



## Student Presentations

### **Richard Holtzclaw: “Translating Social Determinants of Health into Standardized Clinical Entities”**

Social Determinates of Health(SDH) are an important and underutilized source of information in modern clinical spaces. What are Social Determinates of health, why are they important, and how and why should we use translational informatics to gather and store this data in standardized ways. There are two main steps in translating SDH into standardized clinical entities, both requiring collaborative efforts. The first is the establishment of a shared ontology of medical terminology codes (i.e. SNOMED, ICD-10, etc.). The second is to use the FHIR standard to create profiles and extensions which will allow FHIR resources to be used to store the coded SDH as clinical entities. The complexities of these steps must be accounted as make an effort to create interoperable clinical informatics solutions which utilize SDH.

### **Carter Allen: “Statistical Models for Multi-layer Biological Networks”**

Networks are ubiquitous data structures in biology as they allow for simultaneous representation of pairwise interactions among a set of entities (e.g., cells, genes, organisms, etc...), unlike tabular data which does not naturally adapt to data with such a complex and highly dependent correlation structure. With the popularization of experimental technologies to sequence genes at the cellular level, it is becoming crucial to develop robust data analysis pipelines to describe the heterogeneity present in network data. Further, with the rise of multi-omics technologies, a need has arisen for methods that can integrate multiple network structures, so called “multi-layered networks”, where each layer of the network data represents the similarity among entities along a certain biological dimension (e.g., RNA profiles, spatial locations, among others). To this end, we have developed a Bayesian Multi-layer Stochastic Block Model (MLSBM) and accompanying R package suited to the task of identifying sub-populations in multi-layered network data. We first develop a full Bayesian specification for the simpler single-layer SBM, and then make the necessary extension of the model to the multi-layer case. We then show that the Bayesian MLSBM is able to more accurately identify clusters when compared to methods that only consider one dimension of the data. Finally, we present the R package, mlsbm for fitting the model and discuss our current work in adapting this method to cutting edge application areas such as spatial transcriptomics.

### **Nida Viqar: “Exploring the advancement of biomedical imaging informatics through natural language processing (article review)”**

Brain imaging is expensive and time consuming to do for large populations research. Currently, automated reading for common brain phenotypes has not been developed. We will be exploring an article where a rule-based NLP algorithm was developed. Two sources of radiology reports were used to develop and test algorithm. Final model of algorithm had excellent sensitivity, specificity, and positive predictive value for cerebral atrophy, cerebral small vessel disease, and deep old infarcts. In conclusion, an NLP algorithm was made with neuroradiology reports that identified patients with important cerebrovascular phenotypes at a large scale.

### **Xiaofu Liu: “Extracting phenotype of diseases and ADEs with text mining”**

This research focuses on extracting phenotypic information from literature with text mining algorithms. And the purpose of the research is to construct a comprehensive database of disease (and ADE) phenotype. The first part, also what I am working on, is to find sentences, which include phenotypic information of diseases or ADEs, as comprehensively as possible. We manually labeled some sentences that can be used as the training set of the following supervised learning model and mapped some features which can help distinguishing sentences including phenotype from other sentences. We then compared the different machine learning models and evaluated the performance with the same set of data. Now, we got a 0.79 precision, 0.90 recall, and 0.84 f1-score. To prompt the performance, the deep learning algorithm will be considered in the next step of our research.

**Friday, November 6<sup>th</sup>, 11:00am-12:00pm**  
**Carmen Zoom**