Bruce J. Swihart

Department of Biostatistics
615 N. Wolfe Street, E3138
Baltimore, Maryland 21205

bruce.swihart@gmail.com
https://github.com/~swihart
http://www.biostat.jhsph.edu/~bswihart

Research Statement

My philosophy is to work at the interface between various exciting scientific problems and (bio)statistical methods development. I am inspired by cutting edge scientific problems and direct collaboration with scientists; I aspire to build my career around finding, refining, and creating the best analytic methods that address these problems. I also am interested in building a successful research group in a top Biostatistics or Statistics Department. This will require long-term thinking in terms of choice of problems, collaborators, and careful planning for grant applications as well as careful integration with the Department and University. Building my career, contributing to a great research atmosphere, and helping others around me are my main goals.

One of my interests is real-life, applied, **statistical machine learning and competition**. Indeed, I find that competition is one of the best ways to explore and substantiate methods, motivate students, and promote (bio)statistics. After orchestrating a top 1% finish of the 1358 teams in the Kaggle-hosted $3 million **Heritage Health Prize 1**, in which I utilized established approaches and developed new methods in **statistical machine learning and prediction,** I anticipate being invited and am eager to participate in the invitation-only sequel **Heritage Health Prize 2.** The **Heritage Health Prize 2** will be an unprecedented opportunity to compete in predicting with **big healthcare data** free of de-identification. In addition to competing, my research interests will continue to specifically include **functional data** and **high dimensional data analyses**, especially those with a longitudinal component (multiple visits) and an epidemiological framework (at least several hundred subjects studied with the aim of exploring health-related questions). Applications of this research include but are not limited to **Cancer, HIV, Sleep Epidemiology, Accelerometry, Imaging, (Stroke) Kinematics, Electronic Medical Records (EMRs), and Personalized Medicine.**

Furthermore, my graduate thesis research work focused on the connections of **copulas** to **generalized linear mixed models**, and I plan to continue exploring statistical methods and theory in this area. Future work for the **likelihood-based marginal modeling of binary and ordinal outcomes** includes establishing more general properties and producing **open-source software** for ease of model implementation. In the **machine learning and prediction** arena, I plan to develop a linear mixed model implementation of **Bridge Regression** (a generalization of **Lasso** and **Ridge Regression**).

The theme throughout all of my research is incorporating measurement and analysis methodology for the study of a population of individuals over many visits. In this setting of public health research, data generated by one individual at one visit is often high dimensional or "big." Thus, unearthing associations in a study population of individuals over several visits can be daunting not only at the data management phase but also in the actual analysis. In meeting the challenge this era of big data offers, I have found three areas to be essential to my research in public health: **summarization, visualization, and organization**.

Summarization: Models serve as summarization-filters of datasets and the vehicle for answering questions. Too much summarization may result in oversimplification and loss of useful information, whereas not enough may allow the noise to overtake the signal. In one of my main areas of focus, sleep-research, less summarization and more information were needed. A common summarization of percentage-of-night spent in a sleep stage eschewed transition and time-to-transition information. Thus, fundamental differences in sleep between those with and without sleep-disordered breathing were masked. In contrast, in my stroke kinematic research, more summarization and less information was needed. The size and complex structure of the data were overwhelming. Summarization of two-dimensional curves by prototypical scalars was needed, and the information retained by the scalars was highly predictive of stroke. By formulating models to utilize and honor an appropriate amount of information, discriminating health indices of the sleep/sleep-disordered breathing and the stroke/kinematic processes were derived and then were further studied for associations with health outcomes.

Visualization: A fundamental key to acquiring data features and communicating the process under study is visualization. In medicine the measures of an individual's anatomy and physiology are becoming more sophisticated and being collected longitudinally. A sleep hypnogram is a common visualization of an individual's sleep stage trajectory over time. Trying to view several hypnograms at once, in so-called spaghetti plots of longitudinal data analysis, leads to over-plotting. I established lasagna plots, which utilize heatmaps to view several simultaneous sleep trajectories, and developed a sorting framework to further the utility of visualizing clustered repeated measures data, which appear often in public health research.

Organization: The field of (Bio)statistics appreciates organization, largely through generalization. Logistic regression and ANOVA serve different purposes but are instances of the same framework: generalized linear models. In public health dealing with correlated data of individuals while needing inference of population effects has driven research in the area of marginalized random effect models for binary outcomes. My recent research established these models as a special instance of a copula model and constructed a taxonomy of properties to classify the known marginalized random effect models. The taxonomy development enabled the proposition of a very rich class of models utilizing stable distributions.

As I continue in my career as a public health researcher, I broadly intend to develop methodological approaches to appropriately summarize health correlates contained in high-dimensional data, employ insightful and dynamic visualizations, and aid in the organization of the field by making connections between seemingly distinct methods and generalizations of established methodology.