Data frameworks and disciplinary collaborations to aid in the use of electronic health records (EHR) for improved understanding of population health, disease etiology, and treatment effectiveness.

A position paper for the Specialist Meeting on Human Dynamics in the Mobile Age, August 11-12, 2014

Caroline A. Thompson, PhD, MPH
Assistant Professor of Epidemiology, Graduate School of Public Health, San Diego State University

The rapid wide-scale adoption of the electronic health record (EHR) as a result of “meaningful use” initiatives offers great promise for evidence-based medicine to use routinely collected data on patients of all types, rather than relying on randomized controlled trials (RCTs) of selected patients. EHR data for cancer research are especially promising because most cancer care is delivered in healthcare systems, and RCTs often lack representativeness since they enroll only 3% of United States cancer patients. EHR data have important limitations for many types of research, however, due to systematic errors, arising from wide variation in data reliability. Research using such data sources requires rigorous attention to study design. A particular problem with EHR data are instances of non-random completely missing data due to provider use variations or patient migration. Most populations captured in EHR systems are highly dynamic with frequent in- and out- migration based on patient choice, employment, insurance, and geography and this may blind researchers to some types of care. Gaps in care records and poorly defined source populations can not only lead to difficulties for inference, but also pose fundamental challenges in identifying (and comparing) appropriate study and target populations. The impact of these problems is heightened when data “missingness” may be related to the study hypotheses.

Data linkage has been seen as solution to filling the gaps when using incomplete data from a single source. Linkage requires finding data on the same persons in multiple EHRs, or in other population-based data sources such as cancer registries. Cancer registries are particularly useful because they include incidence-based surveillance data and baseline clinical characteristics of disease that are often lacking in, or cannot be easily derived from, EHRs. Cancer registries, however, generally do not include data on repeated interventions, or cancer recurrences, but these details may be derived from a comprehensive EHR. Cancers with high rates of survival, e.g., breast cancer, are often characterized by treatment periods, or episodes of care, that can continue intermittently for months or years. Thus, linking two or more institution EHRs, especially if they serve the same catchment area, can fill in gaps in patient care pathways, and improve studies of the comparative effectiveness of different care settings and interventions. Linking data sets is a necessary, but not sufficient solution to fragmented data; careful consideration of research questions, the appropriateness of available data, and relevance of the population to answer such questions should not be ignored.

Many disease states that are the object of epidemiologic investigation can only be inferred from a constellation of features in a typical EHR. For example, metastatic cancer recurrence is a highly fatal potential long term outcome for any cancer survivor. Identifying recurrent cancer patients from EHR is a substantial challenge, which involves extracting patterns of care that are typical in recurrence, and until recently this has only been implemented using cohort querying techniques primarily defined by physician expert understanding of disease and treatment processes. It is likely that advanced data science techniques such as machine learning for such "clinical phenotyping" will eventually replace cumbersome clinically-driven cohort querying that may miss important features to distinguish patient classes. Indeed, for the topic of breast cancer recurrence, work has already been done using natural language processing (NLP) to improve consistency and accuracy of the information gleaned from the health record.

Data mining techniques cannot replace traditional epidemiologic studies for the purpose of determining disease etiology or comparative effectiveness research (CER), however. Distinguishing signal from noise in big health data is a substantial concern because observational population health research is already fraught with biases such as confounding by indication, lost-to-follow-up bias, and threats to external validity due to inability to generalize to a source population using convenience samples (such as those provided by one or more EHRs.) A medical research community and a "headline thirsty" media presence that are both highly focused on large sample sizes and the statistical significance of findings may result in hurdles and blunders for EHR-based research. In very large non-experimental data, heterogeneity of both measured and (especially) unmeasured variables will contribute to highly significant, but potentially invalid inferences. Validation studies are imperative. Additionally, causal inference theory that underpins much of modern epidemiology provides some important tools to aid in the exploration of assumptions, and robustness of findings. These include directed acyclic graphs (DAGs) for depiction of data generating mechanisms and identification of potential sources of bias, and quantitative bias analysis. Sensitivity analyses should not limited inferential studies (e.g., CER), especially when sample sizes are large and heterogeneous. Even seemingly innocuous descriptive studies can be deceiving if one does not carefully consider the underlying data generating mechanisms (which become complex very quickly when linkage is involved) and the careful definition of the population of interest.

Successful use of EHR for etiologic research will require the cooperation of many disciplines, especially medicine, bioinformatics, epidemiology, and statistics. For hypothesis generation and "signal quenching" data mining is already an important strategy to approach these data. Deep learning techniques to aid in clinical phenotyping are promising. With the aid of validation studies using epidemiologic study design (and randomized controlled trials, when possible), the harmonization of these data sources, and harmony of these scientists may propel our understanding of disease states, treatment options, and provide the knowledge necessary to move into the age of "precision medicine".