Dear Genentech research team,

As a fourth-year Ph.D. student majoring in Biostatistics at the University of Pittsburgh, I was excited to see your posting for 2020 summer internship positions on stattrak. Working as an intern with your research team would be an exciting and unique experience for me.

Over three years of study in the Ph.D. program helped me build a solid foundation. The statistics courses like Bayesian data science, applied mixed model analysis, nonparametric theory, survival analysis and high-dimensional statistics have strengthened my theoretical knowledge in statistics and ability to do statistical modelling and inference. Meanwhile, I also get involved in several research projects, which strengthened my research skills.

In the meta-analysis projects, I mainly focused on developing novel p-value combination methods to address challenges arising in modern high-throughput data analysis. Traditional p-value combination methods mainly focus on combining independent p-values with relatively dense signals to increase statistical power. However, in modern large-scale data analysis, correlation and sparsity are common features which can lead to failure of classical methods. My first project aims to develop methods to combine dependent p-values. One motivating example is SNP-set test in genome-wide association studies (GWAS). Due to privacy and cyber security issues, researchers tend to share only the summary statistics of GWAS instead of the raw data, which is essentially a massive collection of p-values of all the SNPs with unknown dependency structure. Traditional methods fail in this case due to violation of the assumption of independency and lack of knowledge of the dependency structure. To address this problem, we proposed a novel method for the robust testing for combining dependent p-values using transformation of heavy-tailed distribution. As shown in both theory and simulations, our method enjoys robustness for Type I error control under arbitrary dependency structures of p-values. My second project aims to propose a p-value combination method to handle the sparsity of true signals among a large collection of p-values, which is another common challenge in GWAS. By modifying the classical Fisher’s method, we proposed a new approach that is not only optimal for the detection of weak and sparse signals, but also achieves better computational efficiency and better finite sample performance comparing to other existing methods. For my third project, even though sparsity of true signals among p-values is common in GWAS application, I noticed that there are still lots of cases where whether the true signals are dense or sparse is unknown for a given set of p-values. Noting that existing methods designed for detecting weak and sparse signals tend to fail in the dense case while classical methods fail in the sparse case, I started my third project to develop a method that can be powerful for both scenarios. By modifying the method in my second project, we proposed a new approach that are theoretically proven to be optimal no matter the true signals are dense or weak and sparse. We then ran comprehensive simulations to show that our method has good finite sample performance in most cases. (maybe delete project two and three)

One of the primary goal of statistics is to develop methods to integrating and utilizing information. In the meta-analysis projects, I developed methods for the integration of information from a collection of p-values. However, information can also be aggregated by jointly modelling of different types of data. My second (if delete previous project 2 and 3) project aims to propose a method for disease subtyping for high-dimensional omics data. Traditionally, biologists tend to perform classical unsupervised clustering strategy to identify subpopulations that have similar patterns in gene features for disease subtyping. However, as the features corresponding to irrelevant confounders (e.g. gender or age) may dominate the clustering process, the resulting clusters may or may not capture clinically meaningful disease subtypes. To address this problem, we proposed a unified latent generative model to perform outcome-guided disease subtyping constructed from omics data. By jointly modelling the clinical outcomes, baseline covariates and omics data, the method is able to identify outcome associated clusters (disease subtypes) that are otherwise easily masked by other facets of clinically irrelevant cluster structure.

As a Graduate Student Researcher (GSR), I am also involved in many collaborative projects related to biology and immunology of ovarian cancer and endometriosis. The GSR training significantly strengthened my abilities to solve practical statistical problems. In addition, I have developed strong communication and presentation skills through my teaching experience and collaborative experience with several biologists and physicians.

I regard this internship opportunity as a valuable chance. I believe that my solid educational background, extensive research experiences and strong communication skills make me a strong candidate for this position. If my credentials are of interest to you, please feel free to contact me at [yuf31@pitt.edu](mailto:yuf31@pitt.edu) or 914-314-9339. Thank you for your precious time and consideration!

Sincerely,

Yusi Fang