Mining electronic health record data: finding the gold nuggets

Lucila Ohno-Machado, Editor-in-Chief

Electronic health records (EHRs) have been increasingly adopted in the United States because of governmental incentives and the realization that healthcare should not lag behind other industries in deriving knowledge from "big data." There has been much discussion about the use of big data to discover patterns for targeted therapy and disease prevention. However, much less is said about the readiness of EHR data for such data mining initiatives. The existence of data is often equated with the existence of good data, ie, high-quality, standardized data that can be used in sophisticated data analyses to reveal patterns that previously escaped observation. Although the difficulties of preparing data for such analyses are well known to informaticians, biostatisticians, and computer scientists specializing in machine learning, not all stakeholders (ie, administrators, clinicians, researchers, and patients) appreciate the challenges of using data from different health systems for meaningful analyses. It is not uncommon for such stakeholders to assume that, if health systems utilize the same software versions of the same EHR system, then the data will be immediately comparable. Because bringing data from different health systems together is so difficult, due to privacy concerns and institutional policies, many believe that the investment in bringing the data “together” (in a federated or in a centralized way) is sufficient to allow immediate analyses. This is not so. The inconvenient truth is that much needs to be done to EHR data before they can be used for analyses and decision-making – a topic that has been a focus of the informatics community for many years.

The first differences that becomes apparent when data from different health systems are brought together relate to the overall format of those data. Data format in and of itself can be relatively easy to fix, but addressing health systems’ heterogeneous utilization of ontologies and terminology standards and differences in semantics (eg, the various definitions of “uncontrolled diabetes”) requires intervention from data modeling experts as well as biomedical or behavioral domain experts. While the last issue of JAMIA focused on data standards, this issue focuses on the transformation of narrative text into structured data using natural language processing techniques. This fundamental preprocessing step towards data mining extracts concepts from clinical notes and the body of biomedical research literature (see pages 938 to 1020 for articles in our Special Focus on natural language processing).

Once data are harmonized and structured, as well as processed to prevent re-identification (see page 1029, see page 1072), they can potentially be used for various types of data mining, such as malpractice mitigation (see page 1020). More examples of data mining initiatives are included in this issue: associating month of birth (as a proxy for seasonal maternal-infant exposures) with certain diseases (see page 1042), and risk stratification for acute kidney disease (see page 1054).

The current status and future direction of EHR systems should be discussed further. EHR systems are currently used for public health monitoring, eg, tuberculosis contact investigation (see page 1089) and have also been used extensively for quality improvement and clinical decision support (see page 1081), eg, to decrease the rate of adverse events via e-prescribing (see page 1094). However, these types of activities can only be effective across EHR systems if those systems are truly interoperable (see page 1099) and used properly. For this to happen, EHR systems need to be designed with end users in mind (see page 1102).

Our informatics community is working relentlessly to prepare EHR data for data mining and is uniquely positioned to evaluate EHR data’s quality and usefulness in a variety of applications. JAMIA will continue to publish “gold nuggets” found in the course of data mining as well as negative results that significantly contribute to the body of informatics knowledge.