variance estimator (Efron, 1979) or a sandwich variance estimator obtained by stacking the calibration and outcome model estimating equations Boos and Stefanski (2013). In practice, the bootstrap estimator is often implemented because it is fairly simple to apply (Keogh et al., 2020), but simple Wald confidence intervals constructed using bootstrap standard error estimates can suffer from poor coverage (Davison and Hinkley, 1997). For data from a complex survey design, valid application of the bootstrap is much less straightforward. Baldoni et al. (2021) introduced a multiple imputation (MI)-based procedure for variance estimation when applying regression calibration to complex survey data. The sandwich variance estimator has also been extended for data from a complex survey design (Binder, 1983; Lumley and Scott, 2017). Likely due to the lack of software for specialized two-stage setting, the sandwich variance estimation has rarely been applied in the setting of regression calibration and simple random sampling, and we are not aware of any example in existing literature where this approach has been used for regression calibration in a complex survey design.

In this paper, we describe how to apply the sandwich variance estimator in two-stage regression settings. This work is motivated by applications in two cohort studies in the US: the Women's Health Initiative (WHI) and the Hispanic Community Health Study/S-

tudy of Latinos (HCHS/SOL), which we describe in more detail in the next section. We then introduce the two-stage model setting and review the stacked estimating approach of Boos and Stefanski (2013), which provides a sandwich variance estimate for the outcome model regression parameters. We develop a procedure that uses quantities returned from the regression fit and other standard functions in R Software to compute the sandwich estimator, which can be implemented using our functions, available on GitHub at https://github.com/lboe23/sandwich2stage. We use a simulation study to assess how the sandwich compares to its competing variance estimators, considering scenarios in regression calibration where the bootstrap is observed to have coverage problems. Finally, we report the analyses of our data examples and conclude by summarizing our findings.

4.3. Motivating Data Examples

The Women's Health Initiative (WHI), a collection of studies launched in 1993, investigated the major causes of morbidity and mortality in US post-menopausal women aged 50-79 (The Women's Health Initiative Study Group, 1998). It is of interest to assess the association of incident diabetes with dietary energy, protein, and protein density consumption, where these exposures are self-reported and thus error-prone. The WHI also included the Nutritional Biomarker Study, in which objective recovery biomarkers for energy and protein intake were recorded on a subset of participants Neuhouser et al. (2008). Tinker et al. (2011) used regression calibration to correct for error in self-reported dietary energy and protein and then evaluated the association between the calibrated dietary variables and incident diabetes using a Cox proportional hazards model. This study used the bootstrap for variance estimation for the outcome model parameters. We reanalyze data from Tinker et al. (2011) and compare the bootstrap and sandwich variance estimates.

For our complex survey setting, we consider an example from The Hispanic Community Health Study/Study of Latinos (HCHS/SOL), an ongoing, multicenter community-based cohort study of Hispanics/Latino adults aged 18-74 years recruited from randomly selected households at 4 US field centers (Chicago, Illinois; Miami, Florida; Bronx, New

York; San Diego, California). The HCHS/SOL cohort was recruited using a multi-stage, probability-based sampling design. A random subset of HCHS/SOL participants were enrolled in the Study of Latinos: Nutrition and Physical Activity Assessment Study (SOLNAS), and objective recovery biomarkers were collected for several dietary components (Mossavar-Rahmani et al., 2015). Baldoni et al. (2021) used regression calibration to correct for the measurement error in self-reported dietary potassium in the HCHS/SOL and assessed the cross-sectional association between calibrated potassium intake and baseline hypertension-related outcome variables. We reanalyze this data to compare the variance estimates from MI and the sandwich.

4.4. Methods

This section begins by introducing the two-stage model setup. For ease of presentation, we present two-stage regression in the context of regression calibration, where at stage 1, nuisance parameters are estimated from a regression model used to adjust (calibrate) the observed exposure, and in stage 2, an outcome regression model is fit using the plug-in estimator from stage 1. We describe the proposed sandwich estimator and the established competing variance estimators.

4.4.1. Notation and two-stage model setup

We consider the two-stage model setting, where a $j \times 1$ vector $\boldsymbol{\alpha}$ and the $k \times 1$ vector $\boldsymbol{\beta}$ are estimated at stage 1 and stage 2, respectively. Consider a study cohort of N individuals, either from a SRS or a complex survey. For $i = 1, \ldots, N$, let X_i^* be an observed, error-prone covariate, which is assume to be linearly related with the true, unobserved exposure X_i and other error-free covariates Z_i such that

$$X_i = \alpha_0 + \alpha_1 X_i^* + \alpha_2 Z_i + U_i, \tag{4.1}$$

where U_i is a random error term that is independent of all variables and has mean 0 and variance σ_U^2. Suppose there is also a sub-study of size n in which X_i^{**} is observed to follow

the classical measurement error model, i.e.

$$X_i^{**} = X_i + \epsilon_i, \tag{4.2}$$

where ϵ_i is a random error that is independent of all variables, with mean 0 and variance σ_ϵ^2.

4.4.2. Stage 1 Model

Denote the stage 1 model of interest by $f(x^{**}|x^*, z; \boldsymbol{\alpha})$, the conditional probability density function of X_i^{**} given X_i^* and Z_i. Data for the stage 1 model consist of values $\{X_i^{**}, X_i^*, Z_i\}$ for the n individuals in the sub-study. Let $l_1(\boldsymbol{\alpha})$ be the corresponding log-likelihood for the stage 1 model. To apply regression calibration, one builds a calibration model to estimate the average true exposure given the observed covariates, namely $\hat{X}_i(\boldsymbol{\alpha}) = E(X_i|X_i^*, Z_i; \boldsymbol{\alpha})$, which is the stage 1 model. For ease of notation, we often suppress $\boldsymbol{\alpha}$ and use $\hat{X}_i$ to denote both the model $\hat{X}_i(\boldsymbol{\alpha})$ and the fitted value $\hat{X}_i(\hat{\boldsymbol{\alpha}})$, where the plug-in estimator $\hat{\boldsymbol{\alpha}}$ is used. The calibration model parameters may be estimated if, for a subset of individuals, X_i or X_i^{**}, a measure containing independent classical error, is observed.

4.4.3. Stage 2 Model

Suppose we are interested in the relationship between some outcome, Y_i, and the covariates (X_i, Z_i). Let $g(y|x, z; \boldsymbol{\alpha}, \boldsymbol{\beta})$ be the outcome, or stage 2, model of interest, which has corresponding log-likelihood function $l_2(\boldsymbol{\beta})$. The stage 2 model is fit using the N observations $\{Y_i, X_i, Z_i\}$ from the main study. When the exposure of interest is unobserved, one can substitute the estimated $\hat{X}_i$ in for X_i in the outcome model to obtain an estimate of the unknown regression parameter vector, $\boldsymbol{\beta}$ (Prentice, 1982; Keogh et al., 2020). This is accomplished by using the estimated plug-in parameters, $\hat{\boldsymbol{\alpha}}$, from the stage 1 model to estimate $\hat{X}_i$.

4.4.4. Variance Estimation

We now describe the different variance estimators that incorporate the uncertainty added by the estimation of the nuisance parameters in the stage 1 model. We present the formulation of the sandwich variance estimator (Boos and Stefanski, 2013) for the SRS and complex survey design settings. We also review the bootstrap variance estimator (Efron, 1979) and the MI procedure proposed by Baldoni et al. (2021).

Sandwich Variance Estimator

Case 1: Simple Random Sample

The sandwich variance estimator is obtained by stacking the calibration and outcome model estimating equations (Boos and Stefanski, 2013). We are interested in the "stacked" parameter vector, $\theta = (\boldsymbol{\alpha}, \boldsymbol{\beta})$, which includes the parameters from the stage 1 and stage 2 models. We introduce these methods for a subset of outcome models, e.g. the familiar generalized linear model (GLM) and Cox proportional hazards model, but our technique works more generally for any pair of stage 1 and stage 2 models in which a sandwich variance estimator exists. We outline sufficient assumptions for sandwich variance estimation in Section C.1 in the Appendix.

Define $U_i(\theta)$ as the $j + k$-dimensional vector of stacked estimating equations formed for θ, which can be broken down into the estimating equations for the stage 1 model, $U_{i1}(\theta)$, and the stage 2 model, $U_{i2}(\theta)$. For maximum likelihood estimation, $U_{i1}(\theta)$ and $U_{i2}(\theta)$ are the score functions, or the vector of first derivatives of the log-likelihood functions $l_{i1}(\boldsymbol{\alpha})$ and $l_{i2}(\boldsymbol{\beta})$, respectively, with respect to the parameters being estimated. We now write $U_i(\theta)$ as:

$$U_i(\theta) = \begin{bmatrix} U_{i1}(\theta) \\ \\ U_{i2}(\theta) \end{bmatrix} = \begin{bmatrix} \frac{\partial l_{i1}(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} \\ \\ \frac{\partial l_{i2}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \end{bmatrix}. \tag{4.3}$$

Estimates of our vector of unknown parameters, $\hat{\theta}$, can be found by solving the equations

$\sum_{i=1}^{N} U_i(\theta) = 0$. Following Boos and Stefanski (2013), a sandwich estimator for the variance of $\hat{\theta}$ takes the form:

$$V(\hat{\theta}) = A(\hat{\theta})^{-1} B(\hat{\theta}) \left[A(\hat{\theta})^{-1} \right]^T / N, \tag{4.4}$$

where

$$A(\hat{\theta}) = \frac{1}{N} \sum_{i=1}^{N} \frac{\partial U_i(\theta)}{\partial \theta} \Big|_{\theta=\hat{\theta}}, \tag{4.5}$$

and

$$B(\hat{\theta}) = \frac{1}{N} \sum_{i=1}^{N} U_i(\hat{\theta}) U_i(\hat{\theta})^T. \tag{4.6}$$

The $[j+k] \times [j+k]$ matrix $A(\hat{\theta})$ has the following form:

$$A(\hat{\theta}) = \frac{1}{N} \sum_{i=1}^{N} \begin{bmatrix} \frac{\partial U_{i1}(\theta)}{\partial \boldsymbol{\alpha}}|_{\theta=\hat{\theta}} & \frac{\partial U_{i1}(\theta)}{\partial \boldsymbol{\beta}}|_{\theta=\hat{\theta}} \\ \frac{\partial U_{i2}(\theta)}{\partial \boldsymbol{\alpha}}|_{\theta=\hat{\theta}} & \frac{\partial U_{i2}(\theta)}{\partial \boldsymbol{\beta}}|_{\theta=\hat{\theta}} \end{bmatrix} = \frac{1}{N} \sum_{i=1}^{N} \begin{bmatrix} \frac{\partial U_{i1}(\theta)}{\partial \boldsymbol{\alpha}}|_{\theta=\hat{\theta}} & 0 \\ \frac{\partial U_{i2}(\theta)}{\partial \boldsymbol{\alpha}}|_{\theta=\hat{\theta}} & \frac{\partial U_{i2}(\theta)}{\partial \boldsymbol{\beta}}|_{\theta=\hat{\theta}} \end{bmatrix}. \tag{4.7}$$

In the case of maximum likelihood estimation, the $[j \times j]$ upper-left quadrant, $\frac{\partial U_{i1}(\theta)}{\partial \boldsymbol{\alpha}}|_{\theta=\hat{\theta}}$, and $[k \times k]$ bottom-right quadrant, $\frac{\partial U_{i2}(\theta)}{\partial \boldsymbol{\beta}}|_{\theta=\hat{\theta}}$, are the Hessian matrices for the stage 1 and stage 2 models, respectively. All elements of the $[j \times k]$ upper-right quadrant, $\frac{\partial U_{i1}(\theta)}{\partial \boldsymbol{\beta}}|_{\theta=\hat{\theta}}$, are 0 because the stage 2 parameters are not involved in the estimating equation for the stage 1 model. The $[k \times j]$ elements of the bottom-left quadrant, $\frac{\partial U_{i2}(\theta)}{\partial \boldsymbol{\alpha}}|_{\theta=\hat{\theta}}$, are non-zero since the estimated exposure $\hat{X}_i$ in the stage 2 model relies on the $\boldsymbol{\alpha}$ parameters from stage 1. These derivatives can be computed directly if a closed-form solution exists or using numerical derivatives. The $[j+k] \times [j+k]$ matrix $B(\hat{\theta})$ has the following general form:

$$B(\hat{\theta}) = \frac{1}{N}\sum_{i=1}^{N}\begin{bmatrix} U_{i1}(\hat{\theta}) \\ \\ U_{i2}(\hat{\theta}) \end{bmatrix}\begin{bmatrix} U_{i1}^T(\hat{\theta}) & U_{i2}^T(\hat{\theta}) \end{bmatrix} = \frac{1}{N}\sum_{i=1}^{N}\begin{bmatrix} U_{i1}(\hat{\theta})U_{i1}^T(\hat{\theta}) & U_{i1}(\hat{\theta})U_{i2}^T(\hat{\theta}) \\ \\ U_{i2}(\hat{\theta})U_{i1}^T(\hat{\theta}) & U_{i2}(\hat{\theta})U_{i2}^T(\hat{\theta}) \end{bmatrix}. \quad (4.8)$$

In R, this sandwich variance estimate can be computed by (1) directly computing the matrices in equation 4.4, or (2) taking advantage of convenient functions in the survey package and readily available quantities from the fitted stage 1 and stage 2 models (Lumley, 2011). For the latter, one must compute the influence functions of the estimator $A(\hat{\theta})^{-1}\tilde{U}(\hat{\theta})$, where $\tilde{U}(\hat{\theta})$ is the matrix of transposed estimating equation contributions, defined as:

$$\tilde{U}(\hat{\theta}) = \begin{bmatrix} U_{11}^T(\hat{\theta}) & U_{12}^T(\hat{\theta}) \\ U_{21}^T(\hat{\theta}) & U_{22}^T(\hat{\theta}) \\ ... & ... \\ U_{(N-1)1}^T(\hat{\theta}) & U_{(N-1)2}^T(\hat{\theta}) \\ U_{N1}^T(\hat{\theta}) & U_{N2}^T(\hat{\theta}) \end{bmatrix} \quad (4.9)$$

.

In Figure 4.1, we provide sample R code that computes $V(\hat{\theta})$ using the influence function approach. Our R function assumes the stage 1 is a linear model fit using `svyglm` with a gaussian response, with the resulting fit stored in `stage1.model`. We assume the stage 2 model is a generalized linear or Cox model fit using the `svyglm` or `svycoxph` functions, respectively, which return the fitted model `stage2.model`. In Figure 4.1a, we create the SRS survey design and provide sample model fitting statements for the stage 1 and stage 2 models. In Figure 4.1b, we show how to obtain the pieces of the matrix, $A(\hat{\theta})$. The computation of the bottom-left quadrant relies on the function `stage2.alphas.estfuns()`, available on GitHub, which computes the numerical derivatives for the stage 2 estimating

equation with respect to $\boldsymbol{\alpha}$. Next in Figure 4.1c, we show that the matrix $\tilde{U}(\hat{\theta})$ can be obtained by stacking the estimating equation contributions for the stage 1 and stage 2 models. All stage 1 estimating equation contributions for those not in the calibration subset are equal to 0. Finally, in Figure 4.1d we compute the influence functions by multiplying the $[N \times (j+k)]$ matrix $\tilde{U}(\hat{\theta})$ by the $[j+k] \times [j+k]$ matrix $\left[A(\hat{\theta})^{-1}\right]^T$ and dividing by the sample size, N, which are used to compute an estimate of $V(\hat{\theta})$.

Case 2: Complex Survey Design

The two-stage sandwich variance estimator can easily be extended for complex survey data. In this section, we consider designs in which the validation sub-study is nested in the full study cohort, but it is straightforward to extend this approach for other designs. Define π_i as the probability that subject i will be included in the sample, which is known from the survey design. A participant sampled with probability π_i is assumed to represent $1/\pi_i$ participants in the total population, which becomes the sampling weight reflecting unequal probability of selection into the sample (Lumley, 2011). Consider $\breve{U}_i(\theta) = \frac{1}{\pi} U_i(\theta)$, the vector of stacked estimating equations formed for θ in a probability-based sampling design. As before, we can break $\breve{U}_i(\theta)$ down into $\breve{U}_{i1}(\theta)$ and $\breve{U}_{i2}(\theta)$. Define $\breve{l}_1(\boldsymbol{\alpha}) = \frac{1}{\pi} l_1(\boldsymbol{\alpha})$ and $\breve{l}_2(\boldsymbol{\beta}) = \frac{1}{\pi} l_2(\boldsymbol{\beta})$ as the weighted log-likelihood for the stage 1 and stage 2 models, respectively. The $j+k$-dimensional vector of stacked estimating equations is then:

$$\breve{U}_i(\theta) = \begin{bmatrix} \breve{U}_{i1}(\theta) \\ \\ \breve{U}_{i2}(\theta) \end{bmatrix} = \begin{bmatrix} \frac{\partial \breve{l}_1(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} \\ \\ \frac{\partial \breve{l}_2(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \end{bmatrix}. \tag{4.10}$$

To obtain estimates of the unknown, design-based parameters, $\hat{\theta}$, one can solve $\sum_{i=1}^{N} \breve{U}_i(\theta) = 0$. Binder (1983) applied a standard delta method argument to provide a sandwich form for the estimated design-based variance, $V_\pi(\hat{\theta}) = A_\pi(\hat{\theta})^{-1} B_\pi(\hat{\theta}) \left[A_\pi(\hat{\theta})^{-1}\right]^T / N$. Following Lumley and Scott (2017),

$$A_\pi(\hat{\theta}) = \frac{1}{N}\sum_{i=1}^{N} \frac{\partial \breve{U}_i(\theta)}{\partial \theta}\bigg|_{\theta=\hat{\theta}} \tag{4.11}$$

and

$$B_\pi(\hat{\theta}) = \widehat{\text{var}}_\pi\left[\frac{1}{N}\sum_{i=1}^{N} \breve{U}_i(\hat{\theta})\right]. \tag{4.12}$$

In Figure 4.2, we provide sample code illustrating how to compute the sandwich estimator for a complex survey design with stratification, clustering, and unequal probability weighting. We first create a hypothetical survey design object, then fit the stage 1 and stage 2 models. Since `svytotal()` adds the weights, one must provide the unweighted matrix, $\tilde{U}(\hat{\theta})$, which can be accomplished by dividing the estimating functions by the weights.

For the stacked estimating equation, it is important to consider the strata arising from (1) original sampling procedure of the survey design and (2) membership to the calibration subset, if this wasn't one of the original strata by design. In this setting, one can augment the original design strata by cross-classifying the design strata with the subset indicator.

Bootstrap

Bootstrap variance estimation is often implemented in two-stage regression settings due to its simplicity. One step that is often overlooked is ensuring that bootstrap sampling is stratified on subset membership status. For those in the subset, $\{X_i^*, X_i^{**}, Z_i, Y_i\}$ is resampled. Otherwise, $\{X_i^*, Z_i, Y_i\}$ is resampled. The stage 1 model is then fit to the bootstrap sample to obtain a new estimate of the exposure, which is subsequently included in the stage 2 model fit to the bootstrap sample. Section C.2 in the Appendix includes steps for computing bootstrap standard errors using a stratified bootstrap procedure.

Multiple Imputation

In the variance estimation approach proposed by Baldoni et al. (2021), the expected value of the latent exposure variable is multiply imputed for all individuals in the main study by

repeatedly sampling the stage 1 model coefficients required to estimate $\hat{X}_i$. New calibration coefficients can be sampled using either their estimated asymptotic parametric distribution or bootstrap resampling. At each step of the imputation, the outcome model is re-fit using the newly calibrated values. Details on the steps required to use the resampling-based MI approach are provided in Section C.3 of the Appendix.

4.5. Simulation Study

4.5.1. Setup

We use a simulation study to compare the sandwich to competing variance estimators for the SRS and complex survey settings. We let the stage 1 model ($n = 450$) and stage 2 model ($N = 1000$ or $N = 10,000$) be linear and logistic regression models, respectively. We generate two covariates, X_i and Z_i, and vary the correlation between them. We simulate an error-prone covariate X_i^* and vary the error variance σ^2 between 0, 0.25, 0.50, and 1.00 to represent cases of zero, low, moderate, and high measurement error, respectively. We also conduct simulations to mimic complex survey data using code provided by Baldoni et al. (2021) for simulating a superpopulation and drawing samples using a stratified complex survey sampling scheme. Full simulation details are provided in Appendix Section C.4. In all tables, we present median percent (%) biases, median standard errors (ASE), median absolute deviation (MAD), and 95% coverage probabilities (CP). The ASE is the median of the estimated standard errors, while the MAD is the empirical median absolute deviation of the estimated regression coefficients.

4.5.2. Results

In Table 4.1, we present results for the stage 2 model fit to the true exposure X_i, the error-prone exposure X_i^*, and the calibrated exposure $\hat{X}_i$ when the data are simulated from a SRS. For the regression calibration approaches, we compute model-based (naive), sandwich, and bootstrap standard errors. Applying regression calibration reduces the absolute median bias to under 6% in all settings. For the larger samples ($N = 10,000$), using naive standard errors with regression calibration results in CP as low as 84%. Applying sandwich or bootstrap

techniques gives ASEs that more closely resemble the empirical MAD values and CP closer to 95%. In the high correlation and error settings, the bootstrap results in CP that is too high (> 97%), while the sandwich maintains the nominal coverage level (CP= 95%). Similar patterns were observed for simulations using a Cox proportional hazards model in stage 2, where the sandwich performed well compared to the naive ASEs and occasionally outperformed the bootstrap (Appendix Table C.1). In Table C.2 in the Appendix, we compute CP from confidence intervals constructed using the standard Wald procedure with bootstrap standard errors, percentiles of the bootstrap distribution, and the bias-corrected and accelerated (BCA) bootstrap approach. In high correlation and error settings where intervals constructed using bootstrap and sandwich standard errors resulted in CP that was too large, the percentile and BCA bootstrap present an opportunity to improve.

In Table 4.2, we present results for data simulated to be from a complex survey. Once again, the naive model-based standard errors are frequently too small, resulting in $CP < 95\%$. Standard error estimates obtained from the sandwich are generally better-behaved that those estimated from the resampling-based MI approach of Baldoni et al. (2021). The standard errors estimated from MI are oftentimes too large, resulting in CP as high as 99% for the large measurement error and high correlation case when $N = 1000$. These same issues of over-coverage were observed by Baldoni et al. (2021), which the authors attempted to mitigate by using robust estimators for the mean and standard deviation in their calculation of the adjusted variance. While these robust estimators did improve the estimated variances, they did not completely address the over-coverage issues.

4.6. Reanalysis of WHI and HCHS/SOL Data

4.6.1. WHI Data Example

To assess the association between energy, protein, and protein density with the risk of diabetes in the WHI study, we begin by developing our stage 1 models using data from the Nutritional Biomarker Study. The stage 1 model was fit to $n = 356$ sub-study participants as a linear regression of the biomarker value (X^{**}) on the corresponding self-reported value

(X^*) and covariates (Z). The stage 2 model included $N = 77,805$ eligible women and was fit as a Cox proportional hazards model. Following Tinker et al. (2011), we fit the stage 2 model with and without BMI. Details on our models and analytic cohort are provided in Section S5 of the Supplementary Materials.

Incident diabetes was reported in 4278 (5.5%) of the $77,805$ participants in the analytic cohort. Table 4.3 presents hazard ratio (HR) estimates and 95% confidence intervals (CI) for incident diabetes for a 20% increase in consumption of log-energy, log-protein, and log-protein density, as well as estimated β coefficients and standard errors (SEs). In the BMI-adjusted analysis, the HR for a 20% increase in calibrated log-energy intake was 1.54. The naive SE estimate on the log scale, 1.37, is over 40% smaller than the sandwich SE estimates, 2.34, resulting in a 95% CI for the HR for a 20% increase in log-energy intake with the naive SE of (0.94, 2.52), compared to (0.68, 3.51) using the sandwich. The bootstrap SE estimate for this model is 36.97, corresponding to a 95% CI of (0, 842640) for a 20% increase in log-energy consumption. We note that 47 (9.4%) of our 500 bootstrap replicates resulted in $|\hat{\beta}| > 10$ for log-energy, which is not a practical log-HR from an epidemiologic perspective. As an alternative, we present 95% CIs constructed using the percentile bootstrap interval, which is (0.04, 37.17) for a 20% increase in log-energy in the BMI-adjusted model. In this applied example, the sandwich offers a more believable standard error estimate of the regression parameter.

Differences between the sandwich and bootstrap were not as extreme for the non-BMI adjusted energy analysis. Similar patterns were observed for protein, where the HR for a 20% increase in the calibrated log-intake adjusted for BMI was 1.23, with 95% CIs of (1.08, 1.40) for the sandwich, 1.03, 1.46) for the normal bootstrap, and (1.11, 1.52) for the percentile bootstrap. In the BMI-adjusted models with log-protein density, the HR (95% CI) estimated by regression calibration with the naive SE is 1.64 (1.42, 1.88) compared to CIs of (1.14, 2.34), (0.83, 3.24), (1.31, 3.93) estimated by the sandwich, normal bootstrap, and percentile bootstrap, respectively. In this instance, using the bootstrap SE (1.91) resulted in a loss of

statistical significance, which did not occur with the sandwich SE (1.00) or the bootstrap percentile interval. We discuss the differences between our results and those reported by Tinker et al. (2011) in the Supplementary Materials.

4.6.2. HCHS/SOL Data Example

In the HCHS/SOL study, we fit the stage 1 models using data from $n = 310$ SOLNAS subset members by performing a linear regression of the biomarker for log-potassium (X^{**}) on log self-reported 24-hour recall measures (X^{*}) and other covariates (Z). We consider hypertension status and systolic blood pressure outcomes, which are studied using logistic and linear regression stage 2 models, respectively. For each model, the outcome was regressed on the calibrated dietary exposure while adjusting for confounders. All stage 2 outcome models were fit to a subset of $N = 8,176$ participants from the original cohort and accounted for the HCHS/SOL survey design. Additional details on our models and the selection of this subset are included in Section S6 of the Supplementary Materials.

Table 4.4 presents results from our re-analysis of the HCHS/SOL data. The estimated odds ratio (OR) of hypertension for a 20% increase in calibrated log-potassium (95% CI with naive SE) is 0.81 (0.63, 1.03). The naive SE estimate on the log scale of 0.68 is 15% and 44% smaller than the SEs estimated by the sandwich (0.80) and MI (1.22), respectively. Similar patterns were observed for the linear regression of systolic blood pressure on calibrated log-potassium, where the estimated coefficient for a 20% increase was -0.58, with 95% CIs of (-1.26, 0.10) with the naive SE, (-1.37, 0.21) for the sandwich, and (-1.78, 0.62) for MI. The SE estimated by MI, 3.35, was more inflated compared to the sandwich SE, 2.21, for this particular model.

4.7. Discussion

In this manuscript, we increase the practicality of the sandwich variance estimator for standard error estimation in two-stage regression settings, specifically regression calibration and compare it to competing variance estimation methods. We use two data examples to illustrate the straightforward use of the sandwich for data from a SRS and complex sur-

vey design. We have developed R code, which we hope makes computation of the sandwich more accessible and convenient. For obtaining the sandwich in R, one can modify the code provided throughout this paper or use functions from our code on GitHub at https://github.com/lboe23/sandwich2stage, which directly compute sandwich variance estimates for the two-stage setting of regression calibration.

Our numerical study indicated that when the sub-study is a large percentage of the main study, ignoring the uncertainty in the estimation of the stage 1 model parameters does not have a large impact, resulting in well-behaved “naive" standard errors. This point is also discussed by Keogh et al. (2020). We do encourage readers to adjust the naive model standard errors in all two-stage regression model settings, however, as we saw that the naive standard error will be too small in many instances.

In addition to the added computational burden, bootstrap confidence interval estimation can suffer from poor coverage if appropriate bias-adjustment procedures are not applied to departures from normality in the bootstrap estimates (Efron, 1987). Despite the well-known bias and coverage problems of the standard normal Wald bootstrap confidence interval, the BCA procedure is rarely applied in practice due to its complexity. In our data example from the WHI, the SE estimated by the bootstrap in the BMI-adjusted analysis for energy was quite unstable, resulting in extremely inflated 95% CIs for the calibrated energy intake exposure compared to the sandwich estimator. Nonetheless, the sandwich estimator also has some limitations, and occasionally showed coverage problems in settings with substantial measurement error and highly correlated covariates. The BCA bootstrap procedure may be the optimal approach for constructing 95% CIs. However, this technique can be computationally prohibitive in large cohorts like the WHI. In our complex survey simulations, standard errors calculated using MI resulted in overcoverage issues, which may be the result of instability caused by multicollinearity in the simulated data sets (Baldoni et al., 2021). We also observed some inflated standard errors estimated by MI compared to the sandwich in our data example from the HCHS/SOL study. These results suggest that there can occa-

sionally be instability in standard error estimates from the resampling-based bootstrap and MI procedures, which can be easily avoided if a more stable estimator like the sandwich is used for variance estimation.

This paper focused on regression calibration as a motivating example, but the issues discussed apply more broadly to any two-stage regression settings where a plug-in estimate is obtained in stage 1 and used in the outcome model fit in stage 2. Generally speaking, in two-stage regression settings where model-based standard errors are too small and standard errors obtained from resampling based approaches can be unstable, the sandwich variance estimation approach presents a well-behaved, less computationally-intensive alternative that is straightforward to implement. The sandwich may also be extended to multi-stage regression models. One related example in the WHI is a sandwich variance estimator used in regression calibration dietary applications when a third component to the stacked estimating equations is added for biomarker development (Prentice et al., 2021, 2022). Future work will look at extending the software for two-stage regression settings to these multi-stage model settings.

Figure 4.1: Code for obtaining the sandwich matrix using functions from the survey package for a simple random sample

(a) Code for fitting stage 1 and stage 2 models

```
sampdesign <- svydesign(id=~1, data=mydata)
stage1.model<-svyglm(xstarstar~xstar+z,design=sampdesign,
     family=gaussian(),subset=V==1)
sampdesign <- update(sampdesign,xhat =predict(stage1.model,
     newdata=sampdesign$variables) )
stage2.model<- svycoxph(Surv(Time, delta) ~ xhat+z,
     design=sampdesign)
```

(b) Code for obtaining $A(\hat{\theta})$

```
A.upperleft<- -solve(stage1.model$naive.cov)/N
A.bottomright<- -solve(stage2.model$naive.cov)/N
A.upperright<- matrix(0,nrow=j,ncol=k)
A.bottomleft<- stage2.alphas.estfuns(alphas.stage1)
A<-rbind(cbind(A.upperleft,A.upperright),
     cbind(A.bottomleft,A.bottomright))
A.inv<-solve(A)
```

(c) Code for obtaining $\tilde{U}(\hat{\theta})$

```
estfun.stage1<-matrix(0,nrow=N,ncol=j)
is.calibration <- !is.na(xstarstar)
estfun.stage1[is.calibration,] <- estfun(stage1.model)
estfun.stage2<-as.matrix(estfun(stage2.model))
estfun.all<-cbind(estfun.stage1,estfun.stage2)
```

(d) Code for obtaining $V(\hat{\theta})$

```
infl<- as.matrix(estfun.all)%*%t(A.inv)/N
sandwichvar<-vcov(svytotal(infl, sampdesign))
```

Figure 4.2: Code for obtaining the sandwich matrix using functions from the survey package for a complex survey design

(a) Code for fitting stage 1 and stage 2 models

```
sampdesign <- svydesign(id=~PSUid, strata=~strat,
              weights=~myweights, data=mydata)
stage1.model<-svyglm(xstarstar~xstar+z,design=sampdesign,
     family=gaussian(),subset=V==1)
sampdesign <- update(sampdesign,xhat =predict(stage1.model,
     newdata=sampdesign$variables) )
stage2.model<- svyglm(y ~ xhat+z ,
     design=sampdesign,family=binomial())
```

(b) Code for obtaining $A(\hat{\theta})$

```
A.upperleft<- -solve(stage1.model$naive.cov/
                 mean(stage1.model$prior.weights))/N
A.bottomright<- -solve(stage2.model$naive.cov/
                 mean(stage2.model$prior.weights))/N
A.upperright<- matrix(0,nrow=j,ncol=k)
A.bottomleft<- stage2.alphas.estfuns(alphas.stage1)
A<-rbind(cbind(A.upperleft,A.upperright),
     cbind(A.bottomleft,A.bottomright))
A.inv<-solve(A)
```

(c) Code for obtaining $\tilde{U}(\hat{\theta})$

```
estfun.stage1<-matrix(0,nrow=N,ncol=j)
is.calibration <- !is.na(xstarstar)
estfun.stage1[is.calibration,] <- estfun(stage1.model)/
   stage1.model$prior.weights
estfun.stage2<-estfun(stage2.model)/
     stage2.model$prior.weights
estfun.all<-cbind(estfun.stage1,estfun.stage2)
```

(d) Code for obtaining $V(\hat{\theta})$

```
infl<- as.matrix(estfun.all)%*%t(A.inv)/N
sandwichvar<-vcov(svytotal(infl, sampdesign))
```

Table 4.1: The median percent (%) bias, median standard errors (ASE), empirical median absolute deviation (MAD) and coverage probabilities (CP) for 1000 simulated data sets from a simple random sample for a logistic regression stage 2 model fit to true exposure, naive exposure, and calibrated exposure with naive (model-based) standard errors, sandwich standard errors, and bootstrap standard errors. We vary the correlation between X and Z, the sample size (N), and the measurement error variance (σ^2). Sample size of the calibration subset is $n = 450$.

			Low Correlation				High Correlation			
N	$\sigma^{2(1)}$	Method	% Bias	MAD	ASE	CP	% Bias	MAD	ASE	CP
1000	0.00	Truth	-0.04	0.07	0.07	0.94	0.16	0.09	0.10	0.94
	0.25	Naive	-13.78	0.12	0.11	0.91	-42.84	0.12	0.12	0.68
		RC$^{(2)}$ (Naive SE)	-3.76	0.13	0.12	0.94	-3.28	0.21	0.20	0.94
		RC$^{(2)}$ (Sandwich)	—	—	0.12	0.94	—	—	0.20	0.94
		RC$^{(2)}$ (Bootstrap)	—	—	0.12	0.94	—	—	0.21	0.94
	0.50	Naive	-48.54	0.08	0.08	0.36	-68.36	0.09	0.09	0.13
		RC$^{(2)}$ (Naive SE)	-4.61	0.16	0.16	0.93	-2.90	0.28	0.27	0.94
		RC$^{(2)}$ (Sandwich)	—	—	0.16	0.93	—	—	0.27	0.95
		RC$^{(2)}$ (Bootstrap)	—	—	0.16	0.94	—	—	0.29	0.96
	1.00	Naive	-71.90	0.06	0.06	0.01	-83.60	0.06	0.06	0.00
		RC$^{(2)}$ (Naive SE)	-5.31	0.22	0.21	0.94	-4.73	0.37	0.37	0.94
		RC$^{(2)}$ (Sandwich)	—	—	0.21	0.94	—	—	0.37	0.95
		RC$^{(2)}$ (Bootstrap)	—	—	0.22	0.96	—	—	0.41	0.98
10000	0.00	Truth	0.05	0.02	0.02	0.94	0.05	0.03	0.03	0.94
	0.25	Naive	-12.28	0.04	0.03	0.69	-41.87	0.04	0.04	0.01
		RC$^{(2)}$ (Naive SE)	-1.90	0.05	0.04	0.85	-1.15	0.08	0.06	0.89
		RC$^{(2)}$ (Sandwich)	—	—	0.05	0.94	—	—	0.08	0.95
		RC$^{(2)}$ (Bootstrap)	—	—	0.05	0.94	—	—	0.08	0.96
	0.50	Naive	-47.61	0.03	0.03	0.00	-67.45	0.03	0.03	0.00
		RC$^{(2)}$ (Naive SE)	-2.73	0.07	0.05	0.84	-1.68	0.11	0.08	0.89
		RC$^{(2)}$ (Sandwich)	—	—	0.06	0.94	—	—	0.10	0.95
		RC$^{(2)}$ (Bootstrap)	—	—	0.07	0.95	—	—	0.11	0.96
	1.00	Naive	-70.93	0.02	0.02	0.00	-82.74	0.02	0.02	0.00
		RC$^{(2)}$ (Naive SE)	-2.62	0.09	0.07	0.85	-1.61	0.15	0.12	0.88
		RC$^{(2)}$ (Sandwich)	—	—	0.09	0.94	—	—	0.14	0.95
		RC$^{(2)}$ (Bootstrap)	—	—	0.09	0.95	—	—	0.17	0.97

$^{(1)}$ σ^2 = the variance of the random, normally distributed measurement error
$^{(2)}$ RC = Regression calibration

Table 4.2: The median percent (%) bias, median standard errors (ASE), empirical median absolute deviation (MAD) and coverage probabilities (CP) for 1000 simulated data sets from a simple random sample for a logistic regression stage 2 model fit to true exposure, naive exposure, and calibrated exposure with naive (model-based) standard errors, sandwich standard errors, and MI standard errors. We vary the correlation between X and Z, the sample size (N), and the measurement error variance (σ^2). Sample size of the calibration subset is $n = 450$.

			Low Correlation				High Correlation			
N	$\sigma^{2(1)}$	Method	% Bias	MAD	ASE	CP	% Bias	MAD	ASE	CP
1000	0.00	Truth	2.02	0.10	0.10	0.94	3.43	0.13	0.13	0.93
	0.25	Naive	-11.59	0.15	0.15	0.94	-41.68	0.17	0.16	0.82
		RC$^{(2)}$ (Naive SE)	-1.26	0.17	0.16	0.94	-1.43	0.29	0.27	0.95
		RC$^{(2)}$ (Sandwich)	—	—	0.16	0.95	—	—	0.27	0.95
		RC$^{(2)}$ (MI)	—	—	0.17	0.95	—	—	0.29	0.96
	0.50	Naive	-46.80	0.12	0.11	0.62	-67.54	0.12	0.12	0.38
		RC$^{(2)}$ (Naive SE)	-3.93	0.22	0.21	0.94	-1.78	0.37	0.35	0.95
		RC$^{(2)}$ (Sandwich)	—	—	0.21	0.95	—	—	0.36	0.96
		RC$^{(2)}$ (MI)	—	—	0.22	0.97	—	—	0.40	0.98
	1.00	Naive	-70.57	0.09	0.08	0.08	-82.62	0.09	0.09	0.04
		RC$^{(2)}$ (Naive SE)	-1.12	0.29	0.28	0.94	-3.84	0.50	0.49	0.95
		RC$^{(2)}$ (Sandwich)	—	—	0.28	0.96	—	—	0.51	0.97
		RC$^{(2)}$ (MI)	—	—	0.31	0.98	—	—	0.60	0.99
10000	0.00	Truth	0.54	0.03	0.03	0.94	0.90	0.04	0.04	0.94
	0.25	Naive	-11.14	0.05	0.05	0.84	-41.34	0.05	0.05	0.10
		RC$^{(2)}$ (Naive SE)	-1.82	0.06	0.05	0.88	0.15	0.11	0.09	0.90
		RC$^{(2)}$ (Sandwich)	—	—	0.07	0.94	—	—	0.10	0.95
		RC$^{(2)}$ (MI)	—	—	0.07	0.96	—	—	0.11	0.96
	0.50	Naive	-47.31	0.04	0.04	0.00	-67.54	0.04	0.04	0.00
		RC$^{(2)}$ (Naive SE)	-2.85	0.08	0.07	0.86	-1.48	0.14	0.12	0.89
		RC$^{(2)}$ (Sandwich)	—	—	0.09	0.94	—	—	0.14	0.94
		RC$^{(2)}$ (MI)	—	—	0.09	0.95	—	—	0.16	0.96
	1.00	Naive	-71.18	0.03	0.03	0.00	-82.87	0.03	0.03	0.00
		RC$^{(2)}$ (Naive SE)	-3.67	0.12	0.09	0.86	-3.77	0.19	0.16	0.89
		RC$^{(2)}$ (Sandwich)	—	—	0.12	0.94	—	—	0.19	0.94
		RC$^{(2)}$ (MI)	—	—	0.13	0.95	—	—	0.24	0.96

$^{(1)}$ σ^2 = the variance of the random, normally distributed measurement error
$^{(2)}$ RC = Regression calibration

Table 4.3: WHI data analysis (N=77, 805) results from the Cox Proportional Hazards model for incident diabetes with dietary exposures of energy (kcal/d), protein (g/d), and protein density (% energy from protein). Results are shown for each stage 2 model fit to the calibrated exposure with naive (model-based) standard errors, sandwich standard errors, and bootstrap standard errors.

(a) Hazard ratio estimates and 95% confidence intervals (CI) for incident diabetes for a 20% increase in consumption of log-energy , log-protein, and log-protein density.

		HR (95% CI)	
Model[1]	Method	Adjusted for BMI	Not Adjusted for BMI
Log-Energy	RC[(2)] (Naive SE)	1.54 (0.94, 2.52)	2.90 (2.72, 3.08)
	RC[(2)] (Sandwich)	— (0.68, 3.51)	— (2.19, 3.83)
	RC[(2)] (Boot. - Wald)[(3)]	— (0, 842640)	— (2.18, 3.86)
	RC[(2)] (Boot. - Perc)[(4)]	— (0.04, 37.17)	— (2.29, 4.03)
Log-Protein	RC[(2)] (Naive SE)	1.23 (1.12, 1.34)	2.12 (2.00, 2.26)
	RC[(2)] (Sandwich)	— (1.08, 1.40)	— (1.60, 2.82)
	RC[(2)] (Boot. - Wald)[(3)]	— (1.03, 1.46)	— (1.59, 2.84)
	RC[(2)] (Boot. - Perc)[(4)]	— (1.11, 1.52)	— (1.60, 2.87)
Log-Protein Density	RC[(2)] (Naive SE)	1.64 (1.42, 1.88)	1.17 (1.02, 1.33)
	RC[(2)] (Sandwich)	— (1.14, 2.34)	— (0.25, 5.46)
	RC[(2)] (Boot. - Wald)[(3)]	— (0.83, 3.24)	— (0.19, 6.97)
	RC[(2)] (Boot. - Perc)[(4)]	— (1.31, 3.93)	— (0.28, 8.50)

(b) β regression parameter estimates and standard errors estimated by the Cox Proportional Hazards model for incident diabetes for log-energy, log-protein, and log-protein density.

		Adjusted for BMI		Not Adjusted for BMI	
Model[1]	Method	β	SE	β	SE
Log-Energy	RC[(2)] (Naive SE)	2.38	1.37	5.83	0.17
	RC[(2)] (Sandwich)	—	2.34	—	0.78
	RC[(2)] (Bootstrap)	—	36.97	—	0.80
Log-Protein (g/d)	RC[(2)] (Naive SE)	1.12	0.25	4.13	0.17
	RC[(2)] (Sandwich)	—	0.36	—	0.79
	RC[(2)] (Bootstrap)	—	0.48	—	0.82
Log-Protein Density	RC[(2)] (Naive SE)	2.70	0.39	0.84	0.38
	RC[(2)] (Sandwich)	—	1.00	—	4.32
	RC[(2)] (Bootstrap)	—	1.91	—	5.01

[(1)] Each model is adjusted for potential confounders and is stratified on age in 5-year categories, hormone therapy trial participation, and DM-C or OS membership
[(2)] RC = Regression calibration [(3)] Bootstrap with standard normal Wald-based confidence interval [(4)] Percentile bootstrap confidence interval

Table 4.4: HCHS/SOL data analysis ($N = 8,176$) results from the linear regression of baseline systolic blood pressure and the logistic regression of hypertension status each on log-transformed intake of potassium. Results are shown for each stage 2 model fit to the calibrated exposure with naive (model-based) standard errors, sandwich standard errors, and standard errors from the multiple imputation (MI) approach of Baldoni et al. (2021).

(a) Results for a 20% increase in consumption of log-potassium are presented. For linear regression, $\log(1.2)\hat{\beta}$ and 95% confidence intervals (CI) are presented. For logistic regression, odds ratio (OR) estimates of $\exp(\log(1.2)\hat{\beta})$ and 95% CIs are presented.

		Linear Regression	Logistic Regression
Model[1]	Method	β (95% CI)	OR (95% CI)
Log-Potassium	RC[(2)] (Naive SE)	-0.58 (-1.26, 0.10)	0.81 (0.63, 1.03)
	RC[(2)] (Sandwich)	— (-1.37, 0.21)	— (0.61, 1.07)
	RC[(2)] (MI)	— (-1.78, 0.62)	— (0.52, 1.25)

(b) β regression parameter estimates and standard errors estimated by linear and logistic regression models for hypertension-related outcomes for log-potassium.

		Linear Regression		Logistic Regression	
Model[1]	Method	β	SE	β	SE
Log-Potassium	RC[(2)] (Naive SE)	-3.17	1.90	-1.17	0.68
	RC[(2)] (Sandwich)	—	2.21	—	0.80
	RC[(2)] (MI)	—	3.35	—	1.22

[(1)] Each model is adjusted for potential confounders
[(2)] RC = Regression calibration

CHAPTER 5

Discussion

Large cohort studies can serve as a valuable tool for conducting epidemiologic research, such as identifying potential risk factors that drive disease incidence. However, these types of studies often have a mix of methodological issues, including measurement error that arises in complex exposures like self-reported dietary intake and routinely collected outcome data like self-reported disease status or outcomes derived from EHR. Since error-prone observations are increasingly being used in the statistical analyses of data from large cohort studies or those reliant on EHR data, it is crucial that methodology is developed that adjusts for errors in these variables so the resulting exposure-disease associations are estimated without bias. When more precise or gold standard data are available in addition to error-prone outcome data, then statistical methods can be applied to adjust study estimates to avoid bias induced by the error-prone data. Additionally, when there is gold or reference standard data available, the error-prone variables can be used as auxiliary data to augment the likelihood in the analysis of a time-to-event outcome. The primary reason for leveraging auxiliary, error-prone outcome data is to improve statistical efficiency. Reductions in the estimated variance of regression model parameters can be useful in settings where error is present in complex exposures and resulting variance estimates are quite large. Ultimately, efficiency improvements may even help drive cost reductions for future epidemiologic studies of a similar type. Improved variance estimation strategies also needed to be considered in settings like regression calibration, where an estimated exposure is used in place of an error-prone exposure in an outcome model to reduce bias. The standard errors of outcome model regression parameters must account for the uncertainty added by using the estimated exposure, but the most commonly used resampling-based methods for variance estimation can be numerically unstable, thus warranting the consideration of a better approach. Methods in the existing statistical literature do not adequately address errors in both exposure variables and time-to-event outcome variables, nor do they focus enough on improving variance

estimation in the presence of measurement error. In this dissertation, we have proposed several methods to reduce the bias resulting from error-prone data as well as improve variance estimation by leveraging auxiliary outcome data and considering improved strategies for obtaining standard error estimates of regression parameters in two-stage regression settings such as regression calibration.

In Chapter 2, we introduced a method that decreased the bias caused by measurement error in covariates and event classification variables in a discrete time-to-event setting. This method was applied to data from the Women's Health Initiative (WHI) study to evaluate the association between dietary energy and protein and incident diabetes. Through our simulations and data example, we show that we obtain vastly different hazard ratio estimates by adjusting for errors in both the outcome and covariates compared to using the heavily-biased naïve method that ignores all errors. Our method is straightforward to implement and offers a practical approach to achieving nearly unbiased estimates of exposure-disease associations for settings typical of those encountered in practice.

Chapter 3 presents a method that incorporates error-prone, auxiliary data into the analysis of an interval-censored time-to-event outcome. Our approach may be used when (1) a gold standard outcome is available on at least a subset of study participants and (2) auxiliary data for all participants are recorded at one or more fixed time points before or after the gold standard is scheduled to be recorded. By incorporating auxiliary outcome data that is correlated with a gold standard outcome into our analysis, we show that we can improve statistical efficiency in the estimation of exposure-disease associations. We further extend this method to accommodate data from a complex survey sampling design so that it can be applied to our motivating study, the Hispanic Community Health Study/Study of Latinos (HCHS/SOL). HCHS/SOL is a large, community-based cohort study of Hispanic/Latinos in the United States (n= 16,415) consisting of participants recruited using a complex survey design that includes unequal probability sampling, stratification, and clustering. We develop a weighted analogue of our likelihood function and a sandwich variance estimator so that our

method can handle data from a complex survey design. In HCHS/SOL, we are interested in assessing the association between energy, protein, and protein density dietary intakes and the risk of diabetes, when diabetes is recorded using both a self-reported outcome variable (auxiliary data) and a biomarker-based reference standard variable. To address our research question, we apply regression calibration to the proposed method to additionally accommodate covariate error. Our simulations and data example show that by incorporating auxiliary data into our analysis, rather than relying on standard approaches for interval censored survival data that do not leverage auxiliary data, we can drastically reduce the estimated variance.

We present a practical sandwich variance estimation approach for two-stage regression model settings in Chapter 4. We consider regression calibration as the motivating setting for two-stage modeling, where in stage 1, a plug-in estimate of the exposure is obtained, which is subsequently used in the stage 2 outcome model. In settings like regression calibration, model-based standard errors are typically too small, as they do not address the uncertainty added from using an estimated exposure. Thus, variance estimation procedures like the popular bootstrap, or, alternatively, the sandwich variance estimator, are needed to account for this extra uncertainty. In this chapter, we propose the sandwich variance estimator as a more stable, less computationally-intensive variance estimator, which may be used for a large class of problems and for data from either a simple random sample or a complex survey design. We have shown through a numerical study that the sandwich may have advantages over the bootstrap for the simple random sample and the multiple imputation-based approach of Baldoni et al. (2021) for complex survey design settings in terms of achieving nominal 95% coverage probability. Using data examples from the WHI and the HCHS/SOL, we compare the sandwich variance estimator to these alternative estimators to show that the sandwich may often provide a more stable estimate of the standard error. The R code provided throughout this chapter illustrates that it is quite straightforward to compute the sandwich variance estimator using the stacked estimating equation approach from existing functions in the R software. Our expectation is that the code included in this chapter, as

well as the package `sandwich2stage` on GitHub, will make sandwich variance computation more convenient to apply and can be used in many two stage regression analyses, including regression calibration-based applications in nutritional epidemiology.

This dissertation has inspired several areas of future research. The methods proposed in Chapters 2 and 3 both rely on the assumption that, conditional on the true disease status at each visit, each of the error-prone outcomes are independent. We deemed this assumption reasonable for our motivating data examples of interest in this dissertation, since the error-prone, self-reported binary disease status variables in both the WHI and HCHS/SOL are recorded using a questionnaire and outcomes far enough apart in time that many random factors may impact a participant's response to the question. Further, this conditional independence assumption is commonly made in the discrete survival analysis literature where error-prone outcomes are obtained via periodic follow-up (Balasubramanian and Lagakos, 2001, 2003; Meier et al., 2003; Magaret, 2008; Gu et al., 2015). Nonetheless, we note that this may not always be a practical assumption to make in real data settings, and future work will consider a more complex outcome error model that relaxes this assumption. In particular, one could consider letting the sensitivity and specificity of the error-prone outcomes depend on known covariates or previous responses.

Another related extension for the methods in Chapters 2 and 3 relates to the assumption that the sensitivity and specificity of the error-prone outcome data are known constants. For studies like the WHI and the HCHS/SOL with a mix of analytical complications, it may be useful to investigate the possibility of estimating the sensitivity and specificity alongside the regression parameters of interest. In doing so, it would be important to explore how robust the models for sensitivity and specificity are to model misspecification and any potential identifiability issues that may result from estimating these measures of accuracy. Further steps would be required in the variance estimation stage to account for the uncertainty of the sensitivity and specificity estimation. This could be accomplished using techniques similar to the ones discussed throughout this dissertation for addressing the uncertainty added by the

estimation of the nuisance parameters in the calibration model when regression calibration is applied to correct for exposure error. In particular, one could use a stacked estimating equation approach, such as an extension of the technique outlined in Chapter 4, where the equations for estimating the sensitivity and specificity are included as additional components.

Additional areas of future work relate to the extension of the discrete time framework from Chapters 2 and 3. In particular, work is ongoing to extend our estimation methods for interval-censored data from the discrete time setting to the continuous failure time setting which accommodates an outcome error framework. Another potential avenue for future research relates to the extension of our discrete time setup in such a way that the number of possible visit times is not restricted. In our current framework for Chapters 2 and 3, the number of parameters to be estimated can increase with the number of subjects in our analysis if each subjects has a different set of visit times. To avoid any numerical instability, in the data examples, we were required to make the limiting assumptions that a common set of visit times were observed for all participants. While this was a reasonable assumption in the WHI and HCHS/SOL data sets where the observed visit times were, in fact, quite close to the anniversary date for each subject, it is possible that other cohort studies with prospective follow-up may not adhere to this pattern. Thus, we believe a useful extension of our work will consider techniques for the stable estimation of more parameters.

APPENDIX A

Supplementary Material for Chapter 2

A.1. R Code with illustrative example

In this section, we provide an illustrative example showing how we can use R code to apply the proposed method. For this example, we will use a simulated data set with one error-prone covariate (X_1^*) and 4 precisely-recorded covariates (Z_1, Z_2, Z_3, Z_4). Additionally, we have periodic follow-up from 4 visits. The sensitivity, specificity, and negative predictive value of the error-prone outcome are assumed to be 0.60, 0.98, and 0.95, respectively. This data set can be found on GitHub at https://github.com/lboe23/Outcome-Error-RC under the file name Simulated_Data_Example_5Cov_Long.csv.

Before we begin our analysis, we need to load the Rcpp functions that we need to compute the likelihood and the function for the variance calculation. We use the following Rcpp functions, which were developed by Gu et al. (2015) and can be found in the icensmis package on Cran: "loglikC," "gradlikC," "dmat" and "getrids." The variance calculation function can be found on the GitHub site listed above under the file name "Variance_functions.R."

```
Rcpp::sourceCpp('RcppFunc.cpp')

source('Variance_functions.R')
```

As suggested by the name of our data set, the input data is in long form, where each row represents one time point and each subject has multiple rows. Below we read our data into R and then present the first 6 rows of the data.

```
#Read in the simulated data
data_long<-read.csv("Simulated_Data_Example_5Cov_Long.csv",header=TRUE,
sep=",",check.names = FALSE)
head(data_long)
```

```
##   ID subset_ind    x_1_star x_1_starstar      z_1        z_2 z_3 z_4 t y
## 1  1          1 -0.12175921     7.661711 4.168057 -12.385553   0   0 1 0
## 2  1          1 -0.12175921     7.661711 4.168057 -12.385553   0   0 2 0
## 3  1          1 -0.12175921     7.661711 4.168057 -12.385553   0   0 3 1
## 4  2          1  0.02266679     7.576149 1.873360  -7.022348   1   0 1 0
## 5  2          1  0.02266679     7.576149 1.873360  -7.022348   1   0 2 0
## 6  2          1  0.02266679     7.576149 1.873360  -7.022348   1   0 3 0
```

Our input dataset consists of the following variables:

- ID: a unique ID for each subject in the data set.
- $subset\_ind$: an indicator variable representing membership in the calibration subset, which takes the value 1 if the subject is a member of the calibration subset and 0 if they are not. In this data example, $n_C = 500$ of the $N = 10,000$ total study subjects are in the calibration subset.
- x_1_star: the error-prone covariate of interest, prone to both systematic and random error (e.g. self-reported measure of dietary energy).
- $x\_1\_starstar$: the covariate of interest subject to classical measurement error (e.g. biomarker of dietary energy), which is only available for members of the validation subset (those with $subset\_ind = 1$).
- z_1, z_2: precisely measured continuous covariates that we wish to include in our calibration and outcome models.

- z_3, z_4: precisely measured binary covariates that we wish to include in our calibration and outcome models.
- y: the error-prone, binary result where 1 indicates a positive test and 0 indicates a negative test.
- t: the visit time corresponding to each error-prone test result, y.

To apply the proposed method, we will begin by fitting the calibration equations. First, we create a new dataset that only has one row per subject and only includes the members of the calibration subset.

```
unique_data<-data_long[!duplicated(data_long$ID),]
calibration_data<-unique_data[which(unique_data$subset_ind==1),]
```

Next, we will fit the calibration model by regressing the covariate measure with classical measurement error, X_1^{**} on the covariate with prone to more extreme error, X^*, and other covariates, Z_1, Z_2, Z_3, and Z_4. We note that the model below corresponds to equation 2.6 of in main manuscript: $X_i^{**} = \delta_{(0)} + \delta_{(1)} X_i^* + \delta_{(2)} Z_i + V_i$.

```
#Perform calibration - get elements needed to correct for exposure error
calibration_eq<-lm(x_1_starstar~x_1_star+z_1+z_2+z_3+z_4,data=calibration_data)
```

We will now save the summary data from the calibration equation and use this to create our multivariate correction factor from equation 2.8 of the main manuscript, which recall has the following form:

$$\hat{\Delta} = \begin{bmatrix} \hat{\delta}_{(1)p\times p} & \hat{\delta}_{(2)p\times q} \\ 0_{q\times p} & I_{q\times q} \end{bmatrix}.$$

```
calibration_summary<-summary(calibration_eq)
nbeta<-length(calibration_summary$coefficients[-1,1])
Delta_Mat<-rbind(as.matrix(t(calibration_summary$coefficients[-1,1])),
cbind(matrix(0,nrow=nbeta-1,ncol=1),diag(nbeta-1)))
```

Next, we want to save the elements of the variance-covariance matrix from the calibration equation, as this will be used later in the computation of the variance-covariance matrix Σ for $\hat{\beta}$. Note that we do not need the elements of the variance-covariance matrix that correspond to the intercept term for this approach.

```
covla<-vcov(calibration_summary)
covla_fin<-covla[-c(1),-c(1)]
```

We will now begin the process of fitting our outcome models. As we did for the WHI data example in the text, we will consider 3 approaches: (1) the naive method ignoring error in the outcome and covariate, (2) the regression calibration method that corrects for error in the covariate only, and (3) the proposed method. First, let's assign sensitivity, specificty, and negative predictive value.

```
sensitivity<-0.60
specificity<-0.98
negpred<-0.95
```

We will now fit our first outcome model, which corresponds to the naive approach. To estimate regression coefficients for the naive grouped continuous time Cox proportional hazards model, we will fit a generalized linear model with a binomial outcome and assume a complementary log-log link.

```
data_long$t<-as.factor(data_long$t)
fit_naive<-glm(y~t+x_1_star+z_1+z_2+z_3+z_4,family=binomial(link="cloglog"),
data=data_long)
fitsum_naive<-summary(fit_naive)
ntest<-length(unique(data_long$t))
param0b<-fitsum_naive$coefficients[-c(1:ntest),1]
```

Now we will get our data in the format required to use the proposed method. First, we will define the formula that we want to use for our outcome model.

```
formula=y~x_1_star+z_1+z_2+z_3+z_4
```

Now, let's make sure our data is ordered properly before we begin calculating sum of the likelihood components.

```
initsurv = 0.1
id <- eval(substitute(data_long$ID), data_long, parent.frame())
time <- eval(substitute(t), data_long, parent.frame())
y <- eval(substitute(y), data_long, parent.frame())
ord <- order(id, time)

if (is.unsorted(ord)) {
  id <- id[ord]
  time <- time[ord]
  y <- y[ord]
  data <- data[ord, ]}
utime <- sort(unique(time))
timen0 <- (time != 0)
```

Now that our data is in an appropriate form, we can calculate the D matrix, defined in Section 2.3.1 of the main manuscript. Additionally, we calculate J and the number of rows in D.

```
Dm <- dmat(id[timen0], time[timen0], y[timen0], sensitivity,
           specificity, negpred)
J <- ncol(Dm) - 1
nsub <- nrow(Dm)
```

As we get ready to maximize our log-likelihood, we want to think of starting values for our survival parameters. To avoid maximization problems due to the ordered constraint of the survival parameters $1 = S_1 > S_2 > ... > S_{J+1} > 0$, we re-parameterize these terms for optimization. The re-parameterization that we use is a log-log transformation of survival function for S_2, and a change in log-log of the survival function for all other parameters. We consider initial values of 0.1 for our survival parameters, then transform these based on this re-parameterization. Additionally, we define a lower bound of $-\infty$ for the first re-parameterized survival function and 0 for the subsequent $J - 1$ terms.

```
initsurv <- 0.1
lami <- log(-log(seq(1, initsurv, length.out = J +1)[-1]))
lami <- c(lami[1], diff(lami))
tosurv <- function(x) exp(-exp(cumsum(x)))
lowlam <- c(-Inf, rep(0, J - 1))
```

Next, we want to create a matrix version of our covariate data which will be used in the maximization of the log-likelihood.

```
Xmat <- model.matrix(formula, data = data_long)[, -1, drop = F]
beta.nm <- colnames(Xmat)
uid <- getrids(id, nsub)
Xmat <- Xmat[uid, , drop = F]
```

We will now maximize our log-likelihood function that corrects for outcome error only using the "L-BFGS-B" method in the optim function. We will give the lower bound *lowlam* defined above for our survival function parameter estimates and a lower bound of ∞ for our regression coefficient estimates. We will use the *lami* values defined above as our initial values for our baseline survival functions. We will use the estimated regression parameters from the naive method as our starting values for β_{X1}, β_{Z1}, β_{Z2}, β_{Z3}, and β_{Z4} in the proposed method.

```
parmi <- c(lami, param0b)
loglikeoptimize <- optim(parmi, loglikC, gradlikC,
lower = c(lowlam, rep(-Inf, nbeta)), Dm = Dm, Xmat = Xmat,
method = "L-BFGS-B", hessian = T)
```

We can now invert the Hessian matrix to calculate $\hat{\Sigma}_{\beta*}$.

```
cov <- as.matrix(solve(loglikeoptimize$hessian)[-(1:J), -(1:J)])
rownames(cov) <- colnames(cov) <- beta.nm
beta_fit <- loglikeoptimize$par[-(1:J)]
```

It is finally time to apply the proposed method. Below, we calculate our corrected vector of estimated regression coefficients of interest, using equation 2.7 from the main manuscript: $\hat{\beta} = \hat{\beta}^* \hat{\Delta}^{-1}$. Recall that we computed $\hat{\Delta}$ above using regression calibration.

```
corrected_beta<-t(as.matrix(beta_fit))%*%solve(Delta_Mat)
corrected_beta<-as.data.frame(t(corrected_beta))
rownames(corrected_beta) <-beta.nm
```

Lastly, we compute the variance for the proposed approach. To do this, we use the function "Proposed_Var" from the Variance_functions.R file that we imported above. This code for the variance calculation can accommodate 1 error-prone covariate and up to 19 precisely-measured covariates, for a total of 20 covariates in the calibration and outcome models. The input values for this function, in order, are the following: (1) $\hat{\Sigma}_{\beta*}$, the variance-covariance matrix from the method that corrects for outcome error only; (2) the variance-covariance matrix from the calibration model; (3) the estimated multivariate correction factor from regression calibration, $\hat{\Delta}$; and (4) the estimated regression parameters obtained by fitting the model that corrects for outcome error only.

```
corrected_beta_var<-Proposed_Var(cov,covla_fin,Delta_Mat,beta_fit)
SDBeta<-sqrt(corrected_beta_var[1,1])
```

Now, to complete our results table, we will use regression calibration to obtain the results for the method that corrects for covariate error only:

```
corrected_beta_x<-t(as.matrix(param0b))%*%solve(Delta_Mat)
corrected_beta_x<-as.data.frame(t(corrected_beta_x))
cov_cloglog<-(vcov(fitsum_naive)[-c(1:ntest),-c(1:ntest)])
corrected_beta_var_x<-Proposed_Var(cov_cloglog,covla_fin,Delta_Mat,param0b)
SDBeta_x<-sqrt(corrected_beta_var_x[1,1])
```

The last step is to exponentiate our regression parameters and corresponding confidence interval bounds and put them into a table so that we can present the results for all three methods simultaneously. The final results are presented below:

```r
#Put results in table
corrected_HR_myMethod<-cbind(HR=exp(corrected_beta[1,]),
Lower=exp(corrected_beta[1,]-qnorm(0.975)*SDBeta),
Upper=exp(corrected_beta[1,]+qnorm(0.975)*SDBeta))
naive_HR<-cbind(HR=exp(fitsum_naive$coefficients[c("x_1_star"),1]),
Lower=exp(fitsum_naive$coefficients[c("x_1_star"),1]-
qnorm(0.975)*fitsum_naive$coefficients[c("x_1_star"),2]),
Upper=exp(fitsum_naive$coefficients[c("x_1_star"),1]+
qnorm(0.975)*fitsum_naive$coefficients[c("x_1_star"),2]))
corrected_HR_x<-cbind(HR=exp(corrected_beta_x[1,]),
Lower=exp(corrected_beta_x[1,]-qnorm(0.975)*SDBeta_x),
Upper=exp(corrected_beta_x[1,]+qnorm(0.975)*SDBeta_x))
all_results<-rbind(naive_HR,corrected_HR_x,corrected_HR_myMethod)
rownames(all_results)<-c("Naive","Regression Calibration","Proposed")
round(all_results,3)

##                            HR Lower Upper
## Naive                   0.958 0.900 1.021

## Regression Calibration 0.492 0.163 1.479
## Proposed               0.403 0.104 1.560
```

A.2. Derivation of equation (2)

In this section, we follow the notation and logic of Balasubramanian and Lagakos (2003) to show how we derive the likelihood contribution for subject i in equation (2) from equation

(1). The steps are as follows:

$$\begin{aligned}
f(\mathbf{Y_i}, \mathbf{t_i}, n_i) &= \sum_{j=1}^{J+1} \Pr(\tau_{j-1} < T_i \leqslant \tau_j) \Pr(\mathbf{Y_i}, \mathbf{t_i}, n_i | T_i), \\
&= \sum_{j=1}^{J+1} \theta_j \prod_{l=1}^{n_i} \Pr(t_{il}, Y_{il} | t_{i1}, t_{i2}, \ldots, t_{il-1}, Y_{i1}, Y_{i2}, \ldots, Y_{il-1}, T_i) \\
&\quad \times \Pr(n_i | t_{i1}, t_{i2}, \ldots, t_{in_i}, Y_{i1}, Y_{i2}, \ldots, Y_{in_i}, T_i) \\
&= \sum_{j=1}^{J+1} \theta_j \prod_{l=1}^{n_i} \Pr(t_{il} | t_{i1}, t_{i2}, \ldots, t_{il-1}, Y_{i1}, Y_{i2}, \ldots, Y_{il-1}, T_i) \\
&\quad \times \prod_{l=1}^{n_i} \Pr(Y_{il} | t_{i1}, t_{i2}, \ldots, t_{il-1}, t_{il}, Y_{i1}, Y_{i2}, \ldots, Y_{il-1}, T_i) \\
&\quad \times \Pr(n_i | t_{i1}, t_{i2}, \ldots, t_{in_i}, Y_{i1}, Y_{i2}, \ldots, Y_{in_i}, T_i)
\end{aligned}$$

where $\theta_j = \Pr(\tau_{j-1} < T_i \leqslant \tau_j)$.

Now, by the assumption that $\Pr(\mathbf{Y_i} | T_i, \mathbf{t_i}) = \prod_{l=1}^{n_i} \Pr(Y_{il} | T_i, t_{il})$:

$$\begin{aligned}
f(\mathbf{Y_i}, \mathbf{t_i}, n_i) &= \sum_{j=1}^{J+1} \theta_j \prod_{l=1}^{n_i} \Pr(t_{il} | t_{i1}, t_{i2}, \ldots, t_{il-1}, Y_{i1}, Y_{i2}, \ldots, Y_{il-1}, T_i) \times \prod_{l=1}^{n_i} \Pr(Y_{il} | T_i, t_{il}) \\
&\quad \times \Pr(n_i | t_{i1}, t_{i2}, \ldots, t_{in_i}, Y_{i1}, Y_{i2}, \ldots, Y_{in_i}, T_i)
\end{aligned}$$

Finally, following Balasubramanian and Lagakos, Balasubramanian and Lagakos (2003) for the case of a prespecified visit schedule, we have the following:

$$\begin{aligned}
&\Pr(t_{il} | t_{i1}, t_{i2}, \ldots, t_{il-1}, Y_{i1}, Y_{i2}, \ldots, Y_{il-1}, T_i) \\
&= \Pr(n_i | t_{i1}, t_{i2}, \ldots, t_{in_i}, Y_{i1}, Y_{i2}, \ldots, Y_{in_i}, T_i) = 1
\end{aligned}$$

Note, these two probabilities would also drop out of the likelihood if they did not depend on the parameters of interest (β). Now, we arrive at equation (2) for the likelihood contribution for the ith subject:

$$f(\mathbf{Y_i}, \mathbf{t_i}, n_i) \quad = \quad \sum_{j=1}^{J+1} \theta_j \prod_{l=1}^{n_i} \Pr(Y_{il} | \tau_{j-1} < T_i \leqslant \tau_j, t_l) = \sum_{j=1}^{J+1} \theta_j C_{ij}$$

where $C_{ij} = \prod_{l=1}^{n_i} \Pr(Y_{il} | \tau_{j-1} < T_i \leqslant \tau_j, t_l)$.

A.3. Regularity conditions

In this section, we outline sufficient regularity conditions for the proposed estimator, namely asymptotic normality and $\sqrt{N}$-convergence. Recall that we have an approximate estimator that has empirically been observed to have good properties, i.e. have minimal bias and close to nominal coverage, when the event of interest is rare and the true parameter value is of moderate size.

First assume that we have discrete observation times for the failure time that satisfy the following: $0 = \tau_0 < \tau_1 < \tau_2 < ... < \tau_J < \tau_{J+1} = \infty$. Further, define the elements of t_i, the vector of visit times for subject i, to be a subset of $\{\tau_0, \tau_1, \tau_2, ..., \tau_J\}$. Recall that we define $S_j = \Pr(T > \tau_{j-1})$ for $j = 1, ..., J+1$; and $\tau_0 = 0$, and and require that $1 = S_1 > S_2 > ... > S_{J+1} > 0$. The previous two conditions ensure that $0 < \theta_j < 1$ for $j = 1, ..., J$, where $\theta_j = \Pr(\tau_{j-1} < T \leqslant \tau_j)$. Now, assume the following: (1) $\{X_i, X_i^*, Z_i, T_i, Y_i, t_i\}$, $i = 1, ..., N$ are independent and identically distributed, where N is the number of subjects in the main study data; and (2) $\frac{n_C}{N} \to p \in (0, 1)$, where n_C is the number of subjects in the calibration subset.

Assume that $\{X_i, Z_i, Y_i, t_i\}$, for $i = 1, ..., N$ follows the density $f(\mathbf{X_i}, \mathbf{Z_i}, \mathbf{Y_i}, \mathbf{t_i}; \psi^0)$ with the corresponding log-likelihood function $l(\psi) = l(X_i, Z_i, Y_i, t_i; \psi)$; where $\psi = [\beta, \mathbf{S}]$, $\beta = (\beta_X, \beta_Z)^T$, $\mathbf{S} = (S_1, S_2, ..., S_{J+1})^T$, and $l(\psi)$ is as defined in equation (3) of the main text, i.e.

$$l(\psi) = l(\mathbf{S}, \beta) = \sum_{i=1}^{N} \log \left(\sum_{j=1}^{J+1} D_{ij} S_j^{\exp(x_i^T \beta_X + z_i^T \beta_Z)} \right). \tag{A.1}$$

Here, ψ^0 is the vector of regression parameters of interest for the likelihood with the unobserved true data for X. Assume the log-likelihood is twice continuously differentiable and define $l^\star(\psi) = l(X_i^*, Z_i, Y_i, t_i; \psi)$. Let the partially naive score function be denoted $U_N^*(\psi) = (1/N)\partial l^\star(\psi)/\partial\psi$, and let $\hat{\psi}_N^*$ to be the solution to the score equations, $U_N^*(\psi) = 0$. Define ψ^* to be the vector of parameters that solves $E\left[\frac{\partial l(X^*,Z,Y,t;\psi)}{\partial\psi}\right] = E[U^*(\psi)] = 0$. ψ^* will not generally be equal to ψ^0, since the partially naive likelihood does not adjust for the covariate error in X^*. Because $\hat{\psi}_N^*$ is a maximum likelihood estimator (MLE), we can rely on standard regularity conditions to see that with probability going to one as $N \to \infty$, there exists a unique solution to the likelihood equations, $\hat{\psi}_N^*$, that is consistent for ψ^* (Foutz, 1977) and asymptotically normal.(Boos and Stefanski, 2013) Under these regularity conditions, one has

$$\sqrt{N}(\hat{\psi}_N^* - \psi^*) \xrightarrow{d} \mathcal{N}(0, I(\psi^*)^{-1}), \tag{A.2}$$

where $I(\psi^*)^{-1}$ is the Fisher information matrix.

Recall that the proposed estimator $\hat{\beta}$ is defined as $\hat{\beta}^*\hat{\Delta}^{-1}$, where $\hat{\beta}^*$ is the first $p+q$ elements of the vector $\hat{\psi}_N^*$. Since $\hat{\Delta}$ is a linear regression estimator, we can also appeal to standard MLE theory to establish its consistency for the true parameter Δ^0 and asymptotic normality. Finally, we need only satisfy the necessary regularity conditions for the multivariate delta method to establish consistency and asymptotic normality of our proposed estimator. In addition to the established asymptotic normality of $(\hat{\beta}, \hat{\Delta})$, for $g(\beta, \Delta) = \beta\Delta^{-1}$, we need only that its matrix of partial order derivatives be continuous in a neighborhood of (β^*, Δ^0).(Casella and Berger, 2002) Further assume the independence of $\hat{\beta}$ and $\hat{\Delta}$, which holds if the number of subjects in the calibration subset, n_C, is a small percentage of the main study sample size, N. Then, we have:

$$\sqrt{N}\left(\hat{\beta}^*\hat{\Delta}^{-1} - \beta^*(\Delta^0)^{-1}\right) \xrightarrow{d} \mathcal{N}(0, \Sigma), \tag{A.3}$$

where the $(j_1, j_2)^{th}$ element of Σ is defined as $\Sigma_\beta(j_1, j_2) \cong \left(A'\Sigma_{\beta^*}A\right)_{j_1,j_2} + \beta^*\Sigma_{A,j_1,j_2}\beta^{*\prime}$,

with $\Sigma_{\beta*}$ the asymptotic variance of $\hat{\beta}$, and $A = \Delta^{-1}$ and Σ_{A,j_1,j_2} defined similarly as in the main text.

The numerical performance of our proposed estimator has been studied extensively and shown to perform well empirically, as described in the main manuscript. Under these standard regularity conditions, we have illustrated the asymptotic normality of our estimator in the context of an error-prone time-to-event outcome and covariate.

A.4. Supplemental methods and discussion for Women's Health Initiative data example

We adopted exclusion criteria in order to obtain a final analytic data set for our analyses that approximated that used by Tinker et al. Tinker et al. (2011) Applying these exclusion criteria resulted in approximately the same cohort. We excluded anyone who reported diabetes at baseline or during the first year of follow-up for the comparison arm of the WHI Dietary Modification trial (DM-C) participants ($n = 724$) or the first three years of follow-up for the WHI Observational Study (OS) participants ($n = 4109$). We attempted to align characteristics of participants in the DM-C trial with those of participants in the OS by excluding the following participants in the OS: those who had breast, colorectal, or other cancer within 10 years prior to enrollment ($n = 8677$), stroke or myocardial infarction within 6 months prior to enrollment ($n = 155$), body mass index (BMI) < 18 ($n = 678$), hypertension (systolic blood pressure > 200 or diastolic blood pressure > 105)($n = 244$), reported daily energy intake of < 600 kcal or > 5000 kcal ($n = 3571$), $\geqslant 10$ meals prepared away from home each week ($n = 3598$), a special low-fiber diet ($n = 568$), a special malabsorption-related diet ($n = 514$), inadvertent weight loss of > 15 pounds within 6 months of enrollment ($n = 594$), and reported diabetes diagnosis before age 21 at enrollment ($n = 95$). Applying these exclusion criteria and selecting only the participants with no missing data in the calibration and outcome model variables, we arrived at our analytic cohort with 65,358 members. Of these 65,358 participants, 12,121 (18.5%) were from the DM-C and 53,237 (81.5%) were from the OS.

We note that our HRs for the case of correcting for covariate error were substantially different than those originally reported by Tinker et al. Tinker et al. (2011) For example, Tinker et al. (2011) reports that a HR (95% CI) of 2.41 (2.06, 2.82) was associated with a 20% increase in energy intake when BMI was omitted from the outcome model, compared to our 1.421 (1.043, 1.938). There were several differences between these analyses that may have led to this, including slightly different data sets. We reanalyzed our data using a continuous Cox model and found results that were very consistent results with our discrete analysis, so the discrete approach did not explain this difference. First, we investigated the potential discrepancies in results that might arise from the choice of strata. In our original analysis, we stratified our models on age in 10-year categories and DM-C or OS cohort membership, which resulted in 6 strata. We used a continuous Cox model to assess how our results changed when we expanded our strata to (1) age in 5-year categories and DM-C or OS cohort membership (12 strata) or (2) age in 5-year categories, DM-C or OS cohort membership, and hormone therapy trial arm (active estrogen, estrogen placebo, active estrogen plus progestin, estrogen plus progestin placebo, and not randomized) for participants in the DM-C who were also on the hormone trials (36 strata). Table A.7 compares our original results using the discrete proportional hazards model and correcting for covariate error to the results using the continuous time Cox proportional hazards model and allowing for either the 6, 12, or 36 strata described above. When we used a Cox model and applied the post-hoc regression calibration approach to correct for covariate error, we obtained the following HR (95% CI) for a 20% increase in energy intake when the model did not adjust for BMI: 6 strata, 1.333 (0.993, 1.790); 12 strata, 1.334 (0.994, 1.791); 36 strata, 1.328 (0.990, 1.780). Note that these results are fairly consistent with those obtained for the discrete model correcting for covariate error only (HR 1.421; 95% CI 1.043, 1.938). Furthermore, we see that our results were not sensitive to the choice of strata.

One important difference between analyses is that we aligned the covariates between the outcome and calibration models, but Tinker et al. (2011) did not. This alignment is necessary for our approach and in general is recommended for regression calibration in order to

avoid potential sources of bias.(Kipnis et al., 2009) We used a continuous model and a traditional regression calibration approach (non-post-hoc) to show how the results that correct for covariate error only might differ based on the following: BMI is in (1) both the calibration and outcome model, (2) neither model, or (3) the calibration model only. The latter case is an example of not aligning the calibration and outcome model and is not possible for our post-hoc approach used for correcting covariate error. Results comparing these different alignment strategies are presented in table A.8. This table presents HR estimates and 95% confidence intervals for the discrete analysis with the post-hoc correction for covariate error, the continuous Cox model analysis with the post-hoc correction for covariate error, and the continuous Cox model analysis with the non-post-hoc traditional regression calibration correction for covariate error. For analyses that include BMI in both the calibration and outcome models, we obtain similar results for all three approaches for energy, protein, and protein density, indicating that the choice of a discrete analysis or a post-hoc correction does not substantially change our answer. The same is true for analyses that include BMI in neither the calibration nor the outcome model. As we saw in the main manuscript, adjusting for BMI can qualitatively change our answer for methods that adjust for covariate error only, particularly for energy intake. The results from table A.8 suggest that our results can change even more dramatically if we include BMI in the calibration model but exclude it from the outcome model. As an example, we see that this analysis approach results in a HR (95% CI) for a 20% increase in energy intake of 2.768 (2.279, 3.362), suggesting a much stronger association between intake and diabetes than seen previously. The results for protein and protein density intake also change substantially when BMI is included in the calibration model only.

Lastly, we were able to get similar results to Tinker et al. (2011) by adopting a similar analysis approach and adding glycemic load, a covariate that was not in our calibration model, to our outcome model. In this case, the HR (95% CI) for a 20% increase in energy intake was 2.803 (2.314, 3.397) in the continuous model not adjusted for BMI. Finally, we note that discrepancies between results from our proposed approach and those of Tinker et al.

(2011) also stem from the fact that we have both corrected for outcome error and allowed for an imperfect specificity at baseline.

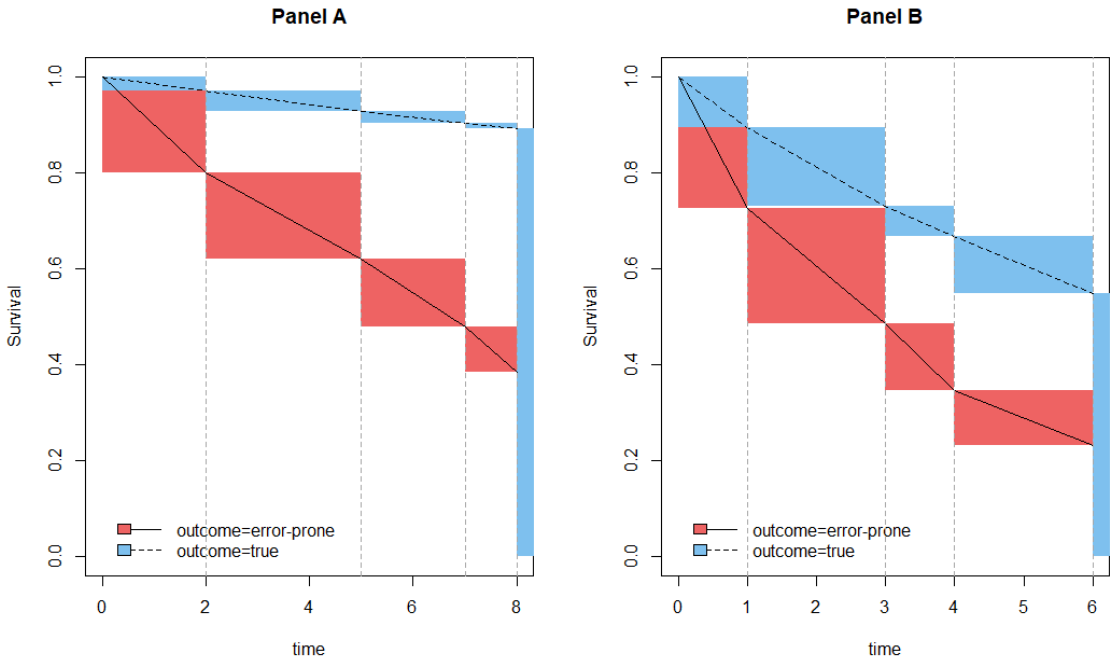

Figure A.1: Estimated nonparametric maximum likelihood estimators (NPMLEs) of the survival distribution for the error-prone outcomes compared to true outcomes for the simulation study, fit using the R package 'interval.'(Fay and Shaw, 2010) Panel A corresponds to censoring rate = 0.90 (baseline hazard = 0.012) with observation times $(2, 5, 7, 8)$. Panel B corresponds to censoring rate = 0.55 (baseline hazard = 0.094) with observation times $(1, 3, 4, 6)$. Vertical lines represent observation times. Simulated from data with $\beta_{X1} = \log(1.5)$, $\beta_{Z1} = \log(0.7)$, $\beta_{Z2} = \log(1.3)$, $e \sim \mathcal{N}(0, 1.31)$, sensitivity = 0.90, and specificity = 0.80.

Table A.1: The mean percent (%) biases, average standard errors (ASE), empirical standard errors (ESE) and coverage probabilities (CP) are given for 1000 simulated data sets for the proposed method, the naive method, a method that corrects for covariate error only, and a method that corrects for outcome error only, with $\beta_{X1} = \log(1.5)$, $\beta_{Z1} = \log(0.7)$, and $\beta_{Z2} = \log(1.3)$; e is normally distributed with mean zero; Sensitivity (Se)=0.80; Specificity (Sp)=0.90.

			Proposed				Naive			
$\hat{\delta}_{(1)}$[1]	CR[2]	β	% Bias	ASE	ESE	CP	% Bias	ASE	ESE	CP
0.60	0.90	β_{X1}	1.616	0.200	0.204	0.950	−88.03	0.046	0.046	0.000
		β_{Z1}	−1.094	0.143	0.142	0.945	−79.22	0.057	0.058	0.002
		β_{Z2}	−3.731	0.143	0.143	0.945	−84.07	0.057	0.054	0.021
	0.55	β_{X1}	−1.231	0.093	0.094	0.949	−68.11	0.038	0.038	0.000
		β_{Z1}	−1.055	0.067	0.066	0.958	−43.46	0.047	0.046	0.079
		β_{Z2}	−3.018	0.066	0.065	0.957	−53.48	0.046	0.045	0.133
0.30	0.90	β_{X1}	1.840	0.283	0.286	0.954	−93.88	0.033	0.033	0.000
		β_{Z1}	−1.233	0.151	0.151	0.947	−82.46	0.054	0.055	0.001
		β_{Z2}	−4.212	0.151	0.150	0.945	−79.74	0.054	0.052	0.025
	0.55	β_{X1}	−2.246	0.131	0.133	0.940	−84.02	0.027	0.027	0.000
		β_{Z1}	−1.967	0.071	0.069	0.951	−52.48	0.045	0.044	0.008
		β_{Z2}	−3.899	0.070	0.068	0.956	−42.08	0.045	0.044	0.306
			Correct Covariate Error				Correct Outcome Error			
$\hat{\delta}_{(1)}$	CR	β	% Bias	ASE	ESE	CP	% Bias	ASE	ESE	CP
0.60	0.90	β_{X1}	−80.15	0.077	0.077	0.015	−38.84	0.120	0.122	0.722
		β_{Z1}	−80.62	0.055	0.056	0.001	6.019	0.146	0.148	0.944
		β_{Z2}	−82.19	0.055	0.053	0.022	−13.31	0.146	0.146	0.936
	0.55	β_{X1}	−47.05	0.064	0.063	0.168	−40.51	0.054	0.056	0.151
		β_{Z1}	−47.15	0.046	0.045	0.042	5.840	0.067	0.068	0.942
		β_{Z2}	−48.47	0.046	0.044	0.192	−12.37	0.066	0.066	0.919
0.30	0.90	β_{X1}	−79.95	0.109	0.108	0.150	−69.05	0.085	0.086	0.109
		β_{Z1}	−80.62	0.058	0.058	0.003	−10.77	0.140	0.141	0.928
		β_{Z2}	−82.31	0.058	0.056	0.030	8.916	0.140	0.140	0.947
	0.55	β_{X1}	−47.49	0.091	0.089	0.419	−70.28	0.038	0.040	0.000
		β_{Z1}	−47.53	0.049	0.047	0.059	−11.21	0.064	0.064	0.892
		β_{Z2}	−48.82	0.048	0.046	0.231	8.673	0.064	0.064	0.938
			Truth							
	CR	β	% Bias	ASE	ESE	CP				
	0.90	β_{X1}	1.038	0.107	0.108	0.951				
		β_{Z1}	2.495	0.107	0.106	0.948				
		β_{Z2}	2.444	0.107	0.108	0.948				
	0.55	β_{X1}	0.517	0.052	0.054	0.942				
		β_{Z1}	1.471	0.052	0.053	0.951				
		β_{Z2}	1.773	0.052	0.052	0.948				

[1] $\hat{\delta}_{(1)}$ = Estimate of attenuation coefficient [2] CR = True censoring rate

Table A.2: The mean percent (%) biases, average standard errors (ASE), empirical standard errors (ESE) and coverage probabilities (CP) are given for 1000 simulated data sets for the proposed method and naive method with $\beta_{X1} = \log(1.5)$, $\beta_{Z1} = \log(0.7)$, and $\beta_{Z2} = \log(1.3)$; e is normally distributed with mean zero. The censoring rate is fixed at 0.90. Here, we vary sensitivity, specificity, and negative predictive value.

Se[1]$= 0.80, Sp$[2]$= 0.90$			Proposed				Naive			
$\hat{\delta}_{(1)}$[3]	η [4]	β	% Bias	ASE	ESE	CP	% Bias	ASE	ESE	CP
0.60	0.98	β_{X1}	3.401	0.219	0.215	0.957	−88.00	0.046	0.047	0.000
		β_{Z1}	3.644	0.157	0.157	0.958	−79.01	0.056	0.058	0.001
		β_{Z2}	1.458	0.156	0.160	0.946	−83.77	0.056	0.058	0.028
	0.90	β_{X1}	5.902	0.270	0.275	0.951	−90.03	0.043	0.044	0.000
		β_{Z1}	4.952	0.194	0.196	0.946	−82.51	0.053	0.054	0.000
		β_{Z2}	−0.967	0.191	0.199	0.947	−86.74	0.053	0.054	0.015
0.30	0.98	β_{X1}	3.994	0.311	0.300	0.960	−93.97	0.033	0.033	0.000
		β_{Z1}	3.631	0.167	0.164	0.956	−82.33	0.054	0.055	0.000
		β_{Z2}	0.826	0.165	0.169	0.945	−79.35	0.054	0.056	0.032
	0.90	β_{X1}	7.238	0.384	0.383	0.963	−95.03	0.031	0.031	0.000
		β_{Z1}	5.193	0.206	0.206	0.947	−85.29	0.051	0.052	0.000
		β_{Z2}	−2.018	0.203	0.210	0.950	−83.03	0.051	0.052	0.018
$Se = 0.90, Sp = 0.80$			Proposed				Naive			
$\hat{\delta}_{(1)}$	η	β	% Bias	ASE	ESE	CP	% Bias	ASE	ESE	CP
0.60	0.98	β_{X1}	2.100	0.231	0.224	0.960	−93.58	0.037	0.037	0.000
		β_{Z1}	3.910	0.166	0.163	0.957	−88.63	0.046	0.046	0.000
		β_{Z2}	3.037	0.164	0.167	0.949	−90.37	0.045	0.046	0.000
	0.90	β_{X1}	3.617	0.283	0.285	0.956	−94.44	0.036	0.036	0.000
		β_{Z1}	5.001	0.203	0.207	0.939	−89.96	0.045	0.045	0.000
		β_{Z2}	1.572	0.200	0.205	0.955	−91.46	0.044	0.045	0.000
0.30	0.98	β_{X1}	1.873	0.327	0.316	0.965	−96.87	0.026	0.027	0.000
		β_{Z1}	3.749	0.175	0.171	0.954	−90.47	0.044	0.044	0.000
		β_{Z2}	2.754	0.173	0.175	0.954	−87.93	0.044	0.044	0.000
	0.90	β_{X1}	3.600	0.401	0.399	0.957	−97.33	0.026	0.026	0.000
		β_{Z1}	5.030	0.215	0.216	0.942	−91.58	0.043	0.044	0.000
		β_{Z2}	1.134	0.212	0.216	0.953	−89.30	0.043	0.043	0.000
$Se = 1, Sp = 1, \eta = 1$			Truth							
		β	% Bias	ASE	ESE	CP				
		β_1	1.962	0.108	0.112	0.936				
		β_2	2.568	0.107	0.109	0.944				
		β_3	1.115	0.107	0.108	0.945				

[1] Se = Sensitivity [2] Sp = Specificity [3] $\hat{\delta}_{(1)}$ = Estimate of the attenuation coefficient
[4] η = Negative predictive value

Table A.3: The mean percent (%) biases, average standard errors (ASE), empirical standard errors (ESE) and coverage probabilities (CP) are given for 1000 simulated data sets for the proposed method and naive method with $\beta_{X1} = \log(1.5)$, $\beta_{Z1} = \log(0.7)$, and $\beta_{Z2} = \log(1.3)$; e is normally distributed with mean zero. The censoring rate is fixed at 0.90. Here, we vary sensitivity, specificity, and probability of missingness at each visit.

Se[1]$= 0.80, Sp$[2]$= 0.90$			Proposed				Naive			
$\hat{\delta}_{(1)}$[3]	P_{Miss}[4]	β	% Bias	ASE	ESE	CP	% Bias	ASE	ESE	CP
0.60	0.10	β_{X1}	−0.416	0.206	0.205	0.952	−87.80	0.048	0.049	0.000
		β_{Z1}	−0.271	0.148	0.152	0.943	−77.60	0.059	0.062	0.004
		β_{Z2}	−2.974	0.148	0.154	0.945	−80.46	0.059	0.061	0.052
	0.40	β_{X1}	−0.031	0.243	0.244	0.955	−85.19	0.056	0.057	0.000
		β_{Z1}	0.579	0.173	0.177	0.940	−73.40	0.068	0.071	0.034
		β_{Z2}	−3.283	0.173	0.180	0.942	−76.70	0.068	0.071	0.168
0.30	0.10	β_{X1}	−1.732	0.292	0.292	0.952	−94.06	0.034	0.034	0.000
		β_{Z1}	−0.774	0.156	0.160	0.954	−81.08	0.056	0.059	0.001
		β_{Z2}	−2.701	0.156	0.162	0.941	−75.82	0.057	0.059	0.063
	0.40	β_{X1}	−1.297	0.344	0.347	0.954	−92.82	0.040	0.040	0.000
		β_{Z1}	0.036	0.183	0.187	0.946	−77.63	0.066	0.069	0.015
		β_{Z2}	−3.146	0.183	0.190	0.945	−71.04	0.066	0.069	0.190
$Se = 0.90, Sp = 0.80$			Proposed				Naive			
$\hat{\delta}_{(1)}$	P_{Miss}	β	% Bias	ASE	ESE	CP	% Bias	ASE	ESE	CP
0.60	0.10	β_{X1}	−1.920	0.218	0.216	0.957	−93.38	0.039	0.037	0.000
		β_{Z1}	−0.451	0.156	0.163	0.949	−87.87	0.047	0.048	0.000
		β_{Z2}	−2.801	0.156	0.164	0.941	−89.04	0.047	0.048	0.000
	0.40	β_{X1}	−2.470	0.264	0.268	0.958	−91.23	0.044	0.044	0.000
		β_{Z1}	−0.637	0.189	0.200	0.944	−84.86	0.054	0.056	0.000
		β_{Z2}	−0.796	0.189	0.200	0.946	−86.60	0.054	0.055	0.012
0.30	0.10	β_{X1}	−3.134	0.308	0.309	0.953	−96.85	0.028	0.026	0.000
		β_{Z1}	−1.012	0.165	0.171	0.953	−89.80	0.045	0.046	0.000
		β_{Z2}	−2.576	0.165	0.174	0.944	−86.46	0.045	0.046	0.000
	0.40	β_{X1}	−4.353	0.374	0.384	0.955	−95.77	0.032	0.032	0.000
		β_{Z1}	−1.383	0.200	0.211	0.944	−87.38	0.052	0.054	0.000
		β_{Z2}	−0.452	0.200	0.213	0.945	−83.22	0.052	0.053	0.015
$Se = 1, Sp = 1, \eta = 1$			Truth							
	P_{Miss}	β	% Bias	ASE	ESE	CP				
	0.10	β_1	0.848	0.108	0.109	0.948				
		β_2	1.458	0.108	0.115	0.925				
		β_3	−1.882	0.108	0.111	0.949				
	0.40	β_1	1.362	0.114	0.117	0.943				
		β_2	1.824	0.114	0.123	0.929				
		β_3	−1.327	0.114	0.117	0.956				

[1] Se = Sensitivity [2] Sp = Specificity [3] $\hat{\delta}_{(1)}$ = Estimate of the attenuation coefficient
[4] P_{Miss} = Probability of missingness at each visit

Table A.4: The mean percent (%) biases, average standard errors (ASE), empirical standard errors (ESE) and coverage probabilities (CP) are given for 1000 simulated data sets for the proposed method, naive method, method that corrects for covariate error only, and method that corrects for outcome error only for a simulated dataset with similar features to the Women's Health Initiative (WHI) data. Here, Sensitivity (Se)=0.61, Specificity (Sp)=0.995, Negative Predictive Value (η) = 0.96, $\beta_{X1} = \log(1.5)$, $\beta_{Z1} = \log(0.7)$, $\beta_{Z2} = \log(1.3)$, e is normally distributed with mean zero, and the censoring rate for the error-prone indicator is fixed at 0.95.

		Proposed				Naive			
$\hat{\delta}_{(1)}$[1]	β	% Bias	ASE	ESE	CP	% Bias	ASE	ESE	CP
0.60	β_{X1}	−0.294	0.057	0.058	0.943	−79.21	0.012	0.013	0.000
	β_{Z1}	0.316	0.041	0.042	0.939	−62.94	0.015	0.015	0.000
	β_{Z2}	−0.578	0.040	0.041	0.941	−68.75	0.015	0.015	0.000
0.30	β_{X1}	−0.186	0.082	0.084	0.950	−89.53	0.009	0.009	0.000
	β_{Z1}	0.366	0.044	0.046	0.940	−68.63	0.014	0.015	0.000
	β_{Z2}	−0.786	0.044	0.044	0.941	−61.04	0.014	0.015	0.000
		Correct Covariate Error				Correct Outcome Error			
$\hat{\delta}_{(1)}$	β	% Bias	ASE	ESE	CP	% Bias	ASE	ESE	CP
0.60	β_{X1}	−65.51	0.022	0.022	0.000	−39.92	0.032	0.032	0.000
	β_{Z1}	−65.26	0.015	0.016	0.000	7.018	0.039	0.040	0.902
	β_{Z2}	−65.55	0.015	0.016	0.000	−9.828	0.039	0.039	0.896
0.30	β_{X1}	−65.48	0.031	0.032	0.000	−69.73	0.023	0.023	0.000
	β_{Z1}	−65.25	0.017	0.017	0.000	−9.444	0.038	0.038	0.853
	β_{Z2}	−65.61	0.017	0.017	0.000	12.408	0.038	0.037	0.867
		Truth							
	β	% Bias	ASE	ESE	CP				
	β_{X1}	−0.208	0.022	0.022	0.961				
	β_{Z1}	0.026	0.022	0.023	0.952				
	β_{Z2}	0.200	0.022	0.022	0.947				

[1] $\hat{\delta}_{(1)}$ = Estimate of attenuation coefficient

Table A.5: Hazard Ratio (HR) and 95% confidence interval (CI) estimates of incident diabetes for a 20% increase in consumption of energy (kcal/d), protein (g/d), and protein density (% energy from protein/d) based on the naive method ignoring error in the outcome and covariate, the method corrected for error in the covariate only, and the proposed method. Here, sensitivity = 0.61, specificity = 0.995, and negative predictive value = 1.

		HR (95% CI)	
Model[1]	Method	Adjusted for BMI[2]	Not Adjusted for BMI
Energy (kcal/d)	Naive	1.002 (0.986, 1.018)	1.024 (1.008, 1.040)
	Regression Calibration	1.041 (0.758, 1.429)	1.421 (1.043, 1.938)
	Proposed	0.973 (0.714, 1.327)	1.314 (0.992, 1.740)
Protein (g/d)	Naive	1.024 (1.010, 1.039)	1.051 (1.035, 1.066)
	Regression Calibration	1.121 (1.036, 1.213)	1.231 (1.130, 1.342)
	Proposed	1.107 (1.025, 1.195)	1.229 (1.128, 1.339)
Protein Density	Naive	1.100 (1.064, 1.137)	1.128 (1.091, 1.167)
	Regression Calibration	1.243 (1.125, 1.374)	1.325 (1.181, 1.486)
	Proposed	1.209 (1.100, 1.329)	1.327 (1.183, 1.490)

[1] Each model is adjusted for potential confounders and is stratified on age (10-year categories) and Dietary Modification trial (DM) or Observational Study (OS) cohort membership. [2] BMI = Body Mass Index (kg/m^2)

Table A.6: Sensitivity Analysis varying sensitivity and specificity of diabetes self-reports across WHI DM-C and WHI OS participants. We consider separate models for dietary energy, protein, and protein density. Each model is adjusted for potential confounders, including BMI, and is stratified on age (10-year categories) and DM or OS cohort membership. We show HR estimates of incident diabetes for a 20% increase in consumption of energy (kcal/d), protein (g/d), and protein density (% energy from protein/d).

Sensitivity		Specificity		HR (95% CI)		
OS	DM	OS	DM	Energy (kcal/d)	Protein (g/d)	Protein Density
0.5800	0.7418	0.9945	0.9972	0.970 (0.713,1.319)	1.113 (1.030,1.202)	1.193 (1.088,1.307)
0.5300	0.9614	0.9945	0.9972	0.954 (0.699,1.302)	1.114 (1.031,1.203)	1.183 (1.081,1.295)
0.6168	0.5800	0.9945	0.9972	0.938 (0.690,1.276)	1.108 (1.027,1.195)	1.199 (1.093,1.314)
0.6282	0.5300	0.9945	0.9972	0.959 (0.700,1.313)	1.106 (1.025,1.194)	1.183 (1.081,1.293)
0.5800	0.7418	0.9951	0.9945	0.974 (0.715,1.326)	1.108 (1.026,1.196)	1.206 (1.099,1.324)
0.5300	0.9614	0.9951	0.9945	0.971 (0.710,1.327)	1.110 (1.028,1.199)	1.193 (1.088,1.308)
0.6168	0.5800	0.9951	0.9945	0.972 (0.709,1.333)	1.105 (1.025,1.191)	1.173 (1.074,1.282)
0.6282	0.5300	0.9951	0.9945	0.960 (0.705,1.306)	1.108 (1.027,1.195)	1.189 (1.087,1.302)

Table A.7: Sensitivity Analysis for different stratification strategies using a modeling approach similar to that of Tinker et al. Tinker et al. (2011) We examine hazard ratio (HR) and 95% confidence interval (CI) estimates of incident diabetes for a 20% increase in consumption of energy (kcal/d), protein (g/d), and protein density (% energy from protein/d) based on discrete proportional hazards analyses and continuous Cox proportional hazards models that correct for error in the covariate (X) only.

		HR (95% CI)	
Model[1]	Method	Adjusted for BMI[2]	Not Adjusted for BMI
Energy (kcal/d)	Discrete, 6 Strata[3]	1.041 (0.758, 1.429)	1.421 (1.043, 1.938)
	Continuous, 6 Strata	0.953 (0.686, 1.323)	1.333 (0.993, 1.790)
	Continuous, 12 Strata[4]	0.953 (0.686, 1.324)	1.334 (0.994, 1.791)
	Continuous, 36 Strata[5]	0.952 (0.685, 1.322)	1.328 (0.990, 1.780)
Protein (g/d)	Discrete, 6 Strata	1.121 (1.036, 1.213)	1.231 (1.130, 1.342)
	Continuous, 6 Strata	1.104 (1.020, 1.194)	1.217 (1.117, 1.325)
	Continuous, 12 Strata	1.103 (1.020, 1.193)	1.216 (1.117, 1.324)
	Continuous, 36 Strata	1.104 (1.021, 1.194)	1.215 (1.116, 1.323)
Protein Density	Discrete, 6 Strata	1.243 (1.125, 1.374)	1.325 (1.181, 1.486)
	Continuous, 6 Strata	1.241 (1.121, 1.374)	1.325 (1.179, 1.489)
	Continuous, 12 Strata	1.241 (1.122, 1.374)	1.324 (1.179, 1.487)
	Continuous, 36 Strata	1.243 (1.123, 1.377)	1.324 (1.179, 1.487)

[1] Each model is adjusted for potential confounders [2] BMI = Body Mass Index (kg/m^2)
[3] 6 strata: age (10-year categories) and Dietary Modification trial (DM) or Observational Study (OS) cohort membership [4] 12 strata: age (5-year categories) and Dietary Modification trial (DM) or Observational Study (OS) cohort membership
[5] 36 strata: age (5-year categories), Dietary Modification trial (DM) or Observational Study (OS) cohort membership, and hormone therapy trial arm.

Table A.8: For model used by Tinker et al., Tinker et al. (2011), we examine the sensitivity of results to choices of how BMI is treated in analyses. We present hazard ratio (HR) and 95% confidence interval (CI) estimates of incident diabetes for a 20% increase in consumption of energy (kcal/d), protein (g/d), and protein density (% energy from protein/d) based on the discrete proportional hazards model with a post-hoc correction for covariate error, the continuous Cox model with a post-hoc correction for covariate error, and the continuous Cox model with the non-post-hoc traditional regression calibration correction for covariate error.

			HR (95% CI)	
Model[1]	Method	BMI in Both[2]	BMI in Neither[3]	Calibration Only[4]
Energy	Disc. PHoc[5]	1.041 (0.758, 1.429)	1.421 (1.043, 1.938)	NA
(kcal/d)	Cont. PHoc[6]	0.953 (0.686, 1.323)	1.333 (0.993, 1.790)	NA
	Cont. Non-PHoc[7]	0.956 (0.650, 1.407)	1.290 (0.967, 1.720)	2.768 (2.279, 3.362)
Protein	Disc. PHoc[5]	1.121 (1.036, 1.213)	1.231 (1.130, 1.342)	NA
(g/d)	Cont. PHoc[6]	1.104 (1.020, 1.194)	1.217 (1.117, 1.325)	NA
	Cont. Non-PHoc[7]	1.099 (1.009, 1.196)	1.208 (1.095, 1.333)	1.790 (1.430, 2.242)
Protein	Disc. PHoc[5]	1.243 (1.125, 1.374)	1.325 (1.181, 1.486)	NA
Density	Cont. PHoc[6]	1.241 (1.121, 1.374)	1.325 (1.179, 1.489)	NA
	Cont. Non-PHoc[7]	1.226 (1.111, 1.352)	1.303 (1.161, 1.463)	1.049 (0.689, 1.597)

[1] Each model is adjusted for potential confounders and stratified on age (10-year categories) and Dietary Modification trial (DM) or Observational Study (OS) cohort membership.
[2] BMI included in both calibration and outcome model (BMI = Body Mass Index (kg/m^2))
[3] BMI included in neither the calibration nor the outcome model
[4] BMI included in the calibration model but not the outcome model
[5] Discrete time model using post-hoc regression calibration
[6] Continuous time model using post-hoc regression calibration
[7] Continuous time model using non-post-hoc regression calibration

APPENDIX B

Supplementary Material for Chapter 3

B.1. R Code with sample data analysis

We now provide R code that illustrates how to apply the proposed method and the standard, no auxiliary data method to a simulated data set. The simulated data mimics the complex survey design of HCHS/SOL that includes unequal probability sampling, stratification, and clustering. To mimic the measurement error of dietary factors in HCHS/SOL, we simulate an error-prone covariate (X^*) and assume that we additionally have 2 error-free continuous covariates, such as body mass index and age (Z_1 and Z_2). The auxiliary data outcome variable is recorded at 8 time points, while the gold standard outcome variable is recorded at year 4. The sensitivity and specificity of the error-prone, auxiliary data outcome are assumed to be 0.61 and 0.98, respectively. This data set is provided on GitHub at https://github.com/lboe23/AugmentedLikelihood with the file name SampleData.RData.

We begin by loading in the functions required to calculate our log-likelihood and gradient. These functions are available on the GitHub site above. This file contains two functions, (1) log_like_proposed() which calculates the log-likelihood for the proposed method and (2) gradient_proposed() which calculates the gradient/estimating function for the proposed method. Both functions require a specification of the function purpose, where the options are "SUM" or "INDIVIDUAL." icensmis package on Cran or on GitHub at https://github.com/XiangdongGu/icensmis/blob/master/src/dataproc.cpp. Below is code that loads all of the required functions:

```
source('PROPOSED_AUGMENTEDLIKELIHOOD_FUNCTIONS.R')
Rcpp::sourceCpp('RcppFunc.cpp')
Rcpp::sourceCpp('dataproc.cpp')
```

Now we assign the sensitivity (Se) and specificity (Sp) values for the auxiliary data. We assume these are known, fixed constants in our analysis. We will also allow for a proportion of

the gold (reference) standard event status variables to be missing, and we fix this missingness rate to be 20% for this analysis.

```
sensitivity<-0.61
specificity<-0.98
prop_m<-0.20
```

Now, we load in sample simulated data. The data we input is in wide form, with one row per subject and each auxiliary data event indicator in a separate column.

```
load(file=paste0('SampleData.RData'))
N<-dim(samp)[1]
```

We now fit the calibration model. Later, for variance estimation in the presence of regression calibration and a complex survey design, we will use the parametric multiple imputation procedure proposed by Baldoni et al. (2021). To do so, we need to save off the estimated calibration model regression coefficients, corresponding estimated covariance matrix, and the design matrix. Finally, we will create our predicted values (xhat) from regression calibration using the "predict" statement.

```
samp.solnas <- samp[(solnas==T),]
lm.lsodi <- glm(xstarstar ~  xstar+z1+z2,data=samp.solnas)
x.lsodi <- model.matrix(lm.lsodi) #X
xtx.lsodi <- t(x.lsodi)%*%x.lsodi #X`X`
ixtx.lsodi <- solve(xtx.lsodi) #(X`X)^-1
samp[,xhat := predict(lm.lsodi,newdata=samp,'response')]
```

We now convert the data to long form, where each row represents one time point and each subject has multiple rows. Recall that each subject in our simulated data set has 8 visits. Then, we are going to sort the data by subject ID.

```
samp_long1<-reshape(data = samp, idvar = "ID", varying =
list(true_result=
c("true_result_1", "true_result_2","true_result_3",
```

```
  "true_result_4", "true_result_5", "true_result_6",
  "true_result_7","true_result_8"),
  result=c("result_1","result_2","result_3",
  "result_4","result_5","result_6",
  "result_7","result_8")), direction="long",
                      v.names = c("true_result","result"),sep="_")

#order long dataset by each subject's ID
  samp_long <- samp_long1[order(samp_long1$ID),]
```

Next, we create a data set with only one row per subject using the duplicated function and applying it to the ID variable. Then, using the data with one row per subject, we save the vector of sampling weights that will be used in the weighted analysis. Additionally, we create a keep statement and apply a function called "after_first_pos" to removes all auxiliary data values of "1" after the first positive from the data in long form.

```
  GS_data<-samp_long[!duplicated(samp_long$ID),c("ID","GS_vis4")]

#Save vector of weights for this dataset
  weights<-as.numeric(unlist(samp_long[!duplicated(samp_long$ID),
  c("bghhsub_s2")]))

keep<-unlist(tapply(samp_long$result,samp_long$ID,after_first_pos))
 datafinal_1<-samp_long[keep,]
```

Suppose we want to simulate missingness in the gold (reference) standard indicator variable, Δ_i. To do so, we first set a seed so that our results are reproducible. Then, we generate N variables called *mcar* from a Uniform(0,1) distribution and let Δ_i be missing for each subject if $mcar < MR$.

```
  set.seed(2548)
  mcar<-runif(N,0,1)
  GS_data$GS_vis4_mis<-ifelse(mcar<prop_m,NA,GS_data$GS_vis4)
```

Next, we create two datasets using the merge function: datafinal, which is the data in long form with one row per visit, and datafinal_GS, which has one row per person for the standard, no auxiliary data analysis.

```
datafinal<-merge(datafinal_1,GS_data,by="ID")
datafinal_GS<-merge(samp_long,GS_data[,c("ID","GS_vis4_mis")],
  by="ID")
```

Now, we write down the formula for our outcome model including the error-prone auxiliary data outcome and three covariates. Recall that we used regression calibration to correct for error in one covariate, so our model includes the predicted value xhat and two precisely recorded covariates, z1 and z2.

```
formula=result~xhat+z1+z2
```

We will now make sure our data is ordered properly before we begin calculating sum of the likelihood components.

```
id <- eval(substitute(datafinal$ID), datafinal, parent.frame())
time <-eval(substitute(datafinal$time), datafinal, parent.frame())
result <- eval(substitute(result), datafinal, parent.frame())
ord <- order(id, time)
if (is.unsorted(ord)) {
  id <- id[ord]
  time <- time[ord]
  result <- result[ord]
  datafinal <- datafinal[ord, ]}
utime <- sort(unique(time))
timen0 <- (time != 0)
```

Next, we will calculate the D matrix and C matrix for our log-likelihood using the Rcpp functions. Additionally, we assign J, the number of auxiliary data visit times.

```
Dm <- dmat(id[timen0], time[timen0], result[timen0], sensitivity,
```

```
            specificity, 1)
Cm <- cmat(id[timen0], time[timen0], result[timen0], sensitivity,
            specificity, 1)
J <- ncol(Dm) - 1
```

In these next steps, we create our covariate matrix with one column per covariate (Xmat). Intially, Xmat will be in long form, with one row per visit. We also assign nbeta (the number of covariates/regression parameters) and uid, a unique indicator for each person's ID. Finally, we redefine our covariate matrix to have just one row per subject.

```
Xmat <- model.matrix(formula, data = datafinal)[, -1, drop = F]
beta.nm <- colnames(Xmat)
nbeta <- ncol(Xmat)
uid <- getrids(id, N)
Xmat <- Xmat[uid, , drop = F]
```

Now, create a unique vector (GSdelta) with only one row per person indicating whether each person had the gold (reference) standard indicator available or not. This will be used to calculate the proposed log-likelihood contribution for each subject based on whether GSdelta=NA, 0 or 1. Additionally, we create the vector of observation times at which the gold (reference) standard is recorded, called GSVis. Lastly, we create a vector of 1's called "noweights" which will be used to fit the proposed estimator in the unweighted analysis.

```
GSdelta <- datafinal[uid,"GS_vis4_mis"]
GSVis<-rep(4,N)
noweights<-rep(1,N)
```

We now finalize the data set for the standard, no auxiliary data analysis using the unique IDs only such that data has N rows. This is the final dataset for standard, no auxiliary data analysis analysis that omits anyone who is missing the gold standard.

```
IC_data<-datafinal_GS[!duplicated(datafinal_GS$ID),c("result",
"true_result","GS_vis4","GS_vis4_mis","BGid", "strat",
```

```
"bghhsub_s2","xstar","xhat","z1","z2")]
IC_GS_datafinal<-IC_data[complete.cases(IC_data$GS_vis4_mis),]
```

Now, we create starting values for our survival parameters. First, to avoid maximization problems due to the ordered constraint of the survival parameters, we re-parameterize these in terms of a log-log transformation of survival function for S_2, and a change in log-log of the survival function for all other parameters $S_3 \dots S_{J+1}$. We consider initial values of 0.1 for our survival parameters, then transform these based on this re-parameterization. We also define lower and upper bounds for our survival parameters. Our lower bound is infinity for the first re-parameterized survival function and 0 for the subsequent $J-1$ terms. Our upper bound is infinity for all terms. Finally, we create a vector *parmi* consisting of a starting value for *beta*, and starting values for re-parameterized survival parameters.

```
initsurv <- 0.1
lami <- log(-log(seq(1, initsurv, length.out = J + 1)[-1]))
lami <- c(lami[1], diff(lami))
tosurv <- function(x) exp(-exp(cumsum(x)))
lowlam <- c(-Inf, rep(0, J - 1))
lowerLBFGS <- c(rep(-Inf, nbeta),lowlam)
upperLBFGS <- c(rep(Inf, nbeta+J))
parmi <- c(rep(0.5,nbeta),lami)
```

Now, we create survey designs using survey package for the two models: the proposed method, with all N subjects, and the standard, no auxiliary data analysis model which excludes subjects missing the gold standard variable.

```
samp_design_reg = svydesign(id=~BGid, strata=~strat,
    weights=~bghhsub_s2, data=IC_GS_datafinal)
samp_design_reg_complete = svydesign(id=~BGid, strata=~strat,
    weights=~bghhsub_s2, data=IC_data)
```

Finally, we fit the proposed method and the standard, no auxiliary data approach with the

weights from the survey design. One may consider the functions `optim()` or `nlminb()` for maximization of the log-likelihood in R.

```
proposed_fit_data_weight<-optim(par=parmi, fn=log_like_proposed,
    gr=gradient_proposed,lower = lowerLBFGS,upper=upperLBFGS,
    method = "L-BFGS-B",N,J,nbeta,Dm,Cm,Xmat,GSdelta,GSVis,
    weights=weights,purpose="SUM",hessian=TRUE)
inverted_hessian<-solve(proposed_fit_data_weight$hessian)

proposed_fit_data_weight_GS<-svyglm(GS_vis4~xhat+z1+z2,
    family=quasibinomial(link="cloglog"),data=IC_GS_datafinal,
    design=samp_design_reg)
```

Next, we calculate a matrix of estimating equation contributions for all individuals. We want the unweighted matrix, because the function `svytotal()` adds the weights later. Then, we compute the influence functions by multiplying this matrix by the inverse of the hessian matrix. Finally, we obtain design-based standard errors using the variance approach of Binder (1983) and functions from the survey package by providing the influence function and survey design to `vcov(svytotal())` (Lumley, 2011).

```
U_prop1<-gradient_proposed(proposed_fit_data_weight$par,N,J,nbeta,
    Dm,Cm,Xmat,GSdelta,GSVis,weights=noweights,purpose="INDIVIDUAL")
infl1<- U_prop1%*%inverted_hessian
mySandVar<- vcov(svytotal(infl1,samp_design_reg_complete,
    influence=TRUE))
```

We now save the estimated parameters, including the estimated regression coefficients and corresponding standard error estimates.

```
beta1est_aux_w<-  proposed_fit_data_weight$par[1]
sandVar<-  sqrt(diag(mySandVar))[1]
fitsum_truth4Year<-summary(proposed_fit_data_weight_GS)
beta1est_truth4Year<-fitsum_truth4Year$coefficients["xhat",1]
```

```
beta1se_truth4Year<-fitsum_truth4Year$coefficients["xhat",2]
```

The final step of our analysis is to use the parametric multiple imputation procedure of Baldoni et al. (2021) to compute the variance that accounts for the extra uncertainty added by the calibration model step. Code for implementing this procedure is available on GitHub at https://github.com/plbaldoni/HCHSsim and involves sampling the calibration model coefficients using their asymptotic parametric distribution M times, resulting in M sets of calibration coefficients. Then, we are able to estimate M predicted covariate (xhat) values, $\hat{X}_i^{(m)}$. Finally, we fit the outcome models each M times, once for each of the newly predicted intake values $\hat{X}_i^{(m)}$ from multiple imputation. We chose $M = 25$ imputations for our analysis. We present full code that applies the procedure of Baldoni et al. (2021) to the proposed estimator fit to simulated data on GitHub at https://github.com/lboe23/AugmentedLikelihood in the code file "Sample_Data_Analysis_Final.R" This full code for running the multiple imputation variance procedure involves running all of the above code in a multiple imputation loop, and thus to avoid redundancy we have chosen to omit it here. On GitHub, we have also provided the output from applying this parametric multiple imputation procedure, titled "SampleAnalysis_MIVarianceResults.RData". Below, we load in this output so that we can obtain our final variance calculations.

```
load(file=paste0('SampleAnalysis_MIVarianceResults.RData'))
output_mi<-as.data.frame(matrix(unlist(list_mi), ncol=5, byrow=T))
colnames(output_mi)<-c("Imp","beta_proposed_mi","se_proposed_mi",
"beta_standard_mi","se_standard_mi")
```

Now, let's use this output to calculate the final estimate of the variance of the regression coefficient $\hat{\beta}$ from each model as: $\hat{V}^* = \frac{1}{M}\sum_{m=1}^{M}\hat{V}^{(m)} + \frac{1}{M-1}\sum_{m=1}^{M}\left(\hat{\beta}^{(m)} - \bar{\hat{\beta}}\right)^2$, where $\bar{\hat{\beta}} = \frac{1}{M}\sum_{m=1}^{M}\hat{\beta}^{(m)}$ and $\hat{\beta}^{(m)}$ and $\hat{V}^{(m)}$ represent the estimated regression coefficient and its estimated variance, respectively, using the m-th completed data set with $m = 1, \ldots, M$. Recall that this formula for computing variance estimates accounting for regression calibration were described in Section 2.3.2 from the main manuscript. Following Baldoni et al. (2021),

we use robust estimators for the mean and standard deviation in this equation, specifically the median and the median absolute deviation:

```
MI_Var<-function(V,betas){
      var<-(median(V)+(mad(betas)^2))
      return(sqrt(var))
}

se_proposed_final<-MI_Var((output_mi$se_proposed_mi)^2,
output_mi$beta_proposed_mi)

se_standard_final<-MI_Var((output_mi$beta_standard_mi)^2,
output_mi$beta_standard_mi)
```

Now, we can use our estimated regression coefficients and their corresponding standard errors to construct a table with estimated hazard ratios (HR) and 95% confidence intervals (CI) associated with a 20% increase in consumption. We will also include relative efficiency calculations in our table, computed as the ratio of the estimated variance of the standard, no auxiliary data approach estimator to the estimated variance of the proposed method estimator, e.g. $\frac{Var(\hat{\beta}_{Standard})}{Var(\hat{\beta}_{Proposed})}$:

```
myfinaltable<-cbind(paste(round(exp(beta1est_aux_w*log(1.2)),2),"(",
round(exp(beta1est_aux_w-1.96*se_proposed_final)^log(1.2),2),",",
round(exp(beta1est_aux_w+1.96*se_proposed_final)^log(1.2),2),")"),
paste(round(exp(beta1est_truth4Year*log(1.2)),2),"(",
round(exp(beta1est_truth4Year-1.96*se_standard_final)^log(1.2),2),
",",
round(exp(beta1est_truth4Year+1.96*se_standard_final)^log(1.2),2),
")"),round((se_standard_final^2)/(se_proposed_final^2),2))
```

Below are our final results from the table we just constructed:

HR (95% CI)		
Proposed	No Auxiliary Data	RE
1.07 (0.99, 1.16)	1.08 (0.98, 1.19)	1.51

B.2. Details of C terms in the likelihood

In Section 2.1 of the main text, we introduce the likelihood contributions for all individuals in terms of $C_{ij} = \left[\prod_{l=1}^{n_i} P(Y_{il}^* | \tau_{j-1} < T_i \leqslant \tau_j, T_{il}^*, \Delta_i)\right]$, which is simply a function of the sensitivity (Se) and specificity (Sp) of the auxiliary data. Recall that we assume constant and known sensitivity (Se) and specificity (Sp), defined as $Se = \Pr(Y_{il}^* = 1 | \tau_{j-1} < T_i \leqslant \tau_j, T_l^* \geqslant \tau_j)$ and $Sp = \Pr(Y_{il}^* = 0 | \tau_{j-1} < T_i \leqslant \tau_j, T_l^* \leqslant \tau_{j-1})$. It is then straightforward to see that $1 - Se = \Pr(Y_{il}^* = 0 | \tau_{j-1} < T_i \leqslant \tau_j, T_l^* \geqslant \tau_j)$ and $1 - Sp = \Pr(Y_{il}^* = 1 | \tau_{j-1} < T_i \leqslant \tau_j, T_l^* \leqslant \tau_{j-1})$In this section, we provide the general form of the C_{ij} terms for the case of no missed visits. The formula introduced below could be modified to allow for missed visits by summing up all terms $\theta_j = \Pr(\tau_{j-1} < T_i \leqslant \tau_j)$ across the $(\tau_{j-1}, \tau_j]$ defining a subject's observational interval. For the case of J total visits for all N subjects, the C_{ij} terms take the following form:

$$
\begin{aligned}
C_{i1} &= Se^{\sum_{j=1}^{n_i} Y_{ij}^*} (1 - Se)^{\sum_{j=1}^{n_i} (1 - Y_{ij}^*)}, \\
C_{i2} &= Sp^{(1 - Y_{i1})} (1 - Sp)^{Y_{i1}} Se^{\sum_{j=2}^{n_i} Y_{ij}^*} (1 - Se)^{\sum_{j=2}^{n_i} (1 - Y_{ij}^*)}, \\
&\ldots \\
C_{i(J+1)} &= Sp^{\sum_{j=1}^{n_i} (1 - Y_{ij}^*)} (1 - Sp)^{\sum_{j=1}^{n_i} Y_{ij}^*}.
\end{aligned}
$$

As an example, consider subject $i = 1$ who has observed auxiliary data vector $\mathbf{Y_i^*} = [0, 0, 0, 1]$ corresponding to annual visit time vector $\mathbf{T_i^*} = [1, 2, 3, 4]$. Suppose the visit times among all N subjects are also $[\tau_0 = 0, \tau_1 = 1, \tau_2 = 2, \tau_3 = 3, \tau_4 = 4, \tau_5 = \infty]$. Then, for subject i with $n_i = J = 4$ and $j = 1$, we have:

$$
\begin{aligned}
C_{11} &= \prod_{l=1}^{4} P(Y_{1l}^{*}|\tau_0 < T_1 \leqslant \tau_1, T_{1l}^{*}, \Delta_1) \\
C_{11} &= P(Y_{11}^{*}|\tau_0 < T_1 \leqslant \tau_1, T_{11}^{*}, \Delta_1) \times P(Y_{12}^{*}|\tau_0 < T_1 \leqslant \tau_1, T_{12}^{*}, \Delta_1) \times \\
&\quad P(Y_{13}^{*}|\tau_0 < T_1 \leqslant \tau_1, T_{13}^{*}, \Delta_1) \times P(Y_{14}^{*}|\tau_0 < T_1 \leqslant \tau_1, T_{14}^{*}, \Delta_1) \\
C_{11} &= P(Y_{11}^{*} = 0|\tau_0 < T_1 \leqslant \tau_1, T_{11}^{*} \geqslant \tau_1, \Delta_1) \times P(Y_{12}^{*} = 0|\tau_0 < T_1 \leqslant \tau_1, T_{12}^{*} \geqslant \tau_1, \Delta_1) \times \\
&\quad P(Y_{13}^{*} = 0|\tau_0 < T_1 \leqslant \tau_1, T_{13}^{*} \geqslant \tau_1, \Delta_1) \times P(Y_{14}^{*} = 1|\tau_0 < T_1 \leqslant \tau_1, T_{14}^{*} \geqslant \tau_1, \Delta_1) \\
C_{11} &= (1 - Se) \times (1 - Se) \times (1 - Se) \times Se
\end{aligned}
$$

Then, following a similar procedure for $j = 2, \ldots, 5$, we see that for subject 1,

$$
\begin{aligned}
C_{11} &= Se(1 - Se)^3, \\
C_{12} &= SpSe(1 - Se)^2, \\
C_{13} &= Sp^2Se(1 - Se), \\
C_{14} &= Sp^3Se \\
C_{15} &= Sp^3(1 - Sp).
\end{aligned}
$$

B.3. Regularity conditions for asymptotic normality

We now provide sufficient regularity conditions for the proposed estimator to be asymptotically normal and achieve a $\sqrt{N}$-convergence rate. We assume the following throughout this section: (1) $\{T_i, \Delta_i, Y_i^*, T_i^*, M_i, X_i, X_i^*, Z_i\}$ is a vector of independent and identically distributed random variables for $i = 1, ..., N$, where N is the number of subjects in the main study data; (2) T_i is the latent, unobserved continuous failure time of interest for subject i; (3) the proportional hazards model holds for the latent true event time, such that $S(t) = S_0(t)^{\exp(x'\beta)}$; (4) the auxiliary outcome status (Y_i^*) is observed at n_i follow-

up times $T_i^* = (t_{i1}, ..., t_{in_i})$ that are a subset of J+1 possible observation times satisfying $0 = \tau_0 < \tau_1 < \tau_2 < ... < \tau_J < \tau_{J+1} = \infty$; and (3) τ_{V_i} is the observation time for the gold standard event status variable Δ_i, where $\tau_{V_i} \in \{\tau_0, \tau_1, \tau_2, ..., \tau_J\}$; (5) τ_{V_i} is the gold standard assessment time for individual i, where $\tau_{V_i} \in \{\tau_0, \tau_1, \tau_2, ..., \tau_J\}$ and Δ_i is the corresponding gold standard event status variable; (6) the binary variable $M_i \in \{0, 1\}$ indicates whether Δ_i is missing; (7) X_i is the p-dimensional true covariate vector of interest with corresponding error-prone vector X_i^*, while Z_i is the additionally observed q-dimensional vector of error-free covariates, where all covariates are random variables with finite variance. In the main text, we define $\theta_j = \Pr(\tau_{j-1} < T_i \leqslant \tau_j)$ and $S_j = \sum_{h=j}^{J+1} \theta_h = \Pr(T > \tau_{j-1})$ for $j = 1, ..., J+1$. Additionally, we require that $1 = S_1 > S_2 > ... > S_{J+1} > 0$, ensuring that $0 < \theta_j < 1$ for $j = 1, \ldots, J$.

B.3.1. Proposed estimator for random sample

First, we consider the case where the data are assumed to be a simple random sample from the population and the covariates of interest are recorded precisely (i.e. error-free). The log-likelihood function $l(\psi) = l(T_i, \Delta_i, Y_i^*, T_i^*, M_i, X_i; \psi)$ is defined in Section 2.1, equation 2.4 from the main text as:

$$
\begin{aligned}
l(\psi) = l(S, \beta) = \sum_{i=1}^{N} l_i(S, \beta) \quad = \quad & \sum_{i=1}^{N} \Bigg[(1 - M_i)\Delta_i \log \left(\sum_{j=1}^{V_i} D_{ij} (S_j)^{\exp(x_i'\beta)} \right) + \\
& (1 - M_i)(1 - \Delta_i) \log \left(\sum_{j=V_i+1}^{J+1} D_{ij} (S_j)^{\exp(x_i'\beta)} \right) + \\
& M_i \log \left(\sum_{j=1}^{J+1} D_{ij} (S_j)^{\exp(x_i'\beta)} \right) \Bigg]. \qquad \text{(B.1)}
\end{aligned}
$$

where $\psi = [\beta, \mathbf{S}]$ and $\mathbf{S} = (S_1, S_2, ..., S_{J+1})'$. Recall that we define the score function as $U_i(\psi) = \frac{\partial l_i(\psi)}{\partial \psi}$. The proposed estimator in this setting, $\hat{\psi}$, is found by solving the score equation $\sum_{i=1}^{N} U_i(\psi) = 0$. Let ψ^0 be the true vector of regression parameters of interest

that solves $E\left[\frac{\partial l(\psi)}{\partial \psi}\right] = E\left[\sum_{i=1}^{N} U_i(\psi)\right] = 0$. Now, further assume that the log-likelihood $l(\psi) = l(T_i, \Delta_i, Y_i^*, T_i^*, M_i, X_i; \psi)$ is twice continuously differentiable with respect to ψ such that there exists an invertible Hessian matrix (Foutz, 1977). We additionally assume the regularity conditions made by Foutz (1977) for establishing consistency and uniqueness of our estimator. Then, appealing to standard maximum likelihood estimation (MLE) theory, with probability going to one as $N \to \infty$, $\hat{\psi}$ is a unique solution to the likelihood equations that is consistent for ψ^0 and asymptotically normal (Boos and Stefanski, 2013). Specifically, one has:

$$\sqrt{N}(\hat{\psi} - \psi^0) \xrightarrow{d} \mathcal{N}(0, \mathrm{V}(\psi^0)), \tag{B.2}$$

where $\mathrm{V}(\psi^0)$ is the Fisher information matrix, denoted by $= I(\psi^0)^{-1}$.

B.3.2. Proposed estimator for complex survey design

We now state the additional regularity conditions needed for the proposed method estimator to accommodate data from a complex survey sampling design by using a weighted log-likelihood function and a sandwich variance estimator to address within-cluster correlation.

Recall from section 2.2 that a sample of N subjects is drawn from a population of size N_{POP} resulting in the sampling probability π_i. Following Lumley and Scott (2017), we assume that $N \to \infty$ and $\frac{N}{N_{POP}} \to p \in (0, 1)$. Additionally assume that π_i is known for the subjects in the sample and is bounded away from 0. The weighted log-likelihood equation of the main text is written as follows: $l_\pi(S, \beta) = \sum_{i=1}^{N} \frac{1}{\pi_i} l_i(S, \beta) = \sum_{i=1}^{N} \breve{l}_i(S, \beta)$. The weighted proposed estimator $\hat{\psi}_\pi$ may be found by solving the weighted score equation, $\sum_{i=1}^{N} \breve{U}_i(\psi_\pi) = \sum_{i=1}^{N} \frac{1}{\pi_i} U_i(\psi_\pi) = 0$. Then, as before, ψ_π^0 is the solution to $E\left[\frac{\partial l_\pi(\psi_\pi)}{\partial \psi_\pi}\right] = E\left[\sum_{i=1}^{N} \breve{U}_i(\psi_\pi)\right] = 0$. Following the same logic applied for the random sample case and since $\hat{\psi}_\pi$ is also a maximum likelihood estimator, we have:

$$\sqrt{N}(\hat{\psi}_\pi - \psi_\pi^0) \xrightarrow{d} \mathcal{N}(0, \mathrm{V}(\psi_\pi^0)), \tag{B.3}$$

where $\mathrm{V}(\psi_\pi^0)$ can be approximated using the implicit differentiation method of Binder (1983). To use this variance estimation approach, assume that $\sum_{i=1}^N \breve{U}_i(\psi_\pi)$ is suitably smooth such that $\hat{\psi}_\pi$ can be implicitly defined as a function of $\sum_{i=1}^N \breve{U}_i(\psi_\pi)$. Further assume that the derivative matrix for $\breve{U}_i$ is full rank, invertible, and a continuous function of ψ_π. Then, we can use a Taylor expansion of U_i at $\hat{\psi}_\pi = \psi_\pi^0$ to arrive at the following estimator for the asymptotic variance of $\hat{\psi}_\pi$: $\hat{\mathrm{V}}[\hat{\psi}_\pi] \approx \left(\sum_{i=1}^N \frac{\partial \breve{U}_i(\hat{\psi}_\pi)}{\partial \psi_\pi}\right)^{-1} \hat{\mathrm{cov}}\left[\sum_{i=1}^N \breve{U}_i(\hat{\psi}_\pi)\right] \left(\sum_{i=1}^N \frac{\partial \breve{U}_i(\hat{\psi}_\pi)}{\partial \psi_\pi}\right)^{-1}$. Details and theoretical justification are provided in Binder (1983).

B.3.3. Proposed estimator for complex survey design and regression calibration

First assume that the error-free covariates X_i and Z_i are available for all sampled individuals, such that our log-likelihood $l_{\pi,X,Z}(\psi) = l(T_i, \Delta_i, Y_i^*, T_i^*, M_i, X_i, Z_i; \psi)$ takes the following form:

$$
\begin{aligned}
l_{\pi,X,Z}(\psi) = \sum_{i=1}^N \frac{1}{\pi_i} l_i^*(S, \beta) \quad = \quad & \sum_{i=1}^N \Bigg[(1 - M_i)\Delta_i \log \left(\sum_{j=1}^{V_i} D_{ij}(S_j)^{\exp(x_i'\beta_X + z_i'\beta_Z)} \right) + \\
& (1 - M_i)(1 - \Delta_i) \log \left(\sum_{j=V_i+1}^{J+1} D_{ij}(S_j)^{\exp(x_i'\beta_X + z_i'\beta_Z)} \right) + \\
& M_i \log \left(\sum_{j=1}^{J+1} D_{ij}(S_j)^{\exp(x_i'\beta_X + z_i'\beta_Z)} \right) \Bigg], \qquad \text{(B.4)}
\end{aligned}
$$

where $\psi = [\beta, \mathbf{S}]$, $\beta = (\beta_X, \beta_Z)'$, and $\mathbf{S} = (S_1, S_2, ..., S_{J+1})'$. Adopting the arguments from sections B.3.1 and B.3.2, the weighted proposed estimator $\hat{\psi}_{\pi,X,Z}$ found by solving the weighted score equation $\sum_{i=1}^N \breve{U}_i(\psi_{\pi,X,Z}) = \sum_{i=1}^N \frac{1}{\pi_i} U_i(\psi_{\pi,X,Z}) = 0$ can also be shown to be consistent for the true parameter $\hat{\psi}_\pi^0$ and asymptotically normal. These arguments only apply to settings in which the true covariate X_i is used in the proposed method instead of $\hat{X}_i$.

We will now make similar arguments of consistency and asymptotic normality for a new version of our log-likelihood that incorporates regression calibration. Begin by assuming

that $\frac{n_C}{N} \rightarrow p \in (0,1)$, where n_C is the number of subjects in the calibration subset. Recall that we assume the classical measurement error model, $X_i^{**} = X_i + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma^2_{\epsilon_i})$, as introduced in Section 2.3.1 from the main manuscript. Assume also that the following linear calibration model from the main manuscript holds: $X_i^{**} = \delta_{(0)} + \delta_{(1)} X_i^* + \delta_{(2)} Z_i + W_i$, where the random measurement error term $W_i \sim N(0, \sigma^2_{W_i})$. Define $\boldsymbol{\delta} = (\delta_{(0)}, \delta_{(1)}, \delta_{(2)})$. Then, since the vector of estimated nuisance parameters $\hat{\boldsymbol{\delta}}$ is a linear regression estimator, we can appeal to standard MLE theory to establish that it is consistent for the true vector of parameters $\boldsymbol{\delta}^0$ and asymptotically normal. To apply regression calibration, the first moment $\hat{X}_i = E(X_i | \boldsymbol{\delta}, X_i^*, Z_i)$ is imputed in place of X_i in our outcome model. To establish asymptotic normality of our regression calibration estimator, we first assume $\boldsymbol{\delta}$ is known and solve the following weighted log-likelihood equation $l^*_{\pi,X*,Z}(\psi) = l(T_i, \Delta_i, Y_i^*, T_i^*, M_i, X_i^*, Z_i; \psi)$:

$$
\begin{aligned}
l^*_{\pi,X*,Z}(\psi) = \sum_{i=1}^{N} \frac{1}{\pi_i} l_i^*(S, \beta) \quad = \quad & \sum_{i=1}^{N} \Bigg[(1 - M_i)\Delta_i \log \left(\sum_{j=1}^{V_i} D_{ij}(S_j)^{\exp(\hat{x}_i' \beta_X + z_i' \beta_Z)} \right) + \\
& (1 - M_i)(1 - \Delta_i) \log \left(\sum_{j=V_i+1}^{J+1} D_{ij}(S_j)^{\exp(\hat{x}_i' \beta_X + z_i' \beta_Z)} \right) + \\
& M_i \log \left(\sum_{j=1}^{J+1} D_{ij}(S_j)^{\exp(\hat{x}_i' \beta_X + z_i' \beta_Z)} \right) \Bigg], \qquad \text{(B.5)}
\end{aligned}
$$

where $\psi = [\beta, \mathbf{S}]$, $\beta = (\beta_X, \beta_Z)'$, and $\mathbf{S} = (S_1, S_2, ..., S_{J+1})'$. We assume that distributions of the variables are such that when X_i is replaced by $\hat{X}_i$, the log-likelihood $l^*_{\pi,X*,Z}(\psi) = l(T_i, \Delta_i, Y_i^*, T_i^*, M_i, X_i^*, Z_i; \psi)$ remains continuously differentiable with respect to ψ and that the Hessian matrix is still invertible. As before, we solve the weighted score equation $\sum_{i=1}^N \breve{U}_i^*(\psi_{\pi,X*,Z}) = \sum_{i=1}^N \frac{1}{\pi_i} U_i^*(\psi_{\pi,X*,Z}) = 0$ in order to obtain our weighted proposed estimator, $\hat{\psi}^*_{\pi,X*,Z}$. Under regularity conditions described previously, ψ_π^* will be a unique, consistent solution to the vector of equations $E\left[\frac{\partial l_\pi^*(\psi_{\pi,X*,Z})}{\partial \psi_\pi}\right] = E\left[\sum_{i=1}^N \breve{U}_i^*(\psi_{\pi,X*,Z})\right] = 0$. In general, ψ_π^* is not the same as ψ_π^0, as the parameter estimates from regression calibration are viewed as an approximation (Buonaccorsi, 2010). Using the techniques of

Boos and Stefanski (2013) once again, we can verify the asymptotic normality our estimator $\hat{\psi}^*_{\pi,X*,Z}$, i.e.:

$$\sqrt{N}(\hat{\psi}^*_{\pi,X*,Z} - \psi^*_\pi) \xrightarrow{d} \mathcal{N}(0, \mathrm{V}(\psi^*_\pi)). \tag{B.6}$$

The regularity in equation B.6 depends on known nuisance parameter vector $\boldsymbol{\delta}$ from the error model. With the additional usual regularity assumptions for linear regression that guarantee a consistent and asymptotically normal estimator for $\hat{\boldsymbol{\delta}}$, the regularity of our calibration estimator will still hold using this plug-in estimator for $\boldsymbol{\delta}$ by appealing to Theorem 5.31 in Van der Vaart (2000). The variance $\mathrm{V}(\psi^*_\pi)$ is estimated using the multiple imputation procedure introduced by Baldoni et al. (2021) described in the main text. When regression calibration is applied to the proposed method to adjust for covariate error, our estimator is only approximate but has been empirically shown to have minimal bias and good coverage probability when the true regression parameter is modest in size and the event of interest under study is rare.

B.4. Supplemental details for HCHS/SOL data example

We adopted the same exclusion criteria used in the ongoing clinical investigation that seeks to understand the relationship between several dietary intake variables and the risk of chronic diseases in the HCHS/SOL cohort. We excluded any participants who reported diabetes or unknown status at baseline ($N = 3428$), had missing covariate data ($N = 373$), or had no auxiliary follow-up ($N = 551$), resulting in $12,317$ eligible participants. To mimic the planned analysis of the clinical investigation, for the 351 subjects in the data who reported a positive diabetes status after one or more missed annual follow-up calls, we imputed that the event happened at the midpoint of the missed follow-up times. For most subjects with a missed call, subjects subsequently reported no diabetes diagnosis had occurred since the last call and so a negative disease status was imputed for the prior follow-up calls. The proposed method is applied to a subset of $8,200$ HCHS/SOL cohort participants, including

all eligible SOLNAS subset participants ($N = 420$) and HCHS/SOL participants from primary sampling units (PSUs) with 4 or fewer members ($N = 282$). The remaining subset members were selected by taking a random sample of 7498 participants that had not yet been selected.

Table B.1: Simulation results are shown for exponential failure times assuming the Cox proportional hazards model with $X \sim Gamma(0.2, 1)$ and $\beta = \log(1.5)$ for (1) the grouped time survival approach that uses the true outcome data from all periodic visits and (2) the standard interval-censored approach that does not incorporate auxiliary data. The median percent (%) bias, median standard errors (ASE), empirical median absolute deviation (MAD) and coverage probabilities (CP) are given for 1000 simulated data sets.

			Gold Standard Every Year				Gold Standard Year 4 Only (No Auxiliary Data)				
MR[1]	CR[2]	N[3]	% Bias	ASE	MAD	CP	% Bias	ASE	MAD	CP	RE[4]
0.0	0.9	1000	-2.015	0.158	0.151	0.953	-1.402	0.160	0.155	0.951	1.019
		10,000	1.025	0.048	0.050	0.948	1.279	0.048	0.051	0.951	1.022
	0.7	1000	0.207	0.100	0.095	0.950	0.614	0.107	0.106	0.950	1.104
		10,000	0.466	0.031	0.031	0.942	0.398	0.033	0.034	0.947	1.137
	0.5	1000	1.061	0.084	0.082	0.938	2.020	0.099	0.102	0.947	1.361
		10,000	0.503	0.026	0.026	0.954	0.382	0.031	0.034	0.951	1.370
0.2	0.9	1000	0.863	0.164	0.160	0.953	-0.378	0.181	0.183	0.951	1.190
		10,000	1.493	0.049	0.052	0.945	0.769	0.054	0.055	0.952	1.197
	0.7	1000	1.966	0.103	0.108	0.939	0.377	0.120	0.116	0.954	1.319
		10,000	1.524	0.032	0.035	0.936	0.332	0.037	0.038	0.946	1.336
	0.5	1000	3.572	0.087	0.095	0.920	2.084	0.111	0.116	0.947	1.607
		10,000	2.453	0.027	0.029	0.914	0.247	0.034	0.036	0.952	1.606
0.4	0.9	1000	1.935	0.176	0.181	0.944	1.178	0.213	0.222	0.959	1.411
		10,000	1.595	0.053	0.053	0.941	2.122	0.062	0.064	0.960	1.411
	0.7	1000	2.100	0.110	0.117	0.927	1.616	0.140	0.138	0.958	1.586
		10,000	2.379	0.034	0.039	0.913	0.758	0.043	0.044	0.946	1.586
	0.5	1000	6.148	0.093	0.106	0.911	3.186	0.130	0.136	0.952	1.911
		10,000	4.344	0.029	0.030	0.885	0.122	0.040	0.043	0.945	1.893

[1] MR = Average probability that the gold standard indicator Δ is missing at year 4

[2] CR = Average censoring rate for the latent true event time at the end of study

[3] N = Sample size for proposed approach; if $MR > 0.0$, sample size for no auxiliary data approach is smaller because of missingness in gold standard indicator Δ.

[4] RE = median relative efficiency, calculated as the median of the ratio of the estimated variance of the standard, no auxiliary data approach estimator to the estimated variance of the proposed method estimator , e.g. $\frac{Var(\hat{\beta}_{Standard})}{Var(\hat{\beta}_{Proposed})}$

Table B.2: Simulation results are shown for exponential failure times assuming the Cox proportional hazards model with $X \sim Gamma(0.2, 1)$, $\beta = \log(1.5)$, and values of $Se = 0.90$ and $Sp = 0.80$ for the auxiliary data. The median percent (%) bias, median standard errors (ASE), empirical median absolute deviation (MAD) and coverage probabilities (CP) are given for 1000 simulated data sets for the proposed method and the standard interval-censored approach that does not incorporate auxiliary data.

			Proposed				No Auxiliary Data				
MR[1]	CR[2]	N[3]	% Bias	ASE	MAD	CP	% Bias	ASE	MAD	CP	RE[4]
0.0	0.9	1000	-0.810	0.160	0.151	0.953	-1.402	0.160	0.155	0.951	1.011
		10,000	1.501	0.048	0.049	0.954	1.279	0.048	0.051	0.951	1.007
	0.7	1000	0.914	0.104	0.099	0.952	0.614	0.107	0.106	0.950	1.043
		10,000	0.391	0.032	0.032	0.945	0.398	0.033	0.034	0.947	1.065
	0.5	1000	1.598	0.091	0.091	0.943	2.020	0.099	0.102	0.947	1.172
		10,000	0.433	0.028	0.029	0.947	0.382	0.031	0.034	0.951	1.178
0.2	0.9	1000	-1.283	0.175	0.177	0.956	-0.378	0.181	0.183	0.951	1.040
		10,000	0.970	0.053	0.052	0.952	0.769	0.054	0.055	0.952	1.050
	0.7	1000	0.273	0.111	0.115	0.957	0.377	0.120	0.116	0.954	1.135
		10,000	0.604	0.034	0.034	0.947	0.332	0.037	0.038	0.946	1.152
	0.5	1000	1.462	0.097	0.097	0.942	2.084	0.111	0.116	0.947	1.306
		10,000	0.478	0.030	0.031	0.948	0.247	0.034	0.036	0.952	1.316
0.4	0.9	1000	-1.054	0.197	0.201	0.958	1.178	0.213	0.222	0.959	1.109
		10,000	1.388	0.059	0.059	0.953	2.122	0.062	0.064	0.960	1.127
	0.7	1000	0.739	0.121	0.120	0.957	1.616	0.140	0.138	0.958	1.278
		10,000	0.762	0.037	0.038	0.952	0.758	0.043	0.044	0.946	1.306
	0.5	1000	2.252	0.103	0.104	0.942	3.186	0.130	0.136	0.952	1.553
		10,000	0.550	0.032	0.033	0.949	0.122	0.040	0.043	0.945	1.549

[1] MR = Average probability that the gold standard indicator Δ is missing at year 4

[2] CR = Average censoring rate for the latent true event time at the end of study

[3] N = Sample size for proposed approach; if $MR > 0.0$, sample size for no auxiliary data approach is smaller because of missingness in gold standard indicator Δ.

[4] RE = median relative efficiency, calculated as the median of the ratio of the estimated variance of the standard, no auxiliary data approach estimator to the estimated variance of the proposed method estimator , e.g. $\frac{Var(\hat{\beta}_{Standard})}{Var(\hat{\beta}_{Proposed})}$

Table B.3: Simulation results are shown for exponential failure times assuming the Cox proportional hazards model with $X \sim Gamma(0.2, 1)$ and $\beta = \log(3)$. The median percent (%) bias, median standard errors (ASE), empirical median absolute deviation (MAD) and coverage probabilities (CP) are given for 1000 simulated data sets for the proposed method and the standard interval-censored approach that does not incorporate auxiliary data. Here, $Se = 0.80$ and $Sp = 0.90$ for the auxiliary data.

			Proposed				No Auxiliary Data				
MR[1]	CR[2]	N[3]	% Bias	ASE	MAD	CP	% Bias	ASE	MAD	CP	RE[4]
0.0	0.9	1000	0.627	0.126	0.121	0.961	0.937	0.135	0.136	0.950	1.155
		10,000	0.032	0.039	0.040	0.940	0.150	0.042	0.042	0.939	1.162
	0.7	1000	1.082	0.114	0.118	0.953	0.657	0.130	0.134	0.949	1.274
		10,000	0.207	0.036	0.035	0.946	0.301	0.041	0.043	0.955	1.281
	0.5	1000	0.393	0.122	0.120	0.951	0.517	0.147	0.151	0.952	1.474
		10,000	0.302	0.038	0.038	0.954	0.164	0.046	0.046	0.950	1.458
0.2	0.9	1000	1.088	0.132	0.130	0.955	1.493	0.151	0.149	0.955	1.286
		10,000	0.175	0.041	0.042	0.939	0.124	0.047	0.046	0.944	1.298
	0.7	1000	1.243	0.119	0.118	0.955	1.213	0.145	0.149	0.950	1.464
		10,000	0.285	0.037	0.038	0.951	0.436	0.045	0.048	0.955	1.473
	0.5	1000	0.505	0.126	0.124	0.950	0.502	0.165	0.179	0.939	1.724
		10,000	0.165	0.040	0.039	0.950	0.041	0.052	0.051	0.948	1.708
0.4	0.9	1000	1.011	0.141	0.140	0.949	1.840	0.175	0.184	0.952	1.515
		10,000	0.049	0.044	0.044	0.950	0.084	0.054	0.053	0.954	1.522
	0.7	1000	1.275	0.125	0.125	0.948	1.897	0.168	0.180	0.944	1.789
		10,000	0.115	0.039	0.041	0.958	0.509	0.052	0.053	0.956	1.792
	0.5	1000	0.692	0.130	0.134	0.944	2.561	0.193	0.205	0.942	2.174
		10,000	0.215	0.041	0.042	0.948	-0.140	0.060	0.058	0.949	2.118

[1] MR = Average probability that the gold standard indicator Δ is missing at year 4

[2] CR = Average censoring rate for the latent true event time at the end of study

[3] N = Sample size for proposed approach; if $MR > 0.0$, sample size for no auxiliary data approach is smaller because of missingness in gold standard indicator Δ.

[4] RE = median relative efficiency, calculated as the median of the ratio of the estimated variance of the standard, no auxiliary data approach estimator to the estimated variance of the proposed method estimator , e.g. $\frac{Var(\hat{\beta}_{Standard})}{Var(\hat{\beta}_{Proposed})}$

Table B.4: Simulation results are shown for data simulated to be from a complex survey with exponential failure times assuming the Cox proportional hazards model with $X \sim Normal(\text{shape}_s + \omega_{gs}, \text{scale}_s + \rho_{gs})$ for an individual in block group g and stratum s and $\beta = \log(1.5)$. The median percent (%) bias, median standard errors (ASE), median absolute deviation (MAD) and coverage probabilities (CP) are given for 1000 simulated data sets for the weighted proposed estimator and the weighted interval-censored approach that does not incorporate auxiliary data when both use a sandwich variance estimator to address within-cluster correlation. Here, $Se = 0.80$ and $Sp = 0.90$ for the auxiliary data.

			Proposed				No Auxiliary Data				
MR[1]	CR[2]	N[3]	% Bias	ASE	MAD	CP	% Bias	ASE	MAD	CP	RE[4]
0.0	0.9	1000	-4.213	0.122	0.139	0.933	-4.642	0.124	0.138	0.938	1.018
		10,000	-3.293	0.042	0.041	0.933	-3.285	0.042	0.041	0.936	1.010
	0.7	1000	-0.778	0.078	0.081	0.936	-0.983	0.079	0.080	0.932	1.024
		10,000	-0.239	0.026	0.028	0.937	-0.331	0.026	0.028	0.937	1.029
	0.5	1000	-0.459	0.062	0.066	0.936	-0.516	0.065	0.067	0.932	1.066
		10,000	-0.210	0.020	0.021	0.938	-0.225	0.021	0.022	0.941	1.087
0.2	0.9	1000	-3.237	0.133	0.152	0.926	-3.294	0.137	0.157	0.937	1.055
		10,000	-3.275	0.046	0.045	0.942	-3.160	0.047	0.047	0.941	1.058
	0.7	1000	-1.547	0.083	0.085	0.929	-1.153	0.088	0.094	0.931	1.117
		10,000	-0.530	0.027	0.028	0.939	-0.194	0.029	0.029	0.937	1.124
	0.5	1000	-0.194	0.066	0.070	0.932	-0.558	0.072	0.076	0.930	1.190
		10,000	-0.143	0.021	0.022	0.946	-0.086	0.024	0.024	0.941	1.230
0.4	0.9	1000	-2.592	0.149	0.167	0.924	-2.799	0.155	0.165	0.930	1.108
		10,000	-2.977	0.051	0.050	0.934	-2.789	0.054	0.054	0.932	1.131
	0.7	1000	-0.665	0.090	0.091	0.930	-0.173	0.101	0.108	0.933	1.266
		10,000	-0.564	0.029	0.031	0.939	-0.337	0.033	0.035	0.947	1.286
	0.5	1000	0.471	0.070	0.072	0.926	0.081	0.082	0.089	0.920	1.407
		10,000	-0.193	0.023	0.024	0.947	-0.063	0.027	0.029	0.941	1.455

[1] MR = Average probability that the gold standard indicator Δ is missing at year 4

[2] CR = Average censoring rate for the latent true event time at the end of study

[3] (N) = Average sample size for proposed approach; if $MR > 0.0$, sample size for no auxiliary data approach is smaller because of missingness in gold standard indicator Δ.

[4] RE = median relative efficiency, calculated as the median of the ratio of the estimated variance of the standard, no auxiliary data approach estimator to the estimated variance of the proposed method estimator , e.g. $\frac{Var(\hat{\beta}_{Standard})}{Var(\hat{\beta}_{Proposed})}$

Table B.5: Sensitivity analysis using HCHS/SOL data on a subset of study participants with visit 2 sensitivity ($Se = 0.77$) and specificity ($Sp = 0.92$) values. Hazard Ratio (HR) and 95% confidence interval (CI) estimates of incident diabetes for a 20% increase in consumption of energy (kcal/d), protein (g/d), and protein density (% energy from protein/d) based on the proposed estimator and the interval-censored approach that does not incorporate auxiliary data.

	HR (95% CI)		
Model[1]	Proposed	No Auxiliary Data	RE[2]
Energy (kcal/d)	1.26 (0.49, 3.24)	1.20 (0.41, 3.82)	1.27
Protein (g/d)	1.37 (0.88, 2.14)	1.37 (0.74, 2.51)	1.85
Protein Density	1.01 (1.00, 1.03)	1.01 (1.00, 1.03)	1.43

[1] Each model is adjusted for potential confounders including age, body mass index (BMI), sex, Hispanic/Latino background, language preference, education, income, and smoking status.
[2] RE = relative efficiency, calculated as the ratio of the estimated variance of the standard, no auxiliary data approach estimator to the estimated variance of the proposed method estimator , e.g. $\frac{Var(\hat{\beta}_{Standard})}{Var(\hat{\beta}_{Proposed})}$

APPENDIX C

Supplementary Material for Chapter 4

C.1. Sufficient Assumptions for Sandwich Variance Estimation

The stacked estimating equation framework outlined in the prior section may be considered for any regular, asymptotically linear estimators. As illustrated in the prior sections, this includes any parametric maximum likelihood estimators, as well as the Cox proportional hazards model estimator. This framework does not apply, however, to a range of machine learning models, including Lasso and Random Forest. Here, we outline mild regularity conditions required for the estimators from stage 1 and stage 2 in order to apply the proposed sandwich variance estimator. Specifically, if we assume that the vector of estimated nuisance parameters, $\hat{\boldsymbol{\alpha}}$, from Stage 1 is a regular, asymptotically linear estimator (e.g. a linear regression estimator), the regularity conditions specified by Foutz (1977) can be used to establish the consistency and uniqueness of this estimator. For the case of a standard maximum likelihood estimator, we may appeal to standard maximum likelihood estimation (MLE) theory to show that $\hat{\boldsymbol{\alpha}}$ is a unique solution to the likelihood equations that is consistent and asymptotically normal (Boos and Stefanski, 2013). Additionally consider an outcome model of interest and suppose all covariates are observed without error. For a generalized linear model, one can appeal to this same standard MLE theory to establish the consistency, uniqueness, and asymptotic normality of the outcome model estimator, $\hat{\boldsymbol{\beta}}$. When the outcome model is a Cox proportional hazards model, the techniques of Andersen and Gill (1982) may be used to establish consistency and asymptotic normality of $\hat{\boldsymbol{\beta}}$.

We can make sufficient, typical regularity assumptions for our stage 1 and stage 2 models to ensure that we have a consistent and asymptotically normal estimator for $\hat{\boldsymbol{\alpha}}$. A common but not necessary assumption is that $\frac{n}{N} \to p \in (0, 1)$, where n is the number of subjects in the calibration subset. More complex assumptions may also be considered (Särndal et al., 2003). The regularity of our estimator $\hat{\boldsymbol{\beta}}$ from stage 2 will still hold using a plug-in estimator

for $\boldsymbol{\alpha}$ by appealing to Theorem 5.31 in Van der Vaart (2000).

C.2. Steps for Computing Bootstrap Standard Errors

Below we have outlined the steps for computing bootstrap standard errors, as commonly applied in the context of regression calibration, using a stratified bootstrap procedure:

1. Choose a number of bootstrap samples to perform (e.g. $B = 500$)

2. For each bootstrap sample,

   (a) Draw a stratified bootstrap sample with replacement of size N: First, draw a sample with replacement of size n from those in the subset only. Then, draw a sample with replacement of size $N - n$ from the non-subset members.

   (b) Fit the regression calibration (stage 1) model on the bootstrap sample.

   (c) Use the calibration model fit to the bootstrap sample to get an estimate of the exposure, $\hat{X}_i^{(b)}$, on all main study participants, for $b = 1, \ldots, B$.

   (d) Fit the outcome regression (stage 2) model using the bootstrap sample with $\hat{X}_i^{(b)}$ to obtain an estimate of the bth regression model parameters, $\beta^{(b)}$.

3. Repeat step (2) B times.

4. For an estimate of the adjusted standard error, compute the standard deviation of the B bootstrap estimates of the regression parameter, $\beta = (\beta^{(1)}, ..., \beta^{(B)})$.

C.3. Steps for Computing Multiple Imputation-Based Standard Errors

Below we have outlined the steps for computing standard errors using the resampling-based multiple imputation approach of Baldoni et al. (2021):

1. Choose a number of imputations to perform (e.g. $M = 25$)

2. For each imputation,

(a) Draw a bootstrap sample with replacement of size n from those in the subset only. Assuming the calibration subset is a simple random sample of the main study, select individuals in the sample with equal probability.

(b) Fit the regression calibration (stage 1) model on the bootstrap sample.

(c) Use the calibration model fit to the bootstrap sample to get an estimate of the exposure, $\hat{X}_i^{(m)}$, on all main study participants, for $m = 1, \ldots, M$.

(d) Fit the outcome regression (stage 2) model using $\hat{X}_i^{(m)}$ and other covariates Z to obtain an estimate of the mth regression model parameters, $\beta^{(m)}$.

3. Repeat step (2) M times.

4. For an estimate of the adjusted standard error of $\hat{\beta}$, compute $\hat{V}^* = \frac{1}{M}\sum_{m=1}^{M}\hat{V}^{(m)} + \frac{1}{M-1}\sum_{m=1}^{M}\left(\hat{\beta}^{(m)} - \bar{\hat{\beta}}\right)^2$, where $\bar{\hat{\beta}} = \frac{1}{M}\sum_{m=1}^{M}\hat{\beta}^{(m)}$ and $\hat{\beta}^{(m)}$ and $\hat{V}^{(m)}$ represent the estimated regression coefficient and its estimated variance, respectively, using the m-th completed data set.

Baldoni et al. (2021) also considered robust estimators for the mean and standard deviation used to compute the adjusted variance in step 4. Specifically, for robustness to skewed estimates, the median and median absolute deviation were used. Further details on the multiple imputation-based variance estimator are described in Baldoni et al. (2021) and code for implementing this procedure is available on GitHub at https://github.com/plbaldoni/HCHSsim.

C.4. Details on Simulation Study to Illustrate Performance of Sandwich Variance Estimator

We conducted a simulation study to show how the sandwich variance approach and competing estimators perform under various different settings. We simulate two covariates, X_i and Z_i, from a multivariate normal distribution with mean 0 and a covariance matrix with all diagonal elements equal to 1. We vary the off-diagonal elements between 0.3 and 0.7 to

represent low and high correlation, respectively, between the two covariates. Next, we assume the logistic regression model, $p_i = P(Y_i = 1) = \frac{\exp(\beta_0+\beta_X X_i+\beta_Z Z_i)}{1+\exp(\beta_0+\beta_X X_i+\beta_Z Z_i)}$ and fix $\beta_0 = 0.2$, $\beta_X = \log(1.5)$, and $\beta_Z = \log(0.7)$. In epidemiologic settings, $\beta_X = \log(1.5)$ represents a log-odds ratio corresponding to the exposure of interest of a moderate size. To simulate the binary outcome Y_i, we generated N variables Un_i from a Uniform(0,1) distribution and then let $Y_i = 1$ if $\text{Un}_i < p_i$ and 0 otherwise.

Later, we run a set of simulations for the random sample case designed to assess the performance of the proposed sandwich variance estimator when a Cox proportional hazards model is considered as our stage 2 model. To simulate this outcome, we generated event times from a continuous time exponential distribution with parameter $\lambda = 0.23\exp(\beta_X X_i + \beta_Z Z_i)$, where $\beta_X = \log(1.5)$ and $\beta_Z = \log(0.7)$. We let our censoring time be 2 such that any subject who experienced the event prior to this time was assumed to have experienced the event of interest. Assigning these parameters resulted in an event rate that approximated that for our logistic regression outcome model, roughly 38%.

Our error-prone covariate X_i^* is simulated to represent a hypothetical dietary intake variable using the following linear measurement error model, $X_i^* = \delta_0 + \delta_1 X_i + \delta_2 Z_i + e_i$, where $\delta_X = 0.20$, $\delta_X = 0.37$ and $\delta_Z = 0.15$. We let $e_i \sim N(0, \sigma^2)$ and considered σ^2 values of 0, 0.25, 0.50, and 1.00 to represent cases of zero, low, moderate, and high measurement error. We assume that our hypothetical biomarker subset is a random sample of $n = 450$ subjects from the main study. Our simulated hypothetical biomarker of interest, X_i^{**}, is generated to follow the classical measurement error model $X_i^{**} = X_i + \epsilon_i$, and we let $\epsilon \sim N(0, 0.2)$.

In our simulation studies designed to mimic the structure the complex survey design data setting, all data generation settings besides the sampling structure were kept the same, including the simulation of X_i, X_i^*, X_i^{**}, Z_i, and Y_i. In a complex survey sampling scheme of this type, it was not possible to fix the total number of individuals selected for a simulated sample exactly, but we were able to obtain sample sizes of approximately $N = 1000$ and $N = 10,000$.

For all settings studied, we conducted 1000 simulation iterations. Simulations for the simple random sample which computed standard error estimates using the bootstrap used $B = 500$ bootstrap samples. Confidence intervals in Tables 1 and S1 are constructed using the typically applied Wald confidence interval computed using bootstrap standard errors. For simulations from the complex survey design, we used $M = 25$ imputations when applying the multiple imputation based procedure of Baldoni et al. (2021).

C.5. Supplementary Details on the WHI Data Analysis

To fit the stage 1 model in the WHI data analysis, we modified calibration models that were previously developed for self-reported intakes of energy, protein, and protein density by Neuhouser et al. (2008) and used by Tinker et al. (2011) to obtain incident diabetes hazard ratios in the WHI cohort. Specifically, since the regression calibration approach requires the calibration model to include the same set of covariates as the outcome model in order to avoid bias (Kipnis et al., 2009), we expanded the set of covariates from the former calibration models to include all confounding variables that will be included in our outcome model. All stage 1 models therefore included body mass index (BMI), age, race-ethnicity, income, education, physical activity in units of metabolic equivalent tasks per week, smoking status, alcohol consumption, hypertension, history of cardiovascular disease, family history of diabetes, and hormone use.

Incident diabetes in the WHI was recorded using a self-reported questionnaire at annual follow-up visits. As in Tinker et al. (2011), we consider the Cox proportional hazards model for our stage 2 model, stratified on age in 5-year categories, hormone therapy trial participation, and DM-C or OS membership. All stage 2 models for diabetes were adjusted by the same set of confounders included in the stage 1 models. We also stratified our stage 2 models on age in 5-year categories, hormone therapy trial participation, and DM-C or OS membership. Following Tinker et al. (2011), we fit two versions of the stage 2 model, one which is adjusted for BMI and the other which is not. This issue, which is discussed by Tinker et al. (2011), relates to the fact that BMI may be a mediator of the relationship

between energy intake and diabetes. We compare model-based standard errors (naive SE) to those estimated by the proposed sandwich estimator introduced in the main paper and the standard bootstrap procedure using $B = 500$ bootstrap samples.

We consider data from women who participated in either the comparison arm of the Dietary Modification trial (DM-C) or the Observational Study (OS) in the WHI (Ritenbaugh et al., 2003; Langer et al., 2003). Note that neither women from the DM-C nor the OS received study interventions. By adopting the same exclusion criteria described by Tinker et al. (2011), we obtained a final analytic data set of $77,805$ participants. These criteria essentially attempt to align the characteristics of participants in the DM-C and OS cohorts as well as exclude any women with missing data or who reported diabetes at baseline. In our analysis, baseline was defined as the time of the first self-reported dietary assessment post-enrollment, year 1 for the DM-C and year 3 for the OS.

Following Tinker et al. (2011), we excluded women who reported diabetes at baseline or during the first year of follow-up for the comparison arm of the WHI Dietary Modification trial (DM-C) participants ($n = 724$) or the first three years of follow-up for the WHI Observational Study (OS) participants ($n = 4109$). In an attempt to align characteristics of women in the DM-C trial with those of women in the OS, the following women in OS were also excluded: those who had breast, colorectal, or other cancer within 10 years prior to enrollment ($n = 8677$), stroke or myocardial infarction within 6 months prior to enrollment ($n = 155$), body mass index (BMI) < 18 ($n = 678$), hypertension (systolic blood pressure > 200 or diastolic blood pressure > 105)($n = 244$), reported daily energy intake of < 600 kcal or > 5000 kcal ($n = 3571$), $\geqslant 10$ meals prepared away from home each week ($n = 3598$), a special low-fiber diet ($n = 568$), a special malabsorption-related diet ($n = 514$), inadvertent weight loss of > 15 pounds within 6 months of enrollment ($n = 594$), and diabetes diagnosis recorded before age 21 at enrollment ($n = 95$). After applying these criteria and including only the participants with no missing data from the stage 1 and stage 2 model variables, we obtained our analytic cohort with $77,805$ participants. Of these $77,805$ women, $19,945$

(25.6%) were from the DM-C and 57,860 (74.4%) were from the OS. Our stage 1 models included 356 eligible women from the Nutritional Biomarker Study who did not have missing data.

Our estimated hazard ratios for all models were somewhat different than those reported by Tinker et al. (2011). Specifically, when BMI was excluded from the outcome model, Tinker et al. (2011) showed that a HR (95% CI) of 2.41 (2.06, 2.82) was associated with a 20% increase in energy intake, which differs slightly from our 2.88 (2.16, 3.85) with adjusted standard errors estimated by the bootstrap. These differences may be explained by a few discrepancies in our analysis and our analytic cohort. Specifically, even after applying the same exclusion criteria and considering only participants with complete data, we had a slightly different data set with 19,968 DM-C and 57,860 OS participants, compared to Tinker et al. (2011) whose analytic data set contained 19,111 DM-C and 55,044 OS participants for a total of 74,155 in the analysis. One major difference between our analysis approach and that of Tinker et al. (2011) related to the variables included in the stage 1 and stage 2 models. We chose to include the same set of confounders, Z, in the stage 1 and stage 2 models, while Tinker et al. (2011) included a set of confounders in the stage 2 outcome model that were not included in the stage 1 model. In general, aligning the set of variables in the stage 1 and stage 2 models is recommended to avoid potential bias and may explain some of the differences in the observed results (Kipnis et al., 2009). Finally, we note that Tinker et al. (2011) included the variables glycemic index and glycemic load in their stage 2 models. We chose to exclude these variables from all models due to numerical instability that they created when added to the stage 1 models, potentially due to correlation with other variables.

C.6. Supplementary Details on the HCHS/SOL Data Analysis

In our reanalysis of the HCHS/SOL data, we fit our stage 1 model to $n = 310$ SOLNAS participants, excluding any SOLNAS participants who were ineligible for the stage 2 analysis from having had a previous diagnosis of high blood pressure or hypertension or making use

of antihypertensive medication. This sample also excluded 11 SOLNAS participants who had extreme biomarker or self-reported values for sodium. Our stage 2 analysis of the HCHS/SOL data included $8,176$ participants from the original HCHS/SOL cohort ($N = 16,415$). We started with the subset used by Baldoni et al. (2021), which was constructed by taking a random sample of $8,208$ HCHS/SOL participants and excluding 83 participants who had missing covariate data. For our sample, we considered these $8,176$ participants with complete data, then added back the remaining 51 eligible SOLNAS ancillary study participants who were not selected by the original random sample.

While the convention in the HCHS/SOL study has previously been to ignore the survey design in the fitting of the stage 1 model, we chose to account for the design in the model, a decision that was made to better capture the variability in the complex survey design. To account for the survey design in the stage 1 model, we subset the parent cohort HCHS/SOL survey design to the participants int the SOLNAS substudy. As described in the main manuscript, attention must be paid to assigning the strata when using data from a complex survey design. For this analysis, we specified the strata as the cross-classification of the subset indicator, V_i, with the strata variables from the HCHS/SOL design. Since SOLNAS is not a nested subset by design, special consideration was required to determine the strata to be used in this data example. The parent HCHS/SOL study included multi-level strata, where the top level was field center, while SOLNAS was only stratified by field center. Thus, our best approximation for determining the strata in this data example was to use the full set of multi-level strata for those not in the SOLNAS subset ($V_i = 0$) and just the field center for those in the SOLNAS subset ($V_i = 1$).

The confounding variables of interest, Z_i used in our stage 1 and stage 2 models are body mass index (BMI), age, Hispanic/Latino background, income, education, physical activity, smoking status, alcohol consumption, field center, language preference, sex, nativity, family history of cardiovascular health disease, and hypercholesterolemia. All analyses used log-transformed biomarkers for sodium and potassium and log-transformed self-reported 24-hour

recall measures. The outcome variables of hypertension and systolic blood pressure were recorded at the HCHS/SOL baseline, in-person clinical examination visit (2008-2011). We consider standard errors estimated by the model (Naive SE), the sandwich, and the MI procedure with $M = 25$ imputations.

C.7. Sandwich Variance Example: Regression Calibration Applied to Logistic Regression

In this section, we use an example to illustrate how one might derive the stacked estimating equations, $U_i(\theta)$, and $A(\hat{\theta})$ and $B(\hat{\theta})$ matrices in order to obtain a sandwich variance estimator. For this example, we consider the setting in which regression calibration is applied to a logistic regression outcome model. Consider the logistic regression outcome model introduced in Section 4.4.3 of the main manuscript. Our parameter vector of interest, which includes the nuisance parameters from the stage 1 model and our stage 2 outcome regression model parameters is then: $\theta = (\alpha_0, \alpha_X, \alpha_Z, \beta_0, \beta_X, \beta_Z)$. We can then use the M-estimation approach of Boos and Stefanski (2013) to obtain the vector of stacked estimating equations for the parameter vector θ as follows:

$$U_i(\theta) = \begin{bmatrix} U_{i1}(\theta) \\ U_{i2}(\theta) \end{bmatrix} = \begin{bmatrix} U_{i1(a)}(\theta) \\ U_{i1(b)}(\theta) \\ U_{i1(c)}(\theta) \\ U_{i2(a)}(\theta) \\ U_{i2(b)}(\theta) \\ U_{i2(c)}(\theta) \end{bmatrix} = \begin{bmatrix} V_i(X_i^{**} - \alpha_0 - \alpha_X X_i^* - \alpha_Z Z_i) \\ V_i(X_i^{**} - \alpha_0 - \alpha_X X_i^* - \alpha_Z Z_i) X_i^* \\ V_i(X_i^{**} - \alpha_0 - \alpha_X X_i^* - \alpha_Z Z_i) Z_i \\ Y_i - p_i \\ \hat{X}_i(Y_i - p_i) \\ Z_i(Y_i - p_i) \end{bmatrix} \tag{C.1}$$

Note that since $\hat{X}_i = E(X_i^{**}|X_i^*, Z_i) = \alpha_0 + \alpha_X X_i^* + \alpha_Z Z_i$, we have:

$$U_i(\theta) = \begin{bmatrix} V_i(X_i^{**} - \alpha_0 - \alpha_X X_i^* - \alpha_Z Z_i) \\ V_i(X_i^{**} - \alpha_0 - \alpha_X X_i^* - \alpha_Z Z_i) X_i^* \\ V_i(X_i^{**} - \alpha_0 - \alpha_X X_i^* - \alpha_Z Z_i) Z_i \\ Y_i - p_i \\ (\alpha_0 + \alpha_X X_i^* + \alpha_Z Z_i)(Y_i - p_i) \\ Z_i(Y_i - p_i) \end{bmatrix}$$

As described in the main manuscript, the estimates $\hat{\theta}$ can be found by solving the equations $\sum_{i=1}^{N} U_i(\theta) = 0$. A sandwich estimator for the variance of $\hat{\theta}$ can then be obtained as:

$$V(\hat{\theta}) = A(\hat{\theta})^{-1} B(\hat{\theta}) \left[A(\hat{\theta})^{-1} \right]^T \tag{C.2}$$

where

$$A(\hat{\theta}) = \sum_{i=1}^{N} \frac{\partial U_i(\theta)}{\partial \theta}|_{\theta=\hat{\theta}} \tag{C.3}$$

and

$$B(\hat{\theta}) = \sum_{i=1}^{N} U_i(\theta) U_i(\theta)^T \tag{C.4}$$

For our example where regression calibration is applied using a linear (stage 1) model to a logistic regression (stage 2) outcome model, we can explicitly define these matrices as follows:

$$A(\theta) = \sum_{i=1}^{N} \begin{bmatrix} \frac{\partial U_{i1(a)}(\theta)}{\partial \alpha_0} & \frac{\partial U_{i1(a)}(\theta)}{\partial \alpha_X} & \frac{\partial U_{i1(a)}(\theta)}{\partial \alpha_Z} & \frac{\partial U_{i1(a)}(\theta)}{\partial \beta_0} & \frac{\partial U_{i1(a)}(\theta)}{\partial \beta_X} & \frac{\partial U_{i1(a)}(\theta)}{\partial \beta_Z} \\ \frac{\partial U_{i1(b)}(\theta)}{\partial \alpha_0} & \frac{\partial U_{i1(b)}(\theta)}{\partial \alpha_X} & \frac{\partial U_{i1(b)}(\theta)}{\partial \alpha_Z} & \frac{\partial U_{i1(b)}(\theta)}{\partial \beta_0} & \frac{\partial U_{i1(b)}(\theta)}{\partial \beta_X} & \frac{\partial U_{i1(b)}(\theta)}{\partial \beta_Z} \\ \frac{\partial U_{i1(c)}(\theta)}{\partial \alpha_0} & \frac{\partial U_{i1(c)}(\theta)}{\partial \alpha_X} & \frac{\partial U_{i1(c)}(\theta)}{\partial \alpha_Z} & \frac{\partial U_{i1(c)}(\theta)}{\partial \beta_0} & \frac{\partial U_{i1(c)}(\theta)}{\partial \beta_X} & \frac{\partial U_{i1(c)}(\theta)}{\partial \beta_Z} \\ \frac{\partial U_{i2(a)}(\theta)}{\partial \alpha_0} & \frac{\partial U_{i2(a)}(\theta)}{\partial \alpha_X} & \frac{\partial U_{i2(a)}(\theta)}{\partial \alpha_Z} & \frac{\partial U_{i2(a)}(\theta)}{\partial \beta_0} & \frac{\partial U_{i2(a)}(\theta)}{\partial \beta_X} & \frac{\partial U_{i2(a)}(\theta)}{\partial \beta_Z} \\ \frac{\partial U_{i2(b)}(\theta)}{\partial \alpha_0} & \frac{\partial U_{i2(b)}(\theta)}{\partial \alpha_X} & \frac{\partial U_{i2(b)}(\theta)}{\partial \alpha_Z} & \frac{\partial U_{i2(b)}(\theta)}{\partial \beta_0} & \frac{\partial U_{i2(b)}(\theta)}{\partial \beta_X} & \frac{\partial U_{i2(b)}(\theta)}{\partial \beta_Z} \\ \frac{\partial U_{i2(c)}(\theta)}{\partial \alpha_0} & \frac{\partial U_{i2(c)}(\theta)}{\partial \alpha_X} & \frac{\partial U_{i2(c)}(\theta)}{\partial \alpha_Z} & \frac{\partial U_{i2(c)}(\theta)}{\partial \beta_0} & \frac{\partial U_{i2(c)}(\theta)}{\partial \beta_X} & \frac{\partial U_{i2(c)}(\theta)}{\partial \beta_Z} \end{bmatrix}$$

$$= \sum_{i=1}^{N} \begin{bmatrix} -V_i & -X_i^* V_i & -Z_i V_i & 0 & 0 & 0 \\ -X_i^* V_i & -X_i^{*2} V_i & -X_i^* Z_i V_i & 0 & 0 & 0 \\ -Z_i V_i & -X_i^* Z_i V_i & -Z_i^2 V_i & 0 & 0 & 0 \\ -\beta_X p_i(1-p_i) & -\beta_X X_i^* p_i(1-p_i) & -\beta_X Z_i p_i(1-p_i) & -p_i(1-p_i) & -\hat{X}_i p_i(1-p_i) & -Z_i p_i(1-p_i) \\ \hat{X}_i(\beta_X p_i^2 - \beta_X p_i) + Y_i - p_i & \hat{X}_i(\beta_X X_i^* p_i^2 - \beta_X X_i^* p_i) + X_i^*(Y_i - p_i) & \hat{X}_i(\beta_X Z_i p_i^2 - \beta_X Z_i p_i) + Z_i(Y_i - p_i) & -\hat{X}_i p_i(1-p_i) & -(\hat{X}_i)^2 p_i(1-p_i) & -\hat{X}_i Z_i p_i(1-p_i) \\ -\beta_X Z_i p_i(1-p_i) & -\beta_X Z_i X_i^* p_i(1-p_i) & -\beta_X Z_i^2 p_i(1-p_i) & -Z_i p_i(1-p_i) & -Z_i(\hat{X}_i) p_i(1-p_i) & -Z_i^2 p_i(1-p_i) \end{bmatrix}$$

and

$$B(\theta) = \sum_{i=1}^{N} \begin{bmatrix} (U_{i1(a)}(\theta))^2 & U_{i1(a)}(\theta)U_{i1(b)}(\theta) & U_{i1(a)}(\theta)U_{i1(c)}(\theta) & U_{i1(a)}(\theta)U_{i2(a)}(\theta) & U_{i1(a)}(\theta)U_{i2(b)}(\theta) & U_{i1(a)}(\theta)U_{i2(c)}(\theta) \\ U_{i1(b)}(\theta)U_{i1(a)}(\theta) & (U_{i1(b)}(\theta))^2 & U_{i1(b)}(\theta)U_{i1(c)}(\theta) & U_{i1(b)}(\theta)U_{i2(a)}(\theta) & U_{i1(b)}(\theta)U_{i2(b)}(\theta) & U_{i1(b)}(\theta)U_{i2(c)}(\theta) \\ U_{i1(c)}(\theta)U_{i1(a)}(\theta) & U_{i1(c)}(\theta)U_{i1(b)}(\theta) & (U_{i1(c)}(\theta))^2 & U_{i1(c)}(\theta)U_{i2(a)}(\theta) & U_{i1(c)}(\theta)U_{i2(b)}(\theta) & U_{i1(c)}(\theta)U_{i2(c)}(\theta) \\ U_{i2(a)}(\theta)U_{i1(a)}(\theta) & U_{i2(a)}(\theta)U_{i1(b)}(\theta) & U_{i2(a)}(\theta)U_{i1(c)}(\theta) & (U_{i2(a)}(\theta))^2 & U_{i2(a)}(\theta)U_{i2(b)}(\theta) & U_{i2(b)}(\theta)U_{i2(c)}(\theta) \\ U_{i2(b)}(\theta)U_{i1(a)}(\theta) & U_{i2(b)}(\theta)U_{i1(b)}(\theta) & U_{i2(b)}(\theta)U_{i1(c)}(\theta) & U_{i2(b)}(\theta)U_{i2(a)}(\theta) & (U_{i2(b)}(\theta))^2 & U_{i2(c)}(\theta)U_{i2(c)}(\theta) \\ U_{i2(c)}(\theta)U_{i1(a)}(\theta) & U_{i2(c)}(\theta)U_{i1(b)}(\theta) & U_{i2(c)}(\theta)U_{i1(c)}(\theta) & U_{i2(c)}(\theta)U_{i2(a)}(\theta) & U_{i2(c)}(\theta)U_{i2(b)}(\theta) & (U_{i2(c)}(\theta))^2 \end{bmatrix}$$

$$= \sum_{i=1}^{N} \begin{bmatrix} V_i(X_i^{**} - \alpha_0 - \alpha_X X_i^* - \alpha_Z Z_i) \\ V_i(X_i^{**} - \alpha_0 - \alpha_X X_i^* - \alpha_Z Z_i)X_i^* \\ V_i(X_i^{**} - \alpha_0 - \alpha_X X_i^* - \alpha_Z Z_i)Z_i \\ Y_i - p_i \\ \hat{X}_i(Y_i - p_i) \\ Z_i(Y_i - p_i) \end{bmatrix} \begin{bmatrix} V_i(X_i^{**} - \alpha_0 - \alpha_X X_i^* - \alpha_Z Z_i) \\ V_i(X_i^{**} - \alpha_0 - \alpha_X X_i^* - \alpha_Z Z_i)X_i^* \\ V_i(X_i^{**} - \alpha_0 - \alpha_X X_i^* - \alpha_Z Z_i)Z_i \\ Y_i - p_i \\ \hat{X}_i(Y_i - p_i) \\ Z_i(Y_i - p_i) \end{bmatrix}^T$$

Table C.1: Simulation results are shown for the Cox proportional hazards regression model (event rate = 0.38) for data simulated to be from a simple random sample. The median percent (%) bias, median standard errors (ASE), empirical median absolute deviation (MAD) and coverage probabilities (CP) are given for 1000 simulated data sets for the outcome model fit to true exposure, naive exposure, calibrated exposure with naive (model-based) standard errors, calibrated exposure with standard errors from the sandwich approach, and calibrated exposure with standard errors from the bootstrap approach ($B = 500$ bootstrap samples). We vary the correlation between the error-prone and precisely-measured covariates (0.3 or 0.7), the sample size (N), and the measurement error variance (σ^2). Sample size of the calibration subset is $n = 450$.

			Low Correlation				High Correlation			
N	$\sigma^{2(1)}$	Method	% Bias	MAD	ASE	CP	% Bias	MAD	ASE	CP
1000	0.00	Truth	0.10	0.05	0.05	0.94	0.21	0.07	0.07	0.94
	0.25	Naive	-12.36	0.08	0.08	0.91	-42.22	0.09	0.09	0.52
		RC$^{(2)}$ (Naive SE)	-2.06	0.09	0.09	0.94	-2.54	0.15	0.15	0.95
		RC$^{(2)}$ (Sandwich)	—	—	0.09	0.95	—	—	0.16	0.95
		RC$^{(2)}$ (Bootstrap)	—	—	0.10	0.95	—	—	0.16	0.96
	0.50	Naive	-48.12	0.07	0.06	0.16	-67.79	0.07	0.07	0.02
		RC$^{(2)}$ (Naive SE)	-2.88	0.12	0.12	0.94	-3.04	0.21	0.20	0.95
		RC$^{(2)}$ (Sandwich)	—	—	0.12	0.95	—	—	0.21	0.96
		RC$^{(2)}$ (Bootstrap)	—	—	0.13	0.95	—	—	0.22	0.96
	1.00	Naive	-71.15	0.05	0.05	0.00	-82.64	0.05	0.05	0.00
		RC$^{(2)}$ (Naive SE)	-2.90	0.16	0.16	0.94	-4.26	0.29	0.28	0.95
		RC$^{(2)}$ (Sandwich)	—	—	0.17	0.95	—	—	0.29	0.97
		RC$^{(2)}$ (Bootstrap)	—	—	0.17	0.96	—	—	0.32	0.98
10000	0.00	Truth	-0.03	0.02	0.02	0.96	-0.29	0.02	0.02	0.96
	0.25	Naive	-12.42	0.03	0.03	0.52	-42.27	0.03	0.03	0.00
		RC$^{(2)}$ (Naive SE)	-2.81	0.04	0.03	0.82	-2.38	0.06	0.05	0.87
		RC$^{(2)}$ (Sandwich)	—	—	0.04	0.94	—	—	0.06	0.96
		RC$^{(2)}$ (Bootstrap)	—	—	0.04	0.95	—	—	0.07	0.96
	0.50	Naive	-47.64	0.02	0.02	0.00	-67.70	0.02	0.02	0.00
		RC$^{(2)}$ (Naive SE)	-3.32	0.05	0.04	0.81	-2.88	0.08	0.06	0.85
		RC$^{(2)}$ (Sandwich)	—	—	0.05	0.95	—	—	0.09	0.96
		RC$^{(2)}$ (Bootstrap)	—	—	0.06	0.95	—	—	0.09	0.96
	1.00	Naive	-70.94	0.01	0.02	0.00	-82.85	0.01	0.02	0.00
		RC$^{(2)}$ (Naive SE)	-3.22	0.08	0.05	0.81	-2.77	0.12	0.09	0.86
		RC$^{(2)}$ (Sandwich)	—	—	0.07	0.95	—	—	0.12	0.96
		RC$^{(2)}$ (Bootstrap)	—	—	0.08	0.95	—	—	0.14	0.97

[1] σ^2 = the variance of the random, normally distributed measurement error
[2] RC = Regression calibration

Table C.2: Simulation results are shown for logistic regression (event rate = 0.38) for the outcome model fit to true exposure, naive exposure, calibrated exposure with naive (model-based) standard errors, calibrated exposure with standard errors from the sandwich approach, and calibrated exposure with standard errors from the bootstrap approach, with bootstrap confidence intervals constructed in 3 ways for $B = 1000$ bootstrap samples. The median percent (%) bias, median standard errors (ASE), empirical median absolute deviation (MAD) and coverage probabilities (CP) are given for 1000 simulated data sets with $N = 1000$ each. We vary the correlation between the error-prone and precisely-measured covariates (0.3 or 0.7) and the measurement error variance (σ^2). Sample size of the calibration subset is $n = 450$.

		Low Correlation				High Correlation			
$\sigma^{2(1)}$	Method	% Bias	MAD	ASE	CP	% Bias	MAD	ASE	CP
0.00	Truth	0.04	0.07	0.07	0.94	0.16	0.09	0.10	0.94
0.25	Naive	-13.78	0.12	0.11	0.91	-42.84	0.12	0.12	0.68
	RC[(2)] (Naive SE)	-3.76	0.13	0.12	0.94	-3.28	0.21	0.20	0.94
	RC[(2)] (Sandwich)	—	—	0.12	0.94	—	—	0.20	0.94
	RC[(2)] (Boot. - Wald)[(3)]	—	—	0.12	0.94	—	—	0.21	0.94
	RC[(2)] (Boot. - Perc)[(4)]	—	—	0.12	0.93	—	—	0.21	0.93
	RC[(2)] (Boot. - BCA)[(5)]	—	—	0.12	0.94	—	—	0.21	0.94
0.50	Naive	-48.54	0.08	0.08	0.36	-68.36	0.09	0.09	0.13
	RC[(2)] (Naive SE)	-4.61	0.16	0.16	0.93	-2.90	0.28	0.27	0.94
	RC[(2)] (Sandwich)	—	—	0.16	0.93	—	—	0.27	0.95
	RC[(2)] (Boot. - Wald)[(3)]	—	—	0.16	0.94	—	—	0.29	0.96
	RC[(2)] (Boot. - Perc)[(4)]	—	—	0.16	0.93	—	—	0.29	0.94
	RC[(2)] (Boot - BCA)[(5)]	—	—	0.16	0.93	—	—	0.29	0.94
1.00	Naive	-71.90	0.06	0.06	0.01	-83.60	0.06	0.06	0.00
	RC[(2)] (Naive SE)	-5.31	0.22	0.21	0.94	-4.73	0.37	0.37	0.94
	RC[(2)] (Sandwich)	—	—	0.21	0.94	—	—	0.37	0.95
	RC[(2)] (Boot. - Wald)[(3)]	—	—	0.22	0.96	—	—	0.41	0.98
	RC[(2)] (Boot. - Perc)[(4)]	—	—	0.22	0.92	—	—	0.41	0.94
	RC[(2)] (Boot. - BCA)[(5)]	—	—	0.22	0.93	—	—	0.41	0.94

[(1)] σ^2 = the variance of the random, normally distributed measurement error
[(2)] RC = Regression calibration
[(3)] Bootstrap with standard normal Wald-based confidence interval
[(4)] Percentile bootstrap confidence interval
[(5)] Bias-corrected and accelerated (BCA) bootstrap interval

BIBLIOGRAPHY

Per Kragh Andersen and Richard D Gill. Cox's regression model for counting processes: a large sample study. *The Annals of Statistics*, pages 1100–1120, 1982.

Michael Baiocchi, Jing Cheng, and Dylan S Small. Instrumental variable methods for causal inference. *Statistics in Medicine*, 33(13):2297–2340, 2014.

Raji Balasubramanian and Stephen W Lagakos. Estimation of the timing of perinatal transmission of hiv. *Biometrics*, 57(4):1048–1058, 2001.

Raji Balasubramanian and Stephen W Lagakos. Estimation of a failure time distribution based on imperfect diagnostic tests. *Biometrika*, 90(1):171–182, 2003.

Pedro L Baldoni, Daniela Sotres-Alvarez, Thomas Lumley, and Pamela A Shaw. On the use of regression calibration in a complex sampling design with application to the Hispanic Community Health Study/Study of Latinos. *American Journal of Epidemiology*, 2021.

Reuben M Baron and David A Kenny. The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51(6):1173, 1986.

David A Binder. On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review/Revue Internationale de Statistique*, pages 279–292, 1983.

Lillian A Boe, Lesley F Tinker, and Pamela A Shaw. An approximate quasi-likelihood approach for error-prone failure time outcomes and exposures. *Statistics in Medicine*, 40 (23):5006–5024, 2021.

DD Boos and LA Stefanski. *Essential Statistical Inference: Theory and Methods*. Springer, New York, NY, 2013.

John P Buonaccorsi. *Measurement Error: Models, Methods, and Applications*. Chapman and Hall/CRC, Boca Raton, FL, 2010.

Raymond J Carroll, David Ruppert, Leonard A Stefanski, and Ciprian M Crainiceanu. *Measurement Error in Nonlinear Models: A Modern Perspective*. Chapman and Hall/CRC, Boca Raton, FL, 2006.

RJ Carroll, Suojin Wang, DG Simpson, AJ Stromberg, and D Ruppert. The sandwich (robust covariance matrix) estimator. *Technical Report*, 1998.

George Casella and Roger L Berger. *Statistical Inference*. Duxbury, Pacific Grove, CA, 2nd edition, 2002.

Centers for Disease Control and Prevention. National diabetes statistics report, 2017. *Atlanta, GA: Centers for Disease Control and Prevention, US Department of Health and Human Services*, 20:1–20, 2017.

ASC Conlon, JMG Taylor, and DJ Sargent. Improving efficiency in clinical trials using auxiliary information: Application of a multi-state cure model. *Biometrics*, 71(2):460–468, 2015.

Anthony Christopher Davison and David Victor Hinkley. *Bootstrap methods and their application*. Cambridge University Press, 1997.

B Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1): 1–26, 1979.

Bradley Efron. Better bootstrap confidence intervals. *Journal of the American statistical Association*, 82(397):171–185, 1987.

Michael P Fay and Pamela A Shaw. Exact and asymptotic weighted logrank tests for interval censored data: the interval R package. *Journal of Statistical Software*, 36(2):1–38, 2010.

Thomas R Fleming, Ross L Prentice, Margaret S Pepe, and David Glidden. Surrogate and auxiliary endpoints in clinical trials, with potential applications in cancer and aids research. *Statistics in Medicine*, 13(9):955–968, 1994.

Robert V Foutz. On the unique consistent solution to the likelihood equations. *Journal of the American Statistical Association*, 72(357):147–148, 1977.

Manfred S Green and Michael J Symons. A comparison of the logistic risk function and the proportional hazards model in prospective epidemiologic studies. *Journal of Chronic Diseases*, 36(10):715–723, 1983.

Rolf HH Groenwold, A Rogier T Donders, Kit CB Roes, Frank E Harrell Jr, and Karel GM Moons. Dealing with missing outcome data in randomized trials and observational studies. *American Journal of Epidemiology*, 175(3):210–217, 2011.

Xiangdong Gu, Yunsheng Ma, and Raji Balasubramanian. Semiparametric time to event models in the presence of error-prone, self-reported outcomes—With application to the Women's Health Initiative. *The Annals of Applied Statistics*, 9(2):714, 2015.

Frank E Harrell Jr. *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. Springer, 2015.

Elizabeth M Hashimoto, Edwin MM Ortega, Gilberto A Paula, and Mauricio L Barreto. Regression models for grouped survival data: Estimation and sensitivity analysis. *Computational Statistics & Data Analysis*, 55(2):993–1007, 2011.

Ruth H Keogh and Ian R White. A toolkit for measurement error correction, with a focus on nutritional epidemiology. *Statistics in Medicine*, 33(12):2137–2155, 2014.

Ruth H Keogh, Pamela A Shaw, Paul Gustafson, Raymond J Carroll, Veronika Deffner, Kevin W Dodd, Helmut Küchenhoff, Janet A Tooze, Michael P Wallace, Victor Kipnis, and LS Freedman. STRATOS guidance document on measurement error and misclassification of variables in observational epidemiology: Part 1—basic theory and simple methods of adjustment. *Statistics in Medicine*, 39(16):2197–2231, 2020.

Victor Kipnis, Douglas Midthune, Dennis W Buckman, Kevin W Dodd, Patricia M Guenther, Susan M Krebs-Smith, Amy F Subar, Janet A Tooze, Raymond J Carroll, and Laurence S Freedman. Modeling data with excess zeros and measurement error: application to evaluating relationships between episodically consumed foods and health outcomes. *Biometrics*, 65(4):1003–1010, 2009.

Robert D Langer, Emily White, Cora E Lewis, Jane M Kotchen, Susan L Hendrix, and Maurizio Trevisan. The Women's Health Initiative Observational Study: baseline characteristics of participants and reliability of baseline measures. *Annals of Epidemiology*, 13 (9):S107–S121, 2003.

Lisa M LaVange, William D Kalsbeek, Paul D Sorlie, Larissa M Avilés-Santa, Robert C Kaplan, Janice Barnhart, Kiang Liu, Aida Giachello, David J Lee, John Ryan, et al. Sample design and cohort selection in the Hispanic Community Health Study/Study of Latinos. *Annals of Epidemiology*, 20(8):642–649, 2010.

Jane C Lindsey and Louise M Ryan. Methods for interval-censored data. *Statistics in Medicine*, 17(2):219–238, 1998.

Thomas Lumley. *Complex surveys: a guide to analysis using R*, volume 565. John Wiley & Sons, 2011.

Thomas Lumley and Alastair Scott. Fitting regression models to survey data. *Statistical Science*, pages 265–278, 2017.

Amalia S Magaret. Incorporating validation subsets into discrete proportional hazards models for mismeasured outcomes. *Statistics in Medicine*, 27(26):5456–5470, 2008.

Karen L Margolis, Lihong Qi, Robert Brzyski, Denise E Bonds, Barbara V Howard, Sarah Kempainen, Simin Liu, Jennifer G Robinson, Monika M Safford, Lesley T Tinker, et al. Validity of diabetes self-reports in the Women's Health Initiative: comparison with medication inventories and fasting glucose measurements. *Clinical Trials*, 5(3):240–247, 2008.

Amalia S Meier, Barbra A Richardson, and James P Hughes. Discrete proportional hazards models for mismeasured outcomes. *Biometrics*, 59(4):947–954, 2003.

Yasmin Mossavar-Rahmani, Pamela A Shaw, William W Wong, Daniela Sotres-Alvarez, Marc D Gellman, Linda Van Horn, Mark Stoutenberg, Martha L Daviglus, Judith Wylie-Rosett, Anna Maria Siega-Riz, et al. Applying recovery biomarkers to calibrate self-report measures of energy and protein in the Hispanic Community Health Study/Study of Latinos. *American Journal of Epidemiology*, 181(12):996–1007, 2015.

Marian L Neuhouser, Lesley Tinker, Pamela A Shaw, Dale Schoeller, Sheila A Bingham, Linda Van Horn, Shirley AA Beresford, Bette Caan, Cynthia Thomson, Suzanne Satterfield, et al. Use of recovery biomarkers to calibrate nutrient consumption self-reports in the Women's Health Initiative. *American Journal of Epidemiology*, 167(10):1247–1259, 2008.

Meng Ning, Qiang Zhang, and Min Yang. Comparison of self-reported and biomedical data on hypertension and diabetes: findings from the China Health and Retirement Longitudinal Study (CHARLS). *BMJ Open*, 6(1):e009836, 2016.

E. J. Oh, B. E. Shepherd, T Lumley, and P. A. Shaw. Raking and regression calibration: Methods to address bias from correlated covariate and time-to-event error. *arXiv preprint arXiv:1905.08330*, pages 1–49, 2019.

Margaret Sullivan Pepe. Inference using surrogate outcome data and a validation sample. *Biometrika*, 79(2):355–365, 1992.

Ross L Prentice. Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika*, 69(2):331–342, 1982.

Ross L Prentice, Mary Pettinger, Marian L Neuhouser, Daniel Raftery, Cheng Zheng, GA Nagana Gowda, Ying Huang, Lesley F Tinker, Barbara V Howard, JoAnn E Manson, et al. Biomarker-calibrated macronutrient intake and chronic disease risk among postmenopausal women. *The Journal of Nutrition*, 151(8):2330–2341, 2021.

Ross L Prentice, Mary Pettinger, Cheng Zheng, Marian L Neuhouser, Daniel Raftery, GA Gowda, Ying Huang, Lesley F Tinker, Barbara V Howard, JoAnn E Manson, et al. Biomarkers for components of dietary protein and carbohydrate with application to chronic disease risk among postmenopausal women. *The Journal of Nutrition*, 2022.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018. URL https://www.R-project.org/.

Cheryl Ritenbaugh, Ruth E Patterson, Rowan T Chlebowski, Bette Caan, Lesley Fels-Tinker, Barbara Howard, and Judy Ockene. The Women's Health Initiative Dietary Modification trial: overview and baseline characteristics of participants. *Annals of Epidemiology*, 13(9):S87–S97, 2003.

Paul Rogers and Julie Stoner. Modification of the sandwich estimator in generalized esti-

mating equations with correlated binary outcomes in rare event and small sample settings. *American Journal of Applied Mathematics and Statistics*, 3(6):243, 2015.

B Rosner, WC Willett, and D Spiegelman. Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. *Statistics in Medicine*, 8(9):1051–1069, 1989.

B Rosner, D Spiegelman, and WC Willett. Correction of logistic regression relative risk estimates and confidence intervals for measurement error: the case of multiple covariates measured with error. *American Journal of Epidemiology*, 132(4):734–745, 1990.

Carl-Erik Särndal, Bengt Swensson, and Jan Wretman. *Model assisted survey sampling*. Springer Science & Business Media, 2003.

Andrea LC Schneider, James S Pankow, Gerardo Heiss, and Elizabeth Selvin. Validity and reliability of self-reported diabetes in the atherosclerosis risk in communities study. *American Journal of Epidemiology*, 176(8):738–743, 2012.

Baiju R Shah and Douglas G Manuel. Self-reported diabetes is associated with self-management behaviour: a cohort study. *BMC Health Services Research*, 8(1):142, 2008.

Pamela A Shaw and Ross L Prentice. Hazard ratio estimation for biomarker-calibrated dietary exposures. *Biometrics*, 68(2):397–407, 2012.

Pamela A Shaw, Veronika Deffner, Ruth H Keogh, Janet A Tooze, Kevin W Dodd, Helmut Küchenhoff, Victor Kipnis, and Laurence S Freedman. Epidemiologic analyses with error-prone exposures: Review of current practice and recommendations. *Annals of Epidemiology*, 27(11):821–828, 2018.

Pamela A Shaw, Paul Gustafson, Raymond J Carroll, Veronika Deffner, Kevin W Dodd, Ruth H Keogh, Victor Kipnis, Janet A Tooze, Michael P Wallace, Helmut Küchenhoff, and LS Freedman. STRATOS guidance document on measurement error and misclassification of variables in observational epidemiology: Part 1—basic theory and simple methods of adjustment. *Statistics in Medicine*, 39(16):2232–2263, 2020.

Bryan E Shepherd and Pamela A Shaw. Errors in multiple variables in human immunodeficiency virus (hiv) cohort and electronic health record data: statistical challenges and opportunities. *Statistical Communications in Infectious Diseases*, 12(s1), 2020.

Bryan E Shepherd and Chang Yu. Accounting for data errors discovered from an audit in multiple linear regression. *Biometrics*, 67(3):1083–1091, 2011.

Paul D Sorlie, Larissa M Avilés-Santa, Sylvia Wassertheil-Smoller, Robert C Kaplan, Martha L Daviglus, Aida L Giachello, Neil Schneiderman, Leopoldo Raij, Gregory Talavera, Matthew Allison, et al. Design and implementation of the hispanic community

health study/study of latinos. *Annals of Epidemiology*, 20(8):629–641, 2010.

Donna Spiegelman, Sebastian Schneeweiss, and Aidan McDermott. Measurement error correction for logistic regression models with an "alloyed gold standard". *American Journal of Epidemiology*, 145(2):184–196, 1997.

The Women's Health Initiative Study Group. Design of the Women's Health Initiative clinical trial and observational study. *Controlled Clinical Trials*, 19(1):61–109, 1998.

Lesley F Tinker, Gloria E Sarto, Barbara V Howard, Ying Huang, Marian L Neuhouser, et al. Biomarker-calibrated dietary energy and protein intake associations with diabetes risk among postmenopausal women from the Women's Health Initiative. *The American Journal of Clinical Nutrition*, 94(6):1600–1606, 2011.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge University Press, 2000.

Grace Y Yi. *Statistical analysis with measurement error or misclassification: strategy, method and application*. Springer, 2017.

Jarcy Zee, Sharon X Xie, and Alzheimer's Disease Neuroimaging Initiative. Assessing treatment effects with surrogate survival outcomes using an internal validation subsample. *Clinical Trials*, 12(4):333–341, 2015.

Donglin Zeng, Lu Mao, and DY Lin. Maximum likelihood estimation for semiparametric transformation models with interval-censored data. *Biometrika*, 103(2):253–271, 2016.

Ying Zhang, Lei Hua, and Jian Huang. A spline-based semiparametric maximum likelihood estimation method for the cox model with interval-censored data. *Scandinavian Journal of Statistics*, 37(2):338–354, 2010.

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