RC25640 (WAT1611-054) November 14, 2016
Computer Science

IBM Research Report

Scoring Disease-Medication Associations Using Advanced NLP, Machine Learning, and Multiple Content Sources

Bharath Dandala, Murthy Devarakonda, Mihaela Bornea
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598 USA

Christopher Nielson
US Department of Veterans Affairs

Research Division
Almaden – Austin – Beijing – Brazil – Cambridge – Dublin – Haifa – India – Kenya – Melbourne – T.J. Watson – Tokyo – Zurich

Scoring Disease-Medication Associations using Advanced NLP, Machine Learning, and Multiple Content Sources

Bharath Dandala IBM Research
Murthy Devarakonda IBM Research
Mihaela Bornea IBM Research
Christopher Nielson US Dept. of Veterans Affairs

Abstract

Effective knowledge resources are critical for developing successful clinical decision support systems that alleviate the cognitive load on physicians in patient care. In this paper, we describe two new methods for building a knowledge resource of disease to medication associations. These methods use fundamentally different content for clinical reasoning and are based on advanced natural language processing and machine learning techniques. One method uses distributional semantics on large medical text, and the other uses data mining on a large number of patient records. The methods are evaluated using 25,379 unique disease-medication pairs extracted from 100 de-identified longitudinal patient records of a large multi-provider hospital system. We measured recall (R), precision (P), and F scores for positive and negative association prediction, along with coverage and accuracy. While individual methods performed well, a combined stacked classifier achieved the best performance, indicating the limitations and unique value of each resource and method. In predicting positive associations, the stacked combination significantly outperformed the baseline (a distant semi-supervised method on large medical text), achieving F scores of 0.75 versus 0.55 on the pairs from the patient records, and F scores of 0.69 and 0.35 on unique pairs.

1 Introduction

Electronic Health Record (EHR) systems have become invaluable repositories of patient information, but their poor design and inadequate functionality make it difficult for physicians to assimilate the vast amounts of data, reducing physician productivity and negatively impacting patient care [1] [2]. Advanced clinical decision support applications can reduce the cognitive load on physicians and improve patient care. These applications need medical knowledge for effective reasoning. One such knowledge is relationships (or more abstractly, associations) between diseases and medications. While Unified Medical Language System (UMLS) [3] semantic network contains manually curated entity relationships, it falls short in a few ways: its coverage is inadequate, the relations are binary, and it is not always clear how far to traverse in the network. An automated association scoring method that provides high coverage and accuracy is highly desirable and can be used to build a useful knowledge resource.

There are many uses for such a method and knowledge resource in clinical applications because these associations are not explicitly maintained in a typical patient record. The method (or the resource) can be used in a patient record summary to show clinicians which medications are related to a patient's active medical conditions. It can also be used in developing cohort models and for predicting disease likelihood and progression using probabilistic graphical models. In this paper, we present two new methods for scoring associations between diseases and medications, and assess their accuracy and coverage.

One of the methods is based on mining ordered medications data in millions of patient records and leveraging the temporality of events, such as disease diagnosis and medication ordering. The data mining produces statistical measures for disease and medication pairs, which are then used as features in a supervised machine learning algorithm for association scoring between a disease and a medication. A learned F1-optimized threshold is then used to classify positive and negative associations.

The second method is based on features obtained with distributional semantics on a large medical text, complemented with features from UMLS. Distributional semantics is used to obtain synonyms,

relations between words, and to develop a taxonomy of the concepts in the domain. From UMLS, semantic types and relations of the entities are retrieved as features. Once again, a supervised machine learning model provides a score for the association, and a threshold is also learned to classify positive and negative associations.

Two aspects of relations need to be considered – type and context. Given two entities, the type of relationship between them can be specific (such as "treats", "prevents", or "causes") or it can be anonymous. Further, a relationship between two given entities can be contextual in that a specific passage may entail a specific relationship between the entities, but this may or may not hold true in a larger corpus. This paper concerns itself with *anonymous* and *context independent* relationships, which we call associations.

We conducted accuracy and coverage analysis using entity pairs from 100 de-identified patient records provided to us by a large multidisciplinary hospital system. The diseases and medications of each patient were paired and these pairs formed the data set for this study. Medical experts manually labeled each unique pair in the data set. We conducted 10 x 10 cross validations, calculated standard precision, recall, F1 scores, and coverage measures, and plotted P-R curves.

Results showed that the distributional semantics method provided higher recall, the data mining approach provided higher precision and the stacked ensemble of the two methods achieved the overall best performance. Both methods outperformed a previously reported baseline method that uses manifold models and distance learning on a large medical corpus.

2 Related Work

In general, entity associations can be found in human readable form in many sources. Medical textbooks, journal papers, and web content include discourse that states or implies associations. Formal documents, such as the FDA drug labels, are more organized textual resources. Patient records themselves are another valuable source. In addition, UMLS contains relationships such as "treats" and "diagnostic-of". There is a need for automated method(s) to leverage these sources.

Many automated methods exist for relation extraction from passages in general text; a recent review [4] summarizes the research. The 2010 i2b2/VA challenge [5] included extraction of a specific set of relations from clinical notes. However, more work is needed to create a knowledge method or resource for clinical applications. One recent study [6], which we use as the baseline, successfully used manifold models and distance learning to extract seven frequent relations (defined in UMLS) from medical text with the intention of creating a knowledge resource, however, its coverage was limited.

Another relevant system and method is MEDI [7], which builds an indication to prescribable medications association resource using four public resources - RxNorm, Side Effect Resource (SIDER) 2, MedlinePlus, and Wikipedia. The resources are treated as separate voting entities in this approach, which led to the conclusion that either the highest accuracy can be achieved with limited coverage (when all resources contain the entities) or that moderate accuracy can be achieved with a higher coverage (when only fewer resources contain the entities). In contrast, the methods described here automatically "learn" optimal use of the underlying resources.

In [8] [9], association rule mining from EHR records was used to extract medication to disease relationships. The fundamental strategy in these studies was to use co-occurrence of medication orders and patient problems as a source to automatically build association rules between medications and problems. In one of the two methods studied here, such a co-occurrence of medications and problems in a patient record is extracted as one of several mined statistics. A systematic review of existing medication to indication (i.e. a symptom or a diagnosis) knowledge bases and their appraisal was presented in [10]. While an abstract appraisal is useful, here, we attempt a quantitative accuracy analysis with clinical decision support applications in mind.

3 Methods and Experiments

3.1 Distributional Relation Extraction (DRE) Method

Distributional Relation Extraction (DRE) is a supervised machine learning method for discovering associations between given entity pairs using distributional semantics and UMLS. Some of its features are derived from distributional semantics applied to a large medical corpus, and the remaining features are

derived from UMLS. Let us assume that DRE is attempting to determine the strength of the association between two arguments, disease *Hyperlipidemia* and medication *Simvastatin*. Figure 1 shows the features generated for the two arguments. Note that the feature space is sparse and high dimensional. The feature space is described below:

UMLS Type Features. The intuition is that the types of the arguments (in a taxonomy) are important constraints for association scoring, as most of the relations hold between the entities of specific types. For example, given the relation *may_treat* between two arguments, it is expected that the type of the first argument is a Medication, Chemical, Drug, etc. and the type of the second argument is a Disease, Syndrome, or Disorder. UMLS taxonomies are used to obtain one set of argument type features (the second set is described below). Since types in UMLS have multiple levels of granularity, DRE uses multi-granular features: semantic groups for coarse granularity, semantic types for medium granularity and MeSH (Medical Subject Heading) types for fine granularity. The type features are binary valued and so they have a value of 1 when present. In Figure 1, notice that for the first argument, *Simvastatin*, T1-C0003277 (with label *Cholesterol Inhibitors*) is the MSH type, ST1-T121 (with label *Pharmacological Substance*) is the semantic type and SG1-CHEM (with label *Chemicals and Drugs*) is the semantic group. We experimented using combinations of UMLS types for the arguments, but did not see significant performance improvement.

Distributional Semantics (DS) Type Features. DRE features also include types induced by distributional semantics using the text corpora for the arguments, and the distributional semantics tool used here is called JoBimText [11], which is an open source project. The JoBimText tool provides a framework for creating a distributional semantics resource from large corpora, from which we obtain relations between words, similar terms or pseudo-synonyms for a word, and a taxonomy for the domain. We built the JoBimText resource as described in [12] by preprocessing the text corpora available for our project. JoBimText uses a dependency parser adapted for the medical domain [13] for identifying syntactic relations, and the baseline relation extraction system mentioned earlier [6]. The JoBimText framework provides an API to access the resource built in this way. Unlike the UMLS types, there is only a single level of granularity for the DS types. But, each term may have multiple types. The DS type features are determined for both arguments, as shown in Figure 1. For Simvastatin, the DS type features are T1-Medication, T1-Treatment and T1-Inhibitor.

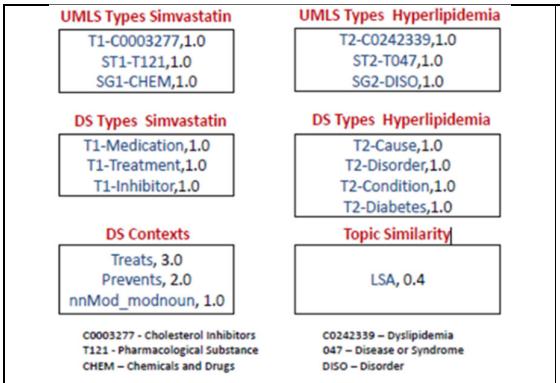

Figure 1. DRE features for Hyperlipidemia and Simvastatin.

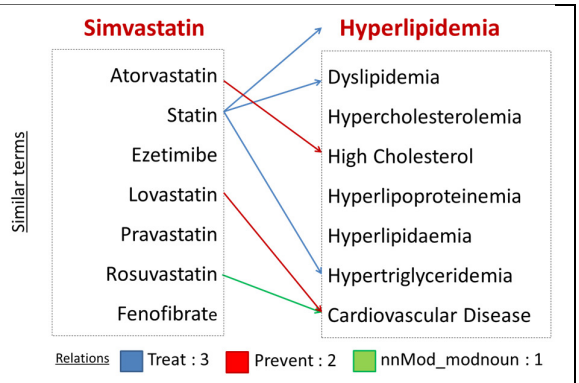

Figure 2. Example of similar terms & relations as used in DRE

Relation Features. The existence of any relations between the two arguments is likely a trigger for other relations. For example, knowing that a drug prevents a disease is an indication for a "treats" relation as well. As mentioned above, the JoBimText tool can be used to identify relations that exist between the arguments in the corpus. However, a specific pair of arguments may not be mentioned together in the corpus (as processed by JoBimText) often enough, but pairs of similar terms may be present. When such expanded term pairs are considered, the relation features may provide a stronger signal. Therefore, in DRE, each relation argument is first expanded to its similar terms. As shown in Figure 2, *Simvastatin* is expanded to seven other medications including *Atorvastatin* and *Statin*, and *Hyperlipidemia* is expanded to seven other diseases including *Dyslipidemia* and *Hypercholesterolemia*. The similar term expansion is done by using the JoBimText tool. The number of expanded terms are limited to the top 10 relevant terms, based on empirical observation of optimal precision-recall trade-off in our study. Among the expanded term pairs, DRE finds three may_treat relations, two may_prevent relations, and one

nnMod_modnoun relation (a syntactic relation obtained by analyzing the parse tree) using JoBimText as shown in Figure 2. The counts for these three relations become feature values for the relation features as shown in Figure 1.

Topic Similarity Features. The topic similarity between arguments can be a useful feature to detect semantic relations between them [14]. Topic similarity does not explain why things are related but does provide an indication of the presence of some relation between them. For example, Cholesterol and Diabetes are related, but their topic similarity is the same regardless of whether the relation is Diagnose or Treat. We scored the topic similarity between the two arguments using Latent Semantic Analysis (LSA) [15] [16]. The value of the LSA feature is equal to the LSA similarity between the two arguments. For the *Simvastatin-Hyperlipidemia* example, the value of the LSA similarity is 0.4.

3.2 Association Data (AD) method

The second association scoring method is based on the intuition that the historical, actual patient care data for a medical problem indicates clinically relevant associations between patients' problems and medications/drug-classes. This method uses a set of statistical measures obtained by mining structured and coded data in approximately six million, longitudinal patient records as features in a supervised machine learning model. The features are described below.

(In the equations below, the following notations are used:

X_D / $\overline{X}_D$ = patient records with/without an order for X and diagnosis of disease D,

$X_{\overline{D}}$ / $\overline{X}_{\overline{D}}$ = patient records with/without an order for X and diagnosis of a disease other than D,

subscript A_D means after diagnosis D, subscript B_D means before diagnosis D,

and the time window for "at", "before", and "after" are specified in the feature definitions.)

Frequency at diagnosis. The fraction of patients who received an order for the given medication or its drug class among the patients who have been diagnosed with a given disease. The time window for the order is three months before or two days after.

$$\boldsymbol{FreqAtDx(D,X)} = \frac{\boldsymbol{X_D}}{\boldsymbol{X_D + \overline{X}_D}}$$

Relative Frequency at diagnosis. The fraction of patients who received an order for the medication or its drug class among the patients who have been diagnosed with a given disease relative to the other diseases. The time window is the same as above.

$$\boldsymbol{RelFreqAtDx(D,X)} = \frac{\boldsymbol{X_D}}{\boldsymbol{X_{\overline{D}}}}$$

After versus Before diagnosis. The ratio of the number of times the medication or its drug class was ordered before the diagnosis of a given disease to the number of times the treatment or test was ordered after the diagnosis. If the after count is zero, then the ratio is set to the maximum value observed. The time window is three months for "after" and three months for "before".

$$\boldsymbol{AfterVsBeforeDx(D,X)} = \frac{\boldsymbol{X_{A\_D}}}{\boldsymbol{X_{B\_D}}}$$

Odds Ratio at diagnosis. The odds ratio of receiving the medication or its drug class at diagnosis relative to the diagnosis of the given disease. The time window for "at" diagnosis is 30 days before and two days after.

$$\boldsymbol{OddsRtAtDx(D,X)} = \frac{\boldsymbol{X_D/\overline{X}_D}}{\boldsymbol{X_{\overline{D}}/\overline{X}_{\overline{D}}}}$$

Odds Ratio Before diagnosis. The odds ratio of receiving the given medication or its drug class relative to before/at the diagnosis of the given disease. The time window is three months for "before" diagnosis and 30 days before and two days after for "at" diagnosis.

$$\boldsymbol{OddsRtBeforeDx(D,X)} = \frac{\boldsymbol{X_{B\_D}/\overline{X}_{B\_D}}}{\boldsymbol{X_D/\overline{X}_D}}$$

Odds Ratio After versus Before diagnosis. The odds ratio of receiving the given medication or its drug class relative to before/after diagnosis of the given disease. The time window is three months before for the "before" diagnosis and three months after for the "after" diagnosis.

$$\boldsymbol{OddsRtAfterVsBeforeDx} = \frac{\boldsymbol{X_{A\_D}/\overline{X}_{A\_D}}}{\boldsymbol{X_{B\_D}/\overline{X}_{B\_D}}}$$

Number of Patients Ordered. Total number of patients who received an order for the medication or its drug class within three months before to three months after the first diagnosis of the disease.

$$\boldsymbol{N(D,X) = X_D}$$

Pearson Product-Moment Correlation. This feature is the Pearson correlation value between the given disease (D) diagnosis and ordering the given medication or drug class (X) at the time of diagnosis. The Pearson product-moment correlation is calculated using the standard formula, using diagnoses data set $\{d_i\}$, $i = 1..m$, where d_i is 1 if D is diagnosed in the patient record i and 0 otherwise, and for each orders data set $\{x_i\}$, $i = 1..m$ where x_i is 1 if X is ordered in the corresponding patient record i and 0 otherwise. The number of patient records in the data set is m for both data sets.

Jaccard Index. This feature is the Jaccard index calculated between the diagnoses set and the medications set for a given disease (D) and medication (X).

$$\boldsymbol{JaccardIndex(D,X) = \frac{\{d_i\} \cap \{x_i\}}{\{d_i\} \cup \{x_i\}} = \frac{X_D}{(X_D + X_{\overline{D}} + \overline{X}_D)}}$$

Arguments Expansion and Feature Aggregation

Since the arguments to relation scoring are words and phrases (terms), it is necessary to map them to a standardized form so that the arguments can be matched with the data in the patient records. Therefore, all terms are first linked to one or more UMLS concept unique identifiers (CUIs) and then diseases are mapped to ICD9 codes and medications to RxNORM. In most cases, these mappings are one to many. So, for a given pair of disease and medication terms, (D, M), the method first generates standardized pairs $\{(D_i, M_j), i = 1..n, j = 1..m\}$, and then the above defined statistical measures for each pair of standardized entities. The next step is to aggregate these n x m feature vectors into a single vector for the (D, M) using the decaying sum (where $decay(p_0 \ldots p_g) = \sum_{i=0}^{g} \frac{p_i}{2^i}$) which produces a single vector, $\{a_1, \ldots, a_k\}$. This process is shown in Figure 3.

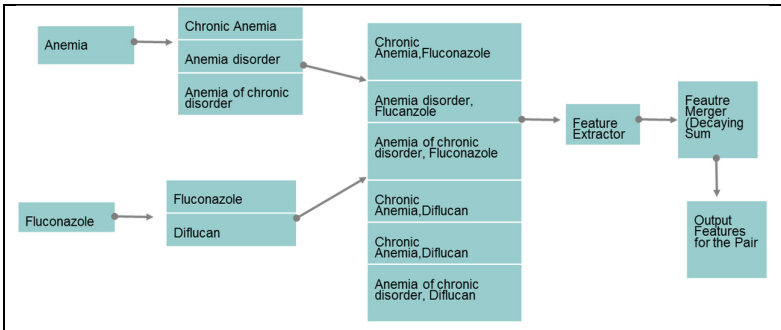

Figure 3. Argument expansion and feature merging in AD; In Feature Extractor, for each expanded pair, the set of features described in Section 3.2 are retrieved from the association data which was previously mined from patient records; In Feature Merger, the feature values of all expanded pairs are aggregated using the decaying sum (see the text) method to create a single feature vector for the given arguments.

3.3 Ensemble Method

A stacked ensemble of DRE and AD methods, designated as DRE+AD, was created. Individual scores are used as features, and a supervised machine learning model learned the optimal way to combine these scores and hence the methods and their sources. This approach is known as stacking. We used stacking (rather than combining features from both sources in a single model) for two reasons: (1) DRE and AD are fundamentally two different approaches and we wanted to study them separately and in combination; (2) DRE and AD achieved optimum performance with different machine learning methods – AD performed well with Random Forest because it uses a small number of features whereas DRE uses a high dimensional feature set which makes Logistic Regression a more suitable and effective approach.

3.4 Models and Training

For AD, a random forest [17] model was built since it provided the best accuracy. For DRE, DRE+AD, and for threshold learning, logistic regression models were accurate and were built with the LIBLINEAR package [18]. The training data set contained positive and negative examples of entity pairs labeled as associated (positive) or not associated (negative). As discussed below, positive and negative associations were imbalanced in the ground truth. We randomly sampled the larger set to create balanced training data.

3.5 Data Preparation

For this assessment, approximately 122,374 disease-medication pairs, were extracted automatically from 100 de-identified, actual patient records that were obtained under an IRB approval from a large, multi-specialty hospital. Data characteristics are summarized in Table 1. The diseases for each patient record are obtained using an automated problem list generation [19] [20] but can be replaced by the diagnostic codes or other means. The medications are taken from the medication orders in each patient record and are represented as text strings and RxNORM codes (as entered in the medication order). Therefore, if a patient record has D diseases in the problem list and M medications in the medication orders then the patient record yields $D \; x \; M$ disease-medication pairs.

Table 1. Description of the data used in this study.

Description	Value
Patient records	100
Disease-medication pairs from the patient records	122,374
De-duplicated disease-medication pairs	25,379
Positive associations in the gold standard (de-duplicated)	1,642
Negative associations in the gold standard (de-duplicated)	23,737

Two characteristics of this data are worth noting: (1) It contains duplicate pairs; (2) Negative examples (entities that don't have an association) are significantly larger than the positive examples. However, the data set is representative of the association scoring system input in realistic clinical applications. To remedy the duplication and asymmetry, we present results first without duplicates and later show the impact of the occurrence frequency on accuracy. Furthermore, we separately report accuracy for positively associated pairs and negatively associated pairs in the gold standard.

3.6 Gold Standard Development

The gold standard required for this study was developed by senior year medical students, who were presented with the pairs of unique entities from the data set and were asked to indicate whether a pair has an association or not. A physician (an MD) gave guidelines and examples to the students for the manual assessment. The students were instructed to identify any direct relationship between a pair. For example, the instructions allowed relationships such as a medication may treat or prevent a disease, or may cause a disease as a side effect. From the initial trials, it became obvious that the association is mostly independent of a patient and therefore any duplicate pairs in the aggregated data were eliminated and the students were asked to assess the relationship independent of the patient record from which the pair was drawn. Each association was assessed by two students and any conflicts were resolved by the MD. The gold standard was later vetted once the automated methods were run on this data, and corrections, if any, were made to the gold standard. The final gold standard contained 25,379 unique disease-medication pairs, including 1,642 positive instances and 23,737 negative instances.

3.7 Experiments, Accuracy Metrics, and Analysis

We used a 10 x 10 cross validation to conduct accuracy analysis. The model was always trained using an equal number of positive and negative pairs, but the model is tested on the imbalanced set. The results are reported for the aggregate of all 10-fold cross validation iterations.

In the experiments, we tested the performance of four methods: the baseline method described earlier, DRE, AD, and a stacked ensemble of DRE and AD. Each experiment involved obtaining the association scores for each of the four methods for the entity pairs in the test set, using a threshold to determine if the association is positive or negative as per the method. Association scores range from 0 to 1, and a scored association is positive if the predicted score is greater than or equal to the threshold, and negative otherwise. Using the gold standard, we then computed true positives, false positives, false negatives, and true negatives, from which we computed standard precision (P), recall (R), and F score (F1) for positive and negative associations. Note that if a method has no coverage for an entity pair, then a zero score is returned by the method and hence results in a negative association.

F1 scores for positive associations were determined and plotted at threshold values from 0.1 to 0.9, in intervals of 0.1. We also plotted the precision-recall curves and compared areas under the curves. The threshold values that optimize F1 for each method were obtained and used in the final performance comparison of the methods.

As one may recall from Section 3.5, the data contains multiple instances of some problem-medication pairs since the entity pairs are extracted from 100 patient records. For example, several patients may

have been diagnosed with Diabetes and many of them may be prescribed Metformin as a treatment, in which case, the entity pair Diabetes-Metformin may occur several times in the data. Using the frequency of such occurrences in the original data as a weighting function, we determined weighted accuracy of the methods.

4 Results and Discussion

4.1 Coverage

Table 2 shows the coverage for the baseline and AD methods. Coverage is the percentage of the entity pairs in the data for which the underlying methods and sources entail a positive or negative association. DRE, as it uses UMLS CUIs and types as some of its features, always returns non-empty values for the features. On the other hand, AD and the baseline method end up with all empty features, for at least some entity pairs. For example, if a medication is never prescribed for a disease, the patient records would never have any data for it.

Table 2. Coverage for the baseline and AD methods

Method	Coverage	
	Positive Associations	Negative Associations
Baseline	43%	12%
AD	88%	41%

For the positive associations, AD has a very high coverage (88%), but the baseline method has only 43% coverage. This reflects in the baseline method's poor accuracy in predicting positive associations. For the negative associations, both have poor coverage but it does not matter as much since the default score of 0.0 would end up being a correct prediction for negative associations.

4.2 Accuracy, Thresholds, and P-R Curves

First, consider the positive association at the optimum threshold values, as shown in Table 3. For the unique (i.e. unweighted) pairs, DRE performed slightly better than AD with an F1 score of 0.60 compared to an F1 score of 0.56 for AD. Both methods performed significantly better than the baseline, which achieved an F1 score of only 0.35. The stacked method of DRE and AD performed better than the individual methods, achieving an F1 score of 0.69. Among the individual methods, AD had higher precision (0.62) and DRE had higher recall (0.66).

Table 3. Accuracy analysis of the methods in predicting the positive associations

Method	Optimum Threshold	Unweighted			Weighted		
		Precision	Recall	F1 Score	Precision	Recall	F1 Score
Baseline	0.26	0.28	0.48	0.35	0.58	0.53	0.55
DRE	0.20	0.56	0.66	0.60	0.68	0.62	0.67
AD	0.26	**0.62**	0.52	0.56	0.72	0.63	0.67
DRE + AD	0.29	0.57	**0.71**	**0.69**	**0.77**	**0.73**	**0.75**

When weighted entity pairs are considered, which represent the frequency of occurrence of the entity pairs in the patient records we used for this study, the performance pattern of the methods generally remained unchanged. DRE and AD performed significantly better than the baseline; each achieving an F1 score of 0.67. The stacked ensemble achieved the highest F1 score of 0.75. All methods achieved higher precision than recall. We handled the imbalanced nature of the dataset by learning a threshold value that optimizes the F1 score for positive associations, which is considered one of the effective ways to deal with imbalanced datasets [21]. The optimum thresholds are shown in Table 3 for the various methods, and Figure 4 shows how the F1 score varies with the threshold for the methods.

Table 4. An ablation study of accuracy for unique (unweighted) positive pairs

Method and Feature Ablation	Precision	Recall	F1 Score
AD – without drug class features	0.60	0.46	0.52
AD – with only drug class features	0.68	0.42	0.52
DRE – with only LSA features	0.50	0.25	0.33
DRE – with only UMLS features	0.21	0.44	0.28
DRE – with only DS features	0.60	0.51	0.55
DRE – with only DS features, but without argument expansion	0.52	0.46	0.42

An ablation study of feature groups for the unique (i.e. unweighted) positive pairs is shown in Table 4. We removed (ablated) a selected group of logically related features and determined the accuracy which would show the importance of the feature group to the model. For AD, using either drug-class or individual drugs alone in calculating feature scores achieved the same accuracy, but when used together they improved the overall

accuracy. For DRE, the distributional semantics (DS) features with argument expansion achieved accuracy close to the best DRE accuracy, and argument expansion by itself contributed significantly to the accuracy of the DS features. However, the UMLS features alone achieved the lowest F1 score, although they were useful in improving recall.

In predicting negative associations, all methods achieved very high performance for both unweighted and weighted cases. This result merely reflects the fact that all methods return 0.0 when the underlying sources provide no information for the arguments, which happens to correctly predict a negative association. It is good to see that the default in these cases does no harm because a knowledge resource or method needs to handle all scenarios well in clinical applications.

Precision-recall curves for the methods are shown in Figure 5. As precision improves, recall reduces, at different rates for the different algorithms, which reflects in the area under the curve (AUC) metrics.

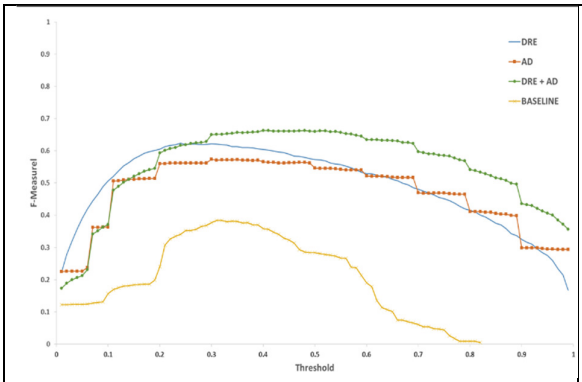

Figure 4. Threshold vs. F1 scores

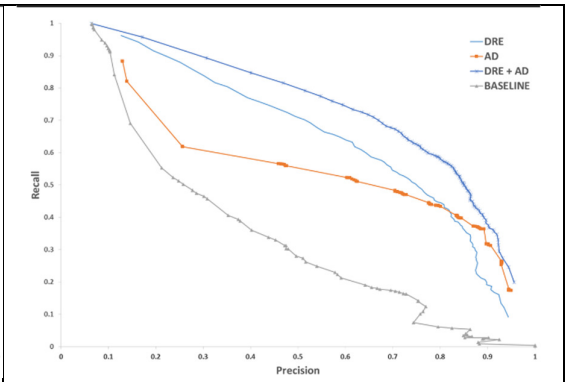

Figure 5. Precision-Recall Curves

The ability of the ensemble method to improve precision while not losing recall as rapidly is helped by AD and resulted in the overall high F1 score.

4.3 Discussion

While each of the two methods we evaluated achieved a reasonable level of accuracy, the ensemble method achieved the best performance. The two sources and methods complement each other, forming a more effective method for association scoring. DRE relies on carefully written medical text and manually curated knowledge resources, while AD relies on statistical measures of patient care data. The intrinsic nature of the resources used by the methods reflects in the performance of the methods on certain types of entities. For example, DRE is better at coverage on rare diseases, such as scoring the pair *hypophosphatemic rickets* and *calcitriol*, and for over the counter (OTC) medications. AD performs better than DRE when not all medications within a class are equally used to treat a problem. For example, for the pair *migraine and headache syndromes* and *Inderal la*, DRE scored 0.007 whereas AD scored 0.773, which is a true positive.

5 Conclusion

To reduce the cognitive load on physicians in using large amounts of data in the modern EHR systems, it is necessary to imbue clinical applications with fundamental medical knowledge, such as the relationships between diseases and medications. This paper presented two new methods that used different ways of extracting features (distributional semantics and data mining) and two different content sources (large medical context and patient records) for the task. We compared the accuracy of these distinctly different approaches and their ensemble with a baseline method published previously. The results showed that an ensemble provides an accurate relation scoring system because of individual methods leveraging different content sources and feature extraction. It can be used as an on-demand scoring system, or as a method to generate association scores for a large set of entities a priori for later use in clinical decision support applications. The methods introduced here are promising, and can be expanded in the future to score specific relations such as “treats” and “prevents”, and to score relations between other types of clinical data.

References

[1] R. Wachter, The Digital Doctor, McGraw-Hill, 2014.

[2] T. D. Shanafelt, L. N. Dyrbye, C. Sinskye, O. Hasan, D. Satele, J. Sloan and C. P. West, "Relationship Between Clerical Burden and Characteristics of the Electronic Environment With Physician Burnout and Professional Satisfaction," *Mayo Clinic Proceedings,* vol. 91, no. 7, pp. 836-848, 2016.

[3] US National Library of Medicine, "UMLS Reference Manual," National Library of Medicine (US), September 2009. [Online]. Available: http://www.ncbi.nlm.nih.gov/books/NBK9675/. [Accessed 15 04 2014].

[4] J. Hirschberg and C. D. Manning, "Advances in natural language processing," *Science,* vol. 349, no. 6245, pp. 261-266, 2015.

[5] O. Uzuner, B. R. South, S. Shen and D. L. Scott, "2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text," *Journal of the American Medical Informatics Association,* vol. 18, no. 5, pp. 552-556, 2011.

[6] C. Wang and J. Fan, "Medical Relation Extraction with Manifold Models," in *The 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014)*, 2014.

[7] W.-Q. Wei, R. M. Cronin, H. Xu, T. A. Lasko, L. Bastarache and J. C. Denny, "Development and evaluation of an ensemble resource linking medications to their indications," *Journal of the American Medical Informatics Association : JAMIA,* vol. 20, no. 5, pp. 954-961, 2013.

[8] A. Wright, E. S. Chen and F. L. Maloney, "An automated technique for identifying associations between medications,," *Journal of Biomedical Informatics,* vol. 43, pp. 891-901, 2010.

[9] F. Severac, E. A. Sauleau, N. Meyer, H. Lefevre, G. Nisand and N. Jay, "Non-redundant association rules between diseases and medications: an automated method for knowledge base construction," *BMC Medical Informatics and Decision Making,* vol. 15, no. 29, 2015.

[10] H. Salmasian, T. H. Tran, H. S. Chase and C. Friedman, "Medication-indication knowledge bases: a systematic review and critical appraisal," *Journal of the American Medical Informatics Association,* vol. 22, no. 6, pp. 1261-1270, 2015.

[11] C. Biemann and M. Riedl, "Text: now in 2D! A framework for lexical expansion with contextual similarity," *Journal of Linguistic Modeling,* vol. 1, no. 1, pp. 55-95, 2013.

[12] A. Gliozzo, "Beyond Jeopardy! Adapting Watson to New Domains Using Distributional Semantics," 2013. [Online]. Available: https://www.icsi.berkeley.edu/icsi/sites/default/files/events/talk_20121109_gliozzo.pdf. [Accessed 18 04 2014].

[13] D. Ferrucci, A. Levas, S. Bagchi, D. Gondek and R. T. Mueller, "Watson: Beyond Jeopardy!," *Artificial Intelligence,* pp. 93-105, 2013.

[14] A. M. Gliozzo, M. Pennacchiotti and P. Pantel, "The domain restriction hypothesis: Relating term similarity and semantic consistency," in *Proceedings of HLT-NAACL*, Rochester, NY, 2007.

[15] S. Deerwester, D. T. Susan, G. W. Furnas, T. K. Landauer and R. Harshman, "Indexing by Latent Semantic Analysis," *Journal of the American Society for Information Science,* vol. 41, no. 6, pp. 391-407, September 1990.

[16] S. Simmons and Z. Estes, "Using latent semantic analysis to estimate similarity," Hillsdale, NJ, 2006.

[17] L. Breiman, "Random Forests," *Machine Learning,* vol. 45, no. 1, pp. 5-32, 2001.

[18] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang and C.-J. Lin, "Liblinear: A library for large linear classification.," *Journal of Machine Learning Research,* vol. 9, p. 1871–1874, 2008.

[19] M. Devarakonda and C.-H. Tsou, "Automated Problem List Generation from Electronic Medical Records in IBM Watson," in *Proceedings of the Twenty-Seventh Conference on Innovative Applications of Artificial Intelligence*, Autin, TX, 2015.

[20] M. V. Devarakonda and N. Mehta, "Cognitive Computing for Electronic Medical Records," in *Healthcare Information Management Systems, 4th Edition*, A. C. Weaver, J. M. Ball, R. G. Kim and M. J. Kiel, Eds., Springer International, 2015.

[21] W. Klement, S. Wilk, W. Michaowski and S. Matwin, "Classifying Severely Imbalanced Data," in *Advances in Artificial Intelligence, Proc. of 24th Canadian Conf on Artificial Intelligence*, St. John's, Springer Berlin Heidelberg, 2011, pp. 258-264.