Bao Han Ngo
DRP Reflection: Introduction to Survival Analysis
Mentor: Antonio Olivas-Martinez
December 16, 2022

I am very grateful I had the opportunity to participate in the Statistics Directed Reading Program this quarter, especially because it was my first quarter in the statistics major. For the first few weeks, we focused on statistics and probability basics, such as conditional probability, PDFs, and CDFs, and where these concepts may show up in survival analysis. All of this complemented what I was learning in my introductory probability class, STAT 340, and helped me appreciate the class material more since I saw how it could be applied to real data and research outside of lecture and homework. Once we had the basics down, we moved into the main focus of the quarter survival analysis.

Survival analysis focuses on analyzing data that measures "time to an event," with many applications in biological/health contexts. Despite being called survival analysis, survival analysis is not only concerned with life or death; the events of interest vary and can range from death to regaining function. Something that makes health data different from other types of data is that it can be censored. Censoring refers to data where the exact survival time (time to the event) is unknown. This can happen for a variety of reasons, but most commonly, censoring is a result of participants withdrawing from the study before the observation time ends, or participants reaching the end of the observation time without experiencing the event. Even though the exact time to event is not known, censored observations can still be used since we know that it took at least the censored time for the participant to have the event.

Within survival analysis, a variety of models can be used to learn more about survival and hazard. First, survival is concerned with the probability that it takes longer than a certain time for the event to happen. A popular way to model survival is with Kaplan-Meier curves. Being non-parametric, the only assumption made is that censoring is random and independent. However, since no assumptions are made about the specific distribution for the data, the Kaplan-Meier curves are not smooth, whereas parametric curves are smoother, and closely match the theoretical. To determine if two Kaplan-Meier curves are significantly different (such as the curves for the treatment and placebo groups), the log-rank test can be used, providing a p-value.

Unlike survival, which focuses on the probability that an event happens after a specified time, hazard is the risk of having the event at the specified time. The Cox Proportional Hazards (Cox PH) model is a semi-parametric model that makes the same assumptions about independent, random censoring and also that hazards are proportional overtime. The Cox PH model yields a hazard ratio, which is essentially a ratio comparing the risk of the event occurring between two groups. If the confidence interval for the hazard ratio does not include 1, it can be concluded that the risk of event differs between the two groups. Both the log-rank test and Cox PH model reveal differences between two groups, but the log-rank test yields only a p-value to

indicate the presence of significant difference, while the Cox PH model describes by how much the two groups differ.

By the end of the quarter, I was able to apply these models and concepts to my own simulated dataset, investigating the use of denosumab on fracture risk in post-menopausal women with osteoporosis. Survival analysis is full of choices and different models, each yielding equally valid results. It is up to the researcher to choose how they want to analyze the real and messy data.