Automated Knowledge Acquisition from Clinical Narrative Reports

Xiaoyan Wang, MA[1], Amy Chused, MD[1], Noémie Elhadad, PhD[1],
Carol Friedman, PhD[1*], Marianthi Markatou, PhD[2*,]
Dept of [1]Biomedical Informatics, [2]Biostatistics, Columbia University, New York, NY

ABSTRACT

Knowledge of associations between biomedical entities, such as disease-symptoms, is critical for many automated biomedical applications. In this work, we develop automated methods for acquisition and discovery of medical knowledge embedded in clinical narrative reports. MedLEE, a Natural Language Processing (NLP) system, is applied to extract and encode clinical entities from narrative clinical reports obtained from New York-Presbyterian Hospital (NYPH), and associations between the clinical entities are determined based on statistical methods adjusted by volume tests. We focus on two types of entities, disease and symptom, in this study. Evaluation based on a random sample of disease-symptom associations indicates an overall recall of 90% and a precision of 92%. In conclusion, the preliminary study demonstrated that this method for knowledge acquisition of disease-symptom pairs from clinical reports is effective. The automated method is generalizable, and can be applied to detect other clinical associations, such as between diseases and medications.

INTRODUCTION

The availability of up-to-date knowledge bases that provide comprehensive disease-specific information is essential for clinical applications ranging from quality of care to hypothesis generation. Knowledge bases that are manually constructed, such as QMR and Micromedex, ensure accuracy and quality[1, 2]. Manual construction, however, is often laborious and costly. Additionally, since biomedical knowledge is constantly evolving as new discoveries are made and practices change over time, the knowledge bases have to be constantly updated. Therefore it would be beneficial to develop automated methods to create and update these knowledge bases.

Biomedical knowledge is often buried in either biomedical literature or narrative clinical reports. As one of the high throughput technologies, natural language processing (NLP) has been applied in biomedicine for decades[3]. NLP can extract and encode massive amounts of text data in literature and medical reports within a relatively short time period. Various NLP systems have been applied in biomedical literature and narrative clinical reports[4-7]. In biomedicine, much of the knowledge acquisition research utilizing text includes: discovering hidden associations between medical entities[8] and knowledge base construction[9]. Collier et al. pointed out that applications in molecular biology and biomedical literature provide for interesting results, but studies focused on clinical corpora and electronic health records (EHR) are still needed in order to bring together the domains[10]. Several studies have demonstrated the use of structured and unstructured data in the EHR for improving quality of health care [11, 12]. While structured data in EHR systems ensures fast and optimal information retrieval, unstructured data, such as narrative clinical reports, allows flexibility of expression and representation of elaborate clinical entities such as symptoms.

In recent years, statistical methods combined with NLP systems have gained ground in knowledge acquisition[13]. These methods focus on extracting and establishing associations between entities from textual data. Determining clinically important associations among entities from a large database is challenging for several reasons: 1) the associations are typically not explicitly stated in the clinical reports, 2) there are multiple simultaneous hypotheses, and 3) associations that are statistically but not clinically significant could be easily selected due to large sample size.

Co-occurrence statistics have proven very effective in acquiring associations between entities[14]. The idea for using co-occurrence statistics is based on the hypothesis that an entity and its associated entities are more likely to appear together than random combinations of the entity and the associated entities.

Our group has been developing automated methods for detecting associations between clinical entities from both the biomedical literature and narrative clinical records using NLP and statistical methods [15-18]. Cao et al. applied a co-occurrence statistical methodology for the automated generation of medical problem lists using structured output generated by processing narrative clinical reports[15, 16]. In this work, the authors analyzed co-occurrence at the level

[*] Carol Friedman and Marianthi Markatou contributed equally as senior authors to this work

of the report. As proof of concept, Chen et al. adapted these results to detect trends in medications associated with certain diseases based on the literature, and also narrative clinical reports [17,18].

The present study builds upon the previous work by developing an automated approach to discover disease-symptom associations from clinical narrative reports. Contextual information associated with entities is incorporated into analyzing co-occurrence data. Fixed margin volume tests are used to determine the strength of association in order to select associations accurately[16, 17]. We found that the system was effective and for the most part, the associations were consistent with expert resources.

MATERIALS & METHODS

1. Materials

The data warehouse in NYPH collects and maintains a variety of structured and unstructured data for patient records. Textual discharge summaries dictated in 2004 were used in this study.

2. Text processing and Data selection

MedLEE is a NLP system that has been used for a range of clinical and research applications [19-21]. In this study, MedLEE was used to parse and transform discharge summaries into a structured representation in extended markup language (XML) format. MedLEE output consists of various types of findings (e.g. problem, medication, procedure) along with various types of modifiers (e.g. certainty, family history, temporal, section of report), and UMLS codes.

Two tables were created based on UMLS semantic types to classify terms already encoded by MedLEE as disease or symptom. *Disease or symptom* [T047], *Mental or Behavioral Dysfunction* [T048], and *Neoplastic Process* [T191], were used for diseases, while *Sign or Symptom* [T184] and *Finding* [T033] for symptom.

All the findings were filtered so that only diseases and symptoms were included, and findings with certain modifier values were excluded (*low certainty, negation, family history,* and *past history*). In order to obtain diseases and symptoms occurring at the time of admission and to avoid those caused by subsequent treatment or procedures, an additional contextual filter consisting of the section where the clinical information occurred was applied to exclude diseases and symptoms occurring during the hospital course.

3. Statistical measurement and knowledge base construction

A disease and a symptom were considered to co-occur if they appeared in the same case report. Co-occurrence tables for each disease-symptom pair across all the possible disease and symptom values were generated. All co-occurrence tables that had a frequency of less than 2 were excluded because they were unlikely to yield meaningful statistical results.

To test the hypothesis of no association between a disease and a symptom, the $\chi 2$ statistic was used. The definition of cutoff point was described by Cao and colleagues[16, 17]. In the present study, because the data are 2 x 2 tables with the same row margins, we computed the adjustment to the chi-square p-value that corresponds to tables with fixed row margins. Fixed row margin tests are partially conditional tests, where the conditioning argument is the variable that describes the row marginals. This conditioning guarantees that the margins of the table do not provide any evidence either in favor or against the null hypothesis of independence (i.e. no association). Fixed row margin volume tests have similar interpretation with the unconditional volume tests, that is, they can be interpreted as a distance from the surface of independence. The larger the distance, the stronger the association. For details on the method of computation of the fixed margin test and cutoffs see [16, 17]. For symptoms associated with a particular disease, a ranked list for disease-symptom pairs was generated based on the strength of the statistics. A no-intercept linear regression model was constructed for the disease to identify the cutoff point [16, 17].

All the pairs with greater $\varepsilon(\chi 2)$ than the cutoff point were acquired and included in the knowledge base of disease-symptom associations.

4. Evaluation.

The diseases were ranked and stratified according to frequency of occurrences in the disease vector. The strata of top1-20 (stratum 1), 40-60(stratum 2), 90-110 (stratum 3) and 140-160 (stratum 4) were considered to represent the "most common", "common", "less common" and "rare" diseases. Three diseases were randomly selected from each stratum and compared with the reference standard for evaluation. We categorized symptoms associated with a particular disease into three classes based on a reference source described below: 1) direct association: symptoms that are manifestation of the disease (i.e. *chest pain* for *coronary heart disease*); 2) indirect association: symptoms that are directly associated with other diseases which are highly associated with the particular disease, or symptoms that are consequences of the disease through a clinically plausible pathway (i.e. *chest pain* for *diabetes*); and 3) no association: symptoms that are not known to be associated with the disease, or symptoms that are conceptually vague or non-informative without more context (i.e. the concept *difficulty*, which was found to be associated with *depressive disorder*).

Reference standard To evaluate our disease-symptom knowledge base, a reference standard was used that consisted of a practicing physician (Expert) and two other reference resources: (a comprehensive consumer health online resource: www.webmd.com (WebMD) and a textbook for medical students and residents (TextBook)) [22]). The Expert classified the disease-symptom information for each disease based on medical knowledge. One of the authors (XW) used the two other reference resources, to manually extract and classify the relevant disease-symptom information. The classifications from each of the three reference resources were then combined to create a reference standard as follows: (1) if an association was agreed on by at least two resources, the majority was chosen as the reference standard; (2) if an association was not agreed on by any of the resources, the response from WebMD was chosen to be the reference standard due to its comprehensiveness.

Quantitative evaluation Two metrics were used to assess the performance of our method. Recall was calculated as the ratio of the number of distinct disease-symptom pairs that were identified by our method over the total number of the corresponding disease-symptom pairs in the reference standard (i.e. TP/(TP+FN)). Precision was measured as the ratio of the number of distinct disease-symptom pairs returned by our method that were correct according to the reference standard divided by the total number of disease-symptom pairs found by our method (i.e. TP/(TP+FP)).

Qualitative evaluation Associations between diseases and symptoms were further analyzed manually. Indirect associations were classified as 'close associations' (i.e., symptoms that are directly associated with diseases which are highly associated with the particular disease) and 'remote associations' (i.e., symptoms that are consequences of the disease through a clinically plausible pathway). Similarly, pairs in the category 'no association' were further classified as 'no association' (i.e., symptoms that are not associated with the disease) and 'vague association' (i.e., symptoms that are conceptually vague or non-informative without context).

RESULTS

1. Data Statistics

The case reports in this study included a total of 25,074 discharge summaries from NYPH. Co-occurrence data in the corpus are summarized in Table 1. There are a total of 1,366 unique disease concepts in the database, and the top 150 diseases accounted for 90% of the occurrences. A total of 1,767 pairs of disease-symptom were selected based on their statistics measurements for disease-symptom knowledge base construction. Table 2 presents some simplified entries in the disease-symptom knowledge base, which can be accessed at the following URL: http://www.dbmi.columbia.edu/~xiw7002/DS-KB/

Data in the corpus	Count
Discharge summaries	25074
Unique disease entities	1366
Unique symptom entities	563
Unique disease-symptom co-occurring pairs	31249
Disease-symptom association pairs selected by χ^2 statistic with fixed margin volume test adjustment	1767

Table 1. Summary of the data that was selected

Disease	Symptom
coronary heart disease	chest pain, angina pectoris, shortness of breath, hypokinesia, sweat, sweating increased, pressure chest, dyspnea on exertion, orthopnea, chest tightness
accident cerebrovascular	Dysarthria, asthenia, speech slurred, facial paresis, hemiplegia, seizure, numbness, unresponsiveness

Table 2. Simplified entries from knowledge base of disease-symptom associations obtained using our methods.

2. Results of Evaluation

2.1. Results of overall evaluation

A total of 183 disease-symptom pairs associated with 3 randomly selected diseases were evaluated from each of the 4 strata (12 in all). Validity results indicated a kappa of 0.82 (WebMD and TextBook), 0.62 (TextBook and Expert), 0.67 (WebMD and Expert), respectively. Recall and precision were 0.77, 0.73 respectively if only direct associations were considered as true positives (TP=134). For different clinical and research applications, however, it may not be necessary to distinguish between indirect and direct associations. If these two types of associations were combined and considered as TP (TP=168), the recall and precision would be 0.90 and 0.92 respectively. The automated method had high recall (0.77-0.90) and high precision (0.73-0.92).

2.2 Results of stratum-specific evaluation

To test if the method was sensitive to frequency of occurrences of diseases, a stratum-specific analysis was performed. The recall and precision for each stratum were computed based on combining direct association and indirect association as TP. The results were summarized in Figure 1. Stratum I representing

Disease	Symptom				
	Direct association (61%)	**Indirect association**		**No association**	
		Close association (17%)	**Remote association (14%)**	**Vague association (6%)**	**No association (2%)**
Diabetes	Polyuria, polydypsia, orthopnea, sweat, sweating increased, vertigo, vomiting, labored breathing	shortness of breath, pain chest , asthenia, nausea	rale, mental status changes	unresponsiveness	
depressive disorder	feeling suicidal, suicidal, hallucinations auditory, feeling hopeless, weepiness, sleeplessness, motor retardation, irritable mood, blackout, mood depressed, hallucinations visual, worry, agitation, tremor, verbal, auditory, hallucinations, nightmare, unable to concentrate		intoxication	difficulty	

Table 3. Examples of Qualitative Evaluation.

"most frequent diseases" had highest recall and precision (0.92, 0.95 respectively). Stratum 4 representing "rare disease" had moderate recall and precision (0.87, 0.89 respectively).

2.3 Results of qualitative analysis

Overall, we determined that 61% of the associations were direct associations, 17% were close associations, 14% were remote associations, 6% were vague associations and 2% were not associated. Examples of the qualitative analysis are shown in table 3.

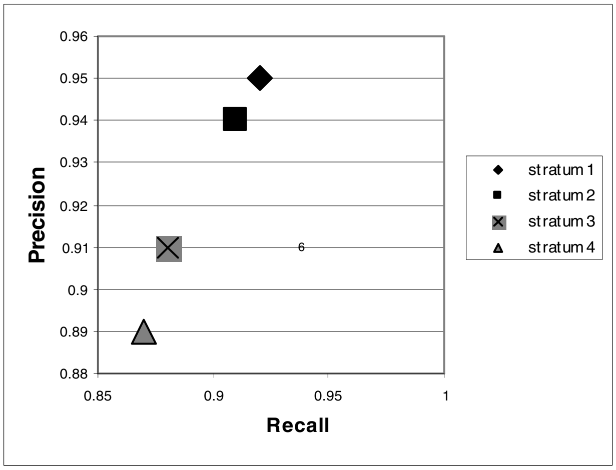

Figure 1. Precision versus recall of stratum-specific evaluation

DISCUSSION

The aim of the present study was to develop an automated tool to acquire knowledge from clinical narrative reports. The results indicate that use of the automated approach is feasible and effective in detecting associations in clinical narrative reports.

Including a contextual filter in our automated method which consists of the sections where the clinical information appears, is important. Our initial tests suggested that a simple strategy without contextual filters affected both recall and precision because symptoms resulting from treatments or procedures were confounded with those associated with the diseases. In particular, some important associations for rare diseases were missed entirely. Generally, symptoms occurring at the time of admission are more likely related to the disease processes than those occurring during the hospital stay. Filtering out information from certain sections seemed to effectively exclude disease-symptom associations where the symptoms were the results of treatment provided during the hospital stay (i.e. symptoms caused by medications or procedures), and the other filters seemed to remove other spurious information (i.e. negation).

Determining the type of association is difficult. The methodology put forward in [16, 17] is not designed to discriminate between the different types of associations. For example, as shown in Table 3, "*chest pain*" and "*short of breath*" were detected by our methods as being associated with the disease "*diabetes*", and the association was determined to be an indirect one using the reference standard. This occurred because *diabetes* itself is closely associated with *coronary heart disease*, and "*chest pain*" and "*short of breath*" are both direct manifestations of *coronary heart disease,* as determined by the reference standard. However "*mental status changes*" is not closely associated with *Diabetes* but is considered to be a consequence of severely uncontrolled *Diabetes*. Vague associations such as "*difficulty*" generally do not provide enough specific information and only represent partial symptoms. Further work would be needed to differentiate between the different types of associations.

Direct associations are important for understanding the manifestation of a particular disease. Indirect associations, however, are valuable for acquiring more complex relationships that occur in clinical settings. For example, if we were interested in detecting if a symptom was caused by a drug, we would need to determine that the symptom is not a direct manifestation of a disease or an indirect manifestation, which is a consequence of another disease or symptom.

For the diseases we studied, it is noteworthy that a majority of the associations for those which are diagnosed partially based on signs and symptoms (e.g. "*depression disorder*" in Table 3) were direct associations. In contrast, a large portion of the associations were indirect associations (close or remote associations) for diseases diagnosed based on lab tests (e.g. "*diabetes*" in Table 3). As a further line of research, it would be desirable to explore whether we can develop methods to distinguish between direct and indirect associations based upon these observations. More sophisticated statistical models should also be explored to achieve this goal.

One of the challenges we encountered in this work is the granularity of diseases and symptoms. Multiple UMLS codes for a disease or symptom were generated by MedLEE when we first initiated the study due to varying amounts of specificity associated with modification. For example, a range of UMLS codes were extracted for cough, such as *barking cough, cough at rest, brassy cough, persistent dry cough, non-productive cough* and *hacking cough*. A few methods have been explored to solve the issue. In this study, we modified MedLEE encoding to exclude some relative anatomical modifiers (e.g. right) and degree modifiers (e.g. severe). This worked well for symptoms. Further investigation, however, is still required to group and identify these entities at a clinically meaningful level of granularity, as that is important for acquiring accurate associations between clinical entities.

One of the limitations in this study is that these associations are based on inpatient reports and therefore may reflect different disease-symptom associations than those that would be acquired using reports from outpatients. In the future, we will combine inpatient and outpatient data to get more comprehensive clinical information.

CONCLUSION

Biomedical knowledge is often buried in text data such as narrative clinical records. In this study, we described an automated method that effectively acquires knowledge from narrative reports in a clinical data warehouse. Methods that enable automated updating of knowledge bases that provide comprehensive disease-specific information are essential for clinical applications ranging from quality of care to hypothesis generation.

Acknowledgments

The authors thank Elizabeth Chen and Lyudmila Shagina for their assistance and contributions. This work is supported in part by grants R01 LM007659 and R01 LM008635 from the National Library of Medicine. The work of Dr Markatou was supported by NSF grant DMS-0504957.

REFERENCES

1. Berner, E.S., et al., *Performance of four computer-based diagnostic systems.* N Engl J Med, 1994. **330**(25): p. 1792-6.
2. Clauson, K.A., et al., *Clinical decision support tools: analysis of online drug information databases.* BMC Med Inform Decis Mak, 2007. **7**(7): p. 7.
3. Baruch, J.J., *Progress in programming for processing English language medical records.* Ann N Y Acad Sci, 1965. **126**(2): p. 795-804.
4. Aronson, A.R., et al., *The NLM Indexing Initiative.* Proc AMIA Symp, 2000: p. 17-21.
5. Witte, R., T. Kappler, and C.J. Baker, *Enhanced semantic access to the protein engineering literature using ontologies populated by text mining.* Int J Bioinform Res Appl, 2007. **3**(3): p. 389-413.
6. Pakhomov, S., et al., *Electronic medical records for clinical research: application to the identification of heart failure.* Am J Manag Care, 2007. **13**(6 Part 1): p. 281-8.
7. Friedlin, J. and C.J. McDonald, *Using a natural language processing system to extract and code family history data from admission reports.* AMIA Annu Symp Proc, 2006. **925**: p. 925.
8. Weeber, M., et al., *Text-based discovery in biomedicine: the architecture of the DAD-system.* Proc AMIA Symp, 2000: p. 903-7.
9. Hahn, U., M. Romacker, and S. Schulz, *Creating knowledge repositories from biomedical reports: the MEDSYNDIKATE text mining system.* Pac Symp Biocomput, 2002: p. 338-49.
10. Collier, N., et al., *Recent advances in natural language processing for biomedical applications.* Int J Med Inform, 2006. **75**(6): p. 413-7.
11. Walker, J.M., et al., *EHR Safety: The Way Forward to Safe and Effective Systems.* J Am Med Inform Assoc, 2008. **28**: p. 28.
12. Simon, S.R., et al., *Electronic health records: which practices have them, and how are clinicians using them?* J Eval Clin Pract, 2008. **14**(1): p. 43-7.
13. Chaussabel, D., *Biomedical literature mining: challenges and solutions in the 'omics' era.* Am J Pharmacogenomics, 2004. **4**(6): p. 383-93.
14. Narayanasamy, V., et al., *TransMiner: mining transitive associations among biological objects from text.* J Biomed Sci, 2004. **11**(6): p. 864-73.
15. Cao, H., et al., *Mining a clinical data warehouse to discover disease-finding associations using co-occurrence statistics.* AMIA Annu Symp Proc, 2005: p. 106-10.
16. Cao, H., G. Hripcsak, and M. Markatou, *A statistical methodology for analyzing co-occurrence data from a large sample.* J Biomed Inform, 2007. **40**(3): p. 343-52.
17. Chen, E., et al., *Automated acquisition of disease-drug associations from biomedical and clinical documents.* 2007.
18. Chen, E.S., et al., *Automated acquisition of disease drug knowledge from biomedical and clinical documents: an initial study.* J Am Med Inform Assoc, 2008. **15**(1): p. 87-98.
19. Friedman, C., et al., *Automated encoding of clinical documents based on natural language processing.* J Am Med Inform Assoc, 2004. **11**(5): p. 392-402.
20. Friedman, C., et al., *Representing information in patient reports using natural language processing and the extensible markup language.* J Am Med Inform Assoc, 1999. **6**(1): p. 76-87.
21. Friedman, C., et al., *A general natural-language text processor for clinical radiology.* J Am Med Inform Assoc, 1994. **1**(2): p. 161-74.
22. Ferri, F., *Ferri's Differential Diagnosis: A Practial Guide to the Differential Diagnosis of Symptoms, Signs, and Clinical Disorders* Mosby Elsevier. 2006.