BIOSTATISTICS SEMINAR

Sarah C. Lotspeich, PhD
Postdoctoral Fellow
Department of Biostatistics

Getting thrifty with data quality: Efficient two-phase designs for error-prone data

Clinically meaningful variables are increasingly becoming available in observational databases like electronic health records (EHR). However, these data can be error-prone, giving misleading results in statistical inference. Data auditing can help maintain data quality but is often unrealistic for entire databases (especially large ones like EHR). A cost-effective solution is the two-phase design: error-prone variables are observed for all patients during Phase I and that information is used to select patients for auditing during Phase II. However, even these partial audits can be expensive. To this end, we propose methods to promote the statistical efficiency of two-phase designs, ensuring the integrity of observational cohort data while maximizing our investment. First, given the resource constraints imposed upon data audits, targeting the most informative patients is paramount for efficient statistical inference. Using the asymptotic variance of the maximum likelihood estimator, we compute the most efficient design under complex outcome and exposure misclassification. Since the optimal design depends on unknown parameters, we propose a multi-wave design to approximate it in practice. We demonstrate the superior efficiency of the optimal designs through extensive simulations and illustrate their implementation in observational HIV studies. Then, to obtain efficient odds ratios with partially audited, error-prone data, we propose a semiparametric analysis approach that uses all information and accommodates many error mechanisms. The outcome and covariates can be error-prone, with correlated errors, and selection of Phase II records can depend on Phase I data in an arbitrary manner. We devise an EM algorithm to obtain estimators that are consistent, asymptotically normal, and asymptotically efficient. We demonstrate advantages of the proposed methods through extensive simulations and provide applications to a multi-national HIV cohort.

November 4, 2021 **133 Rosenau Hall** **3:30-4:30 PM**

Zoom Link:
https://unc.zoom.us/j/93545206596?pwd=N1IKeVZjSFhuM2IhSDJCWjN3c21BUT09 Meeting ID: 935 4520 6596 Passcode: 823321

UNC GILLINGS SCHOOL OF GLOBAL PUBLIC HEALTH