Bridging Statistical Strategies for Censored Covariates

Tanya P. Garcia
University of North Carolina at Chapel Hill

Yanyuan Ma
Penn State University

Joan Hu
Simon Fraser University

January 28–February 2, 2024

1 Overview of the Field of Censored Covariates

Survival analysis, born over 360 years ago from the mortality studies of Graunt and Halley, originally focused on time to death. Since then, it has broadened to include any time to event, such as time to disease onset. Various statistical methods have emerged for this analysis, and have since been applied in fields like actuarial science, biomedicine, and engineering. Yet, all of these methods, and the applications they are used for, are designed for settings where the outcome in the statistical model is censored. Now, with evolving biomedical challenges, a new modeling challenge has emerged: dealing with censored covariates rather than outcomes. This challenge arises, for example, in the design of clinical trials for Huntington disease, a genetically inherited neurodegenerative disease. A key component of that clinical trial design is estimating how symptoms worsen over time—the symptom trajectory—before and after a diagnosis. Yet estimating the symptom trajectory is not easy because Huntington disease progresses slowly over decades, so studies that track symptoms often end before a diagnosis can be made. This leaves researchers with the challenge of estimating the symptom trajectory as a function of a right-censored covariate, time to diagnosis.

Various strategies now exist to tackle the censored covariate problem. One strategy is a complete case analysis, in which we analyze only the subset of data that was not censored.[1,2] While deleting data is generally discouraged in statistics due to potential loss of efficiency, complete case analysis for a censored covariate problem can still provide a consistent but less efficient estimator for model parameters under certain conditions.[3] A second strategy is weighting methods (i.e., inverse probability weighting) which assigns weights to observations to reflect their contribution to the analysis.[4,5] A third strategy is imputation which estimates and fills in missing values due to censoring, and allows for standard statistical analysis techniques to be applied to the completed data set.[6–15] A fourth strategy is likelihood-based methods which handles covariate censoring by placing models on all of the random variables involved and then maximizing the resulting likelihood. Parametric[8,10,16–22] and semiparametric models[5,22–25] have so far been considered in likelihood-based methods. A fifth strategy is Bayesian methods which use joint probability distributions to relate data and unknowns. Prior distributions are used to characterize initial uncertainty about model parameters and the conditional distribution of the censored covariate. Then, posterior distributions are computed for estimation and inference, and the complexity of that estimation and inference varies depending on the chosen priors.[10,26–30] A sixth, and

final, strategy is threshold methods which offer a solution to dealing with covariate censoring by replacing the censored covariate with a dichotomized version based on a threshold.[31] Two main approaches, deletion and complete threshold methods, have so far been considered and have been shown to yield consistent estimators under certain conditions. The choice between the thresholding methods depends on the primary interest of the analysis and efficiency considerations.[31]

A recent review paper[32] led by Dr. Garcia, discusses each method's strengths and weaknesses, along with recommendations for researchers. There have been significant developments in methods to manage censored covariates, but there is no one-size-fits-all solution and many pressing issues still remain to be resolved. This workshop brought researchers together to start tackling those issues.

2 Recent Developments and Open Problems

2.1 The Censored Covariate Problem

Mathematically, a right-censored covariate model in its simplest form is

$$Y = m(X, \boldsymbol{Z}; \boldsymbol{\theta}) + \epsilon, \quad \epsilon \sim \text{Normal}(0, \sigma_\epsilon^2),$$

where the outcome Y represents symptom severity scores, the covariate X represents time to diagnosis, the covariate $\boldsymbol{Z}$ represents other patient measures, and $m(\cdot)$ represents a linear or nonlinear function known up to a parameter $\boldsymbol{\theta}$. Because of right-censoring, rather than observe X, we observe $W = \min(X, C)$ and $\Delta = I(X \leq C)$, where C is a random censoring time and Δ is the censoring indicator. The goal is to derive a consistent estimator for $\boldsymbol{\beta} \equiv (\boldsymbol{\theta}^T, \sigma_\epsilon^2)^T$ using the observed data $(Y, W, \boldsymbol{Z}, \Delta)$.

2.2 Open Problem #1: Noninformative Covariate Censoring

One common censoring mechanism known as *noninformative censoring* is when X and C are statistically independent. modeling where the outcome is the time-to-event, the independence lets researchers drop the distribution for censoring when estimating.[33] Yet when the time-to-event is the *covariate* as in our case, we cannot just drop the distribution for censoring as it impacts model estimation, since the estimation will involve the probabilities of X being censored, $\text{pr}(X > C)$, and not being censored, $\text{pr}(X < C)$, both of which involve the distributions for X and C. Therefore, with a censored covariate problem, we have two models that could be wrong: the distribution for X and the distribution for C. Mismodeling either or both distributions can produce biased estimates. A strategy to remove the bias will help researchers improve the accuracy and reliability of statistical models dealing with censored covariate data.

2.3 Open Problem #2: Informative Covariate Censoring

The vast majority of (right-) censored covariate models assume that censoring is noninformative.[11–13, 15, 20, 23, 26, 34–42] Such an assumption is often defensible, but this assumption is highly suspect when censoring occurs because, for example, a patient drops out due to worsening symptoms. In this type of censoring, known as *informative censoring*, X and C are statistically dependent, and failing to adjust for informative censoring will lead to biased results. An ongoing challenge with this problem is simultaneously modeling

2.4 Open Problem #3: Interval Covariate Censoring

Much of the published literature on censored covariate problems has dealt with right- and left-censored covariate problems, but interval covariate censoring has been studied only to a limited extent. Yet, interval covariate censoring is prevalent in many biomedical studies and addressing this problem could open pathways for researchers dealing with censored covariate data.

2.5 Open Problem #4: Extensions of Censored Covariate Problems

We have so far presented the censored covariate problem in the context of regression models. Yet, a primary interest in the field is adapting techniques used for cross-sectional regression models to longitudinal, survival,

and quantile regression models with censored covariates. Extensions to longitudinal data requires addressing typical challenges, including correlation from repeated measurements. Extensions to survival data entails capturing multiple censoring mechanisms simultaneously—one for the outcome and one for each censored covariate. This may necessitate additional assumptions about how the censoring mechanisms interact. Extensions to quantile regression models require appropriate modeling and adaptation of estimation procedures.

2.6 Open Problem #5: Relationship Between Censored, Missing, and Mismeasured Covariates

Incomplete data are almost inevitable in many applications: data might not be measured (i.e., data are missing), might be mismeasured (i.e., data are measured with error), or might not be captured yet (i.e., data are censored, as in time-to-event analyses). There are unique nuances to how we handle missing data, measurement error, and censoring, each with its own growing body of literature. However, these contexts share similarities that, if unified, could help analysts understand how and when a method for one type of incomplete data can be applied to another type. Identifying such connections is an open problem in these areas.

3 Workshop Structure and Diversity

3.1 Overview of the Workshop

Our workshop was fully in-person with 38 participants. The content of our workshop included:

1. **One opening session** in which participants played the game "Yes, and!" In the game, participants shared fun facts about themselves, and found connections with each other by identifying facts that were common between participants. This game helped initiate a sense of community. To deepen that community-feel, we also established three workshop values: (i) Curiosity: participants were encouraged to ask any and all questions, no matter how basic or out-of-left field, so that we could better understand the challenges in the field of censored covariates and then brainstorm solutions to overcome those challenges; (ii) Open circles: participants were encouraged to invite people into their circles during lunches and breaks so that everyone felt welcomed; and (iii) Presence: participants were encouraged to minimize outside distractions and fully engage in the workshop sessions.

2. **Sixteen research talks** that were delivered by leaders in the fields of covariates that are censored, missing, and measured with error. While there are unique differences between these three areas, all are forms of incomplete data. We sought to identify and understand their connections to evaluate if methods for one type of incomplete data can be applied to another type.

3. **Three "think tank" sessions** gave research participants the opportunity to meet in smaller groups to discuss ideas and refine them based on talks presented during the workshop. Groups were composed of individuals at different career levels and research expertise.

4. **Three "professional development" sessions** were delivered by Dr. Garcia on skills to help the research participants tackle three common issues that pop up in research careers. The sessions were on: (i) Creativity and clarity: participants learned a specific iterative process, rather than a haphazard one, of how to map out clear, articulated research ideas; (ii) How to listen: participants learned how to listen for a person's emotions and values in a conversation so as to form deeper connections with that person; (iii) How to coach: participants learned how to ask questions to help their mentees (or colleagues) navigate difficult situations and find solutions that work for them.

3.2 Diversity of Participants

3.2.1 How Diversity is Defined

Data from the American Statistical Association show that between 2011-2020, the racial demographics of those who pursued a PhD in statistics and biostatistics was approximately 5% for African Americans, 5%

for Hispanics, and 0.4% for American Indians, Alaska Natives, Native Hawaiians, or Pacific Islanders.[43] Women fared better, in that 50% of individuals who earned a PhD in statistics and biostatistics were women.[44] However, there is a hole in the academic career pipeline: despite women representing half of statisticians/ biostatisticians with a PhD, only four out of ten are tenured non-full professors, and only two out of ten are tenured full professors.[44]

These numbers indicate that while people of color continue to be under-represented in statistics and biostatistics, so too are women in tenured positions. We therefore defined under-represented groups based on race, ethnicity, and gender, in addition to factors defined by BIRS's commitment to equity, diversity, and inclusion.

3.2.2 Diversity in the Organizing Committee

The organizing committee was an all-female, tenured-faculty team: Tanya Garcia is an Associate Professor of Biostatistics at the University of North Carolina at Chapel Hill in North Carolina, Yanyuan Ma is a Professor of Statistics at Penn State in Pennsylvania, and Joan Hu is a Professor of Statistics at Simon Fraser University in British Columbia.

The lead organizer, Dr. Garcia, is also a Hispanic, early-career researcher. BIRS defines early-career researchers as individuals within ten years of their doctoral degree. Tanya Garcia earned her PhD in August 2011, making her above the ten-year cut-off. However, since 2011, she had two kids, one of whom was born in 2020, during the global pandemic when childcare options were either unavailable or severely limited. These life circumstances are recognized by the Natural Sciences and Engineering Research Council of Canada as reasons to extend the early-career status, and thus qualify Tanya Garcia as an early-career researcher.

3.2.3 Selection and Diversity of the Workshop Participants

The organizing committee invited participants to span the range of (i) career level (from graduate students to tenured faculty), (ii) diverse populations (i.e., including but not limited to race, ethnicity, and gender), (iii) geographic region, and (iv) sub-specialty expertise for censored covariates.

Of our 38 participants, 50% were from under-represented groups. These participants represented:

- Four career levels: graduate student, tenure-track faculty, tenured faculty, and mathematical statisticians.
- Ten nationalities of citizenship: United States, Canada, Mexico, Spain, Belgium, Congo, Singapore, China, India, and Australia.
- Thirteen geographic locations: United States (Northeast, Midwest, South, and West), Canada (British Columbia, Alberta, Manitoba, and Ontario), Spain (Catalonia), Belgium, Switzerland (Geneva Canton), and Australia (New South Whales.
- Experts in five sub-specialties: censored covariates, noninformative and informative censoring with time-to-event modeling, model-free methods for covariates that are measured with error or have missing data, surrogate markers, and Bayesian methods.

4 Presentation Highlights

4.1 Discussions of Open Problem #1: Noninformative Covariate Censoring

1. **Dr. Roland A. Matsouka** presented a talk titled "Regression with a right-censored predictor using the inverse probability weighting methods. In his talk, Dr. Matsouka shared the effectiveness of inverse probability weighting (IPW) methods in handling right-censored covariates in regression analysis. Dr. Matsouka shared three different IPW approaches—inverse censoring probability weights, Kaplan-Meier weights, and Cox proportional hazards weights—and highlighted that these methods, particularly those based on the Cox proportional hazards model, can significantly improve the estimation accuracy in regression models with right-censored covariates. This improvement is because IPW methods adjust the analysis to account for the censoring, and as a result, IPW methods can help mitigate bias and enhance the robustness of statistical estimates. Dr. Matsouka applied the IPW methods to a case study using data from the Framingham Heart Study, in which he showed ways to estimate the

relationship between the age of onset for cardiovascular events and levels of low-density lipoprotein among cigarette smokers. He concluded his talk encouraging further exploration and development of IPW methods and other innovative approaches to address the challenges posed by censored covariates in regression analysis.

2. **Dr. Sarah Lotspeich** presented a talk titled "Extrapolation before imputation reduces bias when imputing heavily censored covariates." Dr. Lotspeich presented a new technique that focuses on the idea of "extrapolating, then imputing" to reduce bias when imputing censored covariates under noninformative covariate censoring. Dr. Lotspeich highlighted a parametric Weibull model to perform the extrapolation, and several participants shared that nonparametric extrapolation approaches may be considered to enhance the accuracy of the extrapolation component.

3. **Dr. Jing Qian** presented a talk titled "Threshold regression to accommodate a censored covariate." Dr. Qian discussed the challenge of handling noninformative covariate censoring in regression modeling, which is common, for example, in Alzheimer's disease studies when investigating the association between brain amyloid levels and maternal age of dementia onset. Dr. Qian proposed a threshold regression approach for linear regression models with censored covariates. These methods allow for immediate significance testing and unbiased estimation of regression coefficients. He highlighted the asymptotic properties of these estimators and pointed out that the asymptotic properties only require mild regularity conditions. Additionally, he described practical strategies for selecting the threshold. He shared that there is an R package called censCov to implement these methods.

4.2 Discussions of Open Problem #2: Informative Covariate Censoring

Mr. Jesus Vazquez presented a talk titled "Evaluating the robustness and efficiency of estimators for informative covariate censoring." Mr. Vazquez focused on the challenges of developing consistent and efficient estimators with censored covariate problems when the covariate censoring is informative. This problem was tackled by using various estimators, including the Complete Case (CC) estimator, Inverse Probability Weighting (IPW) estimator, Modified Augmented CC Estimator, Modified Augmented IPW Estimator, and Full Likelihood Estimator. Each estimator was analyzed in terms of its advantages and limitations, with theoretical frameworks and equations provided along with discussions on their properties. Mr. Vazquez evaluated these estimators in different simulation studies, where he assessed the performance of different estimators under various scenarios. He then showed results of applying the methods to the ENROLL-HD study, an observational study of Huntington disease.

4.3 Discussions of Open Problem #3: Interval Covariate Censoring

1. **Dr. Guadalupe Gomez Melis** presented a talk titled "Regression and goodness-of-fit with interval-censored covariates." The talk was motivated by the fact that interval-censored observations frequently arise in medical studies, particularly in longitudinal settings where the response variable is the elapsed time until an event of interest. Ad hoc methods like using the midpoint of interval-censored covariates and applying ordinary least squares are generally invalid. Instead, Dr. Melis proposed a likelihood-based approach combined with a two-step algorithm to jointly estimate regression coefficients and the marginal distribution of the covariate. She showed that the method, demonstrated with simulations and data from an HIV/AIDS clinical trial, yields asymptotically normal estimators. Dr. Melis also discussed the importance of residual analysis for evaluating the goodness of a fitted model, ensuring correct specification of regression functions and the consistency of estimators. While residual analysis for uncensored data is standard, it is less developed for censored covariate data. To address this gap, Dr. Melis proposed a new definition of residuals for linear models incorporating interval censored covariates, applicable even when the response variable is interval-censored. Compared to existing methods, these new residuals demonstrate better performance in model checking.

2. **Dr. Ezra Morrison** presented a talk titled "Regression with Interval Censored Covariates: Application to Cross-Sectional Incidence Estimation." The work for this talk was motivated by the interest to estimate HIV incidence rates, which is important for efficient allocation of public health resources

and assessing the impact of interventions. While the traditional approach involves longitudinal cohort studies, which are costly and time-consuming, the alternative approach discussed in the talk is cross-sectional surveys. These surveys involve recruiting a representative subset of the population and testing for biomarkers indicating recent infection. Dr. Morrison explained the process of modeling HIV biomarkers by infection duration using longitudinal calibration datasets with known infection dates. He considered various biomarkers, sequential testing, and dichotomized versions. Additionally, he presented ways to estimate the incidence rates from biomarkers and infection duration, along with the challenges posed by interval-censored data in the study design. He demonstrated the effectiveness of his methods in simulation studies under different scenarios, and compared the performance of estimation methods such as midpoint imputation, uniform imputation, and joint modeling, with joint modeling showing promising results.

4.4 Discussions of Open Problem #4: Extensions of Censored Covariate Problems

1. **Dr. Ingrid Van Keilegom** presented a talk titled "Quantile Regression with Censored Covariates." Dr. Van Keilegom began her talk introducing quantile regression, which estimates conditional quantiles of the response variable across different values of predictors. An advantage of quantile regression is that it provides a broader view of the conditional distribution, and is especially useful for analyzing the tails of the distribution. Her main model of interest was $Y = \boldsymbol{Z}^{\top}\boldsymbol{\beta}(\tau) + \epsilon$ where the quantile function is $Q_{\epsilon|\boldsymbol{Z}}(\tau) = 0$. Quantile regression involves using the check function (i.e., $\rho_\tau(u) = u\{\tau - I(u < 0)\}$), and she explained that there is a one-to-one correspondence between the check function and the asymmetric Laplace density (ALD). Dr. Van Keilegom explained that quantile regression can be viewed as maximizing the likelihood of the asymmetric Laplace density. However, when the outcome or the covariate is censored, the equivalence between the check function and ALD breaks down. Moreover, naively correcting the ALD likelihood for censoring will lead to inconsistent parameter estimation. Instead, Dr. Van Keilegom proposed an enriched Laplace distribution (ELD) as a way to handle censored data in quantile regression. ELD enriches ALD with polynomial expansions to approximate any continuous density. She proposed a two-step likelihood-based estimation procedure that involves estimation of the density function for censored covariates and then uses these estimates to maximize the likelihood for the response variable. She used simulation studies and an application to real data on drug treatment for liver disease to demonstrate the effectiveness of the proposed method. Dr. Van Keilegom highlighted that the methodology can be extended to handle more complex scenarios, including multiple censored covariates and flexible models for covariates.

2. **Dr. Molin Wang** presented a talk titled "Survival analysis when a binary covariate is measured at sparse time points." This work focused on studying the relationship between a time-dependent binary covariate and survival time. The covariate represents a binary exposure or treatment that may change over time, such as aspirin initiation in colorectal cancer (CRC) patients. The motivation stems from the need to understand how aspirin initiation affects survival among CRC patients. The study utilized data from the Nurses' Health Study and Health Professionals Follow-Up cohorts, comprising 1371 CRC cases with 249 CRC-related deaths over 10 years. The binary exposure status, aspirin initiation, was collected biennially, leading to sparse time points for measurement. Dr. Wang used statistical methods, including Cox proportional hazards models, to analyze the data, considering various covariates such as family history of CRC, body mass index, age at diagnosis, sex, tumor stage, and location. She discussed the challenges of handling time-dependent covariates, including interval censoring and sparse measurement points, and proposed a novel calibration framework for survival analysis to address these issues. Simulation studies and real-data analyses demonstrated the effectiveness of the proposed methods in estimating the association between aspirin initiation and survival time, considering different scenarios and subpopulations based on prior clinical research. She concluded her talk with points on model selection, calibration techniques, and potential applications of the proposed framework in other research settings.

4.5 Discussions of Open Problem #5: Relationship Between Censored, Missing, and Mismeasured Covariates

1. **Dr. Layla Parast** presented a talk titled "Surrogacy Validation with Censored or Missing Surrogates." Surrogate markers in clinical trials are variables or measurements that are used as substitutes for primary outcomes. Surrogate markers are particularly valuable because they can potentially expedite the drug approval process by allowing for shorter follow-up periods and quicker assessments of treatment efficacy. However, the use of surrogate markers introduces challenges, especially when dealing with censored or missing data. Strategies to tackle these challenges have yet to be developed.

2. **Dr. Jianxuan Liu** presented a talk titled "Semiparametric approach for causal inference with corrupted covariate." Dr. Liu discussed the challenges in power grid reliability, emphasizing the complexity of handling covariates that are high-dimensional and corrupted by measurement error. The measurement error stems from data that are self-reported, have recall bias, and are misreported because the data are responses to sensitive questions. Typical measurement error problems are addressed using multiple imputation, SIMEX, or a propensity score model. In this work, Dr. Liu used semiparametric methods to find estimators based on an efficient score. She also introduced a dimension reduction technique and a beta regression model for analyzing bounded outcomes.

3. **Dr. Xavier de Luna** presented a talk titled "Uniformly valid inference in high-dimensional settings." His talk began with outlining recent advancements in achieving uniformly valid inference for low-dimensional causal parameters within high-dimensional nuisance models, which is common in causal inference applications. He reviewed the literature on uniformly valid causal inference and discussed the trade-offs involved in using such procedures, highlighting the prevalence of naive and invalid post-model selection inference in statistical practice. He pointed out that uniformly valid inference are particularly crucial in high-dimensional settings where regularization type estimators like lasso are commonly used. He pointed out that approaches yielding uniformly valid inference converge uniformly in distribution over various data generating mechanisms, and often require sparsity conditions for nuisance models. He explored different methods, such as double-selection outcome regression estimators, which ensure uniform asymptotic unbiasedness but may suffer from increased variability compared to naive methods. He also addressed the challenge of selecting instruments without prior knowledge, and discussed the implications on the asymptotic variance achievable by unbiased estimators. He pointed out that while uniformly valid inference is desirable, it may lead to inflated variability compared to naive methods. He showed the trade-offs through a study of a double-selection outcome regression estimator, which strikes a compromise between uniform validity and efficiency.

4. **Dr. Jinbo Chen** presented a talk titled "A constrained maximum likelihood approach to developing well-calibrated risk prediction models." When evaluating the added value of new predictors, the standard practice involves comparing the performance of a base model with standard predictors to an updated model that includes both standard and candidate predictors. Yet, the challenge with candidate predictors is that their data often come from sources that may not represent the target population, leading to differences in predictor distribution and case-mix or outcome prevalence. This challenge arises, for example, in breast cancer risk assessment, where clinicians are interested in evaluating the added value of breast density in assessing breast cancer risk. Dr. Chen pointed out that achieving an unbiased assessment of added value requires unbiased risk estimates from both the standard and updated models. With regards to model calibration, acceptable calibration of risk prediction models ensures unbiased assessment of new predictors. Dr. Chen introduced a constrained maximum likelihood approach for model calibration, which enforces calibration against the base model through constrained minimization of the loss function. She validated her method and applied it to various scenarios, including breast cancer risk prediction and evaluation of different breast density imaging measures and polygenic risk scores.

5. **Dr. Farouk Nathoo** presented a talk titled "Neural Network Feature Extraction and Bayesian Group Sparse Multitask Regression for Imaging Genetics." The motivation for this work was investigating the association between genetic variations and neuroimaging measures as related to Alzheimer's disease. Genetic data were obtained through high-throughput single nucleotide polymorphism (SNP)

data, while brain imaging endophenotypes were derived from magnetic resonance imaging (MRI). The primary statistical challenges included regressing brain imaging data onto genetic data, considering high-dimensionality and adopting the longitudinal nature of the disease. To tackle these challenges, Dr. Nathoo presented a statistical model that included three components: (i) an extension of a spatial model for flexible correlation structure; (ii) feature extraction from MRI using expert-chosen features and neural networks; and (iii) a Bayesian group sparse multi-task regression model for genetic association analysis. The model was then analyzed using multiple techniques. These included Bayesian methods for uncertainty quantification and regularization, variational Bayes as an alternative to MCMC for inference, and an autoencoder neural networks for feature extraction. The methods were validated in simulation studies and applied to the Alzheimer's Disease Neuroimaging Initiative database. Future work includes extending the model for longitudinal imaging data, incorporating genetic data into disease trajectory modeling, and exploring alternative feature extraction approaches using neural networks.

6. **Dr. Jiwei Zhao** presented a talk titled "Doubly Flexible Estimation under Label Shift." Dr. Zhao presented research on domain adaptation and distribution shift, particularly focusing on training versus testing and source versus target domains. The talk explored covariate shift, where differences arise in the distributions of input features between the source and target domains. Additionally, attention was given to label shift, which occurs when the marginal distribution of the outcome varies between domains while the conditional distribution remains consistent across inputs. Existing methods for addressing label shift typically involve estimating the ratio of target to source outcome distributions, which can pose challenges due to limited data availability in the target domain. Dr. Zhao outlined various approaches from the literature, including discrete, moment matching, nonparametric, kernel mean matching, and parametric methods. In contrast to existing methodologies, Dr. Zhao proposed an innovative approach that estimates a general characteristic of the target population without requiring the estimation of the outcome distribution. This approach maintains validity even in scenarios where certain model components are misspecified. The methodology, encompassing doubly flexible estimation and alternative estimators, was discussed.

7. **Dr. Samidha Shetty** presented a talk titled "Robust Estimation under a Semiparametric Propensity Model for Nonignorable Missing Data." This talk focused on addressing missing data in the Korean Labor & Income Panel Study. The dataset comprised information on 2506 regular wage earners, with partially observed monthly income in 2006 and fully observed covariates including monthly income in 2005, gender, age, and education. The missingness mechanism was modeled through a regression model and a propensity model. Various methods from past papers were reviewed, including parametric and nonparametric approaches. Dr. Shetty proposed a method based on decomposing the covariates and assuming a specific structure for the propensity of missingness. She introduced estimating equations for the parameters of interest, focusing on nonparametric methods. Dr. Shetty proposed a new method that avoids estimating the nonparametric component and uses an efficient score function. She showed that this method consistently estimates quantities of interest while improving computational time and complexity compared to existing approaches. The talk also discussed estimation under a misspecified working model and provided results from the Korean Labor & Income Panel Study data set. The talk highlighted the effectiveness of the proposed method in consistently estimating quantities of interest while avoiding the estimation of the nonparametric component in the missingness mechanism. The method showed improvements in computational time and complexity compared to existing approaches, thereby showing the applicability of semiparametric nonignorable propensity models.

8. **Mr. Yin Tang** presented a talk titled "A Nonparametric Test for Elliptical Distribution based on Kernel Embedding of Probabilities." Mr. Tang discussed the significance of differentiating whether data adhered to elliptical distributions prior to employing related statistical and machine learning methods for multivariate data analysis. He used Central to this discussion was the utilization of kernel embedding to characterize the probability distribution, with emphasis placed on the kernel's characteristic nature when the embedding was injective. He proposed a test statistic to estimate the mean and covariance. He showed the asymptotic distribution of this test statistic under the null hypothesis. Through simulation studies, he evaluated the performance of the proposed test across various scenarios.

5 Impact of the Workshop

5.1 Impact on the Field of Censored Covariates

Until now, researchers have been working in silos to tackle the technical and practical challenges with censored covariates. Our workshop broke apart those silos and addressed the challenges in the field head-on. The research talks and think tank sessions highlighted significant scientific progress in addressing complex issues related to censored covariates in statistical analysis. We discussed new methodologies that improve the accuracy and reliability of statistical estimates, particularly with right- and interval-censored covariates and both under noninformative and informative covariate censoring. These methods included extensions of inverse probability weighting, of conditional mean imputation, of thresholding methods, and of likelihood-based methods. These extensions were shown to help mitigate bias and enhance estimation accuracy by adjusting for censoring effectively.

Furthermore, we discussed ways to extend methods for regression models with censored covariates to quantile regression, survival regression, and surrogate marker analysis. The ideas were based on using likelihood-based techniques as those techniques could easily adjust for the different types of censored variables and which variables were censored. These advancements are current works in progress and their theoretical developments along with open-access code will lead to improved data analysis accuracy in diverse research fields.

Additionally, we explored the relationship between censored, missing, and mismeasured covariates, proposing semiparametric and nonparametric approaches to handle the challenges in a way that reduces model misspecification. Overall, the workshop showcased notable advancements in statistical methodology, offering practical solutions to longstanding challenges in data analysis. These developments have implications across various scientific disciplines, enabling more accurate and dependable insights from datasets with censored covariates.

5.2 Impact on Collaborations

During the workshop, participants made the most of the think tank sessions and extended breaks between talks, which ranged from 30 to 120 minutes. These intervals provided opportunities to dive into ongoing challenges within the field of censored covariates. Participants engaged in discussions aimed at weeding out ideas that have not been successful, clarifying what impeded success, and generating new ideas to overcome those impediments.

As a result of these discussions, several collaborations were formed with specific objectives:

1. Introducing Bayesian Methods: The aim is to extend the application of Bayesian methods beyond the current focus on limit-of-detection censoring to encompass random covariate censoring.

2. Enhancing Imputation Techniques: The aim is to augment existing imputation methods for censored covariates with nonparametric techniques. This approach aims to mitigate issues related to model misspecification.

3. Addressing Interval Covariate Censoring: Efforts are underway to develop techniques tailored to interval covariate censoring, particularly in the context of modeling symptom progression for diseases like Huntington's and Alzheimer's.

4. Identifying Surrogate Markers: Participants are interested in working on incorporating covariate censoring techniques to aid in the identification of surrogate markers that are susceptible to censoring and missingness.

5. Integrating Methodologies: There is a push to establish connections between methodologies designed to address covariates that are censored, missing, and/or measured with error.

Individuals in these newly formed collaborations have laid out plans for further development. The workshop organizers have committed to maintaining communication with each group, reaching out every twelve months to monitor progress and provide assistance in advancing these initiatives. This ongoing support aims to ensure that the ideas generated during the workshop are nurtured and translated into tangible advancements within the field of censored covariates.

5.3 Impact on Mentoring and Training

As leader of a Biostatistics lab, an ongoing challenge Dr. Garcia faced when working with her mentees was transitioning them from students to researchers: students tend to rely on existing knowledge to regurgitate answers, whereas researchers tend to question existing knowledge to create new answers. Inspired by her leadership training programs, she developed a workshop called "Create Spicy Science," which consists of modules to transition her mentees into researchers. These modules include: (i) Desire and Trust: mentees learn how to identify what the field wants to solve (i.e., the desire) and what evidence the field needs to be convinced of a solution (i.e., the trust); (ii) Problem Premise and Solution Premise: mentees learn how to give their solutions significance by defining the problem that is important to the field and why their solution is the best one, creating desire and trust; (iii) Problem-Solution Path: mentees learn a step-by-step system to write a compelling and persuasive argument for their ideas—one that makes others want to learn more; and (iv) Feedback That is Heard: mentees learn how to identify critical writing issues, not simply grammar and punctuation mistakes, and how to communicate those issues in a way that is neither confrontational nor seen as a personal attack.

Dr. Garcia not only taught these modules to my mentees but also had them practice the skills with each other. Every two weeks, her mentees held Trust Sessions—an idea modeled after Pixar's system—where one person presented their ideas and the others gave candid, nonjudgmental feedback about what worked and what did not. After just six months of this training, she noticed that her mentees were blossoming into researchers: they were not hesitant to ask questions to gain clarity on the science, and they could more quickly identify gaps in scientific papers, in talks, and in each other's work.

Inspired by these results, Dr. Garcia shared material from the "Create Spicy Science" training with the workshop participants. She led exercises to help the participants use the material on themselves and with their mentees. These sessions were only 30-minutes and designed to help researchers understand that this type of material can be incorporated into everyday life, even for researchers with limited time. The feedback to these sessions was remarkable:

"I learned a lot from your professional development lectures, which is even a bigger part of the workshop! I chatted with a couple of others, and they felt the same!"

"The Professional Development sessions led by you were unique and inspiring. That definitely helped both junior and senior attendees understand the mentor process better."

"Your professional development sessions were lessons for life."

During the workshop, several participants made adjustments to their presentation slides to add more "spice"; i.e., make their talks more engaging and compelling. Additionally, participants discussed their plans to integrate the techniques learned during the workshop into their mentoring practices. The increased creativity, confidence, and high-quality research output Dr. Garcia saw from her mentees are benefits she wants every faculty member and their mentees to experience, and she believes that participants using these techniques is one step towards making that dream a reality.

6 Concluding Remarks

Research on censored covariates remains an ever-growing field that, until this workshop, was addressed by siloed research groups. This workshop created the opportunity to bring those research groups together. The workshop began with 38 participants who seemingly shared little in common, and ended with those same participants having formed new friendships, connections, and collaborations.

The workshop also inspired some participants to propose future sessions dedicated to further exploring censored covariates. One such session is tentatively planned for July 2024 in Thessaloniki, Greece.

The organizers are grateful to the Banff International Research Station for their support in organizing the workshop. The setting in Banff provided an ideal backdrop for scholarly discussions and collaboration. While the workshop may have concluded, the organizers are hopeful for future opportunities to continue sharing advancements in research on censored covariates.

References

[1] R. J. A. Little. Regression with missing X's: A review. *Journal of the American Statistical Association*, 87(420):1227–1237, 1992.

[2] X. Lv, R. Zhang, Q. Li, and R. Li. Maximum weighted likelihood for discrete choice models with a dependently censored covariate. *Journal of the Korean Statistical Society*, 46(1):15–27, 2017.

[3] Marissa C Ashner and Tanya P Garcia. Exploring the validity of the complete case analysis for regression models with a right-censored covariate. *arXiv preprint arXiv:2303.16119*, 2023.

[4] S. Ahn, J. Lim, M. Cho Phik, R.L. Sacco, and M.S. Elkind. Cox model with interval-censored covariate in cohort studies. *Biometrical Journal*, 60(4):797–814, 2018.

[5] R. A. Matsouaka and F. D. Atem. Regression with a right-censored predictor, using inverse probability weighting methods. *Statistics in Medicine*, 39(27):4001–4015, 2020.

[6] D. B. Richardson and A. Ciampi. Effects of exposure measurement error when an exposure variable is constrained by a lower limit. *American Journal of Epidemiology*, 157(4):355–363, 2003.

[7] E Schisterman, A Vexler, B. W. Whitcomb, and A Liu. The limitations due to exposure detection limits for regression models. *American Journal of Epidemiology*, 164(4):374–383, 2006.

[8] L. Nie, H. Chu, C. Liu, S.R. Cole, A. Vexler, and E.F. Schisterman. Linear regression with an independent variable subject to a detection limit. *Epidemiology*, 21(Suppl 4):S17–S24, 2010.

[9] P. W. Bernhardt, H. J. Wang, and D. Zhang. Flexible modeling of survival data with covariates subject to detection limits via multiple imputation. *Computational Statistics and Data Analysis*, 69:81–91, 2014.

[10] P. W. Bernhardt, H. J. Wang, and D. Zhang. Statistical methods for generalized linear model with covariates subject to detection limits. *Statistics in Biosciences*, 7(1):68–79, 2015.

[11] F.D. Atem, J. Qian, J.E. Maye, K.A. Johnson, and R.A. Betensky. Linear regression with a randomly censored covariate: application to an Alzheimer's study. *Journal of the Royal Statistical Society C*, 66(2):313–328, 2017.

[12] F.D. Atem, E. Sampene, and T.J. Greene. Improved conditional imputation for linear regression with a randomly censored predictor. *Statistical Methods in Medical Research*, 28(2):432–444, 2017.

[13] F.D. Atem, R.A. Matsouaka, and V.E. Zimmern. Cox regression model with randomly censored covariates. *Biometrical Journal*, 61(4):1020–1032, 2019.

[14] Yizhuo Wang, Christopher R. Flowers, Ziyi Li, and Xuelin Huang. Condis: A conditional survival distribution-based method for censored data imputation overcoming the hurdle in machine learning-based survival analysis. *Journal of Biomedical Informatics*, 131:104117, 2022.

[15] Sarah C. Lotspeich, Kyle F. Grosser, and Tanya P. Garcia. Correcting conditional mean imputation for censored covariates and improving usability. *Biometrical Journal*, 64(5):858–862, 2022.

[16] H. Lynn. Maximum likelihood inference for left-censored HIV RNA data. *Statistics in Medicine*, 20(1):33–45, 2001.

[17] P. Austin and L. Brunner. Type I error inflation in the presence of a ceiling effect. *The American Statistician*, 57(2):97–104, 2003.

[18] P. Austin and J. Hoch. Estimating linear regression models in the presence of a censored independent variable. *Statistics in Medicine*, 23(3):411–429, 2004.

[19] S.R. Cole, H. Chu, and E.F. Schisterman. Estimating the odds ratio when exposure has a limit of detection. *International Journal of Epidemiology*, 38(6):1674–1680, 2009.

[20] J.V. Tsimikas, L.E. Bantis, and S.D. Georgiou. Inference in generalized linear regression models with a censored covariate. *Computational Statistics and Data Analysis*, 56(6):1854–1868, 2012.

[21] Zhigang Li, Tor D Tosteson, and Marie A Bakitas. Joint modeling quality of life and survival using a terminal decline model in palliative care studies. *Statistics in Medicine*, 32(8):1394–1406, 2013.

[22] F.D. Atem and R.A. Matsouaka. Linear regression model with a randomly censored predictor: Estimation procedures. *Biostatistics and Biometrics Open Access Journal*, 1(1):555556, 2017.

[23] G. Gómez, A. Espinal, and S.W. Lagakos. Inference for a linear regression model with an interval-censored covariate. *Statistics in Medicine*, 22(3):409–425, 2003.

[24] S. Kong and B. Nan. Semiparametric approach to regression with a covariate subject to a detection limit. *Biometrika*, 103(1):161–174, 2016.

[25] Shengchun Kong, Bin Nan, John D Kalbfleisch, Rajiv Saran, and Richard Hirth. Conditional modeling of longitudinal data with terminal event. *Journal of the American Statistical Association*, 113(521):357–368, 2018.

[26] M.L. Calle and G. Gómez. A semiparametric hierarchical method for a regression model with an interval-censored covariate. *Australian & New Zealand Journal of Statistics*, 47(3):351–364, 2005.

[27] H. Wu, Q. Chen, L Ware, and T. Koyoma. A Bayesian approach for generalized linear models with explanatory biomarker measurement variables subject to detection limit: An application to acute lung injury. *Journal of Applied Statistics*, 39(8):33–40, 2012.

[28] B. J. Gajewski, N. Nicholson, and J. E. Widen. Predicting hearing threshold in nonresponsive subjects using a log-normal Bayesian linear model in the presence of left-censored covariates. *Statistics in Biopharmaceutical Research*, 1(2):137–148, 2009.

[29] R. May, J. Ibrahim, and GenIMS Investigators. Maximum likelihood estimation in generalized linear models with multiple covariates subject to detection limits. *Statistics in Medicine*, 30(20):2551–2561, 2011.

[30] R. Wei, J. Wang, E. Jia, T. Chen, Y. Ni, and W. Jia. Gsimp: A gibbs sampler based left-censored missing value imputation approach for metabolomics studies. *PloS Computational Biology*, 14(1), 2018.

[31] J. Qian, S.H. Chiou, J.E. Maye, F. Atem, K.A. Johnson, and R.A. Betensky. Threshold regression to accommodate a censored covariate. *Biometrics*, 74(4):1261–1270, 2018.

[32] S.C. Lotspeich, M.A. Ashner, J. Vazquez, K.F. Grosser, B. Bodek, B. Richardson, and T.P. Garcia. Making sense of censored covariates: Statistical methods for studies of huntington's disease. *Annual Review of Statistics and Its Applications*, 2024. In press.

[33] Tanya P. Garcia and Layla Parast. Dynamic landmark prediction for genetic mixture models. *Biostatistics*, 22(3):558–574, 2021.

[34] F.D. Atem, J. Qian, J.E. Maye, K.A. Johnson, and R.A. Betensky. Multiple imputation of a randomly censored covariate improves logistic regression. *Journal of Applied Statistics*, 43(15):2886–2896, 2016.

[35] S. Lee, S.H. Park, and J. Park. The proportional hazards regression with a censored covariate. *Statistics & Probability Letters*, 61(3):309–319, 2003.

[36] K. Langohr, G. Gómez, and R. Muga. A parametric survival model with an interval-censored covariate. *Statistics in Medicine*, 23(20):3159–3175, 2004.

[37] X. Lu, B. Nan, P. Song, and M.F. Sowers. Longitudinal data analysis with event time as a covariate. *Statistics in Biosciences*, 2:65–80, 2010. doi:10.1007/s12561-010-9021-2.

[38] A. Sattar, S.K. Sinha, and N.J. Morris. A parametric survival model when a covariate is subject to left-censoring. *Journal of Biometrics and Biostatistics*, 3(2), 2012.

[39] S. Hubeaux and K. Rufibach. Survregcenscov: Weibull regression for a right-censored endpoint with a censored covariate. *arXiv: Computation*, 2014.

[40] H. Zhang, H. Wong, and L. Wu. A mechanistic nonlinear model for censored and mismeasured covariates in longitudinal models, with application in aids studies. *Statistics in Medicine*, 37(1):167–178, 2018. doi:10.1002/sim.7515.

[41] L. Wu and H. Zhang. Mixed effects models with censored covariates, with applications in HIV/AIDS studies. *Journal of Probability and Statistics*, 2018:1581979, 2018.

[42] S. Kong, B. Nan, J. Kalbfleisch, R. Saran, and R. Hirth. Conditional modeling of longitudinal data with terminal event. *Journal of the American Statistical Association*, 113(521):357–368, 2018.

[43] E.K.T. Benn, L.B. Tabb, P. Exum, R.H. Moore, K.H. Morales, F.R. Simpson, S.A. Lawrence, and S.L. Bellamy. Creating and sustaining effective pipeline initiatives to increase diversity in biostatistics: The enar fostering diversity in biostatistics workshop. *Journal of Statistics Education*, 2020. doi:10.1080/10691898.2020.1820409.

[44] A.L. Golbeck. Are women underrepresented in the american statistics profession? *Significance*, 2016. doi: 10.1111/j.1740-9713.2016.00892.x.