BRIDGE NEWSLETTER

BIOSTATISTICS RESEARCH & INVESTIGATION DIGEST

UTHSC MEDICINE

Welcome to our 2023 Winter Newsletter

Meet Our Newest Faculty: Zhu Wang

Interviewed by Trish Goedecke

Dr. Zhu Wang joined our biostatistics division in the fall of 2022.

1) Dr. Wang, where were you before you joined biostatistics here in Memphis?

I came here from University of Texas at San Antonio Health Science Center, another UTHSC. I was doing similar work there, working in biostatistics, both teaching and research.

Originally, I am from Sichuan province in southwest China, where the food is renowned for being very flavorful. My home was in a basin of 88,600 sq mi (about twice the area of Arkansas) surrounded by mountains, where the food culture is distinct, as traditionally it had been difficult to get in and out of the basin. Hiking there was not what we have here in Memphis. I once took two days to climb a mountain… it was very steep. The angle was not quite 90 degrees but approaching that.

Coming to the U.S. was an adventure in the beginning. Of course, that was a long time ago. The two cultures were very different then. I had no idea what my life would be.

2) How do you like to enjoy your free time?

I enjoy both playing and watching tennis—I take it fairly seriously, too. San Antonio has some strong competitors. In person, I've seen the U.S. Open, the Canada Open, the Miami Open and the Connecticut Open. For training, I sometimes play against a tennis ball launching machine, which can send the ball in random directions, plus top spin or under spin. It's quite an opponent; it never gets tired.

Here in Memphis, I've come across a new dream: perhaps I would like to be a tennis coach. Tennis is difficult to learn by yourself, and it's definitely a skill that needs to be developed. No one is born with tennis skills. It's also an expensive sport, especially if you hire a private trainer. I would like to be certified as a trainer and offer these skills to young Memphians who might not otherwise have the means to learn.

3) Could you describe the projects you've taken on since arriving in Memphis?

When I came, I developed a new course, *BIOE869: Data science 3*, in which I introduce machine learning along with statistical methods. This course is designed to provide students with a broad range of techniques available to use in their research. Students implement popular machine learning algorithms using the R programming language, applying the software for data analysis and establishing a deep understanding of machine learning.

I teach the course with sensitivity to the skill levels of the students—it is like personalized learning. The spring course opens for registration at the end of October. I also teach *BIOE720: Biostatistics for Public Health*. Some students come with no experience in R programming—I encourage novice programmers to refer to ChatGPT to gain necessary coding skills.

I've been preparing a presentation for the Cancer Center on machine learning in cancer research—both my own work and the broader area. I hope to stimulate more of this research here. Machine learning is impacting how we do research, much as ChatGPT impacts how we learn. We need to embrace these new technologies.

Regarding collaborations here, I have contributed already to several manuscripts and grant proposals including an R01 submitted recently, "Precision medicine for children with ulcerative colitis." I am grateful to have found no shortage of opportunities to collaborate here at UTHSC.

Biostatistics: Evolving Landscape Integrating with Machine Learning for Advanced Biomedical Insights

Written by Hyo Young Choi

When one mentions biostatistics, what springs to your mind? In conversations with biologists, common themes emerge: power calculations, observational studies, experimental design, and a focus predominantly on inference. These areas have long been the mainstreams of biostatistics, maintaining their significance and demand in practical applications. However, just as biology has undergone a remarkable transformation with technological advancements, biostatistics too is experiencing a wave of change and evolution.

This evolution is notably characterized by the synergistic interaction between statistics and machine learning within biomedical science. The common aim of statistics and machine learning is to extract meaningful knowledge from complex data sets, and the boundary between the two fields is becoming increasingly fuzzy, converging under the expansive umbrella of data science. Biostatisticians adeptly incorporate machine learning and AI tools, grounded in statistical reasoning, to uncover important signals and insights from large-scale biological data. This article aims to spotlight some remarkable contributions of statistics to biomedical science, particularly through its interplay with machine learning.

In modern biological research, especially genomics, the challenge of high-dimensionality is prevalent, where datasets feature a large number of variables relative to a limited sample size. Traditional statistical theories often struggle under these conditions, leading to significant efforts within the statistical community to address these challenges in the past decade. The theoretical understanding of high-dimensional data spaces has been a significant contribution in this area. Theories like High-Dimensional, Low Sample Size (HDLSS) and Random Matrix Theory (RMT) have been instrumental in elucidating the behavior of data points in high-dimensional settings. These theories enable statisticians to differentiate meaningful correlations from spurious ones in large, random datasets. This distinction is pivotal for separating signal from noise, a process fundamental to various dimension reduction techniques like Principal Component Analysis (PCA), Independent Component Analysis (ICA), Partial Least Squares (PLS), and Canonical Correlation Analysis (CCA) which have demonstrated their practical utility in genomic studies. These theoretical frameworks have not only deepened our understanding of high-dimensional spaces but have also translated abstract concepts into practical, intuitive solutions in machine learning.

Moreover, feature selection stands as a critical aspect of machine learning, particularly for high-dimensional data. Statisticians have devoted significant resources to developing strategies like regularization and shrinkage operators, which aid in effectively choosing important variables and generating optimal feature subsets. Sparse modeling techniques such as Sparse PCA, Sparse Non-negative Matrix Factorization (NMF), Sparse PLS, and sparse regression tools like Lasso have become popular for their ability to identify the most informative variables, illuminating the underlying biological signals. Regularization also plays a crucial role in preventing overfitting in machine learning models, a common challenge in high-dimensional data analysis. The intersection of high-dimensional statistics and machine learning represents an active area of research, especially in large-scale genomic studies. Advancements in both theoretical understanding and practical tool development have significantly empowered researchers to decipher the intricate tapestry of biological data, pushing the boundaries of what's achievable with these datasets.

Another notable synergy between statistics and machine learning is the enhancement of interpretability in complex datasets. Traditional machine learning methods, known for their high predictive accuracy, often are challenged in uncovering interpretable mechanistic or causal relationships among variables. Biostatisticians are actively bridging this gap by steering machine learning towards approaches capable of inferring causal relationships. Emerging fields such as Mendelian randomization, which employs genetic variants as instruments to infer causality, are integrating causal inference methodologies with deep neural networks. This approach not only identifies trait-associated variants with precision but also substantially enhances our understanding of the complex interplay between high-dimensional genomic sequences and downstream traits.

Furthermore, machine learning-based causal inference models are increasingly gaining popularity in clinical applications like disease biology, drug discovery, and intervention strategies, where accurate identification of cause and effect is crucial. This burgeoning field includes innovative methods such as causal K-nearest-neighbors, causal random forest algorithms, and causal support vector machines. These methods reformulate popular machine learning algorithms for causal discovery and are applicable across a wide range of fields.

Recently emerging is the intersection of biostatistics and machine learning within the domain of dynamic treatment regimes. These regimes represent a systematic approach to therapeutic decision-making over time, informed by progressively accumulating patient data. Reinforcement learning, a branch of machine learning where an agent learns to make decisions in an environment to achieve maximum cumulative reward, is particularly suited for dynamic treatment regimes. In healthcare, the 'agent' is a decision-making algorithm, the 'environment' comprises the patient and their disease progression, and the 'actions' are treatment choices. The optimization process aims to improve patient outcomes, considered the 'reward' in the reinforcement learning algorithm. Integrating reinforcement learning with robust statistical models is essential in dynamic treatment regimes, enabling efficient estimation and evaluation of optimal treatment strategies. This involves navigating complex and diverse data types, such as censored time-to-event outcomes, and sparse, irregularly spaced, and noisy measurements. Dynamic treatment regimes are particularly promising in precision medicine, where the goal is to tailor a set of sequential decision rules to individual patients. By leveraging the synergistic potential of biostatistics and machine learning, they offer pathways to more effective, individualized treatment plans.

This article only highlighted just a few examples of how biostatisticians are changing biomedical science by developing state-of-the-art models and methods for modern biological data. This evolution marks a pivotal shift in the landscape of biostatistics, transitioning from traditional analytical methods to embracing more complex, data-driven approaches. The integration of machine learning and AI into biostatistical practices is enabling researchers to go deeper into biological data, uncovering previously inaccessible insights.

Within the Division of Biostatistics in the Department of Preventive Medicine, our faculty members bring diverse expertise to the forefront of the big data revolution. We are engaged in pioneering work that span from theoretical advancements to the development and application of cutting-edge techniques for complex biological data including genomics and multi-omics approaches across various diseases. Our mission is to transform this wealth of data into actionable insights, thereby propelling the field of precision medicine into a future where personalized care is not just a vision, but a reality.

References:

1. Ij, H. "Statistics versus machine learning." Nat Methods 15.4 (2018): 233.
2. Liu, Yufeng, David Neil Hayes, Andrew Nobel, and James Stephen Marron. "Statistical significance of clustering for high-dimension, low–sample size data." Journal of the American Statistical Association 103, no. 483 (2008): 1281-1293.
3. Choi, Hyo Young, and J. S. Marron. "Theory of high-dimensional outliers." arXiv preprint arXiv:1909.02139 (2019).
4. Pennington, Jeffrey, and Yasaman Bahri. "Geometry of neural network loss surfaces via random matrix theory." In International conference on machine learning, pp. 2798-2806. PMLR, 2017.
5. Liang, Jane W., and Śaunak Sen. "Sparse matrix linear models for structured high-throughput data." The Annals of Applied Statistics 16, no. 1 (2022): 169-192
6. Lecca, Paola. "Machine learning for causal inference in biological networks: perspectives of this challenge." Frontiers in Bioinformatics 1 (2021): 746712.
7. Brand, Jennie E., Xiang Zhou, and Yu Xie. "Recent Developments in Causal Inference and Machine Learning." Annual Review of Sociology 49 (2023).
8. Lagemann, Kai, Christian Lagemann, Bernd Taschler, and Sach Mukherjee. "Deep learning of causal structures in high dimensions under data limitations." Nature Machine Intelligence (2023): 1-11.
9. Song, Rui, Weiwei Wang, Donglin Zeng, and Michael R. Kosorok. "Penalized q-learning for dynamic treatment regimens." Statistica Sinica 25, no. 3 (2015): 901.
10. Zhao, Ying-Qi, Donglin Zeng, Eric B. Laber, and Michael R. Kosorok. "New statistical learning methods for estimating optimal dynamic treatment regimes." Journal of the American Statistical Association 110, no. 510 (2015): 583-598.

Interns Summer 2023

Written by Tristan Hayes, MSc

Our division recruits summer interns each year. The selected interns work with faculty in our division for about 10 weeks. The program started in 2016, and 15 interns have completed the internship so far. All have gone on to pursue careers in data science, statistics, or medicine. We continued to build on the class of 2022’s success with another large class of 4 interns. This article briefly introduces thisyear’s interns and their projects. For the second year in a row, we hired an intern as a full time staff member.

Congratulations to Chenhao, intern class of 2022, and congratulations to our intern class of 2023!

Summaries provided by Mentors

Galvin Li, MSc UVA, (mentors Chi-Yang Chiu and Feng Liu-Smith) is a first-year medical student at UTHSC Memphis. Galvin came to us with an extensive training in statistics, after completing his Masters degree at University of Virginia and participating in their consulting program. He used the NHANES dataset to investigate the association between sex hormones and skin cancers. Galvin built multiple logistic regression to determine the association of E2, T and T/E2 levels with melanoma risk, adjusting for other relevant variables and examining potential gender-based differences. A repeat analysis was performed using non-melanoma skin cancer as the outcome.

Harper Kolehmainen, (mentors Gregory Farage and Saunak Sen) is currently a senior computer science major at Rhodes College. Harper developed an interactive interface for Genome Scans using Pluto.jl and BulkLMM.jl packages in Julia. Additionally, she helped improve the package documentation and contributed methodological improvisations as well as interface/API changes.

Miyeon Yeon, MS Biostatistics Florida State University, (mentor Hyo Young Choi) is currently a PhD Candidate Biostatistics, Florida State University. While at UT, she contributed to an assessment of mRNA degradation in FFPE tissue. Miyeon examined the RNA-seq data paired in FFPE and fresh frozen primary tumor tissue obtained from a subset of TCGA. She performed genome-wide comparisons of gene expression profiles between FFPE tissue and fresh frozen tissue using unsupervised/supervised clustering methods. The project looked at gene-specific as well as sample-specific degradation using SCISSOR and identified severely degraded samples in both cohorts looking for association with multiple factors such as total gene length, 3’ UTR length, GC concentration, etc. All analyses were performed in R/R Markdown. She also helped develop an R package for assessing RNA degradation which will be available on Github.

Siling Liu, (mentor Qi Zhao) is a MSc Computational Science and Engineering student at Rice University. Siling’s summer work focused on analyzing data from the CANDLE project entitled “Placental Epigenome-Wide Association Study of Early Childhood Body Mass Index Growth Trajectories and Overweight/Obesity Risk”. This is a very complex longitudinal dataset and Miss Liu rose to the challenge.

Meet Fred Yu and Chenhao Zhao

Former summer interns and later staff members with the Biostatistics Division
Interviewed by Trish Goedecke

What attracted you to intern with the Biostatistics division at UTHSC?
Fred Yu: I learned about the internship opportunity in the spring of 2021, while doing a masters in biostats at the University of Washington in Seattle, from departmental emails. I sent an application, and was especially interested in the research of Gregory Farage and Saunak Sen. One of the recent studies by Dr. Sen completed with a prior summer

intern addressed matrix structure with sparseness for regularization. The calculations use proximal algorithms, similar to gradient descent. I was learning about gradient descent at the time and sent my code on "coordinate-descent algorithms." I didn't know they were working in Julia programming language; my code was in R.

Chenhao Zhao: I was looking for internships in the spring of 2022, and my roommate told me about the UTHSC position. He had applied an earlier year, but his internship was canceled due to the pandemic. Now he is studying for a doctorate at Dartmouth. My studies are in quantitative biomedical sciences, biostatistics, computing, and epidemiology. My interests are in computing and coding.

What did you work on during your summer internship with the biostats division?
Chenhao Zhao: My project was providing documentation for Dr. Sen's Julia packages *MatrixLM* and *MatrixLMnet* and developing a demonstration for the package. I was also debugging and testing, updating code for a new version of Julia. Julia is a young computing language and still changes quite a bit between versions. It is similar to R but different. It's fast like C but user-friendly.

Fred Yu: If you conceptually like math, you'll enjoy Julia more than Python. Python commands are not as intuitive. Julia commands are designed as if you are writing a mathematical formula; especially with matrix operations, like linear algebra.

Chenhao Zhao: In Julia, it is easy to do vector operations.

Fred Yu: Unlike Chenhao, I worked mostly on *MatrixLMnet*. *MatrixLM* does multivariate bilinear model. *MatrixLMnet* is the extension of the *MatrixLM* model with sparse estimation of the features effects by elastic net regularization. It was initially designed by (previous intern) Jane Wang with multivariate linear regression. Jane developed the Lasso Regression version of the *MatrixLM* model. I extended it to have the Elastic net regression version, which is a more general version of lasso.

I got to meet Jane in a zoom meeting during my summer internship. We were working on something in her code that had an issue with reproducibility. This was due to randomness included to make a matrix semi-positive definite. We were looking into the source code of her version. My major contribution was to extend the code to have more features and to be more general.

We met with Jane with our concerns about the previous package producing unreproducible results. We proposed our solutions and asked her for advice.

Chenhao Zhao: My roommate also got to meet Jane at Harvard. Anyone with large-scale data will find Julia useful, especially high-dimensional data from gene sequencing.

How has your professional career been affected by your internship with us?
Fred Yu: Two aspects:
Technical: I learned version control using GitHub and the general flow of cooperating in a multi-developer software developing team.

Non-technical: I learned to cooperate and communicate with others. Each week we presented our progress to interns working on different types of projects. We learned to summarize work and present it to a non-familiar audience.

Chenhao Zhao:
Technical skills: Github for version control and the pipeline of developing a package to

have reproducible research. Also, after you publish a package, you need to continue to update it and consider user experience.

Non-technical: I was meeting new people with different skills, benefiting me not only professionally but also in life.

Fred Yu: The internship also led me to my current grad program in Data Science and Engineering. It aroused my interest in programming, in software development. My background was in math; I was doing more on paper. It made me more interested in coding.

What are you doing currently?
Chenhao Zhao: Currently, I help with BERD biostats consulting; and I work with Dr. Chiu on Bayesian methodology for meta-analysis with genetics studies, to analyze heterogeneity. Another project is with Dr. Chiu and Dr. Feng Liu-Smith studying Women's Health Initiative data, which is an epidemiological project. Our target is to find associations with melanoma.

Fred Yu, Chenhao Zhao, Dr. Gregory Farage and Dr. Saunak Sen at Shelby Farms

Fred Yu, Chenhao Zhao, Dr. Gregory Farage at Carolina Watershed

How to Save a Lot of Effort and Avoid Mistakes Importing Data from SAS Into R

Tristan Hayes, the Biostatistics Consulting Manager, shares his experience going between SAS and R data formats.

The Problem: User friendly options to import SAS datasets into R have existed for some time; however, without lots of manual coding in R, much information will be lost. For example: the continuous variable Age in a SAS dataset may set the custom format of 999 as the numeric code for missing data. If we naively use the Haven package from R this would import without error, but then our patient dataset would include the oldest humans on the planet! Another example: the dataset may code race and ethnicity as 1-5, but each of those numeric codes represents a different race (levels of the race factor variable in R). Without lots of manual coding, many datasets will not be imported correctly.

The solution: In this practical example, our solution takes hundreds of lines of custom SAS formats from a lung cancer screening dataset and generates 94 lines of R code. Note: this solution does require access to SAS for the very first step, but after that step, it is all open source. We use a SAS Macro to read the custom SAS formats and generate the R script to properly import, format, and label the variables. Our practice dataset comes from the National Lung Screening Trial Public Access Data as a hands-on example for each step of data processing.

Data Pipeline Solution:

1. From SAS, use the SAS-R script to generate R code setting the levels, labels and formatting of each variable à
2. Run the generated R code and correct any errors à
3. Sit back, take a sip of coffee and marvel at the hundreds of lines of code you wrote in the blink of an eye. Now you have a properly read-in, final dataframe ready to analyze in R.

Items We Will Need

1. The SAS to R script from Github. You can download it directly from here: https://github.com/clindocu/sas-r Please be sure to star the developer's project so they can get some credit for their effort.
2. Import_format.sas – This is the SAS custom format script which explains to SAS how to label and display the data.
3. nlst_data_formatted.sas7bdat – this is the SAS format of the National Lung Cancer Screening Trial dataset we are using here.

Step 1 edit: sas-r

Sas-r is a simple SAS macro program. You only need to change three things:

a. Source Dataset Location (where your sas dataset, the sas7bdat file is located)

b. Paste the custom formats directly into the sas-r script

c. Output R Program Name

For part (b), you can also point the macro to a format script file (*.sas), or you may have a custom SAS formats library. The easiest way to do this step is to just paste the custom formats into the sas-r macro (right under where it says: "Or Create SAS Formats here"), as shown below.

Note: for the code shown here from another example, we had to add the option (notsorted), so that R does not sort the formats alphabetically. One would hate for their ordinal Likert scale variable to be resorted arbitrarily!

Step 2

Click run in SAS

Step 3

Paste generated R code into your R program and lean back in your chair. You may have to install the Haven package, which does the heavy lifting of reading in the dataset into R.

Errors, Oh No!!!!!!

In this case the NLST trial has some custom formats for missing values: **.E="Screen date after lung cancer diagnosis"**, **.N="No screen date on record"**, **.W="Wrong Screen Administered"**. Unfortunately, it appears the common approach when reading

SAS datasets into R is to ignore these different types of missing values and just group them together as missing. In R we effectively have NA and NaN. For a biostatistician, the distinction between types of missing is meaningful. For these errors, you will have to decide on your own how to handle it. Maybe it will work fine for you to just set them all as NA, but as they say, your mileage vary.

Final Tips/Questions?/Link

Add the notsorted option to the SAS Proc Format (the default setting for the macro is that categorical variables factors will be sorted alphanumeric).

For ordinal factors, the SAS Macro appears to omit this detail. You will likely need to manually apply this:

```
ordered_vars <- c(34:63)
dataset[ordered_vars] <- lapply(dataset[ordered_vars], as.ordered)
```

Sources:

1. https://github.com/clindocu/sas-r
2. https://wiki.cancerimagingarchive.net/display/NLST

Find Out More