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 Weng[1]

▸ Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/amiajnl-2013-001930).

For numbered affiliations see end of article

Correspondence to
Dr Chunhua Weng, Department of Biomedical Informatics, Columbia University, 622 West 168th St. VC5, New York, NY 10032, USA; cw2384@columbia.edu

Received 16 April 2013
Revised 17 May 2013
Accepted 10 June 2013
Published Online First
9 July 2013

To cite: Overby CL, Pathak J, Gottesman O, *et al. J Am Med Inform Assoc* 2013;**20**:e243–e252.

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 flucloxacilin-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 to100 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) consortium[15 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 GWAS have relied on prospective identification and recruitment of subjects using a protocol for phenotypic detection.[6 17–21] 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 definition 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[27] to summarize the multisite approach (see table 1, and Evaluating the DILI EHR phenotyping algorithm for further details).

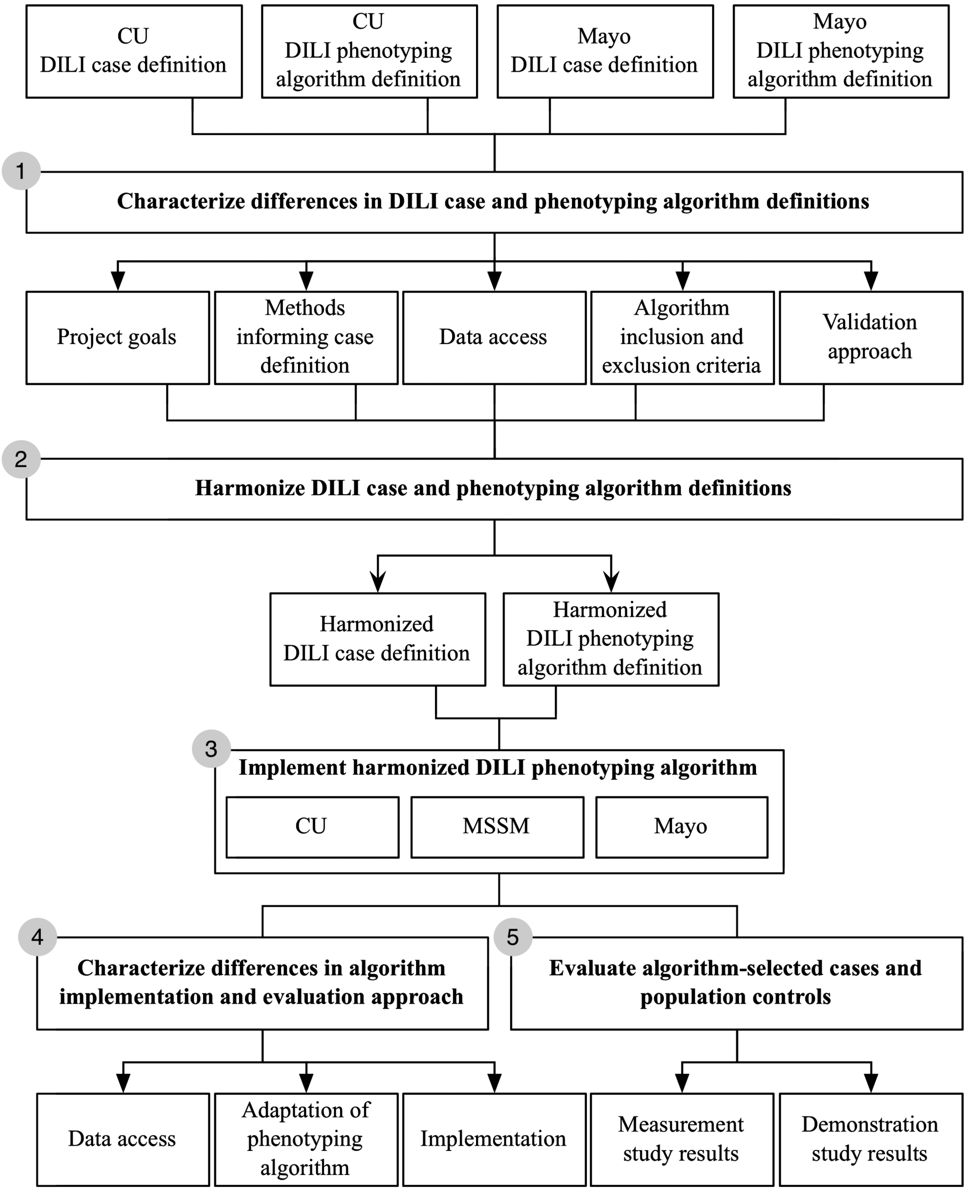
CU
DILI case definition
CU
DILI phenotyping algorithm definition
Mayo
DILI case definition
Mayo
DILI phenotyping algorithm definition
1
Characterize differences in DILI case and phenotyping algorithm definitions
Project goals
Methods informing case definition
Data access
Algorithm inclusion and exclusion criteria
Validation approach
2
Harmonize DILI case and phenotyping algorithm definitions
Harmonized DILI case definition
Harmonized DILI phenotyping algorithm definition
3
Implement harmonized DILI phenotyping algorithm
CU
MSSM
Mayo
4
Characterize differences in algorithm implementation and evaluation approach
5
Evaluate algorithm-selected cases and population controls
Data access
Adaptation of phenotyping algorithm
Implementation
Measurement study results
Demonstration study results

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

Table 1 A summary of the multisite evaluation approach

	Measurement study	Demonstration study
Quantitative results	Number of reviewers	PPV NPV Sensitivity Specificity
Qualitative results	Perceptions of evaluation approach effectiveness: ▸ General evaluation approach ▸ Reviewer expertise	Perceptions of benefit of results: ▸ Themes in FP ▸ Themes in FN

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.
FN, false negative; FP, false positive; NPV, negative predictive value; PPV, positive predictive value.

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 070.0 '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').

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 FPs 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.

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 7864 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

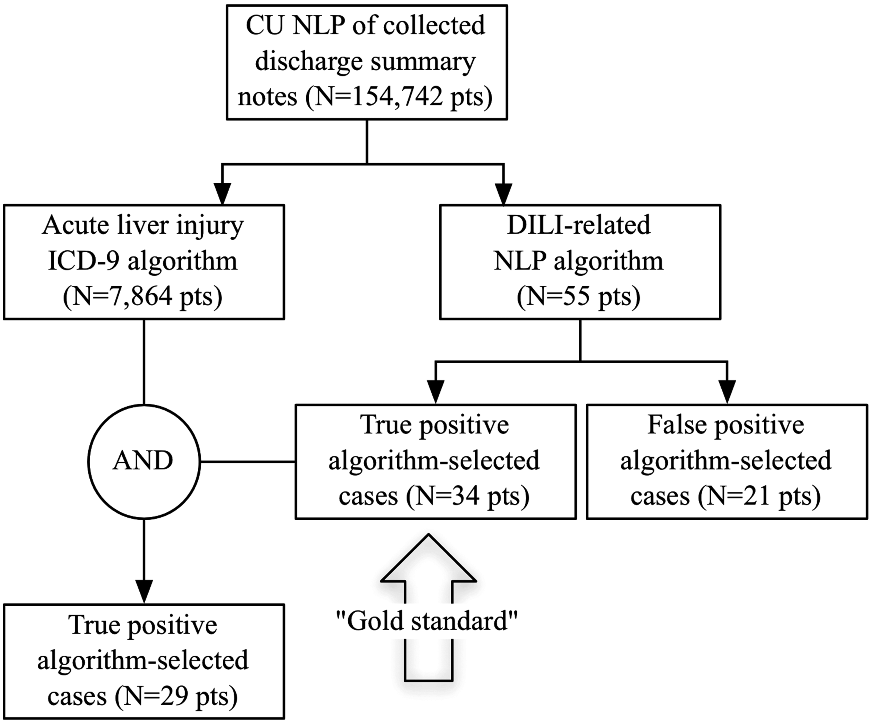
CU NLP of collected discharge summary notes (N=154,742 pts)
Acute liver injury ICD-9 algorithm (N=7,864 pts)
DILI-related NLP algorithm (N=55 pts)
AND
True positive algorithm-selected cases (N=34 pts)
False positive algorithm-selected cases (N=21 pts)
"Gold standard"
True positive algorithm-selected cases (N=29 pts)

Figure 2 Summary of Columbia University (CU) acute liver injury and drug-induced liver injury (DILI)-related natural language processing (NLP) algorithm results. ICD-9, International Classification of Diseases, revision 9.

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.[27] 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.[31 32]

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,[27] 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

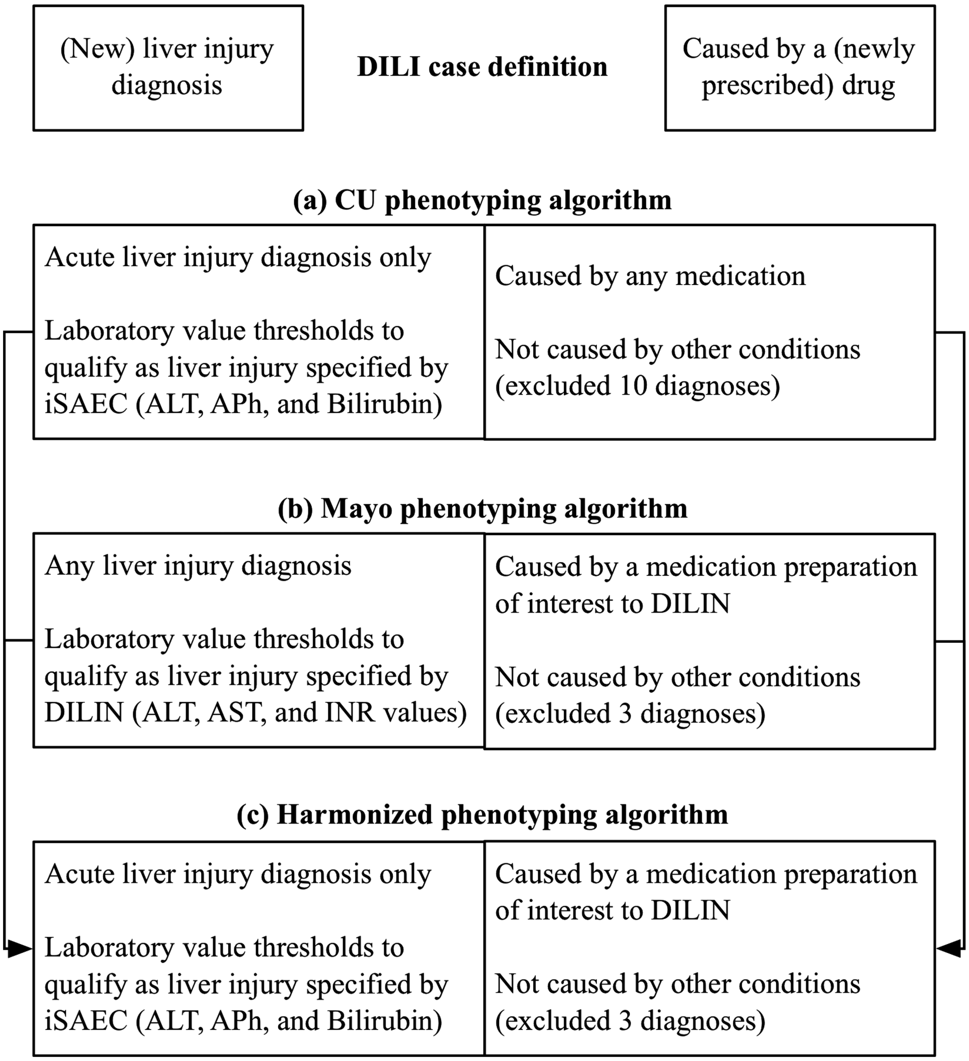
(New) liver injury diagnosis
DILI case definition
Caused by a (newly prescribed) drug
(a) CU phenotyping algorithm
Acute liver injury diagnosis only
Laboratory value thresholds to qualify as liver injury specified by iSAEC (ALT, APh, and Bilirubin)
Caused by any medication
Not caused by other conditions (excluded 10 diagnoses)
(b) Mayo phenotyping algorithm
Any liver injury diagnosis
Laboratory value thresholds to qualify as liver injury specified by DILIN (ALT, AST, and INR values)
Caused by a medication preparation of interest to DILIN
Not caused by other conditions (excluded 3 diagnoses)
(c) Harmonized phenotyping algorithm
Acute liver injury diagnosis only
Laboratory value thresholds to qualify as liver injury specified by iSAEC (ALT, APh, and Bilirubin)
Caused by a medication preparation of interest to DILIN
Not caused by other conditions (excluded 3 diagnoses)

Figure 3 Drug-induced liver injury (DILI) case definition and phenotyping algorithm harmonization. (A) Columbia University (CU) phenotyping algorithm: CU's International Serious Events Consortium (iSAEC)-informed algorithm makes a distinction between acute liver injury and chronic liver injury. CU chose to focus on acute liver injury. With respect to drug exposure, CU considered patients with any drug prescribed within 90 days of an acute liver injury diagnosis. Given iSAEC protocol specifications and access to structured data, CU considered iSAEC-specified threshold values for alanine aminotransferase (ALT), intestinal alkaline phosphatase (APh) and intervascular bilirubin. CU excluded 10 diagnoses initially. (B) Mayo Clinic (Mayo) phenotyping algorithm: Mayo's Drug Induced Liver Injury Network (DILIN)-informed algorithm considered any liver injury-related diagnoses. Mayo also considered a subset of drugs of interest to DILIN; and specified the temporal relationship between drug administration, DILI diagnosis and laboratory measures. Given DILIN protocol specifications and access to clinical notes of recruited patients, Mayo used DILIN-specified thresholds and text terms for ALT, aspartate aminotransferase (AST), and international normalized ratio (INR) for use by the cTAKES natural language processing engine. Specific emphasis was laid on investigating the DILI-related complications and medications in various sections of the clinical notes (including chief complaints, impression report plans). Mayo excluded three diagnoses initially. Given Mayo's focus on a smaller number of medications, exclusion criteria were less stringent than CU in order to optimize recall. (C) Harmonized phenotyping algorithm. See supplementary file 1 (available online only) for more detail.

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

Table 2 DILI EHR phenotyping algorithm adaptations

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

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.

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.

DILI EHR phenotyping algorithm performance and evaluation

The agreed-upon algorithm selected 37 DILI cases at CU, 56 at Mayo and 24 at MSSM. Our measurement study indicated that the methods for reviewing algorithm-selected DILI cases differed by general approach, reviewer expertise, and the number of reviewers (see table 4). Reviewers agreed during post-evaluation discussions that algorithm-selected cases were challenging to evaluate, particularly for the non-clinical reviewers. A contributing factor for all reviewers was the difficulty confirming that a drug was the causal agent for liver injury in patients who have multiple, potentially interacting, conditions and drugs. This is a typical problem in pharmacovigilance, in which all the cases associated with known non-drug causes of DILI are first eliminated and the ones remaining are the possible causes but further confirmation is needed. Furthermore, as Mayo is a large academic referral center, the laboratory records for many patients were only available as scanned documents, either in PDF form or images, and were therefore not readily queryable. For the two non-clinical CU reviewers, a list of terms and phrases used for DILI case spotting did not appear to be sufficient. The main difference between these clinical and non-clinical reviewers appeared to be their ability to exclude non-DILI patients. Possible advantageous factors for clinical reviewers were their familiarity with language used in clinical notes and with clinical cases that lend themselves to quick disqualification (eg, shock liver).

Quantitative results from our demonstration studies at three institutions are summarized in table 4. CU qualitative results indicated themes in FP including missed exclusion diagnoses (32% of FP), unstable laboratory values such as were often seen in cancer patients, and laboratory values elevated before prescribing a medication. Assuming the number of FN was small compared to baseline population sizes, specificity estimates were near 100% for all institutions.

Given that the CU algorithm was implemented with select medications (see 'Reaching consensus on a DILI phenotyping algorithm'), the NLP-derived gold standard dataset was filtered to include only patients who were prescribed one of these medications (18 out of 37 DILI cases). There was one algorithm-selected DILI case. We therefore estimated the sensitivity of the CU algorithm to be 5.5%. The primary theme in CU FN was cases with an indicated drug that was not administered within 90 days before liver injury diagnosis (55% of FN).

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.[33] One study, for example, reports a PPV of 7.5% (CI 6.5% to 8.5%) for their computer search method to identify ADEs.[34] Another reports their highest PPV to be 23% for an ADE monitoring system.[35] 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 into a clinical trial with a low prevalence of eligible patients.[36] The e-screening approach showed an over fivefold improvement on the investigator review approach (13% compared to 2.4% PPV).

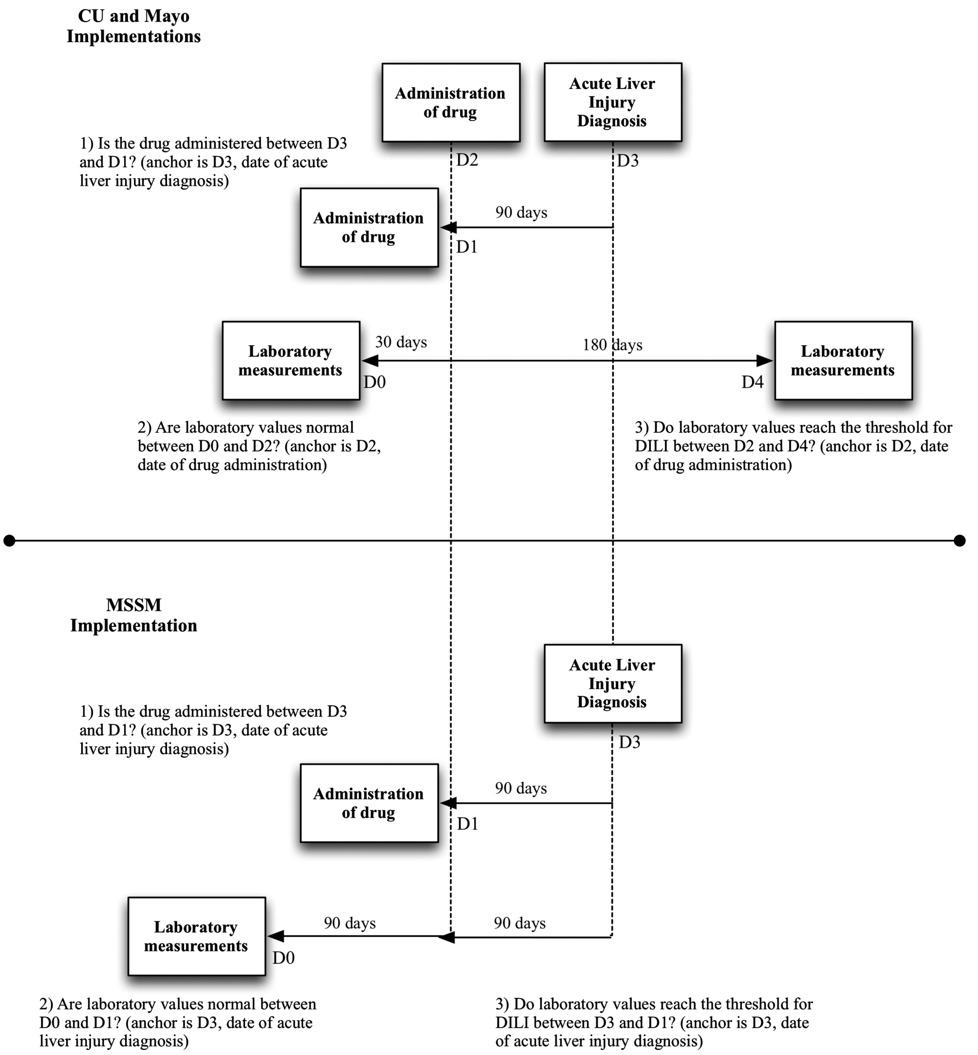
CU and Mayo Implementations
Administration of drug
Acute Liver Injury Diagnosis
D2
D3
1) Is the drug administered between D3 and D1? (anchor is D3, date of acute liver injury diagnosis)
Administration of drug
90 days
D1
Laboratory measurements
30 days
180 days
Laboratory measurements
D0
D4
2) Are laboratory values normal between D0 and D2? (anchor is D2, date of drug administration)
3) Do laboratory values reach the threshold for DILI between D2 and D4? (anchor is D2, date of drug administration)
MSSM Implementation
Acute Liver Injury Diagnosis
D3
1) Is the drug administered between D3 and D1? (anchor is D3, date of acute liver injury diagnosis)
Administration of drug
90 days
D1
Laboratory measurements
90 days
90 days
D0
2) Are laboratory values normal between D0 and D1? (anchor is D3, date of acute liver injury diagnosis)
3) Do laboratory values reach the threshold for DILI between D3 and D1? (anchor is D3, date of acute liver injury diagnosis)

Figure 4 Two implementations of the same electronic health record phenotyping algorithm. Step 1 is implemented the same for both approaches. Step 2 differs by the query anchor and the time frame for which laboratory values are checked to be within normal ranges. The top implementation is anchored on the date of medication administration with laboratory values checked within 30 days before. The bottom implementation is anchored on the date of acute liver injury diagnosis with laboratory values checked between 180 and 90 days before. Step 3 differs by the query anchor and the time frame for which laboratory values are checked for thresholds to qualify as drug-induced liver injury (DILI). The top implementation is anchored on the date of medication administration with laboratory values checked within 180 days following drug administration. The bottom implementation is anchored on the date of acute liver injury diagnosis with laboratory values checked within 90 days before. CU, Columbia University; Mayo, Mayo Clinic; MSSM, Mount Sinai School of Medicine.

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).[34] 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

Table 3 DILI EHR phenotyping algorithm implementation: data access in use

	CU	Mayo	MSSM
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
Data access	CDW. 1 045 125 patients. 2004–10. Limited to inpatients, >1 year old	Mayo Clinic eMERGE cohort. 6916 patients. 2007–13	Biobank clinical datamart. 23 200 patients. 2006–13

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.

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[37] and the KNIME (Konstanz information miner) data analytics platform[38] 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 4 Evaluation approach and results

	CU	Mayo	MSSM
General approach to assign TP and FP status	A. First-pass visualization of temporal data to assign preliminary TP, FP, and unknown status. B. Manual chart review to confirm suspected TP and assign TP or FP for unknown status from previous step	Manual chart review of all algorithm-selected DILI cases	Manual chart review of all algorithm-selected DILI cases
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
[1]Department of Biomedical Informatics, Columbia University, New York, New York, USA
[2]Program in Personalized & Genomic Medicine and Department of Medicine, University of Maryland, Baltimore, MD, USA
[3]Department of Health Sciences Research, Mayo Clinic, Rochester, Minnesota, USA
[4]Division of General Internal Medicine, Mount Sinai School of Medicine, New York, New York, USA
[5]The Charles Bronfman Institute for Personalized Medicine, Mount Sinai School of Medicine, New York, New York, USA
[6]Department of Medicine, Mayo Clinic, Rochester, Minnesota, USA
[7]Department of Systems Biology, Columbia University, New York, NY
[8]Information Technology, Mount Sinai School of Medicine, New York, New York, USA
[9]Division of Nephrology, Mount Sinai School of Medicine, New York, New York, USA
[10]Department 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 Abel Kho); Group Health Cooperative (Jane Grafton and David Carrell); Marshfield Clinic (Peggie Peissig and Jim Linneman); Geisinger Health Center (Jonathan Bock); The Children's Hospital of Pennsylvania (Lyam Vazquez and Frank Mentch); and Vanderbilt University (Josh Denny, Sarah Stallings, and Jacqueline Kirby). The authors would also like to thank Alla Babina, Lyudmila Ena, and Sajed Naseem for Columbia University technical support.

Contributors CLO, JP, CW and GH jointly conceived the study. CLO, JP and CW harmonized two drug-induced liver injury EHR phenotyping algorithms. SM and KB implemented the algorithm at Mayo. OG implemented and assessed the performance of the algorithm at Mount Sinai. CLO, KH, and CF implemented the algorithm at Columbia. KH, AP, CLO, CW and GH assessed the performance of the algorithm at Columbia. JT and SJ assessed the performance of the algorithm at Mayo. All authors reviewed and edited the manuscript.

Funding This work was supported by the NIH National Library of Medicine grant numbers R01 LM006910, R01 LM010016, R01 LM009886, T15 LM007079; by the NIH National Human Genome Research Institute grant numbers U01 HG006380, U01 HG006379; and by the NIH National Institute of Diabetes and Digestive and Kidney Diseases grant number U01 DK082992.

Competing interests None.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement Local EHR phenotyping algorithm implementations discussed in this manuscript are documented and available online at PheKB.org.

REFERENCES

1 O'Grady JG. Acute liver failure. *Postgrad Med J* 2005;81:148–54.
2 Bjornsson E, Jerlstad P, Bergqvist A, *et al*. Fulminant drug-induced hepatic failure leading to death or liver transplantation in Sweden. *Scand J Gastroenterol* 2005;40:1095–101.
3 Russo MW, Galanko JA, Shrestha R, *et al*. Liver transplantation for acute liver failure from drug induced liver injury in the United States. *Liver Transpl* 2004;10:1018–23.
4 Ostapowicz G, Fontana RJ, Schiodt FV, *et al*. Results of a prospective study of acute liver failure at 17 tertiary care centers in the United States. *Ann Intern Med* 2002;137:947–54.
5 Navarro VJ, Senior JR. Drug-related hepatotoxicity. *N Engl J Med* 2006;354:731–9.
6 Daly AK, Donaldson PT, Bhatnagar P, *et al*. HLA-B*5701 genotype is a major determinant of drug-induced liver injury due to flucloxacillin. *Nature Genet* 2009;41:816–19.
7 Urban TJ, Shen Y, Stolz A, *et al*. Limited contribution of common genetic variants to risk for liver injury due to a variety of drugs. *Pharmacogenet Genom* 2012;22:784–95.
8 Aithal GP, Rawlins MD, Day CP. Accuracy of hepatic adverse drug reaction reporting in one English health region. *BMJ* 1999;319:1541.
9 Sgro C, Clinard F, Ouazir K, *et al*. Incidence of drug-induced hepatic injuries: a French population-based study. *Hepatology* 2002;36:451–5.
10 Bell LN, Chalasani N. Epidemiology of idiosyncratic drug-induced liver injury. *Semin Liver Dis* 2009;29:337–47.
11 Chalasani N, Fontana RJ, Bonkovsky HL, *et al*. Causes, clinical features, and outcomes from a prospective study of drug-induced liver injury in the United States. *Gastroenterology* 2008;135:1924–34.
12 Ryan P, Stang P, Hartzema A, *et al*. Observational Medical Outcomes Partnership Health Outcomes of Definitions. Last revised December 2009. http://75.101.131.161/download/loadfile.php?docname=HOI%20Definitions (accessed 16 Apr 2013).
13 Watkins PB. Drug-induced liver injury network. *Am J Gastroenterol* 2008;103:1574–5.
14 Hoofnagle JH. Drug-induced liver injury network (DILIN). *Hepatology* 2004;40:773.
15 Kho AN, Pacheco JA, Peissig PL, *et al*. Electronic medical records for genetic research: results of the eMERGE consortium. *Sci Transl Med* 2011;3:79re1.
16 McCarty CA, Chisholm RL, Chute CG, *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.
17 Fontana RJ, Watkins PB, Bonkovsky HL, *et al*. Drug-Induced Liver Injury Network (DILIN) prospective study: rationale, design and conduct. *Drug Saf* 2009;32:55–68.
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.
19 Shen Y, Nicoletti P, Floratos A, *et al*. Genome-wide association study of serious blistering skin rash caused by drugs. *Pharmacogenomics J* 2012;12:96–104.
20 Molleston JP, Fontana RJ, Lopez MJ, *et al*. Characteristics of idiosyncratic drug-induced liver injury in children: results from the DILIN prospective study. *J Pediatr Gastroenterol Nutr* 2011;53:182–9.
21 Carroll RJ, Thompson WK, Eyler AE, *et al*. Portability of an algorithm to identify rheumatoid arthritis in electronic health records. *J Am Med Inform Assoc* 2012;19:e162–9.
22 Kho AN, Hayes MG, Rasmussen-Torvik L, *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;19:212–18.
23 Kullo IJ, Fan J, Pathak J, *et al*. Leveraging informatics for genetic studies: use of the electronic medical record to enable a genome-wide association study of peripheral arterial disease. *J Am Med Inform Assoc* 2010;17:568–74.
24 Ritchie MD, Denny JC, Crawford DC, *et al*. Robust replication of genotype–phenotype associations across multiple diseases in an electronic medical record. *Am J Hum Genet* 2010;86:560–72.
25 Denny JC, Ritchie MD, Crawford DC, *et al*. Identification of genomic predictors of atrioventricular conduction: using electronic medical records as a tool for genome science. *Circulation* 2010;122:2016–21.
26 Peissig PL, Rasmussen LV, Berg RL, *et al*. Importance of multi-modal approaches to effectively identify cataract cases from electronic health records. *J Am Med Inform Assoc* 2012;19:225–34.
26a Newton KM, Peissig PL, Kho AN, *et al*. Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network. *J Am Med Inform Assoc* 2013;20:e147–54.
27 Overby CL, Weng C, Haerian K, *et al*. Evaluation considerations for EHR-based phenotyping algorithms: a case study for drug-induced liver injury. *AMIA Summit on Translational Bioinformatics*, San Francisco, CA, 18-20 March 2013;2013:130–4.

28 Cimino JJ. From data to knowledge through concept-oriented terminologies: experience with the Medical Entities Dictionary. *J Am Med Inform Assoc* 2000;7:288–97.

29 Cimino JJ, Clayton PD, Hripcsak G, *et al*. Knowledge-based approaches to the maintenance of a large controlled medical terminology. *J Am Med Inform Assoc* 1994;1:35–50.

30 Friedman C, Johnson SB, Forman B, *et al*. Architectural requirements for a multipurpose natural language processor in the clinical environment. *The Annual Symposium on Computer Applications in Medical Care*, New Orleans, LA, 28 Oct-1 Nov, 1995;1995:347–51.

31 Sohn S, Murphy SP, Masanz JJ, *et al*. Classification of medication status change in clinical narratives. *AMIA Annu Symp Proc* 2010;2010:762–6.

32 Pathak J, Murphy SP, Willaert BN, *et al*. Using RxNorm and NDF-RT to classify medication data extracted from electronic health records: experiences from the Rochester Epidemiology Project. *AMIA Annu Symp Proc* 2011; 2011:1089–98.

33 Bates DW, Evans RS, Murff H, *et al*. Detecting adverse events using information technology. *J Am Med Inform Assoc* 2003;10:115–28.

34 Honigman B, Lee J, Rothschild J, *et al*. Using computerized data to identify adverse drug events in outpatients. *J Am Med Inform Assoc* 2001;8:254–66.

35 Jha AK, Kuperman GJ, Teich JM, *et al*. Identifying adverse drug events: development of a computer-based monitor and comparison with chart review and stimulated voluntary report. *J Am Med Inform Assoc* 1998;5:305–14.

36 Thadani SR, Weng C, Bigger JT, *et al*. Electronic screening improves efficiency in clinical trial recruitment. *J Am Med Inform Assoc* 2009;16:869–73.

37 Thompson WK, Rasmussen LV, Pacheco JA, *et al*. An Evaluation of the NQF Quality Data Model for Representing Electronic Health Record Driven Phenotyping Algorithms. *AMIA Annu Symp Proc* 2012;2012:911–20.

38 KNIME (Konstanz Information Miner). (cited 21 February 2013). http://www.knime.org/

39 Chute CG. Medical concept representation. In: Chen H, Fuller SS, Friedman C, Hersh W. eds. *Medical informatics: knowledge management and data mining in biomedicine*. USA: Springer, 2005:163–82.