A collaborative approach to developing an electronic health record phenotyping algorithm for drug-induced liver injury

Casey Lynnette Overby,1,2 Jyotishman Pathak,3 Omri Gottesman,4,5 Krystl Haerian,1 Adler Perotte,1 Sean Murphy,3 Kevin Bruce,3 Stephanie Johnson,6 Jayant Talwalkar,6 Yufeng Shen,1,7 Steve Ellis,5,8 Iftikhar Kullo,6 Christopher Chute,3 Carol Friedman,1 Erwin Bottinger,5,9,10 George Hripcsak,1 Chunhua Weng1

ABSTRACT

Objective To describe a collaborative approach for developing an electronic health record (EHR) phenotyping algorithm for drug-induced liver injury (DILI).

Methods We analyzed types and causes of differences in DILI case definitions provided by two institutions—Columbia University and Mayo Clinic; harmonized two EHR phenotyping algorithms; and assessed the performance, measured by sensitivity, specificity, positive predictive value, and negative predictive value, of the resulting algorithm at three institutions except that sensitivity was measured only at Columbia University.

Results Although these sites had the same case definition, their phenotyping methods differed by selection of liver injury diagnoses, inclusion of drugs cited in DILI cases, laboratory tests assessed, laboratory thresholds for liver injury, exclusion criteria, and approaches to validating phenotypes. We reached consensus on a DILI phenotyping algorithm and implemented it at three institutions. The algorithm was adapted locally to account for differences in populations and data access. Implementations collectively yielded 117 algorithm-selected cases and 23 confirmed true positive cases.

Discussion Phenotyping for rare conditions benefits significantly from pooling data across institutions. Despite the heterogeneity of EHRs and varied algorithm implementations, we demonstrated the portability of this algorithm across three institutions. The performance of this algorithm for identifying DILI was comparable with other computerized approaches to identify adverse drug events.

Conclusions Phenotyping algorithms developed for rare and complex conditions are likely to require adaptive implementation at multiple institutions. Better approaches are also needed to share algorithms. Early agreement on goals, data sources, and validation methods may improve the portability of the algorithms.

INTRODUCTION

Drug-induced liver injury (DILI) is an unexpected adverse hepatic reaction to the pharmacological action of an administered drug. Excluding injury caused by acetaminophen overdose, DILI accounts for up to 15% of liver failure secondary to acute liver injury failure cases,1–4 and is the most frequent cause cited for the withdrawal of approved drugs from the market.5 Previous genome-wide association studies (GWAS) have successfully identified common genetic variants associated with DILI such as an association between rs2395029, a tag single nucleotide polymorphism for HLA-B*5701, and fluvoxacin-induced DILI patients.6 Most variants identified in GWAS analyses to date, however, explain relatively small increases in risk.7 It is difficult to obtain the sample sizes needed to detect these variants of moderate effect size given DILI is a rare condition with an estimated incidence of approximately one case per 10 000 to 100 000.8–11

Consortium efforts such as the International Serious Events Consortium (iSAEC),12 the Drug Induced Liver Injury Network (DILIN)13–14 and the Electronic Medical Records and Genomics (eMERGE) consortium15–16 facilitate pooling data from multiple institutions, and can help produce sample sizes sufficient to identify new susceptibility single-nucleotide polymorphisms for adverse drug events (ADEs) such as DILI. To date, recruitment for iSAEC and DILIN GW AS have relied on prospective identification and recruitment of subjects using a protocol for phenotypic detection.4–7–21 The eMERGE consortium proposes leveraging data in electronic health records (EHRs) linked with DNA biorepositories as an alternative approach to identify subjects for genomics research. This approach has led to the development of several validated algorithms to identify individuals with specific phenotypes (ie, EHR phenotyping algorithms).22–26 Many of these studies have also demonstrated the ability to share EHR phenotyping algorithms among multiple institutions,21,22,26,26a although they usually develop and validate an algorithm at one institution before implementation at other sites. In contrast, in this study, two institutions (Columbia University (CU) and Mayo Clinic (Mayo)) developed DILI EHR phenotyping algorithms separately from one another with project goals and disease case definitions informed by different organizations (eMERGE/iSAEC and DILIN, respectively). The details of Columbia’s algorithm are described elsewhere.27

Given the direct influence of project goals and underlying disease case definitions on EHR phenotyping algorithm design, a major goal for this study was to compare and harmonize two approaches: one informed by eMERGE/iSAEC and the other by DILIN. We also report lessons learned from the subsequent harmonization of the algorithms. We
report the performance of the harmonized algorithm at three institutions and provide an overview of causes of performance differences that may affect its portability. We conclude with a discussion about the complexities in EHR phenotyping for rare conditions such as DILI and recommendations for future work.

METHODS

Distinction between case definition and EHR phenotyping algorithm definition

A case definition describes characteristics that a patient must possess to have a disease from a clinical perspective. An EHR phenotyping algorithm is the translation of the case definition into an executable algorithm that involves querying clinical data elements from the EHR. To illustrate this distinction, a case definition specifying patients with liver injury may translate to a phenotyping algorithm denoting the presence of at least one acute liver injury diagnosis and procedure International Classification of Diseases, revision 9 (ICD-9) code in the EHR. Alternatively, another phenotyping algorithm may define liver injury by a laboratory test that indicates a large decrease in liver function.

Figure 1 defines our five-step methodology for this study. We first characterized differences between DILI case definitions and phenotyping algorithms developed by two institutions (CU and Mayo). Second, we harmonized individual evaluation methods through informal conversations among three authors (CW, JP, and CLO). Our approach to translate the case definition into an operational algorithm is described in the following section (see Harmonizing an operational definition for DILI). Third, the DILI EHR phenotyping algorithm was implemented at three institutions: CU, Mayo, and Mount Sinai School of Medicine (MSSM). Subsequently, we characterized differences in algorithm implementation and evaluation approach. Finally, we used an evaluation framework developed in previous work to summarize the multisite approach (see Table 1, and Evaluating the DILI EHR phenotyping algorithm for further details).

Figure 1  Summary of study methodology. CU, Columbia University; DILI, drug-induced liver injury; Mayo, Mayo Clinic; MSSM, Mount Sinai School of Medicine.
Harmonizing an operational definition for DILI

The CU site defined inclusion and exclusion criteria for case definitions using primarily ICD-9 diagnosis codes assisted with unified medical language system (UMLS) concept codes and/or NewYork Presbyterian medical entities dictionary (MED) codes. The MED contains concepts organized into a semantic network of terms that map to ICD-9 and UMLS codes.\(^{28, 29}\) The general approach taken was therefore first to identify the MED code of a parent concept, then to query the MED hierarchy for all children concepts. We were then able to map the MED codes to ICD-9 and UMLS concept codes. For example, the ‘viral hepatitis’ parent concept has 105 MED children concepts such as 6707 ‘viral hepatitis A with hepatic coma’. Together we mapped these concepts to 23 ICD-9 codes such as 0710.0 ‘viral hepatitis A with hepatic coma’ and seven UMLS codes such as C001959 ‘hepatitis A infection’.

Decisions regarding liver injury inclusion criteria in our algorithm were informed partly by a preliminary assessment of DILI-related diagnoses and acute liver injury diagnoses within the discharge summaries of patients at CU. This assessment was performed using CU’s local natural language processing (NLP) engine, MedLEE,\(^{30}\) to query discharge summary notes for acute liver injury and DILI-related UMLS concepts. Acute liver injury ICD-9 codes were defined according to the observational medical outcomes project,\(^{12}\) then mapped to UMLS codes using the MED. DILI-related UMLS concepts were determined with the use of the medical subject heading (MeSH) browser (http://www.nlm.nih.gov/mesh/). The MeSH heading ‘DILI’ (tree number: C06.552.195) has five entry terms. Three of these terms have corresponding UMLS concept codes: ‘liver injury, drug-induced’ (C086027); ‘toxic liver disease’ (C0348754); and ‘hepatitis, toxic’ (C0019193). We manually reviewed all results. Results confirmed to be a DILI case in a discharge summary note were classified as a true positive (TP), and false positive (FP) otherwise.

We report counts for the NLP-derived acute liver injury, DILI-related diagnoses, and the overlap of both. The way these findings inform phenotyping algorithm specifications is described. TP NLP-derived DILI cases were also used as a gold standard dataset in our algorithm evaluation (see ‘Evaluating the DILI EHR phenotyping algorithm’).

RESULTS

NLP-informed phenotyping algorithm decisions and NLP-selected gold standard

At CU we explored the existence of acute liver injury and DILI-related concepts within the discharge summary notes of patients to inform phenotyping algorithm specifications and to establish a gold standard dataset. A MedLEE query of 265,400 discharge notes from 154,742 unique patients (2004–10) indicated 7,864 adult patients with at least one instance of an acute liver injury concept in their discharge summaries; and 34 TP adult patients with at least one DILI-related concept indicated (out of 55 NLP-selected patients). Twenty-nine TP NLP-selected patients had both acute liver injury and DILI concepts (see figure 2).

Common reasons for FPs in NLP-selected patients were: DILI-related concept(s) in the discussion of patient history (ie, a past medical condition); DILI-related concept(s) in a list of possible diagnoses (ie, an unconfirmed diagnosis); or the note could not be found or was unable to be accessed such as with ‘VIPs’. Overall, given the small number of patients with DILI-related concepts, and to avoid missing undiagnosed DILI cases, we decided to use acute liver injury diagnoses in our harmonized phenotyping algorithm. The 34 NLP-selected cases were used as a gold standard dataset in the sensitivity value estimated by CU (see ‘DILI EHR phenotyping algorithm performance and evaluation’).

Differences in approaches to develop and validate phenotyping algorithms at two institutions

The project goals for developing DILI phenotyping algorithms were initially defined by CU for eMERGE and Mayo for DILIN. The goal for CU was to design an algorithm to identify DILI patients for a broad range of genomic studies of interest to eMERGE institutions. Alternatively, the goal for Mayo was to design an algorithm to identify patients for DILIN-specified cases. As such, the scope of the DILIN-informed algorithm was narrowed, for example, by limiting the number of medications considered. In addition, data access differed initially. CU

Evaluating the DILI EHR phenotyping algorithm

We conducted both measurement and demonstration studies. The three institutions collected qualitative and quantitative data for our measurement study to determine the effectiveness of our evaluation approach. These data included the general evaluation approach and characteristics of reviewers such as expertise and number of reviewers at each institution. For our demonstration study to evaluate our algorithm, all institutions provided estimates for positive predictive value (PPV), negative predictive value (NPV) and specificity. CU also provided estimates for sensitivity. Quantitative data collected from all institutions included: TP and FP counts to estimate PPV; true negative (TN) and false negative (FN) counts to estimate NPV; and FP and TN counts to estimate specificity. All algorithm-selected cases were reviewed to estimate the PPV. To estimate the NPV, each institution reviewed patients from their population who were taking a selected medication suspected to cause DILI in that institution. Mayo reviewed 25 patients; CU and MSSM reviewed 50 patients each. The number of TNs for the specificity calculations were estimated at each institution by subtracting the total of TPs and unknown FNs from the baseline population size, compared to which the number of FP is assumed to be small. The sensitivity was estimated by CU using an NLP-derived ‘gold-standard’ dataset (see ‘DILI case definition and phenotyping algorithm consensus by discussion’). Qualitative data collected for CU demonstration studies included: themes associated with FP results and FN results. On completion of algorithm implementation at each institution, the authors discussed perceptions of site evaluation approaches and the benefit of multisite results.

Table 1 A summary of the multisite evaluation approach

<table>
<thead>
<tr>
<th></th>
<th>Measurement study</th>
<th>Demonstration study</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Quantitative</strong></td>
<td>Number of reviewers</td>
<td>PPV, NPV, Sensitivity, Specificity</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Qualitative</strong></td>
<td>Perceptions of evaluation approach effectiveness:</td>
<td>Perceptions of benefit of results:</td>
</tr>
<tr>
<td></td>
<td>▶ General evaluation approach</td>
<td>▶ Themes in FP</td>
</tr>
<tr>
<td></td>
<td>▶ Reviewer expertise</td>
<td>▶ Themes in FN</td>
</tr>
</tbody>
</table>

Evaluations included measurement studies (to determine the effectiveness of our evaluation approach) and demonstration studies (to demonstrate the value of our algorithm). The results included both quantitative and qualitative data.
primarily utilized structured data of patients in the local clinical data warehouse (CDW). Mayo utilized both structured data and clinical notes of patients recruited to participate in DILIN studies. While DILI case definitions were essentially the same across institutions, the initial phenotyping algorithms developed by both institutions differed significantly. See figure 3A,B for differences in their translation of the DILI case definition into computable form through specifying liver injury diagnoses, medications, laboratory values, and exclusion criteria.

Approaches to validate phenotyping algorithms at the institutions also differed initially. In the absence of a DILI expert, CU developed a protocol to classify TP and FP results. CU randomly selected 100 algorithm-selected cases to evaluate with four reviewers. The details of that evaluation are described elsewhere. A DILI expert assessed the performance of the algorithm on Mayo’s dataset of DILIN-recruited patients.

Reaching consensus on a DILI phenotyping algorithm
A high-level overview of the harmonized DILI phenotyping algorithm is illustrated in figure 3C and described in more detail in supplementary file 1 (available online only). In addition to what is shown in figure 3C, we agreed to utilize both structured and unstructured data to identify patients with diagnoses, medications, and laboratory values of interest. Also the temporal relationship for medication administration, acute liver injury diagnosis, and elevated laboratory values characterizing DILI were specified according to DILI experts interviewed at Mayo. Medications were limited to those of initial interest to DILIN.

Acute liver injury diagnoses were determined primarily by the existence of an ICD-9 diagnosis/procedure code, or by mentioning related concepts within a clinical note. To improve the specificity of our algorithm, we also excluded patients with chronic liver injury diagnosis ICD-9 codes or concepts mentioned in a clinical note, and determined whether laboratory values crossed thresholds for acute liver injury. Acute liver injury-related codes were specified according to the observational medical outcomes project. Chronic liver injury diagnosis ICD-9 and UMLS codes were compiled using the MED.

We examined alkaline phosphatase, alanine aminotransferase, and bilirubin laboratory values to determine whether thresholds for acute liver injury were crossed. Locally defined laboratory codes such as MED codes were used to extract laboratory values within 90 days of acute liver injury diagnosis. We chose to exclude patients who had laboratory values above normal within 30 days before medication administration. Threshold laboratory values for the upper limit of normal (ULN) were defined by either iSAEC expert consensus, or according to the laboratory test manufacturer.

Exposure to a new medication (ie, medication administration) was estimated by a new medication order or mention in a clinical note. Codes were utilized when applicable. Structured data for new medication orders are captured at all institutions, although only in recent years at Mayo. To address this known limitation, Mayo has implemented an NLP-based drug-named entity recognition engine that extracts the drug orders—both outpatient and inpatient—from EHR clinical notes.

For all diagnoses specified as exclusions in either iSAEC or DILIN recruitment protocols, we specified ICD-9 codes and UMLS codes. We then agreed on a subset of high-priority exclusions for this project. These included sclerosing cholangitis, organ transplantation or liver operation, alcohol abuse/liver damage/toxic effects, and viral hepatitis. Given results from preliminary analyses, we also excluded patients with death or overdose ICD-9 codes or concepts mentioned in their clinical notes.

Characteristics of DILI EHR phenotyping algorithm implementation
Experiences implementing the harmonized DILI EHR phenotyping algorithm indicated a need to adapt algorithm specifications for each institution. Table 2 illustrates adaptations of the harmonized algorithm across three institutions. Figure 4 illustrates institution-specific implementation differences. Institutional
Variation was primarily due to baseline population size, availability of structured and unstructured data, and multiple interpretations of the same algorithm.

Variations due to baseline population size
Two institutions queried their biobank-linked EHR (MSSM and Mayo). In the absence of a biobank, CU queried their CDW. As expected, the baseline population for CU’s CDW was much larger than those of the biobank populations. Agreed-upon specifications for medications, exclusion diagnoses, and laboratory temporal relationship were therefore relaxed to optimize algorithm yield for the smaller baseline populations. See table 3 for population sizes at each institution.


Variations due to data access characteristics
Data access characteristics also influenced algorithm implementation (see table 3). For example, at Mayo and MSSM we decided to consider that any medication might be implicated in a DILI case rather than restricting medications to the select few. Given that MSSM was able to leverage medication order data, considering any medication was straightforward. At Mayo, they used common medications that are associated with DILI as a surrogate for any medication. These were medications compiled by DILIN (see LiverTox.nih.gov). Diagnoses that were excluded and the laboratory value temporal relationship specified by institutions also differed. Given the smaller baseline population, Mayo and MSSM included fewer exclusion diagnoses than CU.
and simplified the check for laboratory value temporality. Another difference in institution implementation was the process for identifying relevant laboratory values. ISAEC-specified thresholds for laboratory value ULN were implemented at CU. Manufacturer-specified ULN thresholds were specified for Mayo and MSSM.

Variations due to multiple interpretations of the EHR phenotyping algorithm Interpretation of the DILI EHR phenotyping algorithm led to two main approaches to execute queries. Figure 4 illustrates a side-by-side comparison of the main differences between these approaches. Query steps differ by anchor dates and by time frames within which laboratory values are assessed.

### DISCUSSION
Findings from this work illustrate that EHR phenotyping algorithms may help reduce the burden of traditional approaches to screening for rare conditions. We also highlight several challenges facing EHR phenotyping for complex and rare conditions such as DILI. Moreover, this work suggests that better approaches to share EHR phenotyping algorithms are needed, and early agreement on phenotyping goals, data sources, and validation methods may facilitate the downstream portability of algorithms across institutions.

#### Computerized approaches are an improvement over manual approaches to identify rare adverse drug events
The estimated PPV for this DILI phenotyping algorithm was reasonable compared with other computerized approaches to identify ADEs. One study, for example, reports a PPV of 7.5% (CI 6.5% to 8.5%) for their computer search method to identify ADEs. Another reports their highest PPV to be 23% for an ADE monitoring system. While some level of manual review is still required for rare conditions, computerized screening approaches such as ours may be an improvement over traditional manual processes. This notion is supported by previous work investigating electronic screening (e-screening) methods for clinical trial recruitment. Authors in one study compared an e-screening strategy with investigator review for recruitment for clinical trial recruitment. The estimated PPV for this DILI phenotyping algorithm was reasonable compared with other computerized approaches to identify ADEs. One study, for example, reports a PPV of 7.5% (CI 6.5% to 8.5%) for their computer search method to identify ADEs. Another reports their highest PPV to be 23% for an ADE monitoring system. While some level of manual review is still required for rare conditions, computerized screening approaches such as ours may be an improvement over traditional manual processes. This notion is supported by previous work investigating electronic screening (e-screening) methods for clinical trial recruitment. Authors in one study compared an e-screening strategy with investigator review for recruitment for clinical trial recruitment. The estimated PPV for this DILI phenotyping algorithm was reasonable compared with other computerized approaches to identify ADEs. One study, for example, reports a PPV of 7.5% (CI 6.5% to 8.5%) for their computer search method to identify ADEs. Another reports their highest PPV to be 23% for an ADE monitoring system. While some level of manual review is still required for rare conditions, computerized screening approaches such as ours may be an improvement over traditional manual processes. This notion is supported by previous work investigating electronic screening (e-screening) methods for clinical trial recruitment. Authors in one study compared an e-screening strategy with investigator review for recruitment for clinical trial recruitment.

**Table 2** DILI EHR phenotyping algorithm adaptations

<table>
<thead>
<tr>
<th>Specification</th>
<th>CU</th>
<th>Mayo</th>
<th>MSSM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medication order temporal relationship</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>▶️ Medication order within 90 days prior to acute liver injury diagnosis</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Laboratory value temporal relationship</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>▶️ Laboratory values crossing threshold for DILI within 90 days of acute liver injury diagnosis</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>▶️ All laboratory values below ULN within 30 days before medication administration</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Laboratory value thresholds</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>▶️ Chronic liver injury, sclerosing cholangitis, organ transplantation or liver operation, alcohol abuse/liver damage/toxic effects, viral hepatitis, death, overdose</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Excluded diagnoses</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>▶️ Chronic liver injury cases</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Agreed-upon specifications</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mayo and 24 at MSSM</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Agreed-upon specs with common medications implicated in DILI cases</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Agreed-upon specs</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Manufacturer specified</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Only excluded chronic liver injury cases</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Manufacturer specified</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Only excluded chronic liver injury cases</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

CU, Columbia University; DILI, drug-induced liver injury; EHR, electronic health record; ISAEC, International Serious Events Consortium; Mayo, Mayo Clinic; MSSM, Mount Sinai School of Medicine; ULN, upper limit of normal.
Challenges of identifying rare and complex conditions lead to low yield

Another inevitable challenge for identifying subjects with rare conditions is that the number of algorithm-selected cases, and subsequently confirmed TP cases, will be low. Our algorithm implementation at three institutions collectively yielded 23 DILI cases. While having a higher PPV compared to other computational approaches to identify ADEs may reduce the burden of manual review, the trade-off may be a larger number of missed cases. For example, although we have a higher PPV in comparison to one study’s result, they have produced a higher sensitivity (58; CI 18 to 98) compared to ours (5.5; CI 0.2 to 29). Our low number of DILI cases also illustrates challenges to achieving sample sizes needed to conduct GWAS. To complicate things further, our study considered a range of medications that may be implicated in DILI patients. Previous work illustrates that the mechanism for DILI may depend on the drug implicated, demonstrating a need to stratify DILI patients by medication and leading to further decreases in sample size.

In addition to DILI being a rare condition, it is also complex, as illustrated in our current lack of understanding of the underlying mechanisms of idiosyncratic DILI. The incomplete knowledge of DILI and the heterogeneity in the etiologies of DILI translate into complex or vague expressions for inclusion and exclusion criteria, introducing intricacies that influence algorithm performance. For example, further investigation of algorithm-selected DILI cases at CU indicated that 32% of FP were due to missed exclusion codes. As another example, the majority of FN (55%) at CU were DILI cases with a drug suggested to be causal that was not administered within 90 days before liver injury diagnosis. Given that the CU gold standard dataset was NLP derived, all gold standard patients had mentions of medications in their discharge summary notes. As such, the date of medication administration was approximated by the date mentioned in the discharge summary note. It is possible, however, that this approximation is not accurate and leads to inappropriate exclusion by our phenotyping algorithm. These two examples highlight challenges in defining an algorithm for a

![Figure 4](https://academic.oup.com/jamia/article-abstract/20/e2/e243/710321/1021)
complex condition and limitations to translating complexities into algorithm specifications. In the first example, missed diagnoses are an unavoidable consequence when many excluded diagnoses must be defined—as is common for complex conditions. The second example highlights limitations to defining specific disease manifestations into our phenotyping algorithms (ie, temporality of medication administration in DILI cases). Our findings also illustrate that the complexity of DILI makes the review of algorithm-selected cases challenging. Even so, the complexity of a condition is also what makes it interesting to study.

Adaptation and better approaches to share EHR phenotyping algorithms are needed

We found implementation across institutions requires flexible adaptation of the algorithm due to the differences in baseline population size, availability of structured and unstructured data, and interpretation of the phenotyping algorithm. Our findings illustrate the need for better approaches to share phenotyping algorithms given that our documented DILI EHR phenotyping algorithm was interpreted in different ways. eMERGE institutions are in the process of investigating new approaches for sharing phenotyping algorithms that make use of the quality data model for formal representation and the KNIME (Konstanz information miner) data analytics platform to improve the portability of algorithms.

Despite implementation and interpretation differences, the results are still comparable. On one hand, such implementation differences are inevitable and are necessary algorithm adaptations to accommodate discrepant EHR systems in different institutions. On the other hand, the comparable results can be greatly attributed to the collaborative approach so that the differences in implementations do not degrade the algorithm performance at each site. Our findings suggest that agreement on goals, data sources and validation methods early on is important for sharing those approaches down the road. The similarity in calculated PPV at institutions suggests that early agreement may have improved the portability of our algorithm.

Study limitations and future directions

The small number of institutions represented in this study is a limitation. Two of nine eMERGE institutions completed the manual review of patient records for calculating performance measures reported in this paper. While eMERGE institutions have ample resources dedicated to implement EHR phenotyping algorithms, previous eMERGE-developed algorithms have not gone through the thorough manual review that was performed for this algorithm. The time constraints of individuals with the clinical expertise to review algorithm-selected DILI cases were therefore the primary factor limiting eMERGE institution involvement. After two institutions validate the algorithm, the current practice is for other institutions to complete ‘algorithm verification’ involving manual review of a small number of algorithm-selected cases. To improve on our ability to achieve meaningful sample sizes, we are implementing this algorithm across other eMERGE institutions that are performing such algorithm verification. For the dataset to

| Table 3 DILI EHR phenotyping algorithm implementation: data access in use |
|---------------------------------|-----------------|-----------------|
| Diagnoses | ICD-9 codes and UMLS codes for NLP of discharge summary notes only | ICD-9 codes and NLP of clinical notes | ICD-9 codes |
| Medications | MED codes and UMLS codes for NLP of discharge summary notes only | RxNorm codes (data derived from a structured registry and NLP of drug orders) | Medication codes and text terms of medication fields |
| Laboratory values | MED codes | Laboratory information system | Test codes and text terms of procedure fields |

CDW, clinical data warehouse; CU, Columbia University; DILI, drug-induced liver injury; EHR, electronic health record; eMERGE, Electronic Medical Records and Genomics; ICD-9, International Classification of Diseases, revision 9; Mayo, Mayo Clinic; MED, medical entities dictionary; MSSM, Mount Sinai School of Medicine; NLP, natural language processing; UMLS, unified medical language system.

| Table 4 Evaluation approach and results |
|---------------------------------|-----------------|-----------------|
| General approach to assign TP and FP status | A. First-pass visualization of temporal data to assign preliminary TP, FP, and unknown status. | Manual chart review of all algorithm-selected DILI cases | Manual chart review of all algorithm-selected DILI cases |
| | B. Manual chart review to confirm suspected TP and assign TP or FP for unknown status from previous step | | |
| Number and type of reviewers | 4 reviewers (2 clinical and 2 non-clinical) | 1 reviewer (1 DILI expert) | 1 reviewer (1 clinical) |
| Final decisions | Consensus, determination by 5th clinical reviewer for questionable results | Expert decision | Expert decision, determination by DILI expert for questionable results |
| PPV | 32% (18% to 50%) | 16% (5% to 37%) | 29% (13% to 51%) |
| NPV | 100% (91% to 100%) | 100% (83% to 100%) | 100% (91% to 100%) |
| Sensitivity | 5.5% (0.2% to 29%) | – | – |
| Specificity | ~100% | ~100% | ~100% |

95% CI presented as percentages.
CU, Columbia University; DILI, drug-induced liver injury; FP, false positive; Mayo, Mayo Clinic; MSSM, Mount Sinai School of Medicine; NPV, negative predictive value; PPV, positive predictive value; TP, true positive.

be of most use in future genetic studies, however, all algorithm-selected cases should be reviewed.

Computerized methods to characterize DILI cases by pattern of liver injury and drug causality may be of interest in future EHR phenotyping studies. Furthermore, as a next phase in algorithm development, it is worth exploring data-driven approaches to case identification given the challenges highlighted in this work. However, in order to facilitate scientific communication of phenotyping algorithms, we may not be able to avoid completely textual descriptions of algorithms or algorithm components, which often have inherent ambiguities or imperfect concept granularities that can lead to interpretation and implementation variations. This is the well-known semantic gap challenge.39

CONCLUSION

Phenotyping algorithms developed for rare and complex conditions such as DILI are likely to require a systematic process for local adaptation to maximize consistency when implementing them at multiple institutions. There is also a need for better approaches to share EHR phenotyping algorithms given that our documented algorithm was interpreted in different ways. It appears, however, that our early agreement on goals, data sources, and validation methods may have improved the portability of our algorithm and thus can serve as the best practice for sharing phenotyping algorithms. Despite adaptations, algorithm performance was comparable with other computerized approaches to identify ADEs, and we were able to demonstrate its portability across three institutions.

Author affiliations

1Department of Biomedical Informatics, Columbia University, New York, New York, USA
2Program in Personalized & Genomic Medicine and Department of Medicine, University of Maryland, Baltimore, MD, USA
3Department of Health Sciences Research, Mayo Clinic, Rochester, Minnesota, USA
4Division of General Internal Medicine, Mount Sinai School of Medicine, New York, New York, USA
5The Charles Bronfman Institute for Personalized Medicine, Mount Sinai School of Medicine, New York, New York, USA
6Department of Medicine, Mayo Clinic, Rochester, Minnesota, USA
7Department of Systems Biology, Columbia University, New York, NY
8Information Technology, Mount Sinai School of Medicine, New York, New York, USA
9Division of Nephrology, Mount Sinai School of Medicine, New York, New York, USA
10Department of Pharmacology and Systems Therapeutics, Mount Sinai School of Medicine, New York, New York, USA

Acknowledgements

The authors would like to thank the following institutions and individuals for their feedback on algorithm documentation and implementation at their institutions: Northwestern University (Anna Roberts, Jen Allen Pacheco, and Abeel Kho); Group Health Cooperative (Jane Grafton and David Carrell); Marsh at their institutions: Northwestern University (Anna Roberts, Jen Allen Pacheco, and Abeel Kho); Group Health Cooperative (Jane Grafton and David Carrell); Marsh.

REFERENCES

18 Lucena MI, Molokhia M, Shen Y, et al. Susceptibility to amoxicillin–clavulanate-induced liver injury is influenced by multiple HLA class I and II alleles. Gastroenterology 2011;141:338–47.


