Discovering and Applying Knowledge in Clinical Databases

Columbia University Health Sciences

PI: George M. Hripcsak Grant Number: 3R01LM006910

The long term goal of our ongoing project, "Discovering and applying knowledge in clinical databases," is to learn from data in the electronic health record (EHR) and to apply that knowledge to relevant problems. The advent of the electronic health record (EHR) greatly amplifies the ability to carry out observational research, opening the possibility of covering emerging problems, diverse populations, rare diseases, and chronic diseases in long-term longitudinal studies. Unfortunately, the EHR carries additional challenges. We believe that the biggest challenge comes from the inaccuracy, incompleteness, complexity, and resulting bias inherent in the recording of the health care process. Put another way, EHR data are not simply research data with more noise and missing some values; instead the EHR carries systematic biases that must be addressed before the data can reach their potential. We propose to characterize the effects of the health care process on EHR data, to enumerate the potential biases, and to provide mechanisms to circumvent them. In effect, we propose to study the EHR as an object of interest in itself, using new models, data mining, existing knowledge bases, and innovative algorithms to better understand EHR biases so that we can identify them and correct them or avoid them. We include expertise from two of the nation's major phenotyping projects, eMERGE and OMOP. We hypothesize that we can learn about biases due to the health process through data mining and knowledge engineering and that we can correct or at least avoid those biases, enabling us to better answer informatics and clinical questions. Our aims are as follows: (1) Study health care process biases by correlating raw EHR variables with a panel of health care process-related variables (e.g., admission), using lagged correlation to account for temporal effects, and populating a health care process resource with the correlations and observations. (2) Find associations among raw EHR variables using lagged correlation, information theory, Granger causality, and temporally ordered N-tuples of events, correcting for the health care process biases discovered in Aim 1. (3) Facilitate the definition of higher-level clinical phenotype concepts by applying knowledge resources-including eMERGE and OMOP phenotype definitions and ontologies such as our Medical Entities Dictionary and the UMLS-to the fruit of Aims 1 and 2 to produce semi-automated and automated phenotype query definitions. (4) Develop a high-throughput method to validate phenotype definitions by measuring the ability to uncover known associations, use the generated phenotypes and associations to answer clinical questions, and disseminate the results, including a large knowledge base of correlations that can be used by other researchers to conduct their own studies. PUBLIC HEALTH RELEVANCE: This project studies the electronic health record in order to better understand how health care processes cause problems in the data. By avoiding or correcting those problems, we hope to improve reuse of the data for purposes such as clinical research and quality improvement.