Subodh Rajesh Selukar
Biostatistics PhD Program
Fall 2016

Motivation: About thirty years ago, 3M launched a new drug, Tambocor, prescribed to patients for suppressing abnormal heartbeats. Doctors at the time believed that this suppression would reduce the risk of cardiac arrest in this susceptible population. Thomas Moore's book *Deadly Medicine* documents the clinical trials of the drug and highlights the importance of supervision in clinical trials by biostatisticians. The initial lack of direction from biostatisticians early on led to thousands of patients dying, but later statistical work quickly shut down the prescribing practices of Tambocor as results indicated that the drug proved more harmful than beneficial. The critical impact of biostatisticians in this controversy fascinated me and ignited my interest in the design and analysis of clinical trials during my first undergraduate course in biostatistics.

I entered college, however, utterly unaware of biostatistics. I came to UNC-CH with a desire to pursue medicine and become an oncologist. At the time, I only knew of medicine and biology as avenues for the scientific battle against cancer, so I joined a genetics lab, the Ahmed Lab, in my freshman year. I wanted to study the interplay between telomerase and cancer, and one aspect of this lab explored the genetics of telomerase in the nematode worm, *C. elegans*. My formative project in the lab involved the analysis of RNA-Seq data: I aimed to quantify the number of reads of TERRA, telomeric-repeat containing RNA, in telomere-damaged mutants to determine if any association existed between this read number and telomere damage. I searched for methods online and applied programs like BLAST and Galaxy to no avail. After consultation with a computational biologist, I learned that looking at this specific repeat structure had very little prior study and would require development of new tools. Unfortunately, none in the lab had the statistical training to approach such an endeavor, so we had to stop the project. This lack of understanding, however, sparked a desire in me to learn how to conduct such analyses, so I explored the field of biostatistics and found it to be the perfect combination of mathematical theory, applied concepts and public health impact. As a result, I added the biostatistics major to my previous quantitative biology major.

Training and Experience: Soon afterward, I began work with Dr. Eric Bair at UNC-CH, examining pain measures for a study of temperomandibular disorders (TMD). We are interested in determining if the number of comorbid conditions (conditions present in the patient not related to TMD, such as back pain or obesity) predicts various pain measures in TMD patients. We encountered issues in analysis, however, because we predicted the secondary outcome, pain score, rather than the primary outcome, TMD. Prior research (Monsees et al. 2009) illustrates that since both pain measures and number of comorbid conditions may be associated with disease status, we may encounter inflated Type I error. To adjust for this, I conducted inverse probability weighted (IPW) regression in R to balance the sampling of cases and controls. I have found that the number of comorbid conditions did predict many of the pain measures (after adjustment with a Bonferroni method), so patients with comorbid conditions may experience more severe chronic pain. In the spring, I will present this research as part of my **honors thesis**.

To strengthen my computing skills, I participated in the High-Performance Computing Research Experience for Undergraduates (REU) at UMBC this past summer. At the REU, I took courses in programming that taught students how to utilize powerful data clusters with C and R and then participated in a 6-week research project that applied these skills. My project, in statistical genomics, looked at microarray data from Alzheimer's Disease (AD) patients and tried to draw a set of genes that would best predict AD. My group, composed of three undergraduates and myself, worked under Dr. Kofi Adragni to implement a new methodology he designed. This procedure involved screening the genes based on F-statistics that measured the relationship with AD, sparse sufficient dimension reduction (applying principal-fitted components then sparse estimation to leave only genes strongly associated with AD) followed by hierarchical clustering to group genes. We initially struggled to implement the methodology in R using the data cluster, but eventually we arrived at a set of 49 genes organized into 3 clusters that we found to best predict AD, which we described in a **technical report**. We presented this research at a **poster session at UMBC** over the summer, and we will present it for a broader audience at the **Joint Mathematics Meeting** in January. Finding this set of genes opened an opportunity for collaboration with biologists to explore the importance

of our findings; if biologists find clinically significant information from these results, doctors could make use of these genes in genetics-based personalized medicine in the future. I continue to work on this research with Dr. Adragni, determining the utility of this novel methodology by comparing it to existing methods.

Goals: The growth of genetic data available has given researchers exciting new ways of tailoring medicine to target individual patients. In particular, clinical trial design may be modified to include important genetic information as part of eligibility criteria in so-called "basket trials." Redig and Jänne (2015) discuss the emerging basket trials in oncology, focusing on how useful they can be in precision medicine for rare or difficult to study mutations. Earlier, Noah Simon (2013) proposed a method of adaptive enrichment for Phase III clinical trials, which used single biomarkers as eligibility criteria and illustrated that Type I error was preserved using this design. I want to focus my work in the context of such modern clinical trial designs, both in finding biomarkers and in designing (and analyzing) studies for their use. I also aim to direct this research toward the study of cancer.

My background in computing and statistical genomics prepares me well for identifying biomarkers from large genetic data sets. Specifically, I am familiar with clustering methods from my REU and from working with Dr. Bair, who previously established methods for semi-supervised clustering (2004). As a possible project, I would combine this method with Witten and Tibshirani's (2010) method, sparse clustering, to have "sparse supervised clustering." This technique retains the advantage of associating clusters with a biological outcome (as from Bair's work) and identifies variables responsible for separation between clusters from sparse clustering. Naturally, then, working with **Dr. Daniela Witten** would be perfect for such a project. I would also enjoy collaborating with **Dr. Noah Simon**, whose work above strongly influenced my interests. Regarding clinical trials, I also would like to connect with **Dr. Michael LeBlanc**; I hope that his interest in targeting patient subgroups in Phase II and III clinical trials will align with this basket trial design I am interested in studying. Indeed, it was finding subgroups who responded differently to treatment in late-stage clinical trials that led to such designs. Beyond these professors, I know UW has a wealth of fantastic researchers in broad areas within biostatistics. Knowing that my interests may change as I progress, I plan to reach out to more faculty, and I might find that another area or another professor is ideal for me.

I intend to become a professor in biostatistics after my graduate study, so I plan to gain a deep understanding of the subject while in graduate school. At the REU, I had a firsthand experience in developing methodology, but I also saw the theoretical finesse necessary to work on such projects. Because I want to develop such methods as a professor, I need to have an excellent theoretical background. I am actively constructing a framework for this study with courses in real analysis, and the theoretical rigor of the UW program will complete it. Electives in cancer epidemiology and genetics will supplement my natural science background, so I look forward to leaving UW an expert in biostatistics and its applications, especially in cancer research.

I also want to strengthen other facets necessary to become a world-class professor. For more than two years, I have served as a tutor and mentor in chemistry, genetics and biostatistics, and I have learned how to effectively distill complicated ideas into palatable ones. Through the theoretical education and teaching practice that I will gain at UW, I can improve my teaching. I also want to join within-biostatistics collaboration and promote biostatistics outreach. I have participated in open houses and panels for biostatistics at UNC-CH, and I plan to continue my community outreach into graduate school. At UW, I will join and contribute to groups like the Clinical Trials Working Group to improve my own understanding, while also promoting the exchange of ideas. In these ways, I hope to enrich the UW community, while also acquiring the skills necessary to excel in professorship.

My professors have related biostatistics to the climax of a story. My undergraduate experiences have given me an understanding of the broad story by giving me training in natural science and biostatistics and supplying me with strong skills in programming, teaching and collaboration. But I need graduate study to hone in on the climax and tell the story well. Only study at the University of Washington will provide the theoretical rigor and the facilities to pursue my interests, so I can tell the best story, and it is thus my top school of choice.