Identification and Extraction of Family History Information from Clinical Reports

Sergey Goryachev, MS¹, Hyeoneui Kim, RN, PhD¹, Qing Zeng-Treitler, PhD¹,
¹Decision Systems Group, Brigham and Women’s Hospital, Harvard Medical School, Boston, MA, USA

Abstract

Many clinical reports contain family history, which is valuable information for clinical decision support and research. We developed a simple natural language processing algorithm to identify and extract family histories. The algorithm was tested on a set of discharge summaries and outpatient clinic notes. The sensitivity and specificity of detecting family history information was 97.2% and 99.7%, respectively. The accuracy of exact family member assignment was 97.1%.

Introduction

A number of natural language processing (NLP) applications have been developed to extract key findings such as past and present diagnoses for point-of-care decision support as well as clinical research [1-5]. As part of the National Center for Biomedical Computing, Informatics for Integrating Biology & the Bedside (I2B2) [6], we developed an open-source and modularized natural language processing system: the Health Information Text Extraction (HITEx) System [7]. HITEx is a suite of open source NLP tools, written in Java, which builds on top of the General Architecture for Text Engineering (GATE) framework [8].

We have used HITEx to parse discharge summaries and outpatient visit notes [7]. One challenge which we encountered was that family history is sometimes mixed with a patient’s own history and findings in the same section, paragraph, or sentence. In order not to report a patient’s family history as the person’s own diagnosis, it is necessary for us to differentiate the two types of information.

Also, while family histories may be considered false positives in the context of diagnosis extraction, they provide valuable, and sometime critical information for patient care and scientific research [9]. Breast cancer risk prediction models, for example, often include family history of breast cancer as a key variable and as a surrogate for genetic information [10]. Therefore, it is not only necessary to distinguish family history from personal history, but also to capture the details of family history (e.g., grandmother and mother with hypertension, versus a cousin with hypertension).

To identify and extract family history information, we have developed a simple rule-based algorithm. For evaluation, this algorithm was applied to 1,000 sentences which were randomly selected from the discharge summaries and outpatient clinic notes of the Massachusetts General Hospital (MGH) and Brigham and Women’s Hospital (BWH).

Methods

The family history extraction process consists of three main steps: pre-processing, family member and finding concept identification, and family member/patient assignment.

Pre-processing

First, clinical reports are split into sections (e.g., diagnosis, history, and medication), and section headings are coded using a locally developed taxonomy. Second, content of each section is tokenized and split into sentences. Third, noun phrases are extracted after part-of-speech processing. Fourth, noun phrases are mapped to Unified Medical Language System (UMLS) concepts [11].

In this study, all the pre-processing tasks are conducted using existing HITEx components.

Family member and finding concept identification

Because our main interest is to assign various diseases or findings to the correct person (a family member or the patient), UMLS concepts that fall into these two categories (finding and family member) are tagged as such.

Family member concepts are mainly identified by one UMLS semantic type: family group (T099).
Figure 1. The association rules used to assign findings to family members.
We define Finding as concepts which belong to one or more of 8 UMLS semantic types:

1. congenital abnormality (T019)
2. acquired abnormality (T020)
3. injury or poisoning (T037)
4. disease or syndrome (T047)
5. mental or behavioral dysfunction (T048)
6. cell or molecular dysfunction (T049)
7. anatomical abnormality (T190)
8. neoplastic process (T191).

Family member/patient assignment

In the last step, a set of rules is used to associate the finding or group(s) of findings with the most relevant family member or group(s) of family members. Besides family members and findings, the rules employ 3 other high level annotations:

- **Conjunction** – tokens that may indicate the end of a family history-related phrase, but may also be a part of such phrase. Examples of conjunctions include “,” (comma) and “and”.
- **Sentence boundary** – tokens that identify the sentence boundaries, for example period (“.”).
- **Patient possession** – a token or group of tokens that indicates with a high probability that the sentence describes patient, not family history findings. For example: “patient had” or “patient has”.

Tokens or concepts which are not classified as family member, finding, conjunction, sentence boundary or patient, are ignored by the association rules. The rules are represented in Figure 1.

Evaluation

The family history identification and extraction algorithm is implemented in Java and as a GATE module. For evaluation, this module is used along with existing HITEX modules to form a family history extraction pipeline, shown in Figure 2.

A total of 2000 reports were randomly selected from the Partners Research Patient Data Registry (RPDR) [12]: 500 discharge summaries from Brigham and Women’s Hospital (BWH), 500 discharge summaries from Massachusetts General Hospital (MGH), 500 outpatient notes from BWH, and 500 outpatient notes from MGH. We included these two types of reports from two different hospitals to provide better coverage of different writing styles.

<table>
<thead>
<tr>
<th>Text document</th>
<th>Sectionizer</th>
<th>Noun Phrase Finder</th>
<th>UMLS Concept Finder</th>
<th>Family History Finder</th>
<th>UMLS concepts with assigned family history</th>
</tr>
</thead>
</table>

While most report sentences are fairly short and straightforward, the sample does contain some challenging sentences. Table 1 provides a few examples:

| One of her daughters stated that she had a small blood clot in her urine 2 days ago. |
| As a child, the patient had rheumatic fever and had a residual murmur of mitral regurgitation since that time. |
| OCD, brother diagnosed last year, not currently receiving medication or psychotherapy secondary to noncompliance. |
| Positive for coronary artery disease (father died of an MI at age 75 y/o, mother and brother have had diagnoses of coronary artery disease as well), peripheral vascular disease (brother died of a ruptured abdominal aortic aneurysm less than one year ago), hypertension (parents and brothers), and hyperlipidemia (brothers). |

Table 1. Examples of report sentences; findings and indicators of family/patient history are highlighted.

The records were parsed by the family history extraction pipeline application. In this evaluation,
only sections with history-related titles (e.g., “history”, “family history”, “history of present illness”, or “social history”) were used. (As a part of HITEx development, we have identified more than 1000 section headers and mapped them to section categories, according to an internally developed taxonomy. Dozens of sections were categorized as history-related).

All the findings extracted by the pipeline are assigned to either patient or one or more family members. We randomly selected 1000 findings from the 9725 extracted findings for human review. Reviewers answered two questions:

- Is the finding related to the patient or patients’ family?
- If the finding is family related, who are the family members involved?

Two reviewers (one clinician and one non-clinician) performed the review. Differences between the two reviewers were minimal (< 3%) and resolved through discussion.

Results

Out of the 1000 findings, 7 were removed from the analysis because they were the results of wrong mappings. For example, the word *cut* in the following sentence was incorrectly mapped to the concept *incised wound*: “The patient was able to go about getting a hair cut and did some shopping”.

<table>
<thead>
<tr>
<th>Finding is assigned to</th>
<th>Finding is for</th>
</tr>
</thead>
<tbody>
<tr>
<td>Family member</td>
<td>Patient</td>
</tr>
<tr>
<td>Family member</td>
<td>70 (TP)</td>
</tr>
<tr>
<td>Patient</td>
<td>3 (FP)</td>
</tr>
</tbody>
</table>

Table 2: Confusion matrix for Family History Assignment

The family history extraction pipeline performed well in assigning findings to patients or family members (Table 2). It achieved 97.2% sensitivity and 99.7% specificity in detecting family history findings. We found that errors were likely to occur in the sentences that:

- Are complex, with multiple groups of findings or multiple groups of family members, e.g., “The cardiac risk factors included post-menopausal, hypertension, diabetes mellitus, questionable cholesterol, smoking, but no family history”.
- Contain ambiguous statements, e.g., “The patient notes that there have been ill contacts in the family, including children who have upper respiratory infection symptoms”.

Among those findings which were correctly classified as family history, the accuracy of specific family member assignment was high (97.1%). We encountered two types of errors: failure to include all family members associated with a finding and generalization, i.e., a finding that should have been assigned to a particular family member was generally classified as family history.

Discussion

Family history is an important type of clinical information for patient care as well as scientific research. This paper presents a new algorithm for identifying and extracting family histories from free-text clinical reports.

The algorithm was evaluated using a set of discharge summaries and outpatient notes. Although the algorithm is relatively simple, it demonstrated a good ability to detect family history findings (97.2% sensitivity and 99.7% specificity). It has also achieved 97.1% accuracy in the assignment of findings to specific family members. In this evaluation we didn’t examine the negation status or temporal status of findings, which could be extracted by adding HITEx Negation Finder or Temporal Finder modules to the pipeline.

A related prior study by Friedlin et al [13] reported high accuracy rate (sensitivity = 93% and positive predictive value = 97%) in the extraction of family histories. Friedlin’s study differs from ours in several respects: 1) it focused on sections clearly labeled as family history, thus did not need to differentiate family from patient history; 2) it classified family members as primary, secondary or unknown relatives, while we assign findings to exact family members (e.g., father, mother, or sister); 3) although Friedlin’s algorithm was only described very briefly, it appears to be somewhat different from ours as we consider the use of conjunction words and symbols.

In the BWH and MGH’s discharge summaries and outpatient notes which we have analyzed, family history information is not always documented under sections clearly labeled as “family history”. We suspect this may not be a unique problem of the BWH and MGH reports. The ability to distinguish family history from patient history could help to reduce false positives when extracting a patient’s past and present diagnoses from such reports. It could
provide valuable family history information for data mining and hypothesis testing.

A significant limitation of the algorithm is that it is error-prone when handling complex and ambiguous sentences. It also does not have the ability to resolve co-reference. For example, the algorithm would assume the “she” in the sentences “Mother had MI two years ago. She is also treated for hypertension,” to be the patient, if these sentences were to appear in a general history section. If the patient is male, however, “she” was clearly referring to the patient’s mother.

We have incorporated the family history module into HITEx. For future work, we plan to further test and refine the module.

Conclusion

Family history information found in clinical reports is valuable for many applications such as breast cancer risk prediction. Correct identification and extraction of family history presents a challenge for NLP systems, given the heterogeneity of clinical reports.

We have developed a simple, rule-based algorithm to extract family history from clinical reports. It was tested on 2000 clinical reports from two different institutions and showed good sensitivity and specificity in detecting family history. It also achieved high accuracy in exact family member assignment.

Acknowledgment

This research was funded by NIH grant number U54 LM008748. We thank Drs. Shawn Murphy, Ross Lazarus, Susanne Churchill and Isaac Kohane for their advice and collaboration.

References