Data Science in Biostatistics and the Gillings School of Global Public Heath

Donglin Zeng, Professor of Biostatistics, Co-director of the Carolina Survey Research Laboratory (CSRL)

Michael R. Kosorok, Professor and Chair of Biostatistics

Biostatistics Department Data Science Activities

- We focus on analysis of biomedical data and design of biomedical research studies. We are well recognized as a top biostatistics department in US and our faculty are experts in analyzing data from diverse areas: genetics, precision medicine, cancer, longitudinal studies, medical imaging, survey data, environmental health, AIDS, etc.

- Our data have multiple layers and multi-resolutions: genomics, -omics, medical imaging, electronic health records, clinical data; trial and observational data; population and individual data.

- The major activity in data science includes developing powerful new statistical methods (models, algorithms, theory) to tackle emerging challenges in data analysis, design valid surveys and studies for research, collaborating extensively with fields from public health and medical science.

- We house data resources from Atherosclerosis Risk in Communities (ARIC), Hispanic Community Health Study (HCHS), Randomized Intervention for Children with Vesicoureteral Reflux (RIVUS), Subpopulations and Intermediate Outcome Measures in COPD Study (SPIROMICS), etc.

- We currently hold 4 training grants from BD2K, cancer genomics, imaging and genetics, to environmental health so our students are deeply engaged in most emerging data activities.

Data Science Activities in Other SPH Departments

- **Epidemiology** employs a variety of national and international data resources to address a wide range of research questions in public health and social and biomedical sciences. The unique data resources include cohort studies such as ARIC, CBCS for cancer, PIN for pregnancy and medical health records such as ICISS, MarketScan and Medicare health care claims data.

- **Health Behavior** uses data bases such as Behavioral Risk Factor Surveillance System (BRFSS), and numerous survey datasets for specific research purposes.

- **Maternal and Child Health** uses surveillance data (PRAMS), longitudinal survey (Add Health), National Survey of Family Growth, NICHD Study of Early Child Care, and global data (Demographic and Health Surveys, AIDS Indicator Surveys).

- **Nutrition** integrates genetic, epigenetic, biochemical, epidemiological, behavioral, anthropological, and policy-related data to understand interactions between diet and chronic diseases. They house large international datasets such as the China Health and Nutrition Survey, the Cebu Longitudinal Health & Nutrition Survey and US-based cohorts such as SEARCH for Diabetes in Youth Study..

- **Health Policy Management** utilizes numerous data from different agencies including AHRQ, ADA, AMA, BCBS, CDC, etc.

Machine Learning for Precision Medicine

- One cutting-edge data-driven research activity in biostatistics is to advance powerful machine learning methods for precision medicine.
- A number of faculty members including Michael Kosorok, Donglin Zeng, Jason Fine, Danyu Lin and a number of students as well as many collaborators from at least 3 institutions are involved in this ongoing activity.
- We have published a number of statistical method papers in top tier journals.
- P01 grant (Lead PI: Kosorok; just renewed for another 5 years) has the main theme to develop innovative study designs and analytic methods for precision medicine.
- A number of studies using these methods are being carried out at UNC, including an interesting SMART design for laser treatment of scars from severe burns.

Novel Methods for Data-Driven Decision Making

- We invented outcome-weighted learning which enables data-driven decision making without modeling the data.
- Our work is on the cutting edge of data-driven decision making.
- We tackle a variety of data challenges including missing data, censored data, inconsistency between trial and observational populations, high dimensional data, complex interactions, many kinds of modeling including SEM and other hierarchical models, etc.
- We employ both trial data and observational data, single stage and multiple stage, genomics and biomarker data for precision medicine.
- We are working with Big Data from ICISS and other kinds of -omics data to continue advancing methods.

Concluding Slide

- Biostatistics plays a key role in biomedical data science. Its advance ensures internal validity, reproducibility and generalizability of methods for data science.
- On the other hand, continuously emerging challenges in the era of Big Data drive this field to continue developing powerful new analytic methods.
- Other departments in the School of Public Health (SPH) also develop new data science methods, and the SPH is one of the major users of data science on campus.
- We look forward to sharing ideas/thoughts with other data science groups at UNC.
- It is crucially important for UNC
  - to reinforce communication among different disciplines,
  - to fuse innovative strategies to solve common scientific problems,
  - to facilitate access to diverse data resources, and
  - to build efficient research computing infrastructures.