Validation and Enhancement of a Computable Medication Indication Resource (MEDI) Using a Large Practice-based Dataset

Wei-Qi Wei[1], MMed, PhD, Jonathan D. Mosley[2], MD, PhD, Lisa Bastarache[1], MS, Joshua C. Denny[1,2], MD, MS
[1]Department of Biomedical Informatics, Vanderbilt University, Nashville, TN
[2]Department of Medicine, Vanderbilt University, Nashville, TN

Abstract

Linking medications with their indications is important for clinical care and research. We have recently developed a freely-available, computable medication-indication resource, called MEDI, which links RxNorm medications to indications mapped to ICD9 codes. In this paper, we identified the medications and diagnoses for 1.3 million individuals at Vanderbilt University Medical Center to evaluate the medication coverage of MEDI and then to calculate the prevalence for each indication for each medication. Our results demonstrated MEDI covered 97.3% of medications recorded in medical records. The "high precision subset" of MEDI covered 93.8% of recorded medications. No significant prescription drugs were missed by MEDI. Manual physician review of random patient records for four example medications found that the MEDI covered the observed indications, and confirmed the estimated prevalence of these medications using practice information. Indication prevalence information for each medication, previously unavailable in other public resources, may improve the clinical usability of MEDI. We believe MEDI will be useful for both clinical informatics and to aid in recognition of phenotypes for electronic medical record-based research.

Introduction

Medication and diagnosis data are critical components in patient's healthcare trajectory, and are chronicled in an electronic medical record (EMR). Medications can be treatments (or preventative therapies) for a disease, or cause them through adverse effects. Clinically, the indication for a given medication refers to the use of that medication for treating or preventing a particular disease. For example, hyperlipidemia is an indication for use of cholesterol-lowering statin medications, and type 2 diabetes mellitus is an indication for oral hypoglycemic medications, such as sulfonylureas. Medications can have relatively few indications, such as statins or sulfonylureas, or many possible indications, such as corticosteroids (used for a wide variety are autoimmune and inflammatory conditions) or aspirin (used to alleviate pain, reduce fever, and prevent thrombosis).

Electronically linking medications with their indications can improve routine clinical care and secondary uses of EMR data. For clinical care, the linkage could improve evaluation of treatment outcomes[1,2] and assessment of healthcare quality.[3,4] For secondary uses of EMR data, the linkage could enable clinical and genomic research by enhancing understanding of a patient's longitudinal disease and treatment record.[5,6] Medications often serve as important markers of disease severity, and thus are useful in algorithms to identify clinical phenotypes (typically diseases) for research.[7-11] Medications exposures, and the patient's response to the medications during treatment, can enable pharmacogenomic and comparative effectiveness research.[5,6,12-14] In recent studies, medications have been incorporated with diagnostic codes in EMR phenotype algorithms for type 2 diabetes mellitus (T2DM),[9,11,15] Crohn's disease,[10] rheumatoid arthritis,[8] 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.[16-18] In addition, linkage of medications to their indications may improve the accuracy of the detection of adverse drug reactions[19] and elevate the quality and utility of the EMR problem lists.[20]

We have previously described initial efforts to integrate multiple publicly-available medication resources to create a free, computable medication-indication resource, called MEDI.[21] We included RxNorm[22], SIDER 2[23], MedlinePlus[24], and Wikipedia[25] to form MEDI. We applied natural language processing (NLP) and ontology relationships to extract indications from the four resources. All medications were represented by RxNorm concepts (RxCUIs) and indications were formalized into International Classification of Diseases, 9th edition (ICD9) codes.

Additionally, we identified a "high precision subset" of MEDI (MEDI-HPS) whose indications had a precision of over 92% based on physician review.

In this paper, we continued our work on MEDI with an emphasis on clinical utility. We leveraged a large practice-based dataset to investigate the coverage of MEDI on medications prescribed within clinical practice, and to improve the usability of MEDI by calculating the prevalence of each indication for each medication.

Background

As summarized in Table 1, currently MEDI offers two datasets: the complete MEDI indication dataset, and MEDI-HPS, the high precision subset of MEDI. MEDI contains a total number of 3,112 generic drug ingredients and 63,343 medication-indication pairs found from the four sources. MEDI-HPS was created to optimize recall while maintaining reasonable precision. MEDI-HPS was defined as the indications found within either at least two of the four resources or RxNorm, two simple rules that ensured a precision > 92%. MEDI-HPS contains 13,304 unique indication pairs for 2,136 generic drug ingredients. Based on manually review in our prior study, MEDI-HPS offers much higher precision (>92%, precision) than MedlinePlus (75%), SIDER 2 (67%), or Wikipedia (56%). Compared to RxNorm, MEDI-HPS maintains a similar high precision (>92%) but provides 5,264 more indications. MEDI datasets are freely available at http://knowledgemap.mc.vanderbilt.edu/research/content/MEDI.

Table 1. MEDI characteristics

Dataset	Definition	Medications	Indications	Indications/Medication*
MEDI	Found in RxNorm, SIDER 2, MedlinePlus, or Wikipedia	3,112	63,343	20.35±22.00
MEDI-HPS	Found in either ≥ 2 resources or RxNorm	2,136	13,304	6.22±6.09

* mean ± standard deviation.

MEDI and MEDI-HPS datasets have not been validated using clinical practice data. None of the four resources that were used to develop MEDI was directly derived from clinical data. The indications within RxNorm are manually curated, validated, and stored in a structured format. SIDER 2 was developed using text-mining techniques applied to FDA approved drug labels. Both MedlinePlus and Wikipedia are human created resources. Thus, we were not able to provide an important clinically-relevant feature, i.e. indication prevalence, in previous MEDI. Indication prevalence is critical for clinical studies. It may simplify building of clinical informatics applications and secondary use of EMR data since medications are not all used for indications with equal frequency. Some indications are much more common than others, e.g. type 2 diabetes mellitus is a more common indication of metformin than polycystic ovary syndrome. Without prevalence information, it may be challenging to accurately identify the primary indications.

Methods

Study Setting

This study was conducted at Vanderbilt University Medical Center (VUMC) in Nashille, Tennessee. VUMC has previously constructed a de-identified version of its integrated (combined inpatient-outpatient) EMR for clinical and genomic research, called the Synthetic Derivative (SD).[26] The SD contains the records of over 1.7 million unique individuals, including dense longitudinal clinical data for >1 million individuals (average record size of 106,727 bytes). The SD incorporates clinical data from multiple sources, including diagnostic and procedural codes, as well as provider progress notes, hospital admissions, discharge summaries, laboratory data, and medication histories. Since the SD contains only de-identified data, research using this resource has been determined by Vanderbilt's Institutional Review Board to constitute non-human subjects research.

Data Extraction

We used all records in the SD that contained both medication and International Classifications of Disease, 9th edition, (ICD9) codes, which resulted in 1,371,354 individuals. Diagnostic codes (ICD9s) were retrieved from

administrative claims data. Medication data were obtained from electronic prescription records combined with natural language processing applied to clinical narratives using MedEx [27].MedEx extracts drug names and signature information(e.g., dose, route, and frequency) with F-measures of >0.9 from clinical text. We then filtered medications by those with a specified dose, route, frequency, or strength, a rule that we have found improves the likelihood of finding medications patients prescribed to patients. [28] We collapsed mapped specific medications to their generic ingredients using RxNorm relationships.[29] All ingredients were thus represented by the RxCUI of their generic ingredient. For example, *Tylenol Caplet, 325 mg oral tablet* (RxCUI 209387) was mapped to *Acetaminophen* (RxCUI 161).

We calculated the coverage of MEDI and MEDI-HPS as in equation (1), where *All Medications* represented the total number of unique medication-SD record pairs extracted from EMRs, and *Medications Covered by MEDI (or MEDI-HPS)* is the number of unique medication-SD record pairs, in which the medication can be found within the resource.

$$Coverage = \frac{Medications\ Covered\ by\ MEDI}{All\ Medications} \quad (1)$$

Indication Prevalence

We calculated indication prevalence of medications as the number of individuals exposed the medication with each indication divided by all individuals exposed to the medication with at least one MEDI-defined indication. We restricted the denominator to individuals with at least one MEDI indication since many patients may only be seen by subspecialists and thus not have a complete record of diagnoses in their billing history. In contrast, since medicine reconciliation is a required clinical process at nearly all clinics visits and inpatient admissions, comprehensive medication lists are routinely captured, even at focused outpatient subspecialists visits. Thus, for example, a visit to a sports medicine clinic may capture a medication record for metformin (diabetes) and warfarin (an anticoagulant), but the patient would only be billed for their knee injury. Our previous study results[21] suggest that it is a reasonable assumption that MEDI is a relatively complete indication dataset. We defined the prevalence of the *indication i* of *drug m* as in equation (2), where *Records(m)* is the number of unique SD records with drug *m* and any indication defined in MEDI. and *Records(m,i)* represents the number of unique SD records with *drug m* and *indication i* found within their EMRs. Indications were determined by the presence (or absence) of relevant ICD9 codes.

$$Prevalence_{drug\ m,indication\ i} = \frac{Records\ (m,i)}{Records\ (m)} \quad (2)$$

Evaluation

We randomly selected 50 subjects for each of four commonly used medications: propranolol, metformin, methotrexate, and adalimumab. The full medical records of these subjects were manually reviewed to evaluate the coverage and validate the prevalence of indications. Three practicing physicians (JCD for propranolol and metformin, TAL for methotrexate, and JDM for adalimumab) used clinical experience to determine what disease they were prescribed the medication for. Uncertain results were resolved by a discussion.

Result

Medication Coverage

From the data of 1,371,354 unique records, we retrieved a total number of 15,532,057 unique medication-record pairs. MEDI covered 97.25% of them (93.80%+3.45%, summarized in Table 2). The majority of pairs (93.80%) were also found in MEDI-HPS.

Table lists the 20 most frequently prescribed medications at VUMC. The top three medications were acetaminophen (31.33%), aspirin (18.95%), and ibuprofen (17.53%). This finding is consistent with our clinical experience. All were covered by both MEDI and MEDI-HPS.

Table 2. The summary of MEDI coverage

In MEDI	In MEDI-HPS	Unique Medication Patient Pairs	Coverage
No	No	426,662	2.75%
Yes	No	536,619	3.45%
Yes	Yes	14,568,776	93.80%
Sum		15,532,057	100.00%

Table 3. The most commonly used Medications at VUMC.

RXCUI	Patients (Percentage)	MEDI	MEDI-HPS	Drug Description
161	429,636 (31.33%)	Yes	Yes	acetaminophen
1191	259,932 (18.95%)	Yes	Yes	aspirin
5640	240,386 (17.53%)	Yes	Yes	ibuprofen
6387	220,089 (16.05%)	Yes	Yes	lidocaine
7052	191,837 (13.99%)	Yes	Yes	morphine
26225	158,149 (11.53%)	Yes	Yes	ondansetron
4850	155,993 (11.38%)	Yes	Yes	glucose
8745	153,414 (11.19%)	Yes	Yes	promethazine
3498	147,982 (10.79%)	Yes	Yes	diphenhydramine
4278	144,316 (10.52%)	Yes	Yes	famotidine
435	138,823 (10.12%)	Yes	Yes	albuterol
82003	137,551 (10.03%)	Yes	Yes	dioctyl sulfosuccinate
723	134,393 (9.80%)	Yes	Yes	amoxicillin
9863	131,218 (9.57%)	Yes	Yes	sodium chloride nasal
4337	129,167 (9.42%)	Yes	Yes	fentanyl
6960	128,431 (9.37%)	Yes	Yes	midazolam
4603	117,234 (8.55%)	Yes	Yes	furosemide
5224	114,825 (8.37%)	Yes	Yes	heparin
29046	112,539 (8.21%)	Yes	Yes	lisinopril
36567	106,580 (7.77%)	Yes	Yes	simvastatin

A total of 199 unique medications extracted from EMRs were not covered by MEDI. Table 4 showed the most frequently prescribed medications not in MEDI. Most of them were vaccines (e.g. “streptococcus pneumoniae serotype vaccine” and “poliovirus vaccine inactivated, type 1”), probiotics (e.g. “bifidobacterium infantis”), nutrition (e.g. “calcium ascorbate” and “aspartate”), and other ingredients that are included in a product to improve drug absorption or help eliminate active ingredient from body (e.g. inert ingredients). No common prescription drugs were observed in review of the 199 missing medications. A small portion of medication data from EMRs (3.45%) were covered by MEDI but not MEDI-HPS (Table 5). Among these were over-the-counter medications like vitamins. However, other identified medications may have been false positive detections by MedEx. For instance, “lipase” and “amylase” likely represent laboratory tests and not medications prescribed to the patient.

Table 4. Frequent medications not found in MEDI.

RXCUI	Patients (Percentage)	Drug Description
100213	31,272 (2.28%)	bifidobacterium infantis
798302	10,280 (0.75%)	pertussis, acellular
798232	9,142 (0.67%)	streptococcus pneumoniae serotype vaccine
854960	8,681 (0.63%)	pneumococcal capsular polysaccharide type 33f vaccine
763096	6,725 (0.49%)	poliovirus vaccine inactivated, type 1 (mahoney)
56476	6,412 (0.47%)	sodium gluconate
798264	4,976 (0.36%)	l1 protein, human papillomavirus type 16 vaccine
748794	1,273 (0.09%)	inert ingredients
804179	1,172 (0.09%)	measles virus vaccine live, enders' attenuated edmonston strain
802755	1,143 (0.08%)	liver/stomach concentrate
142407	1,056 (0.08%)	calcium ascorbate
644718	1,041 (0.08%)	ferrous asparto glycinate
42543	640 (0.05%)	aspartate
38624	588 (0.04%)	triethanolamine polypeptide oleate condensate
215831	400 (0.03%)	calcium citrate with vitamin d
353110	249 (0.02%)	fibrinolysis inhibitor
287575	249 (0.02%)	tissel
825006	249 (0.02%)	thrombin(human plasma der)
273888	249 (0.02%)	calcium threonate
32485	224 (0.02%)	orabase
221113	137 (0.01%)	levorotatory alkaloids of belladonna

Table 5. Medications that can be found in MEDI but not MEDI-HPS.

RXCUI	Patients (Percentage)	Drug Description
11253	98,989 (7.22%)	vitamin d
48203	68,870 (5.02%)	clavulanate
6033	45,516 (3.32%)	isoleucine
6406	21,312 (1.55%)	lipase
11251	18,662 (1.36%)	vitamin b
19861	14,982 (1.09%)	butamben
1244014	13,517 (0.99%)	vitamin d3
743	10,946 (0.80%)	amylase
71535	10,859 (0.79%)	vecuronium
19711	9,267 (0.68%)	co-amoxiclav
3024	3,786 (0.28%)	cysteine
8031	3,716 (0.27%)	protease

Indication Prevalence

For the second specific aim of this study, we calculated the prevalence information for each indication, which is now available in MEDI for download. For example, warfarin (RxCUI 11289) had 13 indications documented in MEDI, including “atrial fibrillation” and “unspecified cardiac arrhythmia.” The calculated prevalence of warfarin for atrial fibrillation was 0.69. The prevalence of each indication can also be treated as a positive predictive value of a given patient having a given disease when prescribed the medication. For example, when a patient is prescribed warfarin, there is 69% chance that the patient has also been diagnosed with atrial fibrillation. For some medications, the sum of prevalences exceeds 1 because an individual may have two or more indications for a medication (e.g., consider a beta blocker in a patient with coronary artery disease and hypertension). This situation was common in medications treating for complex chronic diseases. For example, a considerable number of metformin users may be diagnosed with both type 2 diabetic mellitus (T2DM) and obesity (Table 6). Even though, our prevalence data clearly suggest that the primary indication of metformin is T2DM rather than other diseases.

Table 6. The prevalence of warfarin indications.

RXCUI	Medication	ICD9 Code	Disease	Prevalence
11289	warfarin	427.31	Atrial fibrillation	0.69
11289	warfarin	427.9	Cardiac arrhythmia NOS	0.25
11289	warfarin	415.19	Pulmonary embolism and infarction	0.14
11289	warfarin	435.9	Transient cerebral ischemia	0.09
11289	warfarin	434.91	Ischemic stroke	0.08
11289	warfarin	453.9	Embolism and thrombosis of unspecified site	0.07
11289	warfarin	411.2	Myocardial infarction	0.04
11289	warfarin	444.9	Arterial embolism and thrombosis	0.03
11289	warfarin	459.9	Circulatory disease NEC	0.02
11289	warfarin	451.9	Phlebitis and thrombophlebitis	0.02
11289	warfarin	964.1	Anticoagulants causing adverse effects	0.00
11289	warfarin	288	Diseases of white blood cells	0.00
11289	warfarin	279.2	Autoimmune disease NEC	0.00

Table 6. The prevalence of some metformin indications.

RXCUI	Medication	ICD9 Code	Disease	Prevalence
6809	metformin	250.00	Type 2 diabetes mellitus	0.80
6809	metformin	798.79	Malaise and fatigue	0.27
6809	metformin	278.00	Obesity	0.21
6809	metformin	250.01	Type 1 diabetes	0.19
6809	metformin	790.6	Other abnormal blood chemistry	0.11
6809	metformin	256.4	Polycystic ovaries	0.08

Manual Evaluation of Indication Prevalence

We manually reviewed 50 subjects for each medication. Physician review found evidence of at least one indication for all but 17/200 subjects. These 17 subjects had medication lists available only. Therefore, we were not able to retrieve any indication for these subjects. The result from our manual chart review confirmed that the MEDI covered the most observed indications. All indications of adalimumab, metformin, and propranolol found within EMRs were covered by MEDI. For methotrexate, MEDI covered the most common ones, e.g. rheumatoid arthritis, breast cancer, and leukemias/lymphomas. A few indications of methotrexate were missed by MEDI, among them: polymyalgia rheumatic, collagen vascular disorder, and uveitis.

Table 7. Summary of Results from Manually Physician Review.

Medication	Subjects	Indication in MEDI	Indication Not in MEDI	No Evidence of Indication
adalimumab	50	48	0	2
metformin	50	49	0	1
propranolol	50	41	0	9
methotrexate	50	40	5	5

We compared the prevalence based on the results of the manual chart review to the prevalence calculated from the SD in MEDI. Chi-square tests showed that there was no statistical difference between them for most indications, except propranolol for hypertension and metformin for obesity. The estimated prevalence of propranolol based on chart review was slightly lower than the number in MEDI (39% vs. 22%). As for metformin and obesity, 10 of the 50 subjects were also diagnosed with obesity and they were all diabetic. Thus, our calculation of prevalence, which allowed multiple indications for a medication, agrees with the chart review. However, the reviewer marked the type 2 diabetes mellitus as the true indication instead of obesity.

Table 8. Major indications of each medication and their prevalence.

Medication	Indication	Prevalence (MEDI)	Prevalence Based on Chart Review (95% CI)
metformin	type II diabetes mellitus	0.79	0.88 (0.76, 0.94)
	Obesity*	0.21	0.02 (0.00, 0.11)
adalimumab	Rheumatoid arthritis	0.46	0.40 (0.28, 0.54)
	Regional enteritis	0.27	0.24 (0.14, 0.37)
	Psoriatic arthropathy	0.16	0.12 (0.08, 0.29)
	Ulcerative colitis	0.10	0.08 (0.04, 0.21)
propranolol	Hypertension NOS*	0.39	0.22 (0.13, 0.35)
	Tachycardia NOS	0.24	0.12 (0.08, 0.29)
	Anxiety state unspecified	0.13	0.06 (0.02, 0.16)
	Palpitations	0.13	0.04 (0.01, 0.14)
methotrexate	Rheumatoid arthritis	0.44	0.32 (0.21, 0.64)

*Significantly different between the estimated prevalence and prevalence in MEDI using chi-square test.

Discussion

Medications are one of the most important markers of disease and treatment course. Although providers presumably always have a reason (indication) for every medication, medications are rarely explicitly linked to their indications within most EMRs, and research into computational resources to enable such linkage is very limited. Previously, we observed this gap and created MEDI as a computable resource to support clinical research and application development. In this paper, we demonstrated that MEDI and MEDI-HPS have high coverage on practice-based data, and that no common prescription drugs were missed by MEDI. We also calculated the prevalence of each indication for each medication. Such prevalence information is previously unavailable in other public resources. Its availability may improve the clinical usability of MEDI. Manual physician review of random patient records with four example medications further demonstrated that the prevalence estimates are grossly valid.

Several limitations should be noted. First, we only calculated unique mediation-record pairs and no prescription frequencies of an individual patient were used. Therefore, we cannot accurately obtain the usage frequency of a medication. It is possible that some medications were not prescribed to an individual. Second, we did not consider the temporal relationship between medications and diagnoses. In other words, a prescription may occur long before a diagnosis. Taking into account temporal relationships may improve prevalence estimates. However, since medications may be recorded without a given diagnosis, it is also possible that such a restriction could harm performance. Third, we counted each MEDI indication as possible even if another indication was also present. As with the metformin-obesity example (which was a specific indication for one of the 50 reviewed charts), this can lead to overestimates for certain drug-indication pairs.

Conclusion

In summary, this paper continued our work on developing an important medication indication resource—MEDI. By leveraging the complete de-identified EMR data from VUMC—one of the largest patient referral center for the Mid-South, we demonstrated that the excellence coverage of MEDI on medications in EMRs. In addition, we improved the usability of MEDI by adding prevalence for each indication. This work further improved the data quality of MEDI and will facilitate future clinical and pharmacological research.

Funding

This study was supported by NIH/NLM "From GWAS to PheWAS: Scanning the EMR Phenome for Gene-Disease Associations" (R01 LM 010685-01A1).

Acknowledgement

We would like to give special thanks to Dr. Thomas Lasko for having helped with chart review. We also thank him for a friendly and very fruitful discussion.

References

1. Cebul RD, Love TE, Jain AK, Hebert CJ. Electronic health records and quality of diabetes care. *The New England journal of medicine.* Sep 1 2011;365(9):825-833.

2. Ghitza UE, Sparenborg S, Tai B. Improving drug abuse treatment delivery through adoption of harmonized electronic health record systems. *Substance abuse and rehabilitation.* Jul 1 2011;2011(2):125-131.

3. Roth CP, Lim YW, Pevnick JM, Asch SM, McGlynn EA. The challenge of measuring quality of care from the electronic health record. *American journal of medical quality : the official journal of the American College of Medical Quality.* Sep-Oct 2009;24(5):385-394.

4. Roth MT, Weinberger M, Campbell WH. Measuring the quality of medication use in older adults. *Journal of the American Geriatrics Society.* Jun 2009;57(6):1096-1102.

5. Tracy RP. 'Deep phenotyping': characterizing populations in the era of genomics and systems biology. *Curr Opin Lipidol.* Apr 2008;19(2):151-157.

6. Wilke RA, Xu H, Denny JC, et al. The emerging role of electronic medical records in pharmacogenomics. *Clinical pharmacology and therapeutics.* Mar 2011;89(3):379-386.

7. Denny JC. Chapter 13: Mining electronic health records in the genomics era. *PLoS computational biology.* Dec 2012;8(12):e1002823.

8. Carroll RJ, Thompson WK, Eyler AE, et al. Portability of an algorithm to identify rheumatoid arthritis in electronic health records. *Journal of the American Medical Informatics Association : JAMIA.* Jun 1;19(e1):e162-e169.

9. 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. *Journal of the American Medical Informatics Association : JAMIA.* Mar-Apr;19(2):212-218.

10. 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.* Apr 9 2010;86(4):560-572.

11. 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. *Journal of the American Medical Informatics Association: JAMIA.* 2012;19(2):219-224.

12. Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. *Journal of the American Medical Informatics Association : JAMIA.* Sep 6.

13. Robinson PN. Deep phenotyping for precision medicine. *Hum Mutat.* May;33(5):777-780.

14. Delaney JT, Ramirez AH, Bowton E, et al. Predicting clopidogrel response using DNA samples linked to an electronic health record. *Clinical pharmacology and therapeutics.* Feb 2012;91(2):257-263.

15. Wei WQ, Leibson CL, Ransom JE, Kho AN, Chute CG. The absence of longitudinal data limits the accuracy of high-throughput clinical phenotyping for identifying type 2 diabetes mellitus subjects. *Int J Med Inform.* Jul 2.

16. 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.* Nov 16 2010;122(20):2016-2021.

17. 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.* Oct 7;89(4):529-542.

18. 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 Medical Genomics.* 2011;4:13-13.

19. Liu M, Wu Y, Chen Y, et al. Large-scale prediction of adverse drug reactions using chemical, biological, and phenotypic properties of drugs. *Journal of the American Medical Informatics Association : JAMIA.* Jun 1 2012;19(e1):e28-e35.

20. Burton MM, Simonaitis L, Schadow G. Medication and indication linkage: A practical therapy for the problem list? *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium.* 2008:86-90.

21. Wei WQ, Cronin RM, Xu H, Lasko TA, Bastarache L, Denny JC. Development and evaluation of an ensemble resource linking medications to their indications. *In revision.* 2013.

22. RxNorm. https://www.nlm.nih.gov/research/umls/rxnorm/. Accessed 02/01/2013.

23. Kuhn M, Campillos M, Letunic I, Jensen LJ, Bork P. A side effect resource to capture phenotypic effects of drugs. *Molecular systems biology.* 2010;6:343.

24. MedlinePlus. http://www.nlm.nih.gov/medlineplus/.

25. Wikipedia. http://en.wikipedia.org. Accessed 02/01/2013.

26. Roden DM, Pulley JM, Basford MA, et al. Development of a large-scale de-identified DNA biobank to enable personalized medicine. *Clinical pharmacology and therapeutics.* 2008;84(3):362-369.

27. Xu H, Stenner SP, Doan S, Johnson KB, Waitman LR, Denny JC. MedEx: a medication information extraction system for clinical narratives. *Journal of the American Medical Informatics Association : JAMIA.* Jan-Feb 2010;17(1):19-24.
28. Schildcrout JS, Denny JC, Bowton E, et al. Optimizing drug outcomes through pharmacogenetics: a case for preemptive genotyping. *Clinical pharmacology and therapeutics.* 2012;92(2):235-242.
29. Nelson SJ, Zeng K, Kilbourne J, Powell T, Moore R. Normalized names for clinical drugs: RxNorm at 6 years. *Journal of the American Medical Informatics Association : JAMIA.* Jul-Aug 2011;18(4):441-448.