Free Text to Standardized Concepts to Clinical Decisions

F

Eva K. Lee
Georgia Institute of Technology, USA

Brent M. Egan
American Medical Association, USA

INTRODUCTION

Clinical decision making is complicated since it requires physicians to infer information from a given case and determine the best treatment based on their knowledge. Data from electronic medical records (EMRs) can reveal critical variables that impact treatment outcomes and inform the allocation of limited time and resources, allowing physicians to practice evidence-based treatment tailored to individual patient conditions. On a larger scale, realistically modifiable social determinants of health that will improve community health can potentially be discovered and addressed.

Although EMR adoption is spreading across the industry, many providers continue to document clinical findings, procedures, and outcomes with "free text" natural language in their EMRs. They have difficulty (manually) mapping concepts to standardized terminologies and struggle with application programs that use structured clinical data. This creates challenges for (multi-site) comparative effectiveness studies. Standardized clinical terminologies (e.g., SNOMED-CT, LOINC, RxNorm, UMLS) are essential to facilitate interoperability among EMR systems. They allow seamless sharing and exchange of healthcare information for quality care delivery and coordination among multiple sites. However, the volume and number of available clinical terminologies are large and expanding. Further, due to the increase in medical knowledge and the continued development of more advanced computerized medical systems, the use of clinical terminologies has extended beyond diagnostic classification.

This chapter summarizes our work in (1) designing an efficient, robust, and customizable information extraction and pre-processing pipeline for electronic medical records; (2) automatic mapping, standardization, and establishing interoperability; (3) uncovering best practices across multiple sites via machine learning; and (4) optimizing access timing and treatment decisions for chronic kidney disease patients Lee et al., 2016, 2019, 2021, 2022; Lee & Uppal 2020).

The work tackles over 800 clinical sites covering 9,000 providers and de-identified data for over 3.0 million patients with health records spanning the last 26 years. To the best of our knowledge, EMR data analysis across hundreds of sites and millions of patients has not been attempted previously. Such analysis requires effective database management, data extraction, preprocessing, and integration. In addition, temporal data mining of longitudinal health data cannot currently be achieved through statistically and computationally efficient methodologies and is still under-explored. This is a particularly important issue when analyzing outcome, health equity, and health conditions for chronic disease patients.

We first extract cohorts of patients from EMRs by disease / symptoms, and treatment features. Content discovery, concept mapping and interoperability are then established among EMRs from the 800+ clinical sites by developing a system that rapidly extracts and accurately maps free text to concise structured

DOI: 10.4018/978-1-7998-9220-5.ch028

medical concepts. Multiple concepts and contents are extracted, and mapped, including patient diagnoses, laboratory results, medications, and procedures, which allows shared characterization and hierarchical comparison. A mixed integer programming-based machine learning model (DAMIP) is next applied to establish classification rules with relatively small subsets of discriminatory features that can be used to predict treatment outcomes for cardiovascular and chronic kidney diseases. Based on our results, optimal treatment design and associated new clinical practice guideline for chronic kidney disease pre-dialysis initiation is demonstrated. The results facilitate improved outcome, health quality, and cost-reduction for patients. Our findings can speed dissemination and implementation of *best practice* among all sites. Rapid learning across multiple sites show that improvement can be achieved within 12 months with a better health outcome, enhanced quality, and reduced cost. The next step will involve analyzing 60 million patients across the United States.

BACKGROUND

Data Extraction, Encryption, and Concept Standardization

It is challenging to establish an efficient data extraction schema for EMR due to the complexity of data and lack of data standards. A common task in EMR is case detection – identifying a cohort of patients with a certain condition or symptom. Coded data such as International Classification of Diseases (ICD) are often not sufficient or accurate (Birman-Deych et al., 2005, Sonabend W et al., 2020). Informatics approaches combining structured EMR data with narrative text data achieve better performance (Li et al., 2008; Savova et al., 2010). Key clinical items can be extracted from narrative texts with methods such as pattern matching using regular expressions (Long, 2005; Turchin et al., 2006; Friedlin and McDonald, 2006; Ravikumar & Ramakanth Kumar, 2021), full or partial parsing based on morpho-semantems (Baud et al., 1998; Mamlin et al., 2003), and syntactic and semantic analysis (Friedman et al., 1994; Jain and Friedman, 1997). Increasingly, more complex statistical and rule-based machine learning approaches (Bashyam and Taira, 2005; Taira and Soderland, 1999) have been developed to tackle this challenge. Biomedical Named Entity Recognition (NER) – the "task of identifying words and phrases in free text that belong to certain classes of interest" (Settles, 2004), allows users to identify key clinical concepts such as physician visits, referrals, dietary management, and suspected problems normally not present in structured data tables.

Once patient information is extracted, data security and confidentiality must be ensured through de-identification steps. According to the Health Insurance Portability and Accountability Act (HIPAA), patients' Protected Health Information (PHI) must be de-identified or anonymized for commercial and research use. PHI exists in both structured and unstructured clinical records (Zikopoulos and Eaton, 2011). This includes patient names, addresses, phone numbers, etc. Manual and rule-based or lexicon-based methods have been used to achieve PHI de-identification (Sweeney, 1996; Ruch et al., 2000; Taira et al., 2002), but they are extremely time-consuming and can be inaccurate. Machine learning approaches have also been developed (Sibanda and Uzuner, 2006; Wellner et al., 2007; Szarvas et al., 2007; Phuong & Chau, 2016). However, due to the complexity of data schemas and the heterogeneity of data structures, it is very challenging to detect PHI with high sensitivity.

Because EMR data include various types of records for patients, data standardization is essential prior to analytic investigation. With multiple facilities and providers, the problem is compounded as