Family history information is an important part of understanding a patient's total health, and it is spread in text throughout the clinical record; mostly untagged and often very unstructured. We have performed a syntactic analysis of family history in 1274 sample clinical texts and created an algorithm to extract specific family history information. Using the Stanford NLP Parser, we created patterns of dependency relations to map specific family members to specific diseases. Preliminary results are promising with a precision and recall of .82 and .52, respectively. We find that this is of sufficient accuracy to drive meaningful, actionable clinical presentation of this information.

1 INTRODUCTION

One key component to a clinician’s understanding of a patient’s total health is a comprehensive view of the patient’s family history. Numerous studies show that patient history guides choices of what diseases to consider when performing a diagnosis, what tests and labs to run, and what treatments may or may not work as well. In this way family history begins to touch upon the same areas a mature genomic solution will.

While there are broader definitions for family history, for this analysis, we will consider a family history to be a diagnosis that is specific enough to be described by a 3 digits or longer International Classification of Diseases 9 code (ICD9). Furthermore, for our purposes the disease must be attributed to a specific family member. For example, we consider "Mother has MI" (myocardial infarction) to be a family history sentence, but exclude "Mother's side has MI". Likewise a diagnosis of "Grandmother died of old age." would be disregarded as lacking disease specificity. This requirement for disease specificity also excludes indirect means suggestive of a particular diagnosis, for example mentions of medical procedures or medications.

One challenge to employing this information is that a patient's family history may be spread throughout the patient's clinical records. Modern health care systems may include structured entry of such information either through paper forms or on-line programs that patients are required to fill in. However, much information is simply recorded in a progress note by a clinician author after an encounter with a patient where the information comes up.

For this work we sought an openly available source of clinical record examples thus allowing reproducibility of results. For this we have employed the medical transcription sample reports provided at http://www.mtsamples.com. These reports cover a total of 35 medical specialties (e.g., general medicine and surgery) and are more than just discharge summaries. It is clear that the family history information in these reports are often results of basic background information that a clinician might ask about, and thus are somewhat disease specific (e.g., "Do either of you parents have a heart condition?").

Given this spread, we believe that it is important to look at the extraction of family history information from all textual sources in the clinical records. This includes comments that may not be in an official “family history” (or FHX in some clinical reports) section of a report. In the latter case, a patient’s family history and his own symptoms are often mixed together without clear distinction in writing. We devise a simple yet effective strategy to identify family history. Additionally, informal writing and phrases are often utilized to describe family history (e.g., mother with diabetes). We adopt a dependency parser instead of a traditional phrase parser to avoid over-fitting.

Specifically, our approach is a two-step method that selects candidate family history sentences based on the presence of “sign-post” words (e.g., "mother", "brother"), and then employs a set of dependency-based syntactic patterns to both extract appropriate diagnosis and identify the family member referred to. We then map the diagnosis to a clinical taxonomy (ICD9) to allow multilevel display of patient information (see Figure 2). We currently do not perform cross-sentence anaphoric or co-reference resolution, but this is clearly important as well.

In the remainder of the paper we will first review related work in family history extraction, then outline our experimental setup and characterize the prominent linguistic features of family history descriptions. We next describe the algorithm and present the main results. Finally, we conclude with some thoughts about how this technology might be more broadly applied.

2 RELATED WORK

Most previous work in this area focuses on extracting
family history information from specific, focused sections of a single class of report. For example, Freidlin and McDonald [2] performed this extraction by first processing the documents with Health Information Text Extraction (HITEx) [16] to map noun phrases to the Unified Medical Language System (UMLS) concepts of family member and diagnosis [8]. The diagnoses were restricted to 8 concept groups. Family history sentences were identified by titles with a family history type phrase (e.g., “family medical history”). They then used a series of co-occurrence based rules [10, 12, 13] such as “Are family members followed by the diagnosis?” to associate diagnoses to family members. They trained their algorithm on 2000 discharge summaries and outpatient visit notes from two hospitals. Their approach achieved a precision and recall of around 0.92 when considering a relatively low number (28) of family history related diagnosis. In contrast, our analysis included 167 specific family history diagnosis.

Goryachev et al [3] also performed a family history extraction. Their data set was limited to 1337 admissions notes, and they also focused on titled sections of family history. Their “proximity phrase” algorithm yielded a positive predictive value of .96 and sensitivity of .93 for connecting 12 designated diseases.

The proposed approach in this article distinguishes itself from these previous works in the following three aspects: (1) Unlike previous attempts that focused on titled sections of clinical texts, our approach is designed to extract family histories from any section within a clinical report. (2) Our approach does not rely on heuristic rules such as co-occurring patterns to associate a diagnosis with a family member. Instead, it relies on the syntactic and semantic dependencies embedded within a sentence. Finally, (3) we have adopted a much stricter criterion for evaluating the correctness of the extracted family-diagnosis pairs: regardless of the number of diagnoses or family members found in a sentence, each family member must be mapped directly as the object of a particular diagnosis as opposed to grouping primary relatives to a group of diagnosis.

Note that using syntactic or semantic patterns to identify the relationships between two entities has been widely applied in the biomedical field. For instance, in [12], the predicate-argument linguistic structure [15] is employed to find the relationship between drugs and genes. In [9], a framework is described to analyze radiology reports using syntactical patterns constructed through full parsing analysis. Excellent reviews can be found in [5, 7].

3 DATASET

To extract and divide the sentences, the General Architecture for Text Engineering (GATE) ANNIE processors along with simple JAPE grammars were used to identify sentences that contain a family member. To separate sentences according to our specific definition of family history, the sentences were manually separated by a member of this team.

3.1 Family History Distribution

For evaluation we seek to use a publicly available dataset rather than an ‘internal’ one. To this end we have selected 1274 sample medical transcription reports from mtsamples.com.

<table>
<thead>
<tr>
<th>Feature from mtsamples</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total reports</td>
<td>1274</td>
</tr>
<tr>
<td>Percentage of sampled family member sentences that contain a diagnosis</td>
<td>27.9%</td>
</tr>
<tr>
<td>Number of reports that contain a mention of a family member</td>
<td>263</td>
</tr>
<tr>
<td>Number of non duplicate sentences that contain a family member</td>
<td>380</td>
</tr>
<tr>
<td>Percentage of family member sentences that contain a family member</td>
<td>39.6%</td>
</tr>
<tr>
<td>Average number of diagnosis per sentence with a family member</td>
<td>.68</td>
</tr>
<tr>
<td>Average number of diagnosis in sentences with or without family member</td>
<td>.25</td>
</tr>
</tbody>
</table>

Table 1. Dataset Sample

We note that many occurrences of a family member (see Table 2 for a list of family members) are non-diagnostic (e.g., “Mother took the patient to the appointment”). These reports in general have no internal structure (e.g., tables) and have a tendency to terseness.

A total sample of 597 unique sentences was randomly selected from the mtsamples documents, among which 299 contain a family member, and 298 do not. The sentences were further divided into two groups, one where family members were recorded as having a disease, and the other where the patient was. It has been assumed that all diagnoses that are not explicitly referring to a family member are referring to the patient (e.g., "diagnosis: tumor").

With this separation of data, it became apparent that when a family member was mentioned, there is a higher incidence of a diagnosis mentioned. Furthermore, 39.6% of the time a family member is mentioned, the sentences contains a family history. This high coincidence of family members and diagnosis provides justification to base our family history extraction algorithm around sentences containing family members, not just title sections of clinical texts.
3.2 Characteristics of Family Mentions

The simplicity of family history sentence suggests that a syntactic pattern approach would be useful for extracting family histories from clinical text. A dependency and verb analysis was performed using the Stanford NLP Parser [7] and GATE Verb Chunker and can be seen in Table 3. The structure of family history sentences tend to be much simpler than those that are not family history sentences. Family history sentences tend to have the main verb in a simple tense form (e.g., "had" or "has"). Also noted is the prevalence of two verbs in the family history that appear much less in other sentences, mainly "died" (e.g., "mother died of breast cancer), and some form of the verb "to have" (e.g., "Father has diabetes"). Furthermore, when a family member is mentioned, a simple unambiguous diagnosis is more likely to be mentioned if the diagnosis refers to the family member.

<table>
<thead>
<tr>
<th>Observation</th>
<th>Family History</th>
<th>Non Family History</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of collapsed dependencies per sentence</td>
<td>9.5</td>
<td>13.15</td>
</tr>
<tr>
<td>Number of diseases per sentence</td>
<td>1.43</td>
<td>0.18</td>
</tr>
<tr>
<td>% Verbs “have”</td>
<td>32.40%</td>
<td>9.50%</td>
</tr>
<tr>
<td>% of Verbs “died”</td>
<td>21.00%</td>
<td>1.40%</td>
</tr>
<tr>
<td>% Simple tense</td>
<td>95.90%</td>
<td>65.90%</td>
</tr>
</tbody>
</table>

Table 3. Syntactic Analysis

3.3 Disease Analysis

To help understand the restrained way in which clinicians document family history, we also performed a disease distribution analysis. We grouped the diseases mentioned into 7 categories based on our observation of groupings in the family history sentences: diabetes, cancer, hypertension, heart problems, substance abuse, stroke, and others. The distribution can be seen in Figure 1.

4 ALGORITHM

4.1 High Level Description

The tendency of family history sentences to have simple structures and relatively small number of dependencies justifies a syntactic pattern based algorithm for family history extraction. Furthermore, because of the descriptive nature of clinical texts, sentences tend to be ungrammatical and incomplete, so we have selected the Stanford Dependency Parser [15] to facilitate syntactic pattern construction and identification. Given a sentence, this parser statistically establishes dependencies between two words, one as governor and and the other as dependent. For example, the collapsed dependency in “died of cancer” is prep_of(died, cancer), where “died” is the governor, and “cancer" is the dependent. By focusing on the binary relationships between words, we can be less concerned with the grammatical incompleteness of clinical texts. Conversely, if we were using a constituency parser (i.e., one that maps out syntactic phrase structure trees) then the grammatical incompleteness of family history sentences in clinical texts (indeed, perhaps all sentences in clinical texts) would be a major obstacle. Therefore, the main approach of our family history algorithm is to link a family member to its diagnosis through a small chain of collapsed dependencies found with the Stanford Parser. This will allow us to avoid the syntactic structure of the entire sentence and focus on the dependency patterns of the phrases that contain family history.

Currently, we have collected 6 collapsed dependency patterns. These patterns were constructed from observing an annotated training set, in which dependency-links appeared often as a connection between a family member and a disease. The distance between family member and disease was often 3 or less dependencies, and had a simple grammatical explanation (as opposed to obscure and lengthy dependency chains that would lead to overfitting).

One benefit of this approach is that we do not have to know the disease that is mentioned; we only need to start with the family member and use these collapsed dependency paths to find a disease.

4.2 Implementation

The first step in our approach is to find if a sentence
has a family member. Next we determine if a particular dependency exists within the parsed sentence (e.g., nsubj) and determine whether a family member is a governor or dependent of the dependency, and then we proceed to the next link in the chain. The 6 dependency patterns are below. To describe the dependency chains, the first pattern will be described very thoroughly, and the remaining will only show dependencies mapped to an example sentence, attempting to show the link from family member to disease.

**Rule 1: Nsubj – Doj**

This dependency chain extracts simple subject – direct object relationships from sentences. It is useful for extracting the family history from simple phrases with “to have” as the main verb. For this pattern, three conditions must be met: 1) a nominal subject (nsubj) dependency must exist, 2) the dependent of the nsubj dependency must be a family member, and 3) the governor of this dependency must be some form of the verb “to have”. This means that a family member is the head noun of the subject of a clause which is predicated by the verb “to have”, which was chosen because of its strong presence in the verb distribution of family history sentences as seen in Table 3. Next, if a direct object (dobj) dependency exists and the governor of this dependency’s index matches the index of the governor of the nsubj dependency, then we assume that the dependency of the dobj relation is a disease, and it is paired with the family member found initially. Below is a display of the sentence and dependence relation.

```
“Mother has diabetes”
nsubj(had-2, Mother-1)
dobj(had-2, diabetes-3)
```

**Rule 2: Nsubj – Pobj**

This chain is similar to the one above in that it searches for a nsubj dependency. However, it is useful for extracting simple “died of”' phrases. Instead of using a dobj collapsed dependency, a prep_of or prep_from collapsed dependency is used instead.

```
“Father died of cancer”
nsubj(died-2, Father-1)
prep_of(died-2, cancer-4)
```

**Rule 3: Prep-With**

This is another simple dependency chain meant to extract from sentences with simple listings of family members and diseases (e.g., “mother with cancer, father with diabetes).

```
Mother with diabetes  prep_with(mother-1, diabetes-3)
```

**Rule 4: PartMod**

This dependency stands for participle modifier and identifies participle form verbs that modify a noun or verb phrase. This dependency usually surfaces with only a few participle modifiers: associated, suffered, had, died, and assessed. Each modifier usually has a particular prepositional phrase after it (e.g. “associated with”, but never “associated on”). This pattern combines the participle modifier with its appropriate prepositional phrase. Some times the participle modifies the diagnosis, and sometimes the family member. The algorithm looks for both:

```
“..heart attack associated with
died of”
partmod(attack-2, associated-3)
pattWith(associated-3, father-5)
```

**Rule 5: RCMod**

This dependency identifies the head of a relative clause which modifies a noun phrase, usually verbs. Sometimes the verb in the relative clause modifier has direct objects, and sometimes it has prepositional objects

```
“mother who has
died of”
rcmod(mother-1, has-3)
dobj(has-3, diabetes-4)
```

**Rule 6: Conj_And**

This dependency is a post-processing step that is used once the chain of dependencies reaches an object. It helps identify multiple diagnoses for a family member, or multiple family member for a diagnosis.

```
```

5 RESULTS

Our family history extraction algorithm was trained on a data set of 299 sentences that mention one or more family members. Of these sentences, 77 contained family history information specific to our definition above. That is, each sentence contained one or more family members that were unambiguously the object of one or more diagnoses that can be specified to a 3 digit ICD9 code. The total number of these family-diagnosis pairs was 167. All results were hand scored to create a gold standard.

The score metric was based on a strict case insensitive string matching of the complete diagnosis noun phrase,
which may have more than one word (e.g., “Myocardial Infarction”), and the proper family member pair, which is always one word. For example, “Father died of lung cancer, mother with diabetes and hypertension” must return three pairs of strings: “father”-“lung cancer”, “mother”-“diabetes”, and “mother”-“hypertension”.

Preliminary results show promise, but also needed work. The algorithm was run on all 299 sentences. Of the 167 pairs determined, 100 returned true positive, 26 returned false positive, and 70 where completely missed. This resulted in a precision of .79 and recall of .58 on the training set.

Below is a distribution of the dependency patterns.

<table>
<thead>
<tr>
<th>Dependency Pattern</th>
<th>TP</th>
<th>FP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nsubj-Dobj</td>
<td>28</td>
<td>10</td>
</tr>
<tr>
<td>Nsubj-Pobj</td>
<td>25</td>
<td>5</td>
</tr>
<tr>
<td>Prep-With</td>
<td>20</td>
<td>3</td>
</tr>
<tr>
<td>PartMod</td>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td>RCMod</td>
<td>11</td>
<td>4</td>
</tr>
<tr>
<td>Conj_And</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>Total</td>
<td>100</td>
<td>26</td>
</tr>
</tbody>
</table>

Table 4. Distribution of Dependency Patterns

The precision can be explained by the table above. Over half of the dependency patterns extracted are the Nsubj-Dobj, and the Nsubj-Pobj dependency patterns. This follows after our previous syntactic analysis which noted high distribution of the verb “died” and some form of “to have”. The remaining patterns reflected the simple syntactic structure of family history sentences.

The lower recall of this algorithm also reflects the unpredictability and inconsistency in clinical text documents, the fragility of NLP syntax parsers when encounter descriptive text, and the inflexibility of the algorithm. Of the false negatives, 77% contain more complex intermediate syntactic structures between the family member and the diagnosis (e.g., “…sister interestingly had arthritis”, “uncle having problems with asthma”). The clauses have adverbs, gerunds, pronoun references, all of which are completely ignored in our algorithm. The remaining 22% of false negatives had complex noun phrase structures for both diagnosis and family member (e.g., “brother with bone and throat cancer”). While a post processor sought to address these dependencies, it was unable to catch all of them.

The patterns were also run on a test set, which was annotated after all the patterns had been determined and the algorithm fixed. The test set consisted of 77 sentences mentioning one or more family members. This algorithm performed (as expected) less well than on the training set with a precision of .61 and recall of .51.

For comparison, we implemented a version of the family history extraction algorithm from [3] and ran it on our training set. While they tag their diagnosis priori to performing their algorithm, we do no such tagging. Therefore, to run on our scoring metric, we fed the algorithm all of the diagnosis to see how well it would assign them to family members. The results of the family history extraction algorithm from [3] on our test set were a precision of .47 and recall of .71. This result is expected as the algorithm assigns groups of diseases to groups of family members, which will have many false positives when assigning each diagnosis to each family member, decreases the precision.

While our algorithm produces low recall, we view it as promising and are optimistic that we will find many other dependency paths to traverse. Furthermore, we find it a good algorithm because, unlike [2] and [3], it does not need the diagnosis a priori through a disease annotator, it is not dependent on a family history section of the clinical text, and it has a very strict scoring metric.

6 VISUALIZATION

Once family history is spotted and aggregated it needs to be presented in a way that clinicians can rapidly make use of the information. We have experimented with two different visualization schemes – one focusing on diseases and the other on familial relation (Figure 2).

The disease focus allows roll-up of the diseases according to the ICD-9 coding hierarchy. This allows an “at-a-glance” view of what might be important for a particular patient. Once drilled down as far as possible it is important to show the text snippet (sentence in our case) that leads to this conclusion. This allows the clinicians to rapidly validate the spot, and to understand the context around the information.

While useful, this “focus on the disease” methodology is different from the traditional way in which doctors see such information. It is more common for it to be aggregated by family member – showing all the relevant diseases that a family member has been diagnosed with.

We have explored both visualizations and received preliminary (positive) feedback from clinicians. In cases where family history is an important factor to consider in a diagnosis, having it easily assessable is an important feature. Major Electronic Health Record (EHR) systems are including such tracking as a system, but anecdotal evidence suggests that these features are rarely populated with information that preexists in the clinical report.

Either as a supplemental visualization, or as an ETL
7 CONCLUSIONS

Family history information is necessary to understanding the total view of a patient. We have shown that patient history is often scattered throughout the patient records, requiring analytics that can spot such sentences wherever they reside.

In this work, through examining 1274 clinical records, we have identified several salient linguistic features used by clinicians when recording information about a patient’s family. We have developed syntactic pattern-based extractors that operate at a reasonably high precision, and with a sufficient recall that supports further endeavor for exploring the effectiveness of such extraction systems in a clinical setting. Lastly, we have developed visualizations that transform this information into an actionable form for presentation to a clinician.

While there is clearly more work to be done (especially in increasing the recall of such “free-form record” extraction systems), the approach holds much promise for allowing clinicians to better make use of this crucial component of a patient’s overall health picture.

8 REFERENCES