Technical Note

Extracting and utilizing electronic health data from Epic for research

Alex Milinovich, Michael W. Kattan

Department of Quantitative Health Sciences, Cleveland Clinic, Cleveland, OH, USA

Correspondence to: Michael W. Kattan, PhD. Department of Quantitative Health Sciences, Lerner Research Institute, Cleveland Clinic, 9500 Euclid Avenue, JJN3-01, Cleveland, OH 44195, USA. Email: kattanm@ccf.org.

Abstract: Many institutions would like to harness their electronic health record (EHR) data for research. However, with many EHR systems, this process is remarkably difficult. We have been using our vast EHR system for research very effectively, with substantial research support and many publications. Herein we share our process and provide recommendations for others wanting to utilize their EHR data for research.

Keywords: Electronic health records (EHRs); Epic; Unified Medical Language System (UMLS)

Submitted Jan 03, 2018. Accepted for publication Jan 09, 2018.
doi: 10.21037/atm.2018.01.13
View this article at: http://dx.doi.org/10.21037/atm.2018.01.13

Introduction

Raw electronic health record (EHR) data are disorganized and full of uncodified variables. Working directly with EHR data for statistical analysis is a challenge in and of itself. Many data points are duplicated and are reliant on upon a very small set of validation criteria shown to the data entry personnel. Intimate knowledge of the data structure of the EHR is necessary for even the simplest of queries. Extracting EHR data is a difficult, time consuming, and often a pragmatic process.

Cleveland Clinic adopted Epic's EHR system in 1995 in the laboratories, and expanded to include medications in 1998, Epic outpatient in 2000, surgical histories in 2002, and Epic inpatient in 2005. However, at Cleveland Clinic, less than 5% of the EHR data are codified variables. The rest are identifiers, dates, and free-text entries. To provide the cleanest and most robust datasets for statistical analysis, numerous statistical techniques including similarity calculations and fuzzy matching are used to clean, parse, map and validate the raw EHR data. The raw data are extracted from both the EHR and other disparate data sources, mapped to discrete ontologies, cleaned and standardized, and finally deposited into a clinical research data repository. Approximately 185 tables from different data sources are condensed into 18 research-ready tables in the data repository automatically on a weekly basis. Via this process, Cleveland Clinic can do live population exploration as well as produce datasets for analysis faster than it takes most organizations to simply identify their base population.

Methods

For this data repository, we utilize Unified Medical Language System (UMLS) identifiers. The Metathesaurus from the UMLS combines synonymous terms and codes from disparate medical vocabularies into concise terms and identifiers. The UMLS data set is freely available for download after applying for a license and verifying your institution has licenses for the specific vocabularies that require one. A subset of the data is generated by the user encompassing the languages and vocabularies of his or her choosing. The subset can then be loaded into a variety of database types such as Oracle, Microsoft SQL, and MySQL.

By mapping as many variables as possible to UMLS identifiers, a simplified data structure can be implemented to store and query the EHR data. Because the UMLS combines many disparate medical vocabularies into

succinct terms, queries become simpler yet more robust with the UMLS's inclusion of relationships, hierarchies, and synonyms among the various terms. For example, the term of "Heart Failure" (C0018801) has relationships to various medications that may treat heart failure, finding sites of heart & myocardium as well as child diagnoses such as congestive heart failure and left-sided heart failure. These relationships allow for easier querying of the data because researchers can identify top-level terms, and then algorithmically identify any child or related terms for their population definitions.

Results

Simplifying the data structure into UMLS identifiers saves thousands of hours in defining populations and extracting normalized data sets for analysis. See *Supplementary* for an example population definition. This, combined with the 18 research-ready tables, allows Cleveland Clinic to move from study design, to data extraction, to analysis very rapidly. The structure also gives Cleveland Clinic the ability to identify and resolve data issues quickly and easily such as removing test patients and invalid lab values. The data issues can then be resolved permanently in most cases by implementing the fix in the extract, transform, and load (ETL) process which moves the data from the source system into the data repository. Many potential data headaches are solved before the data pulled for analysis.

Conclusions

Having the ability to simplify EHR research has many advantages when dealing the vast amounts of data. At Cleveland Clinic, 6.8 million terms are mapped to UMLS identifiers which accounts for over 35 billion individual data points for over 4 million patients. In addition, approximately half a million custom UMLS identifiers have been added at Cleveland Clinic to include providers, locations, and their relationships between each other. In the end, only 9% (approximately 1,000 data points per patient) of columns in the data repository do not utilize UMLS identifiers. These non-UMLS columns include patient identifiers, dates, and visit identifiers. Ultimately, there are approximately 32,000 discrete data elements per patient comprised of both UMLS and non-UMLS data.

Acknowledgements

The authors would like to thank Stephanie Kocian for her editing of the manuscript.

Footnote

Conflicts of Interest: The authors have no conflicts of interest to declare.

Cite this article as: Milinovich A, Kattan MW. Extracting and utilizing electronic health data from Epic for research. Ann Transl Med 2018;6(3):42. doi: 10.21037/atm.2018.01.13

Supplementary

Sample population

Diabetic patients on a GLP-1 medication with an HbA1c >10

- Diabetics = C0011847-Diabetes
  - Get all “child” concepts
  - 2,666 different diagnoses
    - 147 ICD9 codes
    - 792 ICD10 codes
- GLP-1 = C2916791-Glucagon-like Peptide-1 (GLP-1) Agonists [MoA]
  - Get all concepts that have an active ingredient or are a tradename of
  - 69 different medications
- HbA1c = C0366781-Hemoglobin A1c/Hemoglobin.total:Mass Fraction:Point in time:Whole blood:Quantitative
  - 15 different labs in Epic map to this ConceptID