Developing a Data-driven Medication Indication Knowledge Base using a Large Scale Medical Claims Database

Ying Li Ph.D.[1], **Cao Xiao Ph.D.**[2]
[1]**IBM T. J. Watson Research Center, Yorktown Heights, NY, USA**
[2]**AI for Healthcare, IBM Research, Cambridge, MA, USA**

Abstract

Medication-indication knowledge base (KB) is useful for clinical care and also a key enabler for secondary use of observational health data. Over the years there are several indication KBs being developed, however, they were built based on curated data sources and thus may not reflect actual clinical practice. The longitudinal observational health data contain information about real world practice of medication indication, but were rarely used in KB construction. A major challenge of leveraging them is the confounders in multi-medication multi-diagnoses relations. In this study, we proposed a sampling based approach that could explicitly handle the aforementioned confounders, and consequently detect more accurate medication-indication relations. Based on this method, we created a medication-indication KB that reflects actual clinical practice and has broad medication and indication coverages. Our work represents the first attempt to develop a medication-indication KB from a large scale observational health data in an automated and unsupervised manner.

Introduction

The knowledge of medication-indication plays a critical role in clinical care and also serves as a key enabler for the secondary use of observational health data (OHD) for clinical and pharmaceutical research[1,2]. For example, the cohort design has been advocated as the primary design for drug safety studies and comparative effectiveness analysis using OHD[3,4]. It has been the only methodological approach executed within the pilot projects of the Sentinel Initiative[5]. A critical step for the cohort design is to identify the appropriate reference group that can be used to compare with patients on the target drug of interest. Selecting patients who are on active comparator drugs (e.g., drugs that share the same indications as the target drug) is an effective design strategy.

A medication's indication refers to the use of that medication for treating or preventing a particular sign, symptom, disease or condition. Each medication has various number of known indications, which can be further classified as either on-label or off-label[6]. On-label indications are those that are formally recognized by manufacturers and approved by the Food and Drug Administration (FDA). In contrast, off-label indications generally lack of FDA approval although some off-label practices are supported by sufficient evidence[7]. Multiple disparate knowledge sources have provided complementary medication related knowledge including drug formulation, dosage, indication, side effect, etc[8–10]. Among them, the publicly available resources such as DailyMed[11], MedLinePlus[12] and Wikipedia[13] contain limit knowledge of indication. For example, DailyMed only shows FDA approved indications. Thus, how to automatically extract accurate and comprehensive drug indication knowledge from real world health data becomes a meaningful task.

Learning from multiple data sources is a preferred strategy for generating more accurate and comprehensive drug indication knowledge. In [8], the authors integrated the National Drug File - Reference Terminology (NDF-RT)[14], FDA's adverse event reporting system (FAERS)[15] and Semantic Medline (SemMed)[16], and demonstrated comparable performance as the manually curated gold standard in terms of linking several medications to their indications. In another study[17,18], the authors combined RxNorm[19], SIDER2[20], MedlinePlus[12] and Wikipedia[13] to form a computable medication indication resource (MEDI), and the estimated precision and recall for a single resource and different combinations of resources are ranging at $(20\%, 51\%)$ and $(56\%, 94\%)$ respectively. Additionally, they identified a subset of MEDI with the precision of over 92% based on physician review resulting in a MEDI high precision subset (MEDI-HPS). Although the initial success showed improved performance, simple knowledge combination does not fully reflect the actual physician usage and medical coders' decision making in real health care practice due to practitioners' varying training and experience. For example, the MEDI regards disease *essential hypertension* or *hypertension* (coded as 401 and 401.9 respectively) as two indications for the medication *atenolol* while it is common that *benign essential hypertension* (coded as 401.1, which is not in MEDI for any medication) is also considered as

a valid indication for *atenolol* in the clinical practice and medical reimbursement. Thus, these integrated knowledge bases are not adapted enough to be seamlessly integrated into the EHR system and medical claims database.

To solve such a challenge, observational health data, including the routinely-collected longitudinal electronic health records (EHR) and administrative claims data, bring valuable opportunities to develop various medication related knowledge in an automated and comprehensive manner. To accomplish this, it is ideal that a clinical record can explicitly link the medications in patients' prescriptions to the conditions for which they are prescribed. However, in general such linkage is often not available. Therefore, the computational methods are needed to obtain medication-indication association. It is known that confounding is one of the major challenges to acquire accurate associations when using the vast amount of real world data[21]. In this study, the confounding effect is mainly referred to as comorbid confounders such as a spurious medication-disease association is due to the disease that is a common co-morbidity or symptom of approved indication for the target medication.

Several previous studies proposed to develop the medication-disease knowledge using computational methods. For example, Jung *et al*.[22] applied support vector machine (SVM) to classify whether diseases are indications for particular medications. They extracted features from EHR including co-occurrence for a potential medication indication pair along with many associations based on the two by two contingency table. Other features include medication similarity based on prior knowledge. Chen *et al*.[23] applied text mining technique to identify relationship between disease and drug entities from discharge summaries. In particular, they used chi square statistics along with p-value correction to measure the strength of disease-drug associations. Their results indicated the feasibility of automatically acquiring disease-drug knowledge from EHR data. These tabulation (e.g., contingency table) based methods have been used in numerous studies including detecting prescription patterns, adverse drug reactions (ADRs) and medication off-label use[23–25]. However, they suffer greatly from the confounding effect which could lead to many spurious associations[26]. To alleviate confounding effects, Backenroth *et al*[26] introduced regression methods to identify medications associated with a condition while taking other confounding conditions and medications into account in order to remove false positive associations. These regression methods together with study designs (e.g. case control design or cohort design) are commonly used in studies of detecting ADRs using observational healthcare data, and have shown the promising results[21,27]. However, they usually require human experts to draw up a list of health outcomes of interest, either indications or ADRs, and specify them using diagnosis codes, laboratory values and medication mentions. Thus, these methods are not scalable. Moreover, both tabulation methods and regression based methods could only determine medication disease relation on a population level and none of the above methods is capable of linking medication to its indication in a clinical record.

In this paper, we proposed a sampling based approach to generate more accurate indications for each medication through iteratively down-weighting their associations with irrelevant diagnosis codes in each clinical record. Our method is automatic and scalable. Experimental results showed the proposed method could learn an accurate and comprehensive medication indication knowledge base based on OHD, and link the medication to its indication in each clinical record. In our previous work[28], this method was used in an ADR signal detection task and demonstrated better performance than the disproportionation analysis.

Methods and Materials

Data Description

The study was performed using a cohort extracted from Truven MarketScan Commercial Claims and Encounter (CCAE) database[29]. It comprised of de-identified patient medical claims from 2012 with detailed time-stamped records of patient-level clinical events, including drug prescriptions and refills, disease diagnosis, procedures and other meta-information. Since the available claims database only involves outpatient pharmaceutical claims information, we limited our study to outpatient population. We standardized the diagnosis codes using the International Classification of Disease, Ninth Revision, Clinical Modifications (ICD-9-CM) classification system. We mapped NDCs to their generic names using Redbook[30], and further mapped generic names to RxCUI using RxNorm[19]. Each clinical record consists of diagnosis codes recorded in an outpatient visit, and the associated pharmaceutical claims in the same visit for a particular patient. The rational is that if a patient has an outpatient visit and is prescribed with medications, then the prescription will probably be sent to the pharmacy and filled at the same date. We excluded those outpatient visits

that have diagnosis codes but without relevant prescription.

Methodology

To identify the most relevant indications for each drug, we formulate the task as a multi-instance multi-label problem, and modify our previous signal detection approach proposed in [28] to identify the most relevant indication for the medication. A single patient visit can generate several diagnoses and related medications. Some of the diagnoses directly correspond to the indications for which the medications are prescribed, while the rest could be just some co-morbidities and conditions that do not directly associate with the medications being prescribed. This assumption consequently leads to an improved medication-indication co-occurrence count used in indication detection, and thus yields more accurate results. Improved medication-indication counts are obtained by the Monte-Carlo sampling procedure, where for each medication, diagnoses are sampled with a probability proportional to its disproportionality score, which effectively gives more emphasis on the diagnosis with higher probability to be the indication of the corresponding medication. After the enhanced counts are obtained, we can compute final disproportionality scores for each medication-indication pair.

Transfer the Multi-item Gamma Poisson Shrinker (MGPS) Method to Indication Detection

In this study we modified the multi-item Gamma Poisson Shrinker (MGPS)[31] method to detect medication indication under a multi-instance multi-label setting. The MGPS method[31] is a leading disproportionality analysis method originally proposed to detect potential adverse drug reaction (ADR) signals from the spontaneous reporting systems (SRS). The SRS generally has the following structure: each dataset consists of numerous records and each record contains a set of ADR instances along with drugs that are suspected to cause the ADRs, which essentially describes a very similar setting as medication indication detection from claims data. In the medication indication detection setting, the longitudinal claims data can be decomposed into many records, each of which contains a set of prescribed medications along with diagnoses that could be the indications of the medications. Due to such a similarity, it is reasonable to transfer the MGPS method to the drug indication detection task. Below we provide detailed model description.

Our method focuses on low-dimensional projections of the data, specifically 2-dimensional contingency tables. In particular, we let $n_{00}(i,j)$ denote the N_{ij} entry for the number of events regarding the ith indication and the jth medication. Assume each observation of N_{ij} is drawn from a Poisson distribution with an unknown mean μ_{ij} , then the theoretical value of relative association between the ith indication and the jth medication, λ_{ij}, can be computed using Eq. 1.

$$\lambda_{ij} = \frac{\mu_{ij}}{E_{ij}} = \frac{\mu_{ij}}{[n_{0+}(i,j)n_{+1}(i,j)/n_{++}(i,j)]} \tag{1}$$

where $n_{0+}(i,j) = n_{00}(i,j) + n_{01}(i,j)$, $n_{+1}(i,j) = n_{01}(i,j) + n_{11}(i,j)$, and $n_{++}(i,j) = n_{00}(i,j) + n_{11}(i,j) + n_{10}(i,j) + n_{01}(i,j)$. Here we regard the geometric mean of the posterior distribution for each λ_{ij} as the MGPS scores. And λ_{ij} is assumed to arise from a particular 5-parameter prior distribution, namely a mixture of two gamma distributions as given in Eq. 2.

$$\lambda_{ij} \sim wGa(\alpha_1, \beta_1) + (1-w)Ga(\alpha_2, \beta_2) \tag{2}$$

where Ga indicates the Gamma distribution, and α_1, β_1 and α_2, β_2 are their hyperparameters.

Then posterior distribution of λ_{ij} is iteratively fitted based on the observations of data under the Bayesian framework as is given by Eq. 3.

$$\lambda_{ij}|N_{ij} = n_{00} \sim wGa(\alpha_1 + n_{00}, \beta_1 + E_{ij}) + (1-w)Ga(\alpha_2 + n_{00}, \beta_2 + E_{ij}) \tag{3}$$

Based on the posterior, the criteria for generating a signal of medication-indication pair by their posterior expectation of $\log_2(\lambda_{ij})$ can be expressed by the following Eq. 4, which is similar as the one initially proposed by DuMouchel[31] in ranking ADR signals.

$$E[\log_2(\lambda_{ij}|N_{ij}=n_{00})] = \frac{Q_n[\psi(\alpha_1+n_{00})-\ln(\beta_1+E_{ij})]+(1-Q_n)[\psi(\alpha_2+n_{00})-\ln(\beta_2+E_{ij})]}{\ln 2} \tag{4}$$

where ψ is the diagamma function and Q_n can be computed from Eq. 5 and Eq. 6.

$$Q_n = \frac{wf(n_{00};\alpha_1,\beta_1,E_{ij})}{wf(n_{00};\alpha_1,\beta_1,E_{ij})+(1-w)f(n_{00};\alpha_2,\beta_2,E_{ij})} \tag{5}$$

$$f(n_{00};\alpha_1,\beta_1,E_{ij}) = (1+\beta/E)^{-n_{00}}(1+E/\beta)^{-\alpha}\frac{\Gamma(\alpha+n_{00})}{\Gamma(\alpha)n_{00}}. \tag{6}$$

To conclude, the adoption of MGPS method not only provides a shrinkage estimate of relative probability of association to address the sampling variance issue, but also works efficiently on large-scale data. It is considered an important improvement, and thus becomes the most widely used methods that is in routine use by regulators (e.g. FDA) and pharmaceutical manufacturers worldwide. However, the confounders induced by co-occurring indications still remain and would cause inaccurate detections. In the following, we introduce an Monte Carlo sampling step for alleviating this issue.

Monte Carlo Sampling for More Accurate Indication Detection

To alleviate the confounding effects induced by co-occurring diseases, we borrow the idea of discrete choice models[32] and try to filter out these confounding diseases by assuming for each clinical record, each drug has at most one major associated indication. Although some drugs could have multiple indications, previous study confirmed that in most case a drug is prescribed due to a main indication[6]. Therefore, for most cases the "one major indication" assumption holds, though this procedure could be extended to multiple indications. Based on such assumption, we propose the following Monte Carlo Expectation Maximization (MCEM) procedure.

Here is a practical guide for the MCEM procedure. In an MCEM procedure, the maximizer of the posterior probability is approximated with sampled data in the E-step and the value of the maximizer is optimized in the M-step. In our case, we first compute MGPS scores for all indication-drug pairs across all reports using the whole dataset. Then, for each drug in each report, we normalize the MGPS scores across indications in order to obtain these indications' contribution ratio proportional to their MGPS scores related to the prescription of the drug. And these contribution ratio will be used as the sampling probabilities and we will draw from multinomial distribution to assign the major indication to the target drug in the next step. In the next step, we perform iterative Monte Carlo sampling to sample one indication based on the aforementioned probabilities. In each iteration, the sampled indication is added to the report saved throughout these iterations and MGPS scores are re-calculated for the current states over all reports . In the next iteration, updated MGPS scores for all indication-drug pairs are used and we iterate the process until the difference between the optimal values of the maximizer in consecutive iterations is less than a heuristic threshold (e.g. 10^{-3}, 10^{-5}). Last we compute final MGPS scores for all indication-drug pairs. In summary, the Monte Carlo sampling procedure assigns the major indication for each drug ranked by their contribution to the risk score. After iterations, the procedure will down-weight irrelevant causes (e.g. comorbid diseases) for the prescribing drug and thus generate the improved count of co-occurrence of an indication and drug pair.

To qualify a medication-indication signal, the score such as MCEM EB05 must be larger than a pre-specified cutoff. We used the cutoff of 2, which is commonly suggested in the pharmacovigilance practice[33,34], to generate the overall knowledge base.

Evaluation

Reference Standard: To perform appropriate evaluation of the proposed system, we used reference standards consisting of a set of positive controls, mediation-indication pairs known as true treatment relationships, and a set of negative controls consisting of medication-disease pairs that are highly unlikely to be associated. Based on MEDI and

MEDI-HPS [17] we developed two reference standards respectively. Following common practice in prediction task, we constructed a reference standard using all medication-indication pairs in MEDI-HPS as positive controls. The set of negative controls was created by pairing a medication that appears in the MEDI-HPS with an indication that is also the MEDI-HPS, and then removing each of the pairs that is in the set of positive controls. We applied the same procedure using MEDI. Finally, we acquired two reference standards involving MEDI-HPS reference standard (MEDI-HPS RS) and MEDI reference standard (MEDI RS). Note that negative controls lack scientific support in both reference standards, and they might actually be positive controls.

Baseline Methods: Disproportionality analysis (DPA) methods are primary analytic methods for identifying adverse drug reaction (ADR) signals from spontaneous report systems (SRSs). These are also applied to OHD as a potential analytical method for both identifying ADR signals and assessing medication-indication association. Typically, the lower 5th percentile of DPA output is used for extra conservatism. Thus, we adopted this norm and compared the the lower 5th percentile of scores generated by MECM method (MCEM EB05) with those scores generated by the DPA methods include the multi-item Gamma-Poisson shrinker (MGPS EB05), proportional reporting ratios (PRR05) and reporting odds ratios (ROR05) against the above two reference standards.

Metrics: The primary measures are the area under the receiver operating characteristic curve (ROC-AUC), and the area under the precision-recall curve (PR-AUC). PR curve is a better measure when the number of negative controls greatly exceeds the number of positive controls in the reference standard[35].

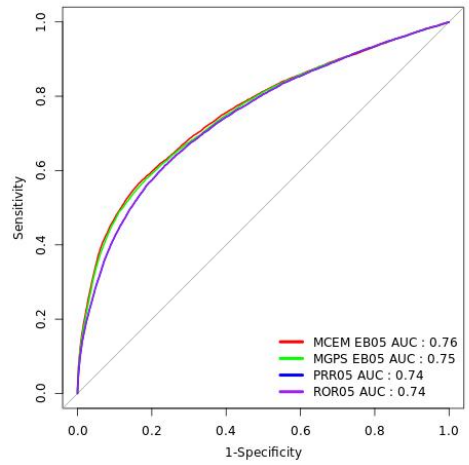
Results

Data characteristics: We acquired $\sim$ 54 million clinical records involving $\sim$ 21 million unique patients, $13,346$ unique ICD9 codes and $2,142$ unique medications. In total, the dataset includes 2.2 million unique medication disease pairs. The number of clinical records mentioning a particular co-occurrent medication disease pair ranges from 1 to $\sim$ 1 million. Among $2,142$ unique medications in the data, $1,637$ of them were mapped to a RxCUI and $1,370$ of them were covered by MEDI. The most frequent prescribed medications not in MEDI include *amphetamine salt combination* and *nitrofurantoin monohydrate/nitrofurantoin, macro*. When intersecting with this dataset, MEDI RS resulted in $14,005$ positive controls and $460,309$ negative controls, and MEDI-HPS RS resulted in $4,528$ positive controls and $274,110$ negative controls.

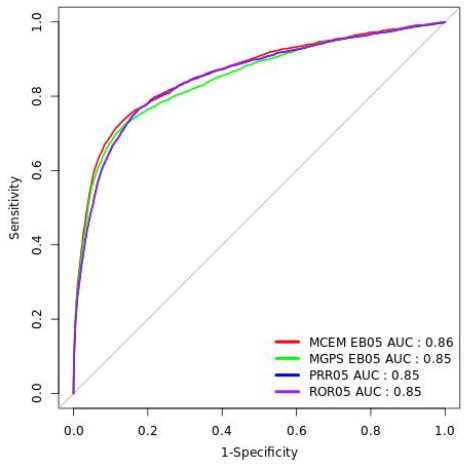
Quantitative evaluation: Figure 1 (a) and (b) show the resulting ROC curves for the MCEM and comparison methods. When evaluating against MEDI RS, the ROC AUC ranges from an AUC of 0.74 using the PRR05 and ROR05 to an AUC of 0.76 using the MCEM EB05. When evaluating against MEDI-HPS RS, the ROC AUCs are 0.85 and 0.86 for baseline methods and the MCEM respectively. The proposed MCEM shows the 1% significant ROC AUC gain compared with the baseline methods using one sided DeLong's test[36]. Figure 1 (c) and (d) show the resulting PR curves. When evaluating against MEDI RS, the PR AUC ranges from an AUC of 0.11 using PRR05 and ROR05 to an AUC of 0.17 using the MCEM EB05. When evaluating against MEDI-HPS RS, the worst AUC is 0.17 using PRR05 and the best AUC is 0.24 using MCEM EB05. The pattern of containment in two PR curve plots between MCEM based results and the baseline methods based results implies that the MCEM method provides greater precision across all levels of recall. In general, Bayesian shrinkage methods including MGPS and MCEM performs better than the frequentest based method including PRR and ROR. All methods have higher ROC AUC and PR AUC when using the MEDI-HPS RS than the MEDI RS.

Medication-indication knowledge base: After establishing the effectiveness of the approach as described above, we proceeded to generate the medication-indication knowledge base. By using cutoff value of 2, we developed a knowledge base including all medication-indication pairs whose MCEM EB05$>$2. Currently, this knowledge base contains a total of $1,820$ medications and $132,722$ medication-indication pairs. The number of indication for each medication ranges from 1 to 2,319. The precisions of this knowledge base are 0.14 and 0.12 based on MEDI RS and MEDI-HPS RS respectively.

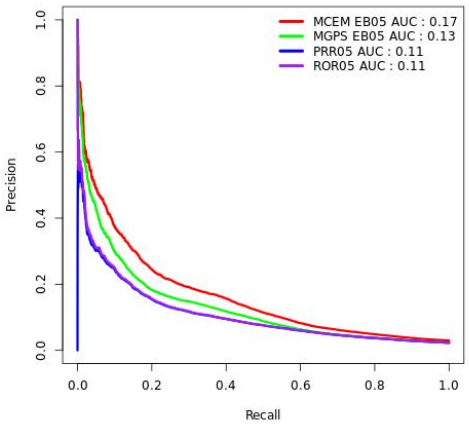
On average, each medication is related to 73 ICD9 coded indications with a standard error of 136. The medium number of ICD9 coded indications is 23. Table 1 shows an example of medication and its top 10 most popular indications in the knowledge base. The most popular indications are those diseases which are assigned as the indications for the medication by the MECM method on most of the clinical records. For example, among all clinical records whereas



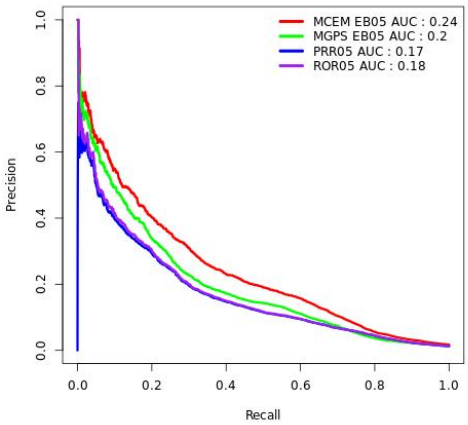
(a) ROC Curve using MEDI RS



(b) ROC Curve using MEDI-HPS RS



(c) PR Curve using MEDI RS



(d) PR Curve using MEDI-HPS RS

Figure 1: Receiver operating characteristic (ROC) curves, precision recall (PR) curves and their related AUCs for the MCEM EB05, MGPS EB05, PRR05 and ROR05 scores.

metformin was prescribed, 31% of them regarded *diabetes mellitus contolled, 250.00* as the reason for prescribing *metformin*. Three of top 10 most popular indications were validated by MEDI. Table 2. shows an example of indication and its most popularly prescribed medications. The most popular medications are those medications which are linked to the indication for most of clinical records. For instance, among all clinical records that have prescription medications for *Alzheimer's disease*, 22% of the them were prescribed with *donepezil hydrochloride*. Five of top 10 medication indication pairs were in MEDI.

We further examined the medication usage in the real world clinical practice using a subset of the knowledge base. We treated the medication-indication pairs that were in MEDI-HPS and marked as *possible label use* as on-label use cases, and the medication-indication pairs that were only in MEDI-HPS but were not regarded as *possible label use* as off-label use. This procedure resulted in a validation data set of 3,018 medication-indication pairs involving 2,128 on-label use cases and 890 off-label use cases. Figure 2 (a) shows the log-scaled score distributions could not be differentiated between on-label use (blue) and off-label use (red) in the validation dataset based on chi-square test. The range of original MCEM EB05 scores is from 2 (min) to 11,266 (max) for on-label use cases, and the range of

original EB05 scores is from 2 (min) to 7,607 (max) for off-label use cases. As demonstrated in Figure 2 (b), the log-scaled count data distributions show the relative separation between on-label use (blue) and off-label use (red) medication-indication pairs in the validation dataset using chi-square statistic. In this plot, the on-label medication-indication pairs are significantly concentrated in higher counts than lower counts. The range of original counts for on-label use is from 2 (min) to 629,236 (max), and the range of original counts for off-label use is from 2 (min) to 898,834 (max).

Table 1: Top 10 most popular indication for metformin with MCEM EB05>2

Medication	ICD9	Indication	Prevalence*	EB05	In MEDI
Metformin	250.00	Diabetes mellitus controlled	0.31	22.16	Y
Metformin	250.02	Diabetes mellitus uncontrolled	0.11	20.49	N
Metformin	272.4	Other and unspecified hyperlipidemia	0.03	2.66	N
Metformin	256.4	Polycystic ovaries	0.02	50.3	Y
Metformin	277.7	Dysmetabolic syndrome X	0.01	34.19	N
Metformin	790.29	Other abnormal glucose	0.01	17.19	N
Metformin	272.2	Mixed hyperlipidemia	0.01	3.07	N
Metformin	272.0	Pure hypercholesterolemia	0.01	2.03	N
Metformin	278.00	Obesity unspecified	0.01	4.61	Y
Metformin	790.21	Impaired fasting glucose	0.01	13.38	N

Prevalence* = # of patients on the medication indication pair/Total # of patients on the medication

Table 2: Top 10 most popular medications for Alzheimer's Disease with MCEM EB05>2

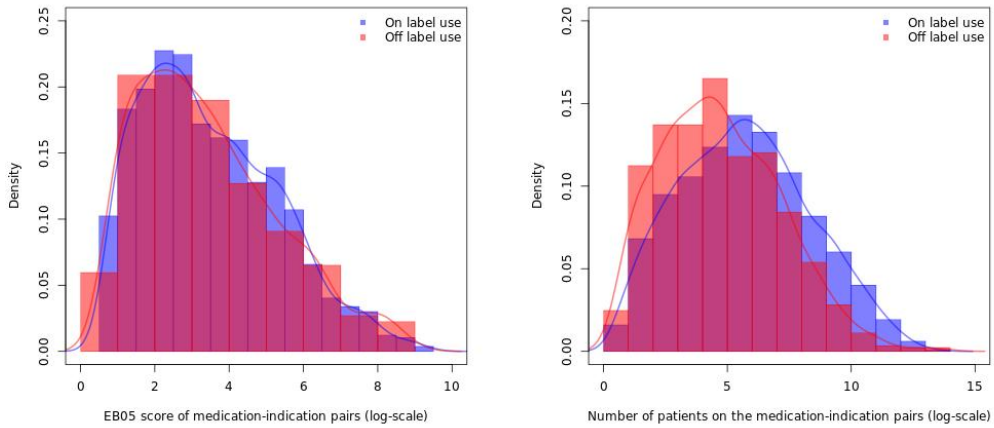
Disease	ICD9	Medication	Prevalence+	EB05	In MEDI
Alzheimer's disease	331.0	Donepezil hydrochloride	0.22	2739.50	Y
Alzheimer's disease	331.0	Memantine hydrochloride	0.13	2517.85	Y
Alzheimer's disease	331.0	Rivastigmine	0.05	3287.00	Y
Alzheimer's disease	331.0	Citalopram hydrobromide	0.03	5.31	N
Alzheimer's disease	331.0	Sertraline hydrochloride	0.03	4.61	N
Alzheimer's disease	331.0	Quetiapine fumarate	0.03	20.22	Y
Alzheimer's disease	331.0	Lorazepam	0.02	3.35	N
Alzheimer's disease	331.0	Valproic acid	0.02	18.91	Y
Alzheimer's disease	331.0	Risperidone	0.02	16.99	N
Alzheimer's disease	331.0	Galantamine hydrobromide	0.01	1468.37	Y

Prevalence+ = # of patients on the medication indication pair/Total # of patients have the ICD9 coded disease

Discussion

The aim of this study is to see the feasibility of automatically developing a medication-indication knowledge base that reflects the actual clinical uses. In particular, we developed a framework that could explicitly handle confounders induced by disease co-morbidities in order to extract improved medication-indication signals from real world patient data. We also performed extensive experiments using real world medical claims data and demonstrated the good performance of the proposed framework.

Overall, our results are promising, but also suggest area of improvement. The precisions of this knowledge base were 0.14 and 0.12 based on MEDI and MEDI-HPS respectively. This could be due to the following reasons: (1) the sub-optimal cutoff value was used in this study. For example, we observed that some of medication-indication pairs, such as *citaloparm hydrobromide - alzheimer's disease* and *metformin - other and unspecified hyperlipidemia*, which were not in MEDI had relatively low EB05 signal scores. By raising the cutoff value may improve the precision of the overall knowledge base, but the recall could be lowered. (2) confounding by co-morbidities and symptom of the indications. For example, *risperiodone - alzheimer's disease* had a high EB05 score that could be due to the reason that *risperiodone* was commonly prescribed to manage *Alzheimer's dementia-associated psychosis*. (3) the absence of



(a) (b)

Figure 2: Results of medication usage analysis based on the validation dataset

variations of ICD9 coded indications in MEDI. For example, *metformin-diabetes mellitus uncontrolled (250.02)* should be considered as a valid medication-indication pair. (4) the medication actually prescribed for the symptom with or without the presence of the disease[37]. For example, the typical treatment for both *other abnormal glucose (790.29)* and *impaired fasting glucose (790.21)* were metformin. (5) the indications were missing in the current clinical records (not shown in tables). For example, the indications for the target medications occurring in the past were filtered out using the same day matching. Most of the missing indications were chronic disease, such as hypertension, or acute disease where a recurrence should be prevented by a treatment, such as acute myocardial infarction. The absence of these indications in current clinical record could be due to the requirement for reimbursement procedure, or limited number of diagnosis codes a claims allows. The maximum number of diagnosis codes allowed for a outpatient visit is four in our dataset. The systematic categorization of the false positive associations is lacking in this study. We plan to work with medical experts to come up a more complete list of reasons for false positive associations. Searching for the optimal cutoff value is also considered in our future work.

The promising applications of this knowledge base include monitoring prescribing patterns in clinical practice, which is currently relying on the National Disease and Therapeutic Index (IQVIA, Durham, NC). This index is consisted of quarterly surveys regarding office-based physicians' recollections of patient encounters. Specifically, the surveyed physicians should provide data on patient demographics, diagnoses/reasons for patients' visits, and the therapies that are prescribed or recommended during the visit. Similar to most of survey studies, the voluntarily reported samples rely on self-selection, and the number of participating physicians is limited[38]. We believe that prescribing patterns can be learned systematically and directly from the observational health data which requires no effort from the physicians and may reflect drug usage more objectively. Therefore, our medication-indication knowledge base may serve as complementary data source to the National Disease and Therapeutic Index. Another application of this knowledge base is that it can facilitate new user cohort design in pharmacovigilance. Using this knowledge base, the medications that share the same indications with the target medication could easily be retrieved. In contrast, the current cohort identification procedure considers each case individually and relies on manual input by the researchers. Theoretically, the proposed method can be applied to detecting ADRs using OHD data, but it is a more difficult task than detecting medication indication association since the ADR outcome is rarely recorded in the structured observational health data. Moreover, the treatment association is much stronger than the association between a medication and its side effect if only consider co-occurrence information.

Claims data quality and validation are key factors in determining the accuracy of the developed knowledge base. These concerns include coding inaccuracies or bias introduced by selection of codes driven by billing incentives rather than clinical care, and coding error caused by excessive or busy workloads in clinician data entry[39]. Additionally,

determining the timing of a diagnosis from medical claims data is also challenging. For example, not every diagnosis was recorded at every visit and for each medication prescription, and therefore the absence of ICD9 code was not always evidence of absence of the disease[40].

Conclusion

Creation and updating of medication-indication knowledge that reflect current clinical practice is challenging. Therefore, it is important to fully automate this task. Our work represents the first attempt to develop medication indication knowledge base from a large scale medical claims database in an automated and unsupervised manner. Our results demonstrate its broad coverage including the entire range of drugs and indications observed in the database, and its reflection of actual clinical practice. This knowledge base may enable many research regarding secondary use of the observational healthcare data. In the future, the methodology could be generalized to other medical claims databases and electronic health records for the development of customized knowledge bases.

References

1. Galanter W, Falck S, Burns M, Laragh M, Lambert BL. Indication-based prescribing prevents wrong-patient medication errors in computerized provider order entry (CPOE). Journal of the American Medical Informatics Association. 2013;20(3):477–481.
2. Radley DC, Finkelstein SN, Stafford RS. Off-label prescribing among office-based physicians. Archives of internal medicine. 2006;166(9):1021–1026.
3. Ryan PB, Schuemie MJ, Gruber S, Zorych I, Madigan D. Empirical performance of a new user cohort method: lessons for developing a risk identification and analysis system. Drug safety. 2013;36(1):59–72.
4. Johnson ES, Bartman BA, Briesacher BA, Fleming NS, Gerhard T, Kornegay CJ, et al. The incident user design in comparative effectiveness research. Pharmacoepidemiology and drug safety. 2013;22(1):1–6.
5. Platt R. FDA's Mini-Sentinel program to evaluate the safety of marketed medical products. Food and Drug Administration. 2012;.
6. Li Y, Salmasian H, Harpaz R, Chase H, Friedman C. Determining the reasons for medication prescriptions in the EHR using knowledge and natural language processing. In: AMIA Annual Symposium Proceedings. vol. 2011. American Medical Informatics Association; 2011. p. 768.
7. Muth C, Gensichen J, Beyer M, Hutchinson A, Gerlach FM. The systematic guideline review: method, rationale, and test on chronic heart failure. BMC Health Services Research. 2009;9(1):74.
8. Wang X, Chase HS, Li J, Hripcsak G, Friedman C. Integrating heterogeneous knowledge sources to acquire executable drug-related knowledge. In: AMIA Annual Symposium Proceedings. vol. 2010. American Medical Informatics Association; 2010. p. 852.
9. Sharp M, Bodenreider O, Wacholder N. A framework for characterizing drug information sources. In: AMIA Annual Symposium Proceedings. vol. 2008. American Medical Informatics Association; 2008. p. 662.
10. Salmasian H, Tran TH, Chase HS, Friedman C. Medication-indication knowledge bases: a systematic review and critical appraisal. Journal of the American Medical Informatics Association. 2015;22(6):1261–1270.
11. DailyMed; 2018. Available from: `https://dailymed.nlm.nih.gov/dailymed/`.
12. MedlinePlus; 2018. Available from: `https://medlineplus.gov/`.
13. Wikipedia; 2018. Available from: `https://www.wikipedia.org/`.
14. National Drug File – Reference Terminology; 2015. Available from: `https://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/NDFRT/`.
15. FDA Adverse Event Reporting System; 2018. Available from: `https://fis.fda.gov/extensions/FPD-QDE-FAERS/FPD-QDE-FAERS.html`.
16. Rindflesch TC, Kilicoglu H, Fiszman M, Rosemblat G, Shin D. Semantic MEDLINE: An advanced information management application for biomedicine. Information Services & Use. 2011;31(1-2):15–21.
17. Wei WQ, Cronin RM, Xu H, Lasko TA, Bastarache L, Denny JC. Development and evaluation of an ensemble resource linking medications to their indications. Journal of the American Medical Informatics Association. 2013;20(5):954–961.

18. Wei WQ, Mosley JD, Bastarache L, Denny JC. Validation and enhancement of a computable medication indication resource (MEDI) using a large practice-based dataset. In: AMIA Annual Symposium Proceedings. vol. 2013. American Medical Informatics Association; 2013. p. 1448.

19. RxNorm;. Available from: https://www.nlm.nih.gov/research/umls/rxnorm/docs/rxnormfiles.html.

20. Kuhn M, Letunic I, Jensen LJ, Bork P. The SIDER database of drugs and side effects. Nucleic acids research. 2015;44(D1):D1075–D1079.

21. Li Y, Salmasian H, Vilar S, Chase H, Friedman C, Wei Y. A method for controlling complex confounding effects in the detection of adverse drug reactions using electronic health records. Journal of the American Medical Informatics Association. 2014;21(2):308–314.

22. Jung K, LePendu P, Chen WS, Iyer SV, Readhead B, Dudley JT, et al. Automated detection of off-label drug use. PloS one. 2014;9(2):e89324.

23. Chen ES, Hripcsak G, Xu H, Markatou M, Friedman C. Automated acquisition of disease–drug knowledge from biomedical and clinical documents: an initial study. Journal of the American Medical Informatics Association. 2008;15(1):87–98.

24. Harpaz R, Vilar S, DuMouchel W, Salmasian H, Haerian K, Shah NH, et al. Combing signals from spontaneous reports and electronic health records for detection of adverse drug reactions. Journal of the American Medical Informatics Association. 2012;20(3):413–419.

25. LePendu P, Iyer SV, Bauer-Mehren A, Harpaz R, Mortensen JM, Podchiyska T, et al. Pharmacovigilance using clinical notes. Clinical pharmacology & therapeutics. 2013;93(6):547–555.

26. Backenroth D, Chase HS, Wei Y, Friedman C. Monitoring prescribing patterns using regression and electronic health records. BMC medical informatics and decision making. 2017;17(1):175.

27. Harpaz R, Haerian K, Chase HS, Friedman C. Mining electronic health records for adverse drug effects using regression based methods. In: Proceedings of the 1st ACM International Health Informatics Symposium. ACM; 2010. p. 100–107.

28. Xiao C, Li Y, Baytas IM, Zhou J, Wang F. An MCEM Framework for Drug Safety Signal Detection and Combination from Heterogeneous Real World Evidence. Scientific reports. 2018;8(1):1806.

29. Truven Health;. Available from: https://truvenhealth.com/.

30. Redbook; 2015. Available from: http://micromedex.com/products/product-suites/clinical-knowledge/redbook.

31. DuMouchel W. Bayesian data mining in large frequency tables, with an application to the fda spontaneous reporting system. American Statistician. 1999;16(1):177–190.

32. Train K. Discrete Choice Methods with Simulation. Cambridge University Press.; 2009.

33. Szarfman A, Machado SG, O'neill RT. Use of screening algorithms and computer systems to efficiently signal higher-than-expected combinations of drugs and events in the US FDA's spontaneous reports database. Drug Safety. 2002;25(6):381–392.

34. Desphande G, Gogolak V, Smith S. Data mining in drug safety: review of published threshold criteria for defining signals of disproportionate reporting [J]. Pharm Med. 2010;24(1):37–43.

35. Davis J, Goadrich M. The relationship between Precision-Recall and ROC curves. In: Proceedings of the 23rd international conference on Machine learning. ACM; 2006. p. 233–240.

36. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC bioinformatics. 2011;12(1):77.

37. Wittich CM, Burkle CM, Lanier WL. Ten common questions (and their answers) about off-label drug use. In: Mayo Clinic Proceedings. vol. 87. Elsevier; 2012. p. 982–990.

38. Kolassa E, Bynum LA, Holmes E. Limitations and potential misinterpretation of the National Disease and Therapeutic Index. International Journal of Pharmaceutical and Healthcare Marketing. 2013;7(1):34–44.

39. Cowie MR, Blomster JI, Curtis LH, Duclaux S, Ford I, Fritz F, et al. Electronic health records to facilitate clinical research. Clinical Research in Cardiology. 2017;106(1):1–9.

40. Hersh WR, Weiner MG, Embi PJ, Logan JR, Payne PR, Bernstam EV, et al. Caveats for the use of operational electronic health record data in comparative effectiveness research. Medical care. 2013;51(8 0 3):S30.