Data from the EHR provides the opportunity to use real-world and real-time information to assess outcomes and improve predictive models. It is estimated that 80% of healthcare data are unstructured sources, such as clinical notes. Extracting relevant features from unstructured sources is a complex process, and descriptive statistics about the content are not well-described. The ability to copy-forward notes within the EHR potentially introduces outdated, inaccurate, or unnecessary information.

Describe the content and diversity of clinical notes within a large, academic healthcare system. Identify potential features that can be used for feature engineering in downstream models, such as note similarity, frequency, and distribution.

All clinical notes from Yale New Haven Hospital from January 2014 through December 2015. Notes extracted in delimited file and converted to JSON, then analyzed with custom Python scripts. Assessed basic descriptive statistics and lexical content stratified by note type, encounter type, and author specialty. Calculated the similarity for each combination of H&P, ED notes, and progress notes for two patients to assess the feasibility of similarity analysis in free-text notes.

These data provide a comprehensive, descriptive assessment of the diversity in unstructured notes. Multiple features can be rapidly extracted which may be beneficial in downstream analytic models. Future work will apply these foundational data and results to predictive models, such as operative risk scores, to assess whether unstructured content can improve predictive accuracy.