Integrating Multiple On-line Knowledge Bases for Disease-Lab Test Relation Extraction

Yaoyun Zhang, PhD, Ergin Soysal MD, PhD, Sungrim Moon PhD, Jingqi Wang MS, Cui Tao PhD, Hua Xu PhD
School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX, USA

Abstract

A computable knowledge base containing relations between diseases and lab tests would be a great resource for many biomedical informatics applications. This paper describes our initial step towards establishing a comprehensive knowledge base of disease and lab tests relations utilizing three public on-line resources. LabTestsOnline, MedlinePlus and Wikipedia are integrated to create a freely available, computable disease-lab test knowledge base. Disease and lab test concepts are identified using MetaMap and relations between diseases and lab tests are determined based on source-specific rules. Experimental results demonstrate a high precision for relation extraction, with Wikipedia achieving the highest precision of 87%. Combining the three sources reached a recall of 51.40%, when compared with a subset of disease-lab test relations extracted from a reference book. Moreover, we found additional disease-lab test relations from on-line resources, indicating they are complementary to existing reference books for building a comprehensive disease and lab test relation knowledge base.

Introduction & Background

Lab tests are one of the principal sources of information for decision making in diagnosis of medical problems, determination of prognosis and monitoring disease progression. Linking lab tests with their associated diseases electronically could improve differential diagnosis of diseases (1), enhance healthcare quality (2), reduce adverse events and liability claims (3) and facilitate computational research applications (e.g., phenotype extraction from electronic health records (EHR) data (4)). However, building a computable knowledge base containing comprehensive disease-lab test relations is not a trivial task and it requires substantial efforts. Current research on this topic is very limited.

Although many knowledge bases or ontologies of diseases and lab tests exist, few of them containing disease-lab test relations. For example, the Unified Medical Language System (UMLS)(5) is a comprehensive resources containing various types of biomedical concepts including diseases and lab tests; but explicit relations between diseases and lab tests in the Semantic Network of UMLS are limited. Disease ontology(6), as a comprehensive knowledge base of inherited, developmental and acquired human diseases, does not provide linkages to lab tests either. Similarly, available lab test ontologies such as Logical Observation Identifiers Names and Codes (LOINC)(7) have no linkage to diseases either. Expert systems such as Quick Medical Reference (QMR) (8) contain relations between diseases and lab tests, in order to support decisions of differential diagnosis. Nevertheless, maintenance of such knowledge mainly depends on domain experts, which is time consuming and costly. Furthermore, such knowledge is not freely available for publics.

Several studies have attempted to extract disease and lab test relations from biomedical literature or clinical data. Halil et al. (2012) (9) used rule-based linguistic methods to extract semantic predications (i.e., semantic relations) of diseases and lab tests from biomedical literature, following relation types defined in the semantic network of UMLS(10). Wright et al. (2010) (11) employed the data mining technique of association rule mining to identify associations between problems and lab test results from noisy EHR data (4) and achieved an accuracy of 55.6%. The 2010 i2b2/VA Workshop on Natural Language Processing Challenges for Clinical Records (12) held a relation classification task mainly focused on assigning relation types between medical problems and lab tests occurred in clinical text. Despite of some success of these initial efforts, challenges still exist for automatically extracting disease-lab test relations from these resources. First, lab test names are dynamic (e.g., new lab tests emerging over the time) and accurate diagnosis of diseases need a dynamic list of lab tests (13). Second, the distribution of relations between diseases and lab tests could be skewed. For example, an analysis of EHR data by Wright and Bates (2010) (14) shows that lab tests exhibit a power law distribution, with the top 4.5% of lab tests accounting for 80% of all lab results. Furthermore, no single sources (even standard reference books) could cover all lab tests and their associated diseases (11); therefore, it is necessary to develop approaches that can extract and compile disease-lab test relations from multiple existing resources.

As emerging recently, many on-line publicly available resources provide rich medical information for both consumers and healthcare providers. Those resources have the potential to be leveraged for building clinical knowledge bases, which are attracting increasing research interests. For example, Wei et al (2013) (15) used public resources including Wikipedia and MedlinePlus to extract drug and indication relations and achieved promising results. Jona et al. (2014) compared the accuracy and completeness of drug information in Wikipedia with standard textbooks of pharmacology. Their study suggested that Wikipedia is an accurate and comprehensive source of drug-related information for undergraduate medical education.
As an initial step towards building a comprehensive, computable knowledge base of disease and lab tests relations, this paper describes our efforts to integrate three on-line resources, LabTestsOnline, MedlinePlus and Wikipedia, to create a freely available, computable disease-lab test resource. The aim of this study has two-fold: (1) to examine whether it is feasible to automatically extract disease-lab test relations from online healthcare information resources with high performance; and (2) to investigate the coverage of disease and lab tests relations of online resources, in comparison with a gold standard reference book.

Method

Data sources

We selected three on-line resources for disease and lab test relation extraction: (1) LabTestsOnline (http://labtestsonline.org/) — a website regarding Clinical Chemistry produced by American Association for Patients or Family Caregivers to better understand clinical lab tests; (2) MedlinePlus (http://www.nlm.nih.gov/medlineplus) — a website that offers consumer health information for health consumers, maintained by NLM; (3) Wikipedia (http://en.wikipedia.org/wiki/Main_Page) — an online free encyclopedia edited collaboratively.

To extract relation of diseases and lab tests from the three on-line resources, relevant web pages of diseases and lab tests need to be retrieved first. LabTestsOnline provides an index of web pages about lab tests. All the web pages are downloaded based on the index. To retrieve web pages of diseases from MedlinePlus and Wikipedia, Application Programming Interfaces (API) provided by these two resources are used, with disease terms from UMLS as query input. The html formats of web pages retrieved from LabTestsOnline and Wikipedia were analyzed and stored in text files. Semi-structured XML files were retrieved from MedlinePlus, which could be analyzed directly.

Disease and lab test concept recognition

We used MetaMap(16) to process the free-text of each resource to recognize diseases and lab tests and mapped them to UMLS concepts. The disease concepts are restricted to certain semantic types in the UMLS semantic groups(17) such as disorder. The lab test concepts are restricted to the two UMLS semantic types: “laboratory procedure” and “Laboratory or Test Result”.

Disease and lab test concept normalization

As mentioned previously, all disease and lab tests extracted from LabTestsOnline, MedlinePlus or Wikipedia by MetaMap were represented as UMLS CUIs. The diseases were further normalized to International Classification of Diseases, Ninth Revision, Clinical Modification (ICD9-CM) codes, which are commonly used within most EHR systems. Lab tests were normalized into LOINC where possible using the tool Regenstrief LOINC Mapping Assistant (RELMA®) provided by LOINC1, as a reference.

Disease – lab test relation extraction

If a disease concept and a lab test concept co-occurred in the same web page, they are treated as a candidate relation pair. Furthermore, to enhance the precision, we further developed source-specific rules to add more constraints to relation extraction process, e.g., considering specific sections in the source web pages. The processing of each resource is described below:

LabTestsOnline: LabTestsOnline usually describes the usage of multiple related lab tests and the indications of test results in one lab test web page. Diseases indicated by the lab test results are listed in the section ‘What does the test result mean’. Hence, diseases occurred in this section are extracted and used to form pairs with the given lab test.

MedlinePlus: The information about diseases in MedlinePlus is well-organized as semi-structured XML files. A basic description of a disease or one type of diseases is first summarized, followed by sections of “Diagnosis&Symptom”, “Treatment”, “Prevention&Screening”, et al. Usually, each entry in these sections is one single named entity of lab test, symptom or therapy et al, which could be recognized by MetaMap straightforwardly. An information box of Medical Encyclopedia with related entries about the disease is also included, most of which are lab tests. Therefore, relation pairs are formed between diseases recognized in the title (or disease summary) and lab tests in the Medical Encyclopedia section.

Wikipedia: As a collaboratively edited open encyclopedia, the coverage of information provided in Wikipedia is unbalanced across different diseases. The sections in disease webpages are often more diverse, when compared with LabTestOnline and MedlinePlus. To extract précised relations, we limited our analysis of Wikipedia to the section

1 http://loinc.org/downloads/relma
‘Diagnosis’, including all the sub-sections within it. Only diseases in webpage titles and lab tests recognized from the ‘Diagnosis’ section are considered as associated.

Evaluation
The extracted disease lab test relations are evaluated using precision and recall, which are defined as follows: 

**Precision**, as calculated in Equation (1), measures the percentage of correct relation pairs among all relations extracted by our systems from each source. In this study, 100 relation pairs were randomly chosen from each resource and were manually checked by reviewing the original text.

\[ \text{Precision} = \frac{\sum \text{correctly extracted relation pairs by the system}}{\sum \text{all relation pairs extracted by the system}} \]  

(1)

**Recall**, as calculated in Equation (2), measures the percentage of system-extracted relation pairs among all relations defined in the gold standard. Here we used the disease lab test relation pairs in the book *Laboratory Tests and Diagnostic Procedures* by Chernecky and Berger (2007)(17) as gold standard, which contains information on common laboratory test results and their uses. To evaluate our approach, we created a subset of known disease-lab test relations from the above reference book, by randomly selecting five common diseases (from common health topics in WebMD²): Alcohol dependence syndrome, Acute myocardial infarction, Congestive heart failure, Rheumatoid arthritis, and Urinary Tract Infection, as well as five rare diseases (from the rare diseases list provided by the Office of Rare Disease Research of National Institute of Health³): Glucocorticoid deficiency, Multiple sclerosis, Malaria, Multiple myeloma, and Cystic fibrosis. In total, 112 relation pairs (64 for five common diseases and 48 for five rare diseases) were selected as gold standard for measuring the recall of the system.

\[ \text{Recall} = \frac{\sum \text{system - extracted relation pairs in the reference standard}}{\sum \text{all relation pairs in the reference standard}} \]  

(2)

In addition, we also compared overlaps among all disease-lab test relations extracted from each source.

Results
Table 1 shows the precision of each resource using one hundred randomly chosen relations. Wikipedia had the highest precision of 87%, while LabTestsOnline demonstrated a lowest precision of 76%. In contrast, recalls (displayed in Figure 1) shows the converse result. LabTestsOnline obtained the highest recall for both common diseases (38.10%) and rare diseases (34.09%). and Wikipedia was the lowest, with 23.81% for the common disease and 22.73% for the rare disease. The precision (81%) and recall (common disease 26.98%; rare disease 25%) of MedlinePlus were in the middle of the other two resources. As illustrated in Figure 1, the combination of all the three sources enhanced the recall sharply, with 57.14% for the common disease and 45.45% for the rare diseases, demonstrating that the relations of the three sources are complementary to each other.

|                | LabTestsOnline | MedlinePlus | Wikipedia |
|----------------|----------------|-------------|-----------|
| Precision      | 76%            | 81%         | 87%       |

Figure 1. Recall of LabTestsOnline, MedlinePlus and Wikipedia and their combination for common and rare diseases.

---

² [http://www.webmd.com/a-to-z-guides/common-topics/](http://www.webmd.com/a-to-z-guides/common-topics/)

³ [http://rarediseases.info.nih.gov/](http://rarediseases.info.nih.gov/)
Figure 2 illustrates the total numbers of diseases, lab tests and relations extracted from each resource and their overlaps. In total, 3,114 relation pairs consisting of 447 diseases and 300 lab tests were extracted from LabTestsOnline; 573 relation pairs consisting of 165 diseases and 193 lab tests were extracted from MedlinePlus; 2,023 relation pairs consisting of 604 diseases and 719 lab tests were extracted from Wikipedia. Examination of overlap among the three sources showed that 227 diseases, 238 lab tests, and 193 relations were extracted from at least two resources.

![Venn diagrams of diseases, lab tests and relations extracted from LabTestsOnline, MedlinePlus and Wikipedia.](image)

**Discussion**

Linking diseases and associated lab tests is valuable for clinical care and a wide range of clinical applications. This study is an initial effort to build a computable knowledge base for disease and lab test relations by integrating three on-line resources. Experimental results demonstrate a high precision for relation extraction, with Wikipedia achieving the highest precision of 87%. Combining the three sources achieved a recall of 51.40% as evaluated by the gold standard generated from a reference book. Moreover, among the ten diseases we explored, the three on-line resources also contributed five disease-lab test relations that are not included in the gold standard, e.g., the test of ‘Aldosterone, Urinary’ for Addison’s Disease extracted from MedlinePlus. These findings indicate that it is feasible to extract disease-lab test relations from such online sources and more importantly, relations in the on-line resources, as well as the gold standard reference book are complementary to each other for building a comprehensive disease and lab test relation knowledge base. Furthermore, our automatic method for relation extraction has advantages in terms of cost and time, as constantly updating an expert-curated knowledge base is time consuming and costly.

Based on our manual review, one of the major reasons for the false positive relations is the fact that the concept of disease or lab test is too general to make a clinically valuable relation. For example, when relations between Heart Diseases and lab tests are extracted from MedlinePlus, Heart Diseases describe a broad range of conditions, such as coronary artery disease, arrhythmias and congenital heart defects, to name a few. The lab test should be linked to a specific disease, instead of general categories. Similarly, some tests are expressed in broader terms in some well-known clinical cases. Blood smear required for malaria disease may just be referred as the general concept microscopy when mentioned later in the text, which in turn may not be properly traced back and encoded into the specific microscopic examination (i.e., blood smear). This issue of granularity is also one of the major causes for the low relation overlap between different resources. The general concepts of diseases and lab tests should be expanded or filtered out of the relation sets in the continuing studies. Mistakes in recognition of diseases and lab tests as appropriate named entities are another major cause for errors in both precision and recall. For instance, different sources use varying terminologies and prefer common language based on their target audiences. For example, the phrase “blood film” in Wikipedia refers to the test of blood smear. However, MetaMap failed to recognize it. Improving the performance of disease and lab test recognition may lead to a better relation extraction performance.

One limitation of this study is that currently the relations between diseases and lab tests are not defined in a fine granularity. A lab test could be used for differential diagnosis, prognosis, progress monitoring of a disease, etc. Our future plan is to refine the relation types in terms of the functional association of lab tests with the disease. In the next stage, we also plan to expand resources for relation extraction, including existing reference books and clinical manual and textbooks for medical education. Furthermore, machine learning based methods will be exploited to improve the performance of relation extraction.
Conclusion

In this study, three public on-line resources are integrated to create a freely available, computable disease-lab test resource using an automated text mining method. Experimental results indicate that it is feasible to automatically extract disease-lab test relations from the three online resources, and such online knowledge sources, as well as the gold standard reference book are complementary to each other for building a comprehensive disease and lab test relation knowledge base. The extracted disease-lab test relations are available for download from https://sbmi.uth.edu/ccb/resources/disease_labtest.htm.

Acknowledgement

This study was supported in part by National Institute of General Medical Sciences grants R01GM102282 and R01GM103859, National Library of Medicine grant R01LM010681, National Cancer Institute grant R01CA141307 and Cancer Prevention and Research Institute of Texas grant R1307.

References

1. Peterson MC, Holbrook JH, Von Hales D, Smith NL, Staker LV. Contributions of the history, physical examination, and laboratory investigation in making medical diagnoses. Western journal of medicine 1992;156(2):163
2. Van Walraven C, Naylor CD. Do we know what inappropriate laboratory utilization is?: A systematic review of laboratory clinical audits. Jama 1998;280(6):550-58
3. Singh H, Thomas EJ, Sittig DF, et al. Notification of abnormal lab test results in an electronic medical record: do any safety concerns remain? The American journal of medicine 2010;123(3):238-44
4. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nature Reviews Genetics 2012;13(6):395-405
5. Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. Nucleic acids research 2004;32(suppl 1):D267-D70
6. Schriml LM, Arze C, Nadendla S, et al. Disease Ontology: a backbone for disease semantic integration. Nucleic acids research 2012;40(D1):D940-D46
7. McDonald CJ, Huff SM, Suico JG, et al. LOINC, a universal standard for identifying laboratory observations: a 5-year update. Clinical chemistry 2003;49(4):624-33
8. Miller RA, McNeil MA, Chal linor SM, Masarie Jr FE, Myers JD. The INTERNIST-1/quick medical REFERENCE project—Status report. Western Journal of Medicine 1986;145(6):816
9. Halil K, Shin D, Fiszman M, Rosemb latt G, Rindflesch TC. SemMedDB: a PubMed-scale repository of biomedical semantic predications. Bioinformatics 2012;28(23):3158-60
10. Schulze-Kremer S, Smith B, Kumar A. Revising the UMLS semantic network. Medinfo 2004;1700-4
11. Wright A, Chen ES, Maloney FL. An automated technique for identifying associations between medications, laboratory results and problems. Journal of biomedical informatics 2010;43(6):891-901
12. Uzuner O, South BR, Shen S, DuVall SL. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. Journal of the American Medical Informatics Association
13. Ledley RS, Lusted LB. Reasoning Foundations of Medical Diagnosis Symbolic logic, probability, and value theory aid our understanding of how physicians reason. Science 1959;130(3366):9-21
14. Wright A, Bates DW. Distribution of problems, medications and lab results in electronic health records: the pareto principle at work. Appl Clin Inf 2010; 1: 32–37, 2010.
15. Wei W-Q, 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:amiajnl-2012-001431
16. Aronson AR, Lang F-M. An overview of MetaMap: historical perspective and recent advances. Journal of the American Medical Informatics Association 2010;17(3):229-36
17. McCray AT, Burgun A, Bodenreider O. Aggregating UMLS semantic types for reducing conceptual complexity. Studies in health technology and informatics 2001(1):216-20
18. Chernecky CC, Berger BJ. Laboratory tests and diagnostic procedures: Elsevier Health Sciences, 2007.