FOSSProF Final Report Template

Project Overview

- Project Title
  Development and Education of the Use of R Language for Implementing Statistical Methods in Clinical Research and Beyond

- Project Summary: Briefly describe the project, its objectives, and the open-source software it focused on.
  This project focuses on the development of useful statistical methods and several statistical software programs to assist statisticians' research in planning and analyzing clinical trials. For this project, we focus on the open-source statistical programming language R. This project's first major objective is to develop novel statistical methods and software programs for simulating correlated multivariate count data that follow zero-inflated generalized Poisson (ZIGP) distributions. The motive is to provide tools to statistical programs for simulating such data that mimic real-life clinical trial data of such nature. For example, counts of observed maximum grade adverse event across treatment cycles and number of hospital visits can be modeled and simulated by the proposed tools. The second objective is to program and summarize useful statistical software programs that are often used in producing analysis reports and plans for oncology trials. Besides, we provide educational posts, on our division's GitHub page, to publish examples of implementing such programs in real clinical oncology settings.

- Target Audience: Who are the primary users or beneficiaries of the project? Are there secondary groups that benefit?
  Target audience is biostatisticians who work on design, analysis, and reporting of clinical trials. Secondary groups include statisticians and statistical students who are interested in multivariate zero-inflated count data generation and research of oncology clinical trials.

- Code Repository: Please include links to publicly available code repositories.
  https://oncologyqs.github.io/

Our division's GitHub page displays publicly available data and code. The repositories are stored within The Quantitative Sciences Division's GitHub Organization (OncologyQS) to ensure that no oncology study-related code is shared before being carefully de-identified and reviewed.

Project Activities and Progress

- Work Completed: Clearly outline the activities undertaken during the grant period. Did you achieve all your planned goals? If not, explain why and what was accomplished instead.

Regarding the first major objective, the PI (Ruizhe Chen) developed three novel methods for simulating data following ZIGP distributions. The activities include literature search, derivations of statistical methods, R programming to implement the developed methods, and writing of the manuscript. First, we achieved the planned goal of drafting a scholarly article describing the method. Two co-authors other than the PI were also involved in the writing process. Secondly, we wrote R programs that implement the proposed method in simulating data from target distributions (ZIGP). These R codes are made available for open-source access on our division's GitHub page.

Regarding the second major objective, Hanfei Qi has assembled, published, and updated 8 posts showcasing useful examples of: (1) how to plot Kaplan-Meier plots in R; (2) How to make nice summary tables in R; (3) How to prepare a Data and Safety Monitoring Boards (DSMBs) using R; (4) How to design a clinical trial with the feature of Bayesian futility monitoring rules in R; (5) R codes to generate summary table of uni-/multi-variate Cox PH models results; (6) R codes to draw forest plots; (7) R codes to draw swimmer plots; (8) R codes to show how to save plots in different formats.

- Technical Milestones: Discuss any specific technical milestones achieved, such as code releases, documentation updates, or bug fixes. Quantify accomplishments with metrics where possible (e.g., user growth, code contributions).

The biostatistics posts on our GitHub page have been released and updated from May 2024 to early 2024. The R codes for complete simulation of ZIGP distributed data are finished recently in September 2024.

- Challenges and Solutions: Share any challenges encountered during the project and how you addressed them.

Regarding the first major objective (statistical method development and manuscript writing of random numbers generator for ZIGP distributed data), the biggest hurdle was method development. It is very difficult to innovate new statistical methods, let along writing codes to achieve it and demonstrate its merits. We are lucky to have 2 senior statisticians' help in advising. Dr. Hakan Demirtas is an expert on simulating pseudo random numbers, and he gave key advice in moving the paper forward. Dr. Qian Shi is an expert clinical trial biostatistician, and she guided the application of the proposed methods in simulating N9741 study's adverse events data. Most importantly, the PI was determined to develop and publish the method and made great efforts in finishing the product.

Regarding the second major objective, it has been challenging in several ways:

1) Finding good sample data.
2) Generalizing the functions for broader use.
3) Ensuring that everything uploaded is shareable.

These challenges arise because many of our studies are small and have unique requirements. The data and code for certain projects can be easily identified, and we're not allowed to share them with people outside the study team, even within our division.

To solve the challenges, we've reviewed past projects, discussed with others in the division about their needs, made improvements based on user feedback, and taken extra time to carefully check and proofread before uploading anything.

Outcomes and Impact

- Project Impact: Describe the positive impact of the project on the Hopkins community, the open-source software ecosystem, or any other relevant groups. Use concrete examples and data to support your claims where possible.

As a group of statisticians, everyone has their own way of working and coding. While it's essential to be creative in our analyses and experiment with new statistical methods, it's equally important to standardize our routine tasks. It's key that we share a consensus on how we organize documents and code, as well as how we perform the most common statistical analyses.

In industry, they usually have a team of programmers to handle all programming, code review, and double-coding. However, we don't have such resources. It has often happened in the past that our colleagues wanted to share their code but couldn't find it when they searched through their old files. Therefore, we established a platform (where we can store code and instructions, allowing contributions from everyone) for common statistical tests, summary tables, and figures that are often applied across most projects.

Our website serves as a repository for posts and instructions on these topics and is hosted on our division's GitHub. It's useful because it streamlines repetitive tasks and centralizes the information we rely on. The platform is iterative, collaborative, and open-source, allowing anyone in our division to contribute new insights or code fixes. Over time, these posts will improve as more team members contribute. This tool ensures that we can consistently deliver high-quality analysis reports on time.

Currently, it includes code for summary tables (e.g., Table 1), summary tables for Cox models, and proportion/rate summaries, along with figures like Kaplan-Meier curves, forest plots, and swimmer plots.

One of the most highly rated resources is the guidance for generating a Data and Safety Monitoring Board (DSMB) report. It has been instrumental in training new employees and has even helped disease groups outside the Department of Oncology. The guidelines of project management benefit us by allowing another team member to seamlessly step in and continue

the work when someone is out. The code for efficacy and toxicity monitoring using Bayesian predictive probabilities is also widely used in our division.

As both biostatisticians and developers, we also have learned a lot in the process. The vignettes saved us a lot of time on repetitive tasks, which are unavoidable for collaborative biostatisticians, and allow me time to learn and try more innovative/advanced methods.

- Community Engagement: Did you actively engage with the open-source community through contributions, conferences, or workshops? Share details and metrics of Participation.

  Dr. Chen is invited to present the statistical paper on ZIGP data generation at the statistics department of UMBC on October 11th, 2024.

  For the seminar info, please see: Stat Colloquium: Dr. Ruizhe Chen · mathweb · myUMBC

- Sustainability / Future Plans: Explain how the project will be sustained beyond the grant period. Are there plans for future development, funding, or community support? Is there potential for further impact?

  Sustainability: The Quantitative Sciences Division will be updating and improving existing resources based on user comments. More vignettes will be added, as we describe in the future plans below.

  Future Plans: One of the most urgent needs is to provide educational resources for fellows and coordinators from the School of Medicine. For example, they are eager to learn what constitutes a good dataset under our current study setting, covering topics such as how to name each variable, the units and format of the variables, and the overall structure of the dataset. This is also beneficial for our division because we won't need to do as much data wrangling if we all reach a consensus.

  Another urgent need is to provide general ideas on available statistical methods for oncology studies, along with their corresponding recommended R packages. There are no wrong models, but there are always better models or methods that best suit the current study. Having such a dictionary of methods will help us think through the options and choose the best available ones. It is also common for us to dig into the package's source code to ensure it performs as expected and meets the claims it makes.

- Lessons Learned: Share key takeaways and insights gained from the project.
  Peer review for coding has been discussed for a long time in our division, but nothing has been done before. Once you recognize the need and started acting, people will follow your lead and appreciate what you've accomplished.

- Attachments: Include relevant documents such as screenshots, code samples, documentation updates, or presentations.

  All codes and examples are available on: [Biostatistics Posts | OncologyQS]