Statement of Purpose

I was motivated to pursue higher education to conduct research in healthcare as I witnessed many children going through their battle with cancer while serving as a volunteer at K-hospital, the largest cancer hospital in Vietnam. This passion has propelled me through my time at Mount Holyoke College where I equipped myself with a solid mathematical and statistical background and strong computational skills for my research projects. The Biostatistics PhD program at Harvard University aligns with my academic and career goals: to broaden my statistical toolbox, to refine my research skills, and to deepen my knowledge of biostatistics, epidemiology and public health. In the long term, I want to pursue a career as a researcher and to work closely with clinical collaborators to best address the application of interest.

My interest in Biostatistics grew enormously when I started my research journey on copula modeling with Professor Evan Ray. 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. This experience 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. Our goal 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 curves and spaghetti plots and ran log-rank tests to compare the group survivals. I presented my work at the MSK departmental symposium, the MHC Learning through Applications symposium, and the Electronic Undergraduate Statistics Research Conference 2020. These opportunities greatly improved my ability to present and communicate scientific and statistical findings to different audiences.

Motivated by applications of statistics in healthcare, I am expanding the cirrhosis project from MSK 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. For my thesis, I have delved more deeply into the missingness mechanism in longitudinal studies and censoring types in survival analysis. During the literature review, I became fascinated by the pattern mixture models (PMM) and its advantages over the linear mixed-effects models (LME), particularly in reducing 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.

My exposure to both statistical methodologies and applications allowed me to learn one key thing: Statisticians can offer an enormous amount to the study of biomedicine. We have the ability to research the development of certain diseases and human vulnerability, and advance public health through developing treatments to these diseases. Through the Biostatistics PhD program at Harvard, I plan to concentrate on the topics of joint modeling and competing risk analyses. With rigorously implemented computation and applied statistical methodologies to clinical data, I would like to investigate the impact of informative censoring and missing not-at-random data.

I appreciate the great variety in Harvard faculty who work on both collaborative and methodological fronts, as well as the opportunity to learn about different research areas through rotations. Within the Department of Biostatistics, I hope to work with Professor Robert Gray and Professor Long Ngo who have published research in the field of survival analysis and longitudinal data analysis. Professor Robert Gray's research in methods for design and analysis of studies with competing risk endpoints directly connects with my research interest. I am incredibly impressed by his work on the Fine-Gray competing risks and his research contributions to the studies of survival analysis. Nevertheless, I am aware that it is important for PhD students to be open-minded and to broaden their horizon. I'm also interested in Professor Lee-Jen Wei's work in multivariate Cox procedures to handle multiple event times, and Professor Alkes Price's research in statistical genetics and the genetic basis of human diseases.

During my time at MSK, I learned about the daily responsibility of biostatisticians and their roles in complex, ethical clinical trials. I hope to contribute my statistical knowledge in collaborative interdisciplinary research through Harvard's large network of hospitals and research centers, particularly Dana-Farber Cancer Institute where I can pursue my passion for cancer research. Furthermore, with a strong theoretical and computational training through the PhD coursework, I will acquire the essential knowledge to conduct research in collaboration with public health officials, medical doctors, and other scholars in this inspiring journey.

Harvard University offers the necessary professional training for me to become a leading scholar in the field of biostatistics. With my strong academic record, research experience and determination to unravel complicated health problems, I will make a positive contribution to the University's research and receive excellent preparation for my future intended career.