Evaluation considerations for EHR-based phenotyping algorithms: A case study for drug-induced liver injury

Casey Lynnette Overby1; Chunhua Weng1; Krystl Haerian1; Adler Perotte1; Carol Friedman1; George Hripcsak1
1Department of Biomedical Informatics, Columbia University, New York, NY

Abstract
Developing electronic health record (EHR) phenotyping algorithms involves generating queries that run across the EHR data repository. Algorithms are commonly assessed within demonstration studies. There remains, however, little emphasis on assessing the precision and accuracy of measurement methods during the evaluation process. Depending on the complexity of an algorithm, interim refinements may be required to improve measurement methods. Therefore, we develop an evaluation framework that incorporates both measurement and demonstration studies. We evaluate a baseline EHR phenotyping algorithm for drug induced liver injury (DILI) developed in collaboration with electronic Medical Records Genomics (eMERGE) network participants. We conduct a measurement study and report qualitative (i.e., perceptions of evaluation approach effectiveness) and quantitative (i.e., inter-rater reliability) measures. We also conduct a demonstration study and report qualitative (i.e., appropriateness of results) and quantitative (i.e., positive predictive value) measures. Given results from the measurement study, our evaluation approach underwent multiple changes including the addition of laboratory value visualization and an expanded review of clinical notes. Results from the demonstration study informed changes to our algorithm. For example, given the goal of eMERGE to identify patients who may have a genetic susceptibility to DILI, we excluded overdose patients.

Introduction
A common process for developing electronic health record (EHR) phenotyping algorithms involves an iterative approach to generate queries that run across the EHR data repository. Such algorithms may specify structured data that can be efficiently extracted from EHRs including laboratory values, diagnostic codes, and computerized provider order entry (CPOE) records. In addition, natural language processing (NLP) algorithms are useful for extracting information from clinical notes. These extraction strategies can be scaled and incorporated into phenotyping algorithms with use of terminology systems such as the Unified Medical Language System (UMLS). While NLP for clinical note mining may improve overall algorithm performance, it is common practice to first assess performance with structured data alone.

The eMERGE (Electronic Medical Records and Genomics) Network [3] has made significant contributions towards developing validated EHR phenotyping algorithms for a range of disease phenotypes that can be shared across institutions.[4] Once an algorithm is completed, it is common for investigators to conduct demonstration studies evaluating and reporting quantitative performance measures such as positive predictive value (PPV) and negative predictive value (NPV). Given the variable complexity of eMERGE EHR algorithms [5], however, there may be challenges to defining algorithm criteria that in turn affects performance.

During the EHR phenotype development process, investigators conduct informal measurement studies (e.g., assessing how accurately elevated liver enzymes measures the presence of acute liver injury by reviewing one or two patients). These studies, however, are insufficient and there is a need for formal approaches to conduct both measurement and demonstration studies together. In collaboration with eMERGE network colleagues; we developed an EHR phenotyping algorithm to identify patients with drug induced liver injury (DILI). Although the benefit of medications commonly associated with DILI outweighs the risk in the general population, we want to steer patients at greatest risk away from particular drugs. Studies investigating the underlying genetic and environmental factors associated with DILI may facilitate characterizing these patients. The goal for our EHR phenotyping algorithm is to facilitate identifying patients who have experienced DILI for recruitment into such studies. Given the complexity of how DILI is represented in the EHR, this algorithm provides a good case study for exploring an evaluation framework that incorporates both measurement and demonstration studies. Here we describe lessons from an early, baseline, algorithm that informed changes in measurement approaches and algorithm design (validation of that algorithm is in progress).

Background
There are several challenges to identifying DILI cases that influence the precision and accuracy of measurement methods. We draw from existing methods for evaluation[6] to define an evaluation framework that involves both measurement and demonstration methods.
Evaluations of EHR phenotyping algorithms developed within eMERGE to date primarily assess the PPV. However, there are characteristics of DILI that lend itself to a broader evaluation approach. Compared to previous conditions investigated within eMERGE (e.g., diabetes[7], cataracts[8]), DILI is a rare condition with added complexity given its dependence on the pharmacological action of an indicated drug.

DILI accounts for around 20% of all hospital admissions due to severe liver injury and 50% of acute liver failure cases in the United States.[9] Even so, it remains a rare condition with an estimated incidence ranging from 4 to 15 cases per 100,000 in the UK.[10, 11] When developing EHR phenotyping algorithms to identify cases with rare conditions of low prevalence and incidence such as DILI, it may be more important to optimize the NPV compared to the PPV.[12] Although this would require follow-up screening, optimizing the NPV will reduce the number of missed cases. Also, applying EHR phenotyping for electronic screening reduces the otherwise screening burden, and avoids unnecessary manual review of ineligible patients. Moreover, while it is possible to produce more accurate results by optimizing the PPV and narrowing the cohort, it can introduce biases due to insufficient documentation.[13]

For DILI, the potential to introduce biases are greater given the need to determine a drug as the causal agent of liver injury. To add to the complexity, many different drugs can cause DILI and the pattern of injury varies between drugs. The pattern of liver injury is based on liver enzyme elevations[14-16], and while there are tools to assess the causality of drugs[17-20] there are no specific tests to confirm causality. Given these challenges, we draw from existing evaluation philosophy to define an evaluation framework for EHR phenotyping algorithms.

**Methods**

**A framework for evaluating EHR phenotyping algorithms**

In the context of EHR phenotype algorithms, we describe the study purpose as the major axis of evaluation. The purpose might be to (a) improve measurement approaches utilized within the algorithm (measurement study e.g., inter-rater reliability), or (b) demonstrate the value of the algorithm (demonstration study e.g., PPV calculation). A demonstration study “establishes a relation – which may be associational or causal – between a set of measured variables.”[6] The goal for a measurement study is “to determine the extent and nature of the errors with which a measurement is made using a specific instrument.”[6] As a secondary axis, we describe the data recorded during the evaluation as either qualitative or quantitative.

**A baseline electronic health record phenotype algorithm for drug induced liver injury**

We initially adapt a DILI case definition and algorithm informed by the International Serious Adverse Events Consortium (iSAEC).[1] Each portion of the case definition (Box 1, 1a-2b) is translated into a baseline executable EHR algorithm (Figure 1, A1-D). The algorithm leveraged primarily structured data from the NewYork-Presbyterian Hospital (NYP) clinical data warehouse (CDW). In the first step of the algorithm (Figure 1, A1; Box 1, 1), liver injury ICD-9 diagnosis and procedure codes were specified as inclusion criteria according to Observational Medical Outcomes Partnership (OMOP).[21] Exposure to a drug was specified as any medication ordered within 90 days prior to liver injury diagnosis (Figure 1, A2; Box 1, 1a). We excluded patients with a history of the same diagnosis from Step A2 within 5 years previous (Figure 1, B; Box 1, 1b). We also leveraged the NYP Medical Entities Dictionary (MED) that contains over 60,000 concepts organized into a semantic network of terms.[22, 23] We queried the MED hierarchy to compile laboratory codes including intestinal alkaline phosphatase (ALP), alanine aminotransferase (ALT), and serum bilirubin indirect (Bilirubin). We assess whether peak laboratory values within 180 days following a medication order were above thresholds specified by iSAEC (Figure 1, C1-C4, Box 1, 1a). We also used the MED to compile ICD-9 codes for other diagnoses (Figure 1, D; Box 1, 2b) for exclusion from algorithm results. We report counts at each step of the algorithm.

**Baseline algorithm evaluation approaches**

Our primary evaluation goal was to assess whether the results from the baseline EHR phenotyping algorithm were appropriate given the DILI case definition (as opposed to appropriate given the phenotype algorithm criteria). This evaluation involved both demonstration and measurement studies.

We first defined a review protocol for assessing the alignment of results with the DILI case definition. It specified manual review of the following artifacts: laboratory values (APH, ALT and Bilirubin), medication orders, and discharge summary notes. This protocol described a general approach to classify each result as a true positive (TP), false positive (FP) or not applicable (NA) case. If a patient did not have a discharge summary,
they were classified as NA. Four independent reviewers with informatics training (two physicians and two non-physicians) used the protocol to guide the review of algorithm results. We randomly selected 100 cases from algorithm results for manual review. Each reviewer assessed 40 cases, 20 were overlapping (i.e., reviewed by all reviewers). Both quantitative and qualitative data to assess our measurement approach and to demonstrate the accuracy of the algorithm were collected using CAT (Coding Analysis Toolkit).[24] The reviewers used it to code results and to record notes regarding reason for TP/FP/NA assignments. The CAT ‘Kappa Tool’ was used to run a comparison of inter-coder reliability measures using Fleiss’ Kappa[25] for the 20 overlapping cases (quantitative data for assessing the measurement approach). Quantitative data for the demonstration study were TP, FP, and NA counts that were used to estimate the PPV. Qualitative notes were assessed by identifying themes associated with FPs. Upon completion of the initial review, the reviewers met to discuss perceptions of evaluation approach effectiveness and appropriateness of results. We used themes associated with FPs to understand how the algorithm should be refined to improve the proportion of appropriate results. The term “appropriate” is used given the result may be correct for an EHR phenotype algorithm definition, but inappropriate given the DILI case definition. We report conclusions drawn from our initial assessment and subsequent decisions regarding our measurement approach and algorithm design.

Results

The baseline DILI EHR phenotyping algorithm (Figure 1) was executed as a series of filtering steps beginning with the NYP CDW population between 2004 and 2012 (N=1,045,125). Counts from each filter are shown in Box 2. After applying all algorithm filters, 560 patients met the criteria for DILI.

Results from evaluating the baseline DILI EHR phenotyping algorithm are summarized in Table 1. The measurement study indicated moderate agreement among reviewers (kappa = 0.5). During the post-evaluation discussion regarding perceptions of evaluation approach effectiveness, we identified two main issues: (a) the evaluation platform influenced the manual review process, and (b) indications of DILI were not always reported in discharge summary notes. The two platforms available to review algorithm results were WebCIS (Web-based clinical information system)[26] and nypx.nyp.org. We determined that two reviewers used WebCIS and two used nypx.nyp.org. While we found the technology provided very similar views of patient data, there were two differences that influenced this evaluation: (a) laboratory value visualization capabilities differed between the two systems; and (b) the available notes differed between the two systems. Through further discussion, we also found that there were instances where indications of DILI were found in documents other than discharge summaries. One reviewer, suggested that admission, resident and consult notes were also important to consider. To improve agreement among reviewers and to better systemize the evaluation process, we adjusted future measurement approaches to include a first pass consensus meeting. The consensus meeting involves visualizing the pattern of laboratory values with indications of when a medication was prescribed and when the patient was diagnosed with liver injury with the team of reviewers. Together, the reviewers classify patients as TP, FP, or unknown. Following, one reviewer uses nypx.nyp.org to investigate and confirm or change preliminary classifications. These classifications are coded using CAT to facilitate validation or consensus adjudication.
The demonstration study indicated that algorithm results consisted of 27% TPs, 42% FPs, and 30% NAs. Our PPV was estimated to be 39%. In addition, many results were correct for the algorithm but inappropriate given the DILI case definition. Relevant FP results were often babies, patients who died, overdose patients, or patients who had a liver transplant. Patients who were babies, died or had a liver transplant were inappropriate primarily because their liver enzymes were too unstable to clearly categorize them as DILI cases. Overdose patients were inappropriate because the cause of DILI was already known for these patients. Therefore, to improve the number of appropriate results (and ultimately our PPV) we specified these conditions as exclusion criteria in our updated algorithm. In addition, given many patients in our result set did not have discharge summaries, we decided to specify the inclusion of only inpatients. By doing this we are more likely to have patients with similar medical record artifacts and therefore will further normalize our evaluation process.

**Box 2. Baseline DILI algorithm counts**

| Measurement study | Demonstration study |
|-------------------|---------------------|
| Kappa coefficient: 0.50 | TP: 27  |
|                    | FP: 42  |
|                    | NA: 30  |
|                  | PPV: TP/(TP+FP) = 27/(42+27) = 39%  |

**Table 1. Summary of results from baseline algorithm analysis**

| Quantitative results | Perceptions of evaluation approach effectiveness:  |
|----------------------|--------------------------------------------------|
|                      | • Differences between evaluation platforms        |
|                      | o Visualizing lab values                           |
|                      | o Availability of notes                            |
|                      | • Discharge summary vs. other notes                |

| Qualitative results | Perceptions of benefit of results (themes in FPs):  |
|---------------------|---------------------------------------------------|
|                      | • Babies                                           |
|                      | • Patients who died                                |
|                      | • Overdose patients                                |
|                      | • Patients who had a liver transplant              |

**Discussion**

This evaluation highlights several challenges for evaluating EHR-based phenotyping algorithms including: (a) insuring reliable measures, and (b) drawing appropriate inferences about performance from demonstration studies. Combining measurement and demonstration studies can help address these challenges.

To insure reliable measures, it is important to distinguish between the case definition and phenotype algorithm definition. For example, in the initial demonstration study assessment, several of our FP results were transplant patients. In many cases these results were correct for the algorithm, however, liver enzyme laboratory values were elevated as a result of the procedure rather than from a medication. Therefore, to improve the reliability of our measurements, we refined our algorithm by excluding transplant patients. As an important note, this approach will introduce a bias toward less critical DILI cases given many DILI cases must have liver transplants as a result of their condition - also emphasizing the need to understand the impact of algorithm design decisions. Rather than relying solely on top-down knowledge engineering to make these difficult design decisions, it may be possible to incorporate bottom-up learning from the data.[27] For example, if we can apply data mining methods to characterize patient data, it may be possible to quantify biases and use these data to better inform design decisions or to determine alternative data-driven approaches that avoid introducing biases all together.

In order to draw appropriate inference about performance, it is important to understand the influence of measurement approaches on the results of demonstration studies. While we were able to calculate a PPV for our algorithm, this calculation relies heavily on the algorithm criteria. For example, our measurement study assessing these criteria indicated moderate agreement among reviewers. Thus, the interpretation of our PPV result was unclear. Our next step, therefore, was to make improvements to both our measurement approach and algorithm design so that performance results can be better interpreted. Moreover, it is important to understand characteristics of the condition under investigation. Given DILI is a rare condition of low prevalence, it may be more (or equally) important to optimize NPV compared to PPV. Both measures are currently being investigated.
Acknowledgements

We are grateful for the interest and contributions of eMERGE network collaborators at Mount Sinai (Drs. Erwin Bottiger and Omri Gottesman) and Mayo Clinic (Drs. Jyotishman Pathak, Christopher G. Chute and Ifthikar J. Kullo). We also thank our International Serious Adverse Events Consortium (iSAEC) collaborator, Dr. Yufeng Shen. This research is supported in part by the Columbia Training in Biomedical Informatics (NIH NLM #T15 LM007079) and the eMERGE network, U01 HG006380-01 (Mount Sinai).

References

1. Aithal GP, Watkins PB, Andrade RJ, Larrey D, Molokhia M, Takikawa H, Hunt CM, Wilke RA, Avigan M, Kaplowitz N et al: Case definition and phenotype standardization in drug-induced liver injury. Clin Pharmacol Ther 2011, 89(6):806-815.

2. Friedman C: A broad-coverage natural language processing system. Proc AMIA Symp 2000:270-274.

3. McCarty CA, Chisholm RL, Chute CG, Kullo IJ, Jarvik GP, Larson EB, Li R, Masys DR, Ritchie MD, Roden DM et al: The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. BMC Med Genomics 2011, 4:13.

4. Kho AN, Pacheco JA, Peissig PL, Rasmussen L, Newton KM, Weston N, Crane PK, Pathak J, Chute CG, Bielinski SJ et al: Electronic medical records for genetic research: results of the eMERGE consortium. Sci Transl Med 2011, 3(79):79re71.

5. Conway M, Berg RL, Carrell D, Denny JC, Kho AN, Kullo IJ, Linneman JG, Pacheco JA, Peissig P, Rasmussen L et al: Analyzing the heterogeneity and complexity of Electronic Health Record oriented phenotyping algorithms. AMIA Annu Symp Proc 2011, 2011:274-283.

6. Friedman CP, Wyatt J: Evaluation methods in medical informatics, 2nd edn. New York: Springer; 2006.

7. Kho AN, Hayes MG, Rasmussen-Torvik L, Pacheco JA, Thompson WK, Armstrong LL, Denny JC, Peissig PL, Miller AW, Wei WQ et al: Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study. J Am Med Inform Assoc 2012, 19(2):212-218.

8. Peissig PL, Rasmussen LV, Berg RL, Linneman JG, McCarty CA, Waudby C, Chen L, Denny JC, Wilke RA, Pathak J et al: Importance of multi-modal approaches to effectively identify cataract cases from electronic health records. J Am Med Inform Assoc 2012, 19(2):225-234.

9. Lee WM: Drug-induced hepatotoxicity. N Engl J Med 2003, 349(5):474-485.

10. Aithal GP, Rawlins MD, Day CP: Accuracy of hepatic adverse drug reaction reporting in one English health region. Bmj 1999, 319(7224):1541.

11. Sgro C, Clinard F, Ouazir K, Chanay H, Allard C, Guilleminet C, Lenoir C, Lemoine A, Hillon P: Incidence of drug-induced hepatic injuries: a French population-based study. Hepatology 2002, 36(2):451-455.

12. Thadani SR, Weng C, Bigger JT, Ennever JF, Wajngurt D: Electronic screening improves efficiency in clinical trial recruitment. J Am Med Inform Assoc 2009, 16(6):869-873.

13. Hripcsak G, Knirsch C, Zhou L, Wilcox A, Melton G: Bias associated with mining electronic health records. J Biomed Discov Collab 2011, 6:48-52.

14. Watkins PB, Seligman PJ, Sears JS, Avigan ML, Senior JR: Using controlled clinical trials to learn more about acute drug-induced liver injury. Hepatology 2008, 48(5):1680-1689.

15. Fontana RJ, Watkins PB, Bonkovsky HL, Chalasani N, Davern T, Serrano J, Rochon J: Drug-Induced Liver Injury Network (DILIN) prospective study: rationale, design and conduct. Drug Saf 2009, 32(1):55-68.

16. Bjornsson E: The natural history of drug-induced liver injury. Seminars in liver disease 2009, 29(4):357-363.

17. Danan G, Benichou C: Causality assessment of adverse reactions to drugs—I. A novel method based on the conclusions of international consensus meetings: application to drug-induced liver injuries. J Clin Epidemiol 1993, 46(11):1323-1330.

18. Maria VA, Victorino RM: Diagnostic value of specific T cell reactivity to drugs in 95 cases of drug induced liver injury. Gut 1997, 41(4):534-540.

19. Miljkovic MM, Dobric S, Dragojevic-Simic V: Accuracy and reproducibility of two scales in causality assessment of unexpected hepatotoxicity. Journal of clinical pharmacy and therapeutics 2012, 37(2):196-203.

20. Rockey DC, Seef LB, Rochon J, Preston J, Chalasani N, Bonacini M, Fontana RJ, Hayashi PH: Causality assessment in drug-induced liver injury using a structured expert opinion process: comparison to the Roussel-Uclaf causality assessment method. Hepatology 2010, 51(6):2117-2126.

21. Ryan P, Stang P, Hartzema A, Overhage M, Reich C, Racoosin J: Observational Medical Outcomes Partnership Health Outcomes of Definitions, in Observational Medical Outcomes Partnership. 2009.

22. Cimino JJ, Clayton PD, Hripcsak G, Johnson SB: Knowledge-based approaches to the maintenance of a large controlled medical terminology. J Am Med Inform Assoc 1994, 1(1):35-50.

23. Cimino JJ: From data to knowledge through concept-oriented terminologies: experience with the Medical Entities Dictionary. J Am Med Inform Assoc 2000, 7(3):288-297.

24. Lu CJ, Shulman SW: Rigor and flexibility in computer-based qualitative research: Introducing the Coding Analysis Toolkit. International Journal of Multiple Research Approaches 2008, 2:105-117.

25. Fleiss JL: Measuring Nominal Scale Agreement among Many Raters. Psychol Bull 1971, 76(5):378-382.

26. Hripcsak G, Cimino JJ, Sengupta S: WebCIS: large scale deployment of a Web-based clinical information system. Proc AMIA Symp 1999:804-808.

27. Hripcsak G, Albers DJ: Next-generation phenotyping of electronic health records. J Am Med Inform Assoc 2012.