Development and evaluation of an ensemble resource linking medications to their indications

Wei-Qi Wei,[1] Robert M Cronin,[2] Hua Xu,[3] Thomas A Lasko,[1] Lisa Bastarache,[1] Joshua C Denny[1,2]

▶ Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/amiajnl-2012-001431).

[1]Department of Biomedical Informatics, Vanderbilt University, Nashville, Tennessee, USA
[2]Department of Medicine, Vanderbilt University, Nashville, Tennessee, USA
[3]School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, Texas, USA

Correspondence to
Dr Joshua Denny, Department of Biomedical Informatics, Vanderbilt University, Eskind Biomedical Library, Room 448, 2209 Garland Ave, Nashville, TN 37232, USA; josh.denny@vanderbilt.edu

Received 21 October 2012
Revised 25 February 2013
Accepted 18 March 2013
Published Online First 10 April 2013

To cite: Wei W-Q, Cronin RM, Xu H, *et al*. *J Am Med Inform Assoc* 2013;**20**:954–961.

ABSTRACT

Objective To create a computable MEDication Indication resource (MEDI) to support primary and secondary use of electronic medical records (EMRs).

Materials and methods We processed four public medication resources, RxNorm, Side Effect Resource (SIDER) 2, MedlinePlus, and Wikipedia, to create MEDI. We applied natural language processing and ontology relationships to extract indications for prescribable, single-ingredient medication concepts and all ingredient concepts as defined by RxNorm. Indications were coded as Unified Medical Language System (UMLS) concepts and International Classification of Diseases, 9th edition (ICD9) codes. A total of 689 extracted indications were randomly selected for manual review for accuracy using dual-physician review. We identified a subset of medication–indication pairs that optimizes recall while maintaining high precision.

Results MEDI contains 3112 medications and 63 343 medication–indication pairs. Wikipedia was the largest resource, with 2608 medications and 34 911 pairs. For each resource, estimated precision and recall, respectively, were 94% and 20% for RxNorm, 75% and 33% for MedlinePlus, 67% and 31% for SIDER 2, and 56% and 51% for Wikipedia. The MEDI high-precision subset (MEDI-HPS) includes indications found within either RxNorm or at least two of the three other resources. MEDI-HPS contains 13 304 unique indication pairs regarding 2136 medications. The mean±SD number of indications for each medication in MEDI-HPS is 6.22±6.09. The estimated precision of MEDI-HPS is 92%.

Conclusions MEDI is a publicly available, computable resource that links medications with their indications as represented by concepts and billing codes. MEDI may benefit clinical EMR applications and reuse of EMR data for research.

INTRODUCTION

Medication and diagnosis data are vital to clinical care and are core features of electronic medical records (EMRs). Medications are prescribed to treat disease (ie, the medication's intended indication), but they can also cause disease (ie, an adverse effect). Linking medications with their diagnoses electronically could improve evaluating treatment outcomes,[1 2] assessing healthcare quality,[3 4] and performing clinical and genomic research by enhancing understanding of a patient's longitudinal disease and treatment record.[5 6] However, medications are not explicitly linked to their indications within most EMRs, and research into computational resources to enable such linkage is limited. In this paper, we integrated four medication resources to create a freely available, computable MEDication-Indication (MEDI) resource, and describe its initial evaluation to assist in computational linkage of medications to their indications.

BACKGROUND

A medication's indication is the disease or condition for which it was prescribed in a given instance. Medications are typically prescribed without any structured record of indication in the EMR. In some cases, such as for medications prescribed using explicit order sets designed for a given diagnosis, a human can infer the indication from the order or clinical documentation. However, in general, computational inference of a medication's indication from EMR data is difficult. For example, disease-specific order sets are primarily found only in the inpatient setting, are not comprehensive for all diagnoses, and do not assert the diagnosis with certainty (eg, a provider may use the pneumonia order set for convenience when in fact the patient has a different infection).

Each medication can have many indications, and indications can be classified as either on-label or off-label. On-label indications are proposed in the early process of drug development by the manufacturer and later approved by the Food and Drug Administration (FDA) after demonstrating efficacy through clinical trials. These on-label indications appear on the package insert for the medication. For example, metformin is FDA-approved to treat type 2 diabetes, and ampicillin may be prescribed to treat urinary tract infections, otitis media, or pneumonia. Many on-label indications can be retrieved freely from the FDA's DailyMed website.[7] DailyMed currently contains drug labels for about 40 994 brand and generic medications for both humans and animals. Off-label indications are conditions for which the medication is used, but which have not been approved by the FDA and do not appear on the package insert. Many medications have common off-label indications.[8] For example, metformin is used off-label to treat polycystic ovarian disease,[9] and ampicillin is used off-label for diverticulitis.[10]

Typically, off-label indications are based on scientific evidence found subsequent to the FDA approval process and collective physician experience.[11] By nature, off-label indications can be controversial, such as the use of statins (a class of cholesterol-lowering medications) for diabetes, regardless of the patient's cholesterol levels.[12 13] Evidence for off-label use may be scattered among various drug resources. Although some proprietary resources list both on-label and off-label indications (eg, Epocrates,

FirstDataBank, and LexiComp), these resources are not freely available. Thus, it can be difficult to obtain a complete list of medication indications, and using a single resource (especially DailyMed) may miss important or common indications. These resources are generally formatted as free text, and require extra processing to convert them into a computable format.

Various medication resources have been created by leveraging either the EMR or literature for pharmaceutical research (eg, new drug discovery and adverse drug detection).[14] For instance, the Therapeutic Target Database (TTD) contains information about medications and their therapeutic targets and provides corresponding cross-links from the ClinicalTrials.gov database.[15] However, TTD is designed for new drug discovery; most of its data are oriented for drug development instead of clinical use. Another resource, the Side Effect Resource (SIDER), was developed using text-mining techniques applied to FDA-approved drug labels. SIDER provides a list of FDA-approved indications on marketed medications mined from FDA drug labels obtained from DailyMed.[16] Since its major focus is on side effects rather than indications, the indication list has not been thoroughly evaluated. Another important and relevant source is RxNorm, developed and maintained by the National Library of Medicine (NLM).[17] RxNorm is an ontology designed for exchanging medication information among clinical systems. It maintains a comprehensive list of commonly used medications (both generic and branded, with structured linkages between them), along with their forms, ingredients, and dosages. The integration of RxNorm with the National Drug File–Reference Terminology (NDF-RT) from the Veterans Health Administration has added significant indication information between single-ingredient medications and diseases through 'may_treat' and 'may_prevent' therapeutic relationships.[17] NDF-RT includes both on-label and off-label indications, but its performance on indications has not been previously reported. Preliminary work with earlier versions of RxNorm and NDF-RT demonstrated that a number of medications were lacking indications.[18 19]

In this paper, we proposed a novel ensemble approach that embraces multiple commonly used medication resources to create a computable drug resource, called MEDI. We believe that MEDI may assist in clinical applications within EMRs and the secondary use of EMR data.

METHODS

Data sources

We selected four medication resources as inputs into MEDI. The four resources included: (1) RxNorm (downloaded on June 4, 2012); (2) SIDER 2 (released on March 16, 2012)—a public medication knowledge base targeting adverse drug reactions extracted from FDA drug labels; (3) MedlinePlus (http://www.nlm.nih.gov/medlineplus)—an NLM-maintained website that offers consumer health information for patients, families, and healthcare providers; and (4) Wikipedia—an online collaboratively edited encyclopedia. RxNorm and SIDER 2 maintain indication information within a formal table structure with structured (ie, coded) medication and indication information. MedlinePlus and Wikipedia are free-text based and required further processing (figure 1).

Medication indication extraction

We retrieved all single-ingredient medication concepts (represented by RxNorm concept unique identifiers (RxCUIs) and defined as having only one 'has_ingredient' relationship), including clinical drugs and brand names from the prescribable subset of RxNorm. We also retrieved all ingredient RxCUIs from RxNorm, which were determined by term type (Term Type in Source (TTY)='IN' or TTY='MIN', or TTY='PIN') and included both single-ingredient and multi-ingredients. RxNorm covers almost all prescription medications currently marketed in the USA. Based on relationships within RxNorm,[17] all concepts were then collapsed into groups by their ingredients. For example, 'Tylenol Caplet, 325 mg oral tablet' (RxCUI 209387) was mapped to 'Acetaminophen' (RxCUI 161).

To obtain indications of a medication from RxNorm, we retrieved all diseases that connect with the medication through either 'may_be_treated_by' or 'may_be_prevented_by' relationships. For SIDER 2, we mapped medications to corresponding RxCUIs where their brand names or drug names matched the terms associated with the RxCUIs. The mapping involved two steps: the first step was looking for exact matches and the second step was searching for partial matches if no exact match was found in step 1 —for example, 'salbutamol sulfate' was mapped to 'salbutamol'. We then retrieved all disease indications documented within the SIDER 2 indication table, which are mapped to Unified Medical Language System (UMLS) concepts.

To obtain medication indications from MedlinePlus, we first retrieved the webpage for each medication through the MedlinePlus Application Programming Interface (API) using the medication's RxCUI as the query input. The resulting Hypertext Markup Language (HTML) pages were parsed and stored as text files. MedlinePlus maintains a consistent document structure for its drug monographs, although the text within each section is free-text. We found that certain sections of MedlinePlus frequently contained drug indication information. We limited our analysis of the MedlinePlus description to the sections 'Why is this medication prescribed', 'About your treatment', and 'Other uses for this medicine', thus ignoring sections such as 'What side effects can this medication cause' and 'Precautions'. We used the KnowledgeMap Concept Indexer (KMCI) to parse the free-text to obtain all non-negated 'disease and finding' concepts, as mapped to UMLS concepts. KMCI is a general-purpose natural language processing (NLP) engine that maps free-text documents to UMLS concepts and includes negation detection through an adaptation of the NegEx algorithm.[20] KMCI has performed favorably in comparison with MetaMap[21] for medical school curriculum documents and has been validated in a variety of clinical and education contexts.[22–25]

To identify medication pages in Wikipedia, we queried the Wikipedia API with medication strings derived from RxNorm (querying with both brand and generic names for each drug). We used KMCI to identify non-negated disease concepts from the resulting Wikipedia pages as we did for MedlinePlus. However, since Wikipedia does not contain a formal structure clearly annotating medication indications, we employed heuristic rules. In perusal of Wikipedia entries, we noted that most medication entries listed indications before side effects, which were often listed in separate sections. Thus, we excluded any concepts found after a 'side-effect', 'safety', or 'toxicology' section. For entries that were just text based without being separated into sections, all content was parsed.

All disease concepts extracted from MedlinePlus or Wikipedia were initially represented as UMLS concepts. After processing by KMCI, concepts were restricted to those that could be mapped to International Classification of Diseases, Ninth Revision, Clinical Modification (ICD9) codes using UMLS relationships (as defined in MRREL), and thus included diseases, syndromes, symptoms, and other clinical findings. Both the original concept and the resultant ICD9 concept were kept for MEDI. ICD9 codes were chosen since these are commonly

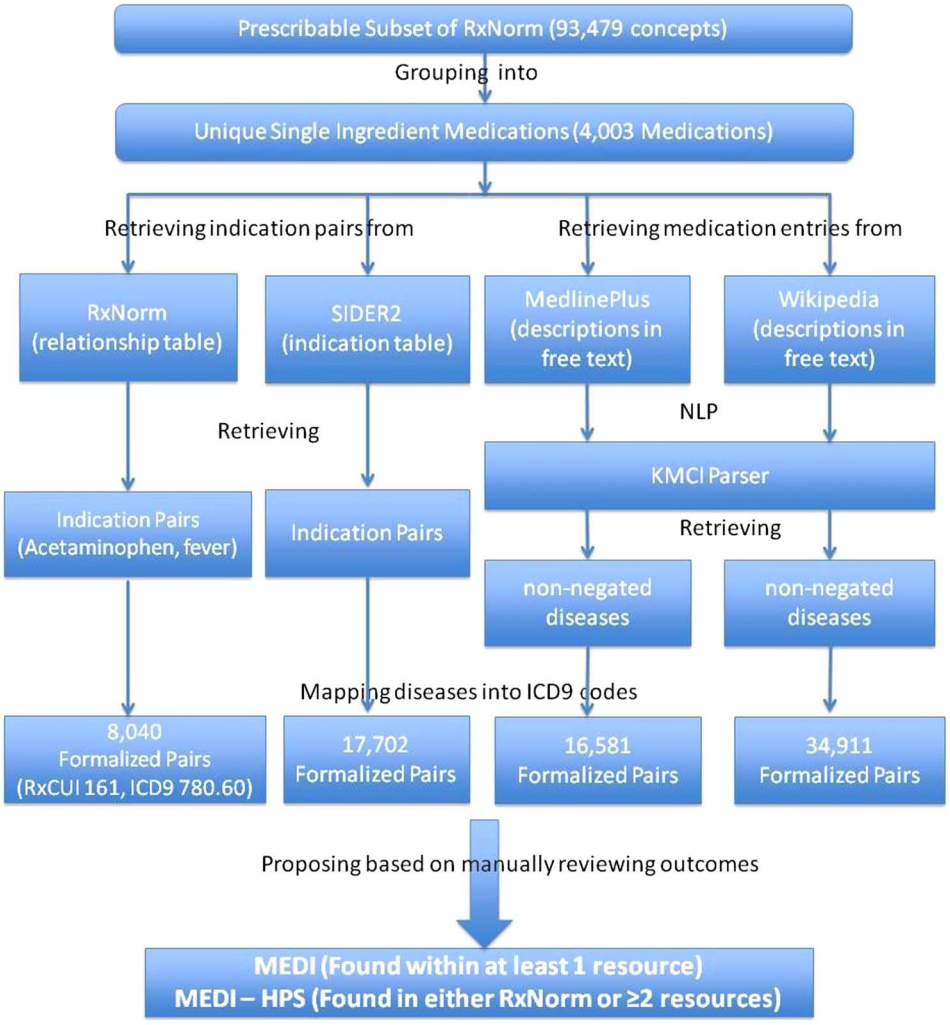

Figure 1 Flowchart for MEDication–Indication (MEDI) creation. HPS, high-precision subset; ICD9, International Classification of Diseases, 9th edition; KMCI, KnowledgeMap Concept Indexer; RxCUI, RxNorm concept unique identifier; SIDER, Side Effect Resource.

available codes within most EMR systems. For those SNOMED-CT concepts that could not be directly mapped to ICD9 codes through UMLS relationships, we used the SNOMED-CT ICD9 CrossMap[26] to map them into corresponding ICD9 codes, where possible. We only used relationships with map advice equal to 1 (one-to-one SNOMED-CT to ICD9 map) or 2 (Narrow to Broad SNOMED-CT to ICD9 map).

Evaluation

We categorized each medication–indication pair by the combination of resources in which it was found (RxNorm alone, RxNorm and SIDER 2, etc). Each category, represented by a row in table 1, is one of 15 possible combinations of our four sources. Each source is positive for eight of these combinations, meaning that all medication–indication pairs in that category were found in the resource (these are indicated by a 'Y' in table 1).

We calculated the true positive rate for each category by manually evaluating 50 randomly selected medication–indication pairs per category. Two practicing physicians (JCD and RMC) each reviewed the indications independently, and differences were resolved by consensus. Physicians used clinical experience, search of drug resources and medical references, and web and PubMed searching to determine the veracity of medication–indication pairs.

We estimated precision and recall of each resource, r, using equations (1) and (2), where C(r) is the set of eight categories for which the resource, r, is positive, size(n) is the number of medication–indication pairs in category n, and TPR(n) is the true positive rate for category n.

$$\text{Precision}(r) = \frac{\sum_{n \in C(r)} \text{size}(n) \cdot \text{TPR}(n)}{\sum_{n \in C(r)} \text{size}(n)} \quad (1)$$

$$\text{Recall}(r) = \frac{\sum_{n \in C(r)} \text{size}(n) \cdot \text{TPR}(n)}{\sum_{\text{all } n} \text{size}(n) \cdot \text{TPR}(n)} \quad (2)$$

These equations estimate the standard precision and recall measures for each resource, r, but they do so using stratified sampling over the categories.

To demonstrate that MEDI has a broader coverage than RxNorm, we compared the indications in MEDI with the indications in RxNorm within the context of cancer. Cancer was chosen because it is a broad group of important diseases easily identified through a single set of ICD9 codes (140–239) covered primarily by prescription medications with well-defined indications. We compared MEDI with RxNorm, and then randomly selected a few medication–indication pairs for validation. We validated through use of general medicine resources such as UpToDate and PubMed searches.

Table 1 Validation results by two reviewers

RxNorm	MedlinePlus	Wikipedia	SIDER 2	Size	Pairs reviewed	False positives	Precision (%)
1 resource							
N	N	N	Y	10880	49	24	51
N	N	Y	N	28323	45	24	47
N	Y	N	N	10836	46	16	65
Y	N	N	N	3683	35	4	89
2 resources							
N	N	Y	Y	1592	35	8	77
N	Y	N	Y	1464	49	4	92
N	Y	Y	N	1142	47	5	89
Y	N	N	Y	813	40	1	98
Y	N	Y	N	868	31	0	100
Y	Y	N	N	381	39	2	95
3 resources							
N	Y	Y	Y	1066	48	1	98
Y	N	Y	Y	603	46	2	96
Y	Y	N	Y	375	56	2	96
Y	Y	Y	N	408	39	1	97
All 4 resources							
Y	Y	Y	Y	909	65	0	100

N, no; SIDER, Side Effect Resource; Y, yes.

RESULTS

Medication indication extraction

From 93 479 unique concepts in the RxNorm prescribable table, we retrieved 61 450 medication concepts that could be mapped to a single ingredient concept (via having only one 'has_ingredient' RxCUI relationship) in RxNorm. Thus, these 61 450 single-ingredient medication concepts, which include brand names and various clinical drug forms, were then grouped into 4003 unique RxNorm medication ingredients (ie, TTY='IN' or TTY='MIN', or TTY='PIN'). Of these 4003 medication ingredients, 3112 (78%) had at least one indication extracted from at least one of the four resources, and 2114 (53%) had indication extracted from at least two (figure 2, left). The 891 medication ingredients without any indication extracted from any resource were typically not medications — for example, 'kiwi allergenic extract' (RxCUI 1010926), 'lime' (RxCUI 1011060), and 'sugar cane extract' (RxCUI 1014711).

From the 3112 medications with indications, we identified 3009 unique ICD9 codes and 63 343 indication pairs (one RxCUI and one ICD9 code) (table 2). The mean±SD number

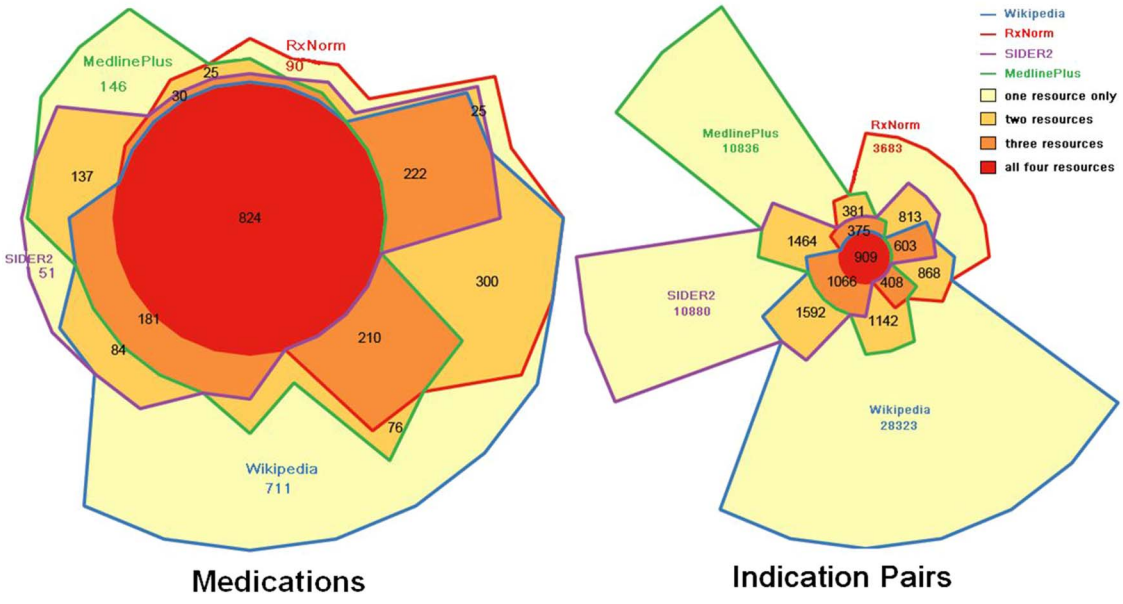

Figure 2 Weighted Venn diagram of the distribution of 3112 medications (left) and 63 343 indication pairs (right) within the four resources. Each border color represents a resource. Different colored areas represent medications–indications that were found within different combinations of resources. The area sizes surrounded by border color(s) are proportional to the number of medications–indications that were found within the corresponding resource(s). SIDER, Side Effect Resource.

Table 2 Number of unique medications, ICD9 codes, and indication pairs extracted from each resource

Resource	Medications (% of total)	ICD9 codes (% of total)	Indication pairs (% of total)
RxNorm	1726 (56)	999 (33)	8040 (13)
SIDER 2	1554 (50)	1703 (57)	17702 (28)
MedlinePlus	1629 (52)	869 (29)	16581(26)
Wikipedia	2608 (84)	2624 (87)	34911 (55)
Union of all resources	3112	3009	63343

SIDER, Side Effect Resource.

of extracted indications for each medication was 20.35±22.00, and mode was 3. Of these 63 343 indication pairs, 53 722 (85%) were found within a single resource, and 9621 (15%) were found within two or more resources (figure 2, right).

Medication indication validation

We initially chose ~700 medication–indication pairs for review (~50 pairs from each category) before we noticed that some of the disease concepts retrieved by NLP were older CUIs that overlapped with RxNorm concepts but appeared as duplicates in the initial review. We then mapped older (deleted) UMLS concepts to current UMLS concepts using the UMLS history files; therefore, the number of medication–indication pairs reviewed in each category was not equal. A total number of 689 medication–indication pairs were finally reviewed for validation. Among these 689 indications, 19 were marked as uncertain (eg, ondansetron and irritable bowel syndrome, albumin and dehydration, estradiol and other malaise and fatigue, etc) because both reviewers agreed that they really were not representative of what the true indication was but yet were not really false indications. Therefore, we ignored these uncertain ones in subsequent analyses. As shown in table 1, the precision was 100% for indications found within all four resources; precisions were above 95% for indications found within any three of the four resources. Precision was near 90% for indications found within two resources, except for indications found only in SIDER 2 and Wikipedia (a precision of 77%). For indications found within only Wikipedia, SIDER 2, or MedlinePlus, the precision dropped to 47–65%. In contrast, indications found only with RxNorm still had high precision (89%). Table 3 shows a random selection of some of the errors from each resource.

Table 4 presents the estimated recall and precision for each resource as defined by equations 1 and 2. We found that RxNorm gave a remarkably high precision (94%) but a relatively low recall (20%). Compared with RxNorm, other resources had lower precision (56–75%) but higher recall (31–51%). Wikipedia achieved the best recall (51%) among the four resources. The differences between RxNorm and the other three resources in terms of recall and precision may be explained by the fact that the RxNorm indications were already curated and stored in a structured format, while the indications from the other three resources were based on either NLP concept retrieval or text mining techniques.

Currently, MEDI contains a total of 3112 medications and 63 343 medication–indication pairs found within the four sources. From our validation results, we observed that high precision could be achieved with indications found within any two (or more) of the four resources. Indications solely within RxNorm also had a high precision. Thus, to optimize recall while maintaining reasonable precision, we defined the MEDI high-precision subset (MEDI-HPS) as the indications found within either at least two of the four resources or RxNorm. The current version of MEDI-HPS contains 13 304 unique indication pairs for 2136 medications. The mean number of indications for each medication is 6.22±6.09. The mode for each medication is 2, while the median is 4. Examples of four commonly used medications from MEDI are provided in online supplementary table S1.

MEDI-HPS offers a comparable number of indications but a much higher precision (92%) than MedlinePlus (75%), SIDER 2 (67%), or Wikipedia (56%). Wikipedia, as the only uncontrolled resource we utilized, contributes the largest number of indications, but also has the lowest precision. MEDI-HPS has slightly lower precision (92%) than either RxNorm (94%) or ≥2 resources (93%) because the precision of medication–indication pairs found only within RxNorm is 89%.

Compared with RxNorm, MEDI-HPS maintains a similar high precision (92%) but provides 5264 (66%) more indications. To demonstrate this advantage, we compared the coverage of MEDI-HPS with RxNorm within the context of cancer medications. We retrieved all medications that have cancer (ICD9 codes

Table 3 Selected example errors from each resource

Resource	Medication (RxCUI)	Disease (ICD9)	Comment
RxNorm	Captopril (1998)	Rheumatoid arthritis (714.0)	The indication was supported by a small case series in 1984[27] but has not been widely accepted thereafter
	Isosorbide (6057)	Esophageal reflux (530.81)	Isosorbide can be used for esophageal spasm, but may cause reflux
MedlinePlus	Sildenafil (136411)	Other malaise and fatigue (780.79)	NLP falsely identified a concept that is irrelevant to an indication: 'Sildenafil is used to improve the ability to exercise in people with pulmonary arterial hypertension (PAH; high blood pressure in the vessels carrying blood to the lungs, causing shortness of breath, dizziness, and *tiredness*)'
	Dexmethylphenidate (352372)	Other specified visual disturbances (368.8)	Mismatched disease concept by NLP: 'Dexmethylphenidate is used as part of a treatment program to control symptoms of attention deficit hyperactivity disorder (ADHD; *more difficulty focusing*,…) in adults and children'
Wikipedia	Ciprofloxacin (2551)	Cystic fibrosis (277.0)	However, the fluoroquinolones are licensed to treat lower respiratory infections in children with *cystic fibrosis* in the UK
	Guaifenesin (5032)	Asthma (493)	Guaifenesin is claimed to be effective in the treatment of the thickened bronchial mucosa characteristic of *asthma*
SIDER 2	Dobutamine (3616)	Atrial fibrillation (427.31)	Contraindication/side effect
	Ephedrine (3966)	Hypertension NOS (401.9)	Contraindication/side effect

NOS, not otherwise specified; NLP, natural language processing; RxCUI, RxNorm concept unique identifier; SIDER, Side Effect Resource.

Table 4 Estimated precision and recall for different resources

	Medications	Indication pairs	Precision (%)	Recall (%)
RxNorm	1726	8040	94	20
MedlinePlus	1629	16581	75	33
Wikipedia	2608	34911	56	51
SIDER 2	1554	17702	67	31
4 resources	433	909	100	2
≥3 resources	1108	3361	98	9
≥2 resources	1847	9621	93	23
MEDI (≥1 resource)	3112	63343	60	100
MEDI-HPS (≥2 or RxNorm)	2136	13304	92	30

HPS, high-precision subset; MEDI, medication indication resource; SIDER, Side Effect Resource.

140–239) as an indication. MEDI-HPS included 269 cancer medications while RxNorm only had 166. A total of 103 (38%) medications were absent in RxNorm, including plerixafor, romidepsin, raloxifene, pralatrexate, and eribulin—all valid cancer drugs whose indications listed in MEDI-HPS were validated through literature review.

MEDI is available in a comma-separated values file format. The file consists of medications represented by RxCUIs and indications mapped to UMLS CUIs and ICD9 codes, as well as other metadata including a column called 'possible_label_use'. The value of 'possible_label_use' is 1 when the indication is mentioned in SIDER 2 and 0 when it is not. Our assumption is that, since SIDER 2 is extracted from drug labels, indications mentioned within SIDER 2 are highly likely to be on-label (ie, FDA-approved) uses.

DISCUSSION

By leveraging existing public resources, ontologies, and NLP, we created a computable medication indication resource for both on- and off-label indications that is mapped to standard billing codes and structured vocabularies. The current version of MEDI contains 63 343 medication–indication pairs for 3112 medications. MEDI-HPS, the high precision subset of MEDI, provides 13 304 indication pairs for 2136 medications. The precision or recall of MEDI-HPS was better than RxNorm, SIDER 2, MedlinePlus, and Wikipedia by themselves. MEDI (and future resources like it) may facilitate computational linkage of prescriptions with their indications, enabling both clinical and research use of EMR data.

The adoption of EMRs has been rapidly expanding in the USA since 2008, especially after the passing of the Health Information Technology for Economic and Clinical Health (HITECH) Act.[28] Requirements such as maintaining structured lists of problems, medications and allergies, and electronic prescribing are key components of meaningful use stage one.[29] The continuing accumulation of EMR data will present unprecedented opportunities for clinical research. However, the 'information gap' between medications and diseases precludes the efficient use of these practice-based medication data, hindering the primary and secondary use of EMRs. MEDI and tools like it may begin to fill that gap.

MEDI may be useful in current phenotype algorithms[30–33] or for future deep phenotyping,[5 6 34 35] both of which require detailed clinical data to accurately classify patients into subpopulations with respect to a disease, a phenotypic subclass of a disease, or a response to a treatment. For instance, medications were used in addition to ICD9s in EMR phenotype algorithms for type 2 diabetes mellitus,[31 33 36] Crohn's disease,[32] rheumatoid arthritis,[30] and many of the other algorithms deployed in the Electronic Medical Records and Genomics (eMERGE) Network to identify cases and controls for genome-wide association studies.[37–39] MEDI may also improve the accuracy of the detection of adverse drug reactions[40] and elevate the quality and utility of the EMR problem lists.[41] In addition, tools such as MEDI may also improve the precision of phenome-wide EMR phenotyping methods, such as the ICD9-based phenome-wide association studies (PheWAS) method, by allowing integration of two axes of clinical information.[38 42]

Wikipedia is one of the most commonly visited websites in the world, but it has rarely been evaluated in the medical literature, owing possibly in part to being an uncontrolled source with uncontrolled structure, challenging its use in medical applications. A small study in 2005 reported that Wikipedia had similar accuracy to Encyclopedia Britannica.[43] Our study shows that Wikipedia's recall on indications is significantly higher than that of RxNorm, SIDER 2, and MedlinePlus. In addition, we noted that Wikipedia contains a number of homeopathic/alternative medications and treatments (eg, parsley and nephrolithiasis) that are not in other resources. Homeopathic medications, because they are not RxNorm prescribable medications, are not included in the current version of MEDI, but may be included in the future.

One source of error and possible area for improvement is the mapping of indications from free-text resources. NLP-induced errors were largely caused by a mismatched disease concept, a failure to recognize negation, or a failure to identify that a concept was actually a side effect/complication. For instance, in the sentence 'the process is called starch gelatinization', KMCI falsely identified a disease concept—CALL (precursor B-cell lymphoblastic leukemia, CUI C1292769) based on the normalization of the past participle 'called' to 'call', and the mapping of 'call' to the acronym 'CALL.' This category of error may be resolved by disallowing mappings between normalized strings and UMLS acronyms. In another sentence, 'Adefovir will not cure hepatitis B and may not prevent complications of chronic hepatitis B such as cirrhosis of the liver or liver cancer', KMCI failed to recognize that the 'liver cancer' was negated because of the distance between the subject and the target term. We hope to correct these classes of error in future work.

MEDI does not replace existing commercially available resources. Many commercially available resources (eg, Epocrates[44] and LexiComp[45]) provide not only drug–indication pairing but also dose guidance that can be tailored to indication, drug formulations, international brand names, safety warnings, and adverse reactions; none of this information is currently provided in MEDI (nor is it the goal of MEDI to be a comprehensive prescribing guide). However, medication and indications are represented by formal concepts and billing codes in MEDI rather than embedded in free text as in Epocrates or LexiComp. Conceptual formalization should facilitate research and application creation. A random review of a few common medications noted that, for some, MEDI may include more indications than commercial resources. For example, MEDI-HPS shows that propranolol can be used for congestive heart failure, panic disorder, and thyrotoxicosis (each clinically valid uses not listed in Epocrates and only one of which, thyrotoxicosis, is listed in LexiComp) in addition to hypertension, migraines, angina, myocardial infarction, pheochromocytoma, arrhythmias, and essential tremor, all of which are listed by all resources. However, MEDI shows that metformin can be used to

treat obesity, which has been supported in research trials[46 47] but is not a common indication. Thus, since MEDI is not a curated resource, some of its listed indications will likely be false positives compared with commercial resources.

Several limitations regarding the creation and evaluation of MEDI should be clarified. First, MEDI is limited to medications and indications found in those four resources. Although the resulting precision is encouraging, the addition of other resources may improve both recall and precision. Second, MEDI currently primarily includes medications composed of a single ingredient (97.7%); only 2.3% were multi-ingredient concepts. Thus, MEDI probably does not include all prescription medications on the market today, which may be especially true for combination medications. Users should be careful about this limitation when they utilize MEDI or MEDI-HPS to conduct research or create applications. Third, we have not made any judgments about the strength of evidence for off-label uses when building MEDI; this information is not easily found. In addition, we estimated recall and precision for each vocabulary, but they could be skewed because of the resources we chose and the artifacts of NLP. Finally, we lacked a complete list of indications or a true gold standard to evaluate recall. In our analysis, recall was estimated using the assumption that the true positive drug–indication pairs from all sources represented the universe of all possible drug–indication pairs. Thus, true recall is likely lower since our method could not detect indications not listed in one of the four resources.

This paper introduces our initial efforts to create MEDI. Future work can make this resource more complete and more robust. For example, further elucidation of off-label status and drug form information would improve MEDI's clinical usability. In addition, medications are not all used with equal frequency. Evaluation of indication prevalence using EMR data may aid identification of new medication–indication pairs and improve accuracy.

CONCLUSION

In summary, MEDI is a freely available, computable medication indication resource that is more comprehensive than existing freely available resources. Because it utilized UMLS concepts and ICD9 codes, MEDI can be easily used in conjunction with billing codes or concepts extracted from free-text using NLP. Our results demonstrate its broad coverage and high accuracy. MEDI may enable research and clinical EMR applications. MEDI is freely available at http://knowledgemap.mc.vanderbilt.edu/research/content/MEDI. We plan to update MEDI periodically as component resources are revised.

Contributors Study initialization: W-QW and JCD. Study design: W-QW, HX, TAL, and JCD. Acquisition of data: W-QW, RMC, and LB. Analysis and interpretation of data: W-QW, RMC, HX, TAL, LB, and JCD. Drafting of the manuscript: W-QW, RMC, HX, TAL, LB, and JCD. All authors contributed to refinement of the manuscript and approved the final manuscript. Grant holder: JCD.

Funding The project was supported by NIH grant 1 R01 LM 010685.

Competing interests None.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1 Cebul RD, Love TE, Jain AK, *et al*. Electronic health records and quality of diabetes care. *N Engl J Med* 2011;365:825–33.
2 Ghitza UE, Sparenborg S, Tai B. Improving drug abuse treatment delivery through adoption of harmonized electronic health record systems. *Subst Abuse Rehabil* 2011;2011:125–31.
3 Roth CP, Lim YW, Pevnick JM, *et al*. The challenge of measuring quality of care from the electronic health record. *Am J Med Qual* 2009;24:385–94.
4 Roth MT, Weinberger M, Campbell WH. Measuring the quality of medication use in older adults. *J Am Geriatr Soc* 2009;57:1096–102.
5 Tracy RP. 'Deep phenotyping': characterizing populations in the era of genomics and systems biology. *Curr Opin Lipidol* 2008;19:151–7.
6 Wilke RA, Xu H, Denny JC, *et al*. The emerging role of electronic medical records in pharmacogenomics. *Clin Pharmacol Ther* 2011;89:379–86.
7 DailyMed 2012; dailymed.nlm.nih.gov.
8 Epstein RS, Huang SM. The many sides of off-label prescribing. *Clin Pharmacol Ther* 2012;91:755–8.
9 Bates GW, Legro RS. Longterm management of Polycystic Ovarian Syndrome (PCOS). *Mol Cell Endocrinol* 2012, pii: S0303-7207(12)00481-9. doi: 10.1016/j.mce.2012.10.029. [Epub ahead of print 20 Dec 2012].
10 Tursi A. Acute diverticulitis of the colon–current medical therapeutic management. *Expert Opin Pharmacother* 2004;5:55–9.
11 Psaty BM, Ray W. FDA guidance on off-label promotion and the state of the literature from sponsors. *JAMA* 2008;299:1949–51.
12 Kearney PM, Blackwell L, Collins R, *et al*. Efficacy of cholesterol-lowering therapy in 18,686 people with diabetes in 14 randomised trials of statins: a meta-analysis. *Lancet* 2008;371:117–25.
13 Preiss D, Seshasai SR, Welsh P, *et al*. Risk of incident diabetes with intensive-dose compared with moderate-dose statin therapy: a meta-analysis. *JAMA* 2011;305:2556–64.
14 Duke JD, Han X, Wang Z, *et al*. Literature based drug interaction prediction with clinical assessment using electronic medical records: novel myopathy associated drug interactions. *PLoS Comput Biol* 2012;8:e1002614.
15 Zhu F, Shi Z, Qin C, *et al*. Therapeutic target database update 2012: a resource for facilitating target-oriented drug discovery. *Nucleic Acids Res* 2012;40:D1128–136.
16 Kuhn M, Campillos M, Letunic I, *et al*. A side effect resource to capture phenotypic effects of drugs. *Mol Syst Biol* 2010;6:343.
17 Nelson SJ, Zeng K, Kilbourne J, *et al*. Normalized names for clinical drugs: RxNorm at 6 years. *JAMIA* 2011;18:441–8.
18 Bodenreider O, Peters LB. A graph-based approach to auditing RxNorm. *J Biomed Inform* 2009;42:558–70.
19 Warnekar PP, Bouhaddou O, Parrish F, *et al*. Use of RxNorm to exchange codified drug allergy information between Department of Veterans Affairs (VA) and Department of Defense (DoD). *AMIA Annu Symp Proc* 2007:781–5.
20 Denny JC, Peterson JF. Identi QT prolongation from ECG impressions using natural language processing and negation detection. *Stud Health Technol Inform* 2007;129 (Pt 2):1283–8.
21 Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. *JAMIA* 2010;17:229–36.
22 Denny JC, Bastarache L, Sastre EA, *et al*. Tracking medical students' clinical experiences using natural language processing. *J Biomed Inform* 2009;42:781–9.
23 Denny JC, Peterson JF, Choma NN, *et al*. Extracting timing and status descriptors for colonoscopy testing from electronic medical records. *JAMIA* 2010;17:383–8.
24 Denny JC, Smithers JD, Miller RA, *et al*. "Understanding" medical school curriculum content using KnowledgeMap. *JAMIA* 2003;10:351–62.
25 Denny JC, Spickard A III, Miller RA, *et al*. Identifying UMLS concepts from ECG Impressions using KnowledgeMap. *AMIA Ann Proc* 2005:196–200.
26 SNOMED CT ICD9 CrossMap 2012; http://www.nlm.nih.gov/research/umls/licensedcontent/snomedctarchive.html (accessed 1 Jan 2013).
27 Martin MF, Surrall KE, McKenna F, *et al*. Captopril: a new treatment for rheumatoid arthritis? *Lancet* 1984;1:1325–8.
28 Shea S, Hripcsak G. Accelerating the use of electronic health records in physician practices. *N Engl J Med* 2010;362:192–5.
29 CMS. EHR Incentive Programs 2012; https://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/index.html (accessed 1 Jan 2013).
30 Carroll RJ, Thompson WK, Eyler AE, *et al*. Portability of an algorithm to identify rheumatoid arthritis in electronic health records. *JAMIA* 2012;19:e162–9.
31 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. *JAMIA* 2012;19:212–18.
32 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.
33 Wei W-Q, Leibson CL, Ransom JE, *et al*. Impact of data fragmentation across healthcare centers on the accuracy of a high-throughput clinical phenotyping algorithm for specifying subjects with type 2 diabetes mellitus. *JAMIA* 2012;19:219–24.
34 Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. *JAMIA*.
35 Robinson PN. Deep phenotyping for precision medicine. *Hum Mutat* 2012;33:777–80.

36 Wei WQ, Leibson CL, Ransom JE, *et al*. The absence of longitudinal data limits the accuracy of high-throughput clinical phenotyping for identifying type 2 diabetes mellitus subjects. *Int J Med Inform* 2013;82:239–47.
37 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.
38 Denny JC, Crawford DC, Ritchie MD, *et al*. Variants near FOXE1 are associated with hypothyroidism and other thyroid conditions: using electronic medical records for genome- and phenome-wide studies. *Am J Hum Genet* 2011;89:529–42.
39 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. doi: 10.1186/1755-8794-4-13. http://www.ncbi.nlm.nih.gov/pubmed/21269473
40 Liu M, Wu Y, Chen Y, *et al*. Large-scale prediction of adverse drug reactions using chemical, biological, and phenotypic properties of drugs. *JAMIA* 2012;19:e28–35.
41 Burton MM, Simonaitis L, Schadow G. Medication and indication linkage: a practical therapy for the problem list? *AMIA Ann Symp Proc* 2008:86–90.
42 Denny JC, Ritchie MD, Basford MA, *et al*. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics* 2010;26:1205–10.
43 Giles J. Internet encyclopaedias go head to head. *Nature* 2005;438:900–1.
44 epocrates. 2013; http://www.epocrates.com/ (accessed 1 Jan 2013).
45 lexicomp 2013; http://www.utdol.com (accessed 1 Jan 2013).
46 Glueck CJ, Fontaine RN, Wang P, *et al*. Metformin reduces weight, centripetal obesity, insulin, leptin, and low-density lipoprotein cholesterol in nondiabetic, morbidly obese subjects with body mass index greater than 30. *Metabolism* 2001;50:856–61.
47 Seifarth C, Schehler B, Schneider HJ. Effectiveness of metformin on weight loss in non-diabetic individuals with obesity. *Exp Clin Endocrinol Diabetes* 2013;121:27–31.