Research Statement

My interest in Biostatistics grew enormously when I started my research on copula modeling with Professor Evan Ray. Copula is a joint function used to measure the dependence between random variables and is widely used in times series modeling. I contributed to the *ncopula* package in R by implementing S3 object-oriented programming to assist Professor Ray in developing his forecasting models. The goal of the project was to construct Archimedean copula trees with different nesting structures to develop nested Archimedean random forest models. Specifically, I calculated the probability density function and cumulative distribution function to estimate parameters of nested Archimedean copulas with maximum likelihood estimation. In addition, I included supplementary functions to transform the parameters so that they fall within the given bounds for different copula families.

I encountered a challenge when the estimation ran into convergence issues in the optimization process because each copula family has its own bounds. To address this problem, I applied a wide range of mathematical transformations such as softmax and multiple conditional scrutinization for each copula family in the transform functions. I examined these transform functions with the comprehensive unit tests to make sure that the estimation was stable and consistent. The future direction of this project is to come up with an algorithm to produce multiple copula trees with different nesting structures from the same set of covariates. The stronger correlated covariates are usually grouped closer to the bottom of a nested copula tree. Nevertheless, it is possible to introduce variation by changing the number of covariates in each copula node and by assigning different families to each node depending on the dependency property. This experience has strongly prepared me for developing statistical software for the computational area of biostatistics and ensuring that methods are reproducible with careful documentation.

With this research in statistical methodology under my belt, I continued my journey in the Biostatistics Department at Memorial Sloan Kettering Cancer Center summer 2020 under the mentorship of Dr. Audrey Mauguen. I investigated the association between biomarker serum bilirubin and survival in Primary Biliary Cirrhosis (PBC) with the Cox Proportional Hazards Model, the Time-Dependent Cox Model, and the Joint Model for longitudinal and time-to-event data. The goal of this project was to compare the estimated hazards ratios from these three different approaches and to evaluate the benefits and drawbacks of the Joint Models. To prepare the data for the analysis, I wrote functions in R to extract the interval endpoints and the corresponding event statuses from the patients' enrollment time. Moreover, I produced data visualizations with Kaplan Meier survival curves and spaghetti plots, ran log-rank tests to compare the group survivals. I concluded that for applications where sample size is large and computational resources are available, the Joint Model should be used because it reduces the bias in parameter estimation relative to the other two Cox models in survival analysis.

In this project, I ran into a challenge when the sets of statistically significant covariates for each model from the stepwise variable selection were different. To assess the validity of the results and to make them more comparable, I reran the multivariate analysis for the three models, each with three different sets of covariates. For the Joint Model, in addition to assessing the proportional hazards assumption in the survival submodel, I examined the correlation structure in the longitudinal submodel to detect any suspicious multicollinearity. Moreover, I ran the sensitivity analysis to examine the impact of potential outliers and influential observations. I found that the estimated hazard ratios from each model were consistent despite different covariate adjustments. I presented my work at the MSK departmental symposium, the MHC Learning through Applications symposium, and the Electronic Undergraduate Statistics Research (eUSR) Conference 2020. These opportunities have greatly improved my presentation skills.

Motivated by applications of statistics in healthcare, I am expanding the cirrhosis project and incorporating it into my senior thesis under Professor Marie Ozanne's supervision. In addition to documenting my prior work, we plan to revisit the association between serum bilirubin and survival in PBC with the cause-specific hazard model to account for the competing risk. We also hope to explore elastic net regularization as a method for variable selection in survival analysis. When writing my thesis, I have delved more deeply into the missingness mechanism in longitudinal studies and censoring types in survival analysis. In this project, I was fascinated by the pattern mixture models (PMM) and its advantages over the linear mixed-effects models (LME) when doing literature review. PMM reduces potential bias in the estimation by accounting for situations where data are missing not-at-random. Currently, I would like to understand more deeply the possible implementation of PMM in joint modelings and explore if the issue of under-identification in PMM raises any estimation challenges for the joint models.

To deepen my knowledge of survival analysis, I expanded my prior work on Cox regression models and joint modeling, to assess the proportionality assumption of the Cox model for my final project in my Advanced Data Analysis class at Amherst College with Professor Nicholas Horton. Specifically, I investigated different techniques to examine the proportionality using the graphical visualization with survival curves, the formal test-based approach with Schoenfeld residuals, and the time-covariate interactions. In my report, I illustrated these techniques in a simulation study for censored data, followed by a careful discussion of the usefulness of these approaches and the bias of the results if this assumption is violated. I introduced ways to address the non-proportionality such as shortening the follow-up time, stratifying the Cox model, and using the extended Cox model to account for the time-varying effect of covariates, depending on the study objectives. From my simulation study, I learned that the conclusions from the goodness-of-fit test and the time-covariate interactions can be contradictory, and the visualization approach is often subjective. This project has helped me comprehend in depth the importance of combining multiple techniques to assess the appropriateness of the Cox model and the responsibilities of biostatisticians to apply their statistical knowledge properly.

I am determined to obtain my PhD in Biostatistics through the program at University of Pennsylvania where I plan to concentrate on the topics of joint modeling and competing risk analyses. I would like to investigate the impact of informative censoring and missing not-at-random data in joint modeling with rigorously implemented computation and statistical methodologies. With the excellent education and training program at Penn, I will become a leading researcher in the field of biomedical research.