# ACMS Statistics Seminar

## Xiaochun Li

**Indiana University School of Medicine**
**Tuesday, September 26, 2023**
**154 Hurley Hall**
**3:30 PM – 4:30 PM**

# Linkage of Big Data of Electronic Medical Records in the Presence of Missing Data

**Background**: Quality patient care requires comprehensive health care data from a broad set of sources. Electronic medical records (EMR) are increasingly distributed across many sources as our nation moves into an era of electronic health record systems. But EMR data are often from independent databases without a common patient identifier, the lack of which impedes data aggregation, causes waste (e.g., tests repeated unnecessarily), affects patient care and hinders research. Record Linkage is the first requite step before effective and efficient patient care and research. Absent a unique universal patient identifier, linkage of patient records is a non-trial task. In addition, the ubiquity of missing data in EMR poses further challenges in record linkage.

**Objectives:** We address the real-world challenges of missing data and matching field selection in linking medical records and evaluate the extent to which incorporating the missing-at-random assumption in the Fellegi-Sunter model and using data-driven selected fields improve patient matching accuracy using real-world use cases.

**Methods:** We incorporated the missing data in the Fellegi-Sunter model using the missing-at-random assumption and compared the proposed approach to the common strategy of treating missing values as disagreement, with matching fields specified by experts or selected by data-driven methods. We used four use cases, each containing a random sample of record pairs with match status ascertained by manual reviews. Use cases included health information exchange (HIE) records deduplication, linkage of public health registry records to HIE, linkage of Social Security Death Master File records to HIE, and newborn screening records deduplication, representative of real-work clinical and public health scenarios. Matching performance was evaluated using sensitivity, specificity, positive predictive value, negative predictive value and F-score.

**Results:** Incorporating the missing-at-random assumption in the Fellegi-Sunter model maintained or improved F-scores whether matching fields were expert-specified or selected by data-driven methods. Combining the missing-at-random assumption and data-driven fields produced the best F-scores in the four use cases.

**Conclusions:** Missing-at-random is a reasonable assumption in real-world record linkage applications: it maintains or improves F-scores regardless of whether matching fields are expert-specified or data-driven. Data-driven selection of fields coupled with MAR achieves the best overall performance, which can be especially useful in privacy-preserving record linkage.

The Department of Applied and Computational
Mathematics and Statistics

Please visit acms.nd.edu to view the full list of speakers.