Towards Comprehensive Longitudinal Healthcare Data Capture

Delroy Cameron*, Varun Bhagwan† and Amit P. Sheth*

* Kno.e.sis Center, Wright State University, Dayton, OH 45435, USA
† IBM Research, San Jose CA 95120, USA
{
delroy, amit}@knoesis.org, {vbhagwan}@us.ibm.com

Abstract—The ability to connect the dots in structured background knowledge and also across scientific literature has been demonstrated as a critical aspect of knowledge discovery. It is not unreasonable therefore to expect that connecting-the-dots across massive amounts of healthcare data may also lead to new insights that could impact diagnosis, treatment and overall patient care. Of critical importance is the observation that while structured Electronic Medical Records (EMR) are useful sources of health information, it is often the unstructured clinical texts such as progress notes and discharge summaries that contain rich, updated and granular information. Hence, by coupling structured EMR data with data from unstructured clinical texts, more holistic patient records, needed for connecting the dots, can be obtained.

Unfortunately, free-text progress notes are fraught with a lack of proper grammatical structure, and contain liberal use of jargon and abbreviations, together with frequent misspellings. While these notes still serve their intended purpose for medical care, automatically extracting semantic information from them is a complex task. Overcoming this complexity could mean that evidence-based support for structured EMR data using unstructured clinical texts, can be provided.

In this work therefore, we explore a pattern-based approach for extracting Smoker Semantic Types (SST) from unstructured clinical notes, in order to enable evidence-based resolution of SSTs asserted in structured EMRs using SSTs extracted from unstructured clinical notes. Our findings support the notion that information present in unstructured clinical text can be used to complement structured healthcare data. This is a crucial observation towards creating comprehensive longitudinal patient models for connecting-the-dots and providing better overall patient care.

Keywords—Text Analytics, Text Mining, Unstructured Text, Semantic Type Extraction

I. INTRODUCTION

The notion of connecting-the-dots has been discussed extensively in the literature in terms of the utilization of background knowledge for detecting Conflict of Interest relationships [1], Insider Threat [2] and ranking of Semantic Associations [3]. In scientific literature, connecting-the-dots has been illustrated in the context of finding undiscovered public knowledge [4], largely due to seminal work by Don R. Swanson, which led to the field of Literature-Based Discovery (LBD). Swanson postulated that while two concepts (A,C) may appear unrelated in the literature, some intermediate concept (B) between them may provide new insights into unknown yet important associations. Recent research efforts [5], [6] have built on this notion by suggesting the use background knowledge to bridge important knowledge gaps in the literature, when the literature alone is insufficient.

In the healthcare domain, connecting-the-dots has implications on overall patient care, and could require an interplay among structured EMR data, unstructured clinical text as well background knowledge. Since the use of background knowledge in healthcare is still in its infancy however, the immediate focus is on the correlation between structured and unstructured data. Interestingly, while EMRs provide easy access to structured elements of patient records, a vast variety of medical data pertinent to patient care is of unstructured form, available as progress notes, discharge summaries, radiology reports, cardiac reports etc.

A knowledge of detailed information from such sources could be crucial. For example, consider PatientA who has been residing with chronic smokers for four years, who all smoke indoors. Consider also PatientB, mainly exposed to cigarette smoke on weekends when socializing with friends. Both patients are likely to be classified as passive smokers in the EMR. However, clearly PatientA is at much higher risk for cardiovascular disease than PatientB. Not only is it important to have an overall sense of a patient’s smoker status, but also to analyze the evidence and context from which such classification was derived. Structured EMRs alone are largely insufficient for providing this context.

In this work therefore, we address the problem of extracting Smoker Semantic Types (SST) along with supporting evidence (called type indicators) from unstructured clinical text. Such information enable complementing SSTs from structured EMRs, thereby facilitating the creation of more holistic patient records. Further, temporal analysis of SSTs from both structured and unstructured data, assist the construction of longitudinal patient views, which are critical for connecting-the-dots and providing the knowledge need to deliver better overall patient care. Our specific contributions are therefore as follows:

- We present a pattern-based method for extracting Smoker Semantic Types from unstructured clinical text.
- We leverage the SSTs obtained from the unstructured texts to bring information to structured EMRs to create more holistic patient records
- We also discuss the utilization of SSTs from both structured and unstructured sources for the creation of longitudinal patient views.
II. RELATED WORK

Automatic SST classification using unstructured clinical text has been studied extensively in the literature, mainly in the i2b2 smoking challenge [7]. The vast majority of classification techniques exploit syntactic and lexical features, together with various rules and heuristics, including machine learning algorithms to identify smoker references and then perform classify patients. Aramaki et al. [8] describe a two-step classifier that uses heuristics to extract sentences containing smoker references. A combination of the Okapi-BM25 and k-Nearest Neighbor (k-NN) classifiers are then used for overall classification, based on word similarity between input sentences and a training set. Only the last sentence is used when a document contains multiple sentences with smoking references.

Savova et al. [9] present a layered approach that first identifies lexical smoker features from sentences, then systematically applying Support Vector Machine (SVM) classifiers to filter out nonsmokers. By eliminating nonsmokers, Savova treated the remaining classification task as a temporal classification task to identify past and current smokers. Heuristics determine the overall document classification based on perceived strengths of the various SST categories in a document.

Cohen [10] also discuss a four-layered approach which first identifies text in the vicinity of smoker references which he calls ‘hot spots.’ Using fixed a length window ±100 characters before and after hot spots, to delimit the boundary containing actual smoker phrases, the tokenized and normalized text is then classified using an ensemble of SVM classifiers which perform error correcting and heuristics to generate the final SST classification.

Szarvas et al. [11] describe a syntactic-based approach in which high frequency smoking phrases are first identified then several classifiers including k-NN, SVM, AdaBoost, Decision Tree etc, are used for classification. This approach exploits text linguistic features such as part-of-speech (POS) tags, verbs and negations as input to the classifiers.

Clark et al. [12] present two techniques based on SVM classifiers. In the first approach, the classifiers classify documents based on all smoking phrases, while in the second, smoking phrases are classified individually, then heuristics are used for overall document classification.

Wicentowski et al. [13] reported on a rule-based approach for SST classification, but limited provide limited details on the approach in the manuscript. Instead they discuss the performance of a Naive Bayes classifier smoke-blind dataset, in which all smoking terms have been removed.

Heinze et al. [14] discuss a four-component classification system called LifeCode, also based on lexical and syntactic features such as segmentation, lexical analysis, phrase parsing. In the fourth component vector analysis is used to assign overall SST classification to smoker phrases.

In this work, we develop a pattern-based approach for SST classification. We use a declarative information extraction system called SystemT and its Annotation Query Language (AQL) specification (discussed in Section III-B1). AQL enables building complex SQL-like queries and views that to extract complex patterns from unstructured text, requiring only a knowledge of frequently occurring smoker and temporal phrases within the corpus. Using this framework, we are able to identify, associate and maintain provenance for SSTs based on flexible, adaptable and extensible rules.

Our work is also significant because not only do we develop techniques for type indicator extraction and SST classification, but we also compare the extracted SSTs to structured EMR SSTs for the same patient records. Hence we are able to complement the EMR data in cases where SSTs align, and provide evidence for resolution in cases when they do not. This is particularly important in scenarios where either a patient’s SST is unknown or a patient claims to have quit smoking (or to have never smoked), when the unstructured text provide conflicting evidence.

Finally, a natural consequence of aligning SSTs from the two data sources is the ability to analyze temporal SST fluctuations over time, using comprehensive longitudinal patient records. Such analysis brings into bear an ability to recognize temporal event identifiers, which we discuss under the topic of Event Dating in Section III-D. In the next section we discuss our approach.

III. APPROACH

Our pattern-based approach to type indicator extraction, SST classification, holistic SST record creation and longitudinal patient view creation from unstructured clinical text involves four steps: 1) data pre-processing; 2) query pattern specification and semantic type extraction; 3) ambiguity resolution and patient classification; and 4) event dating. We begin with data pre-processing in the next subsection.

A. Data Pre-processing

One important challenge with research in the healthcare domain is the availability of data, primarily due to regulations such as HIPAA [15]. We have access to data subsets from a large healthcare provider in Northern California, but are unable to disclose or share the data due to HIPAA. From this subset, we select a random 1000 patient set, such that we have access to both the structured and the unstructured progress notes for all 1000 patients. We limit the set to a thousand to ensure that results can be humanly verified while still surfacing the heterogeneity within the data. Note that while the structured data is available for the duration 2005-2011 and the unstructured progress notes provided by our partner are from 2004-2011, the vast majority of the progress notes were actually from 2010-2011, reflective of emerging healthcare practices. However, the relative disparity in the timeline does not affect our experiments or the ability to
demonstrate the efficacy of incorporating progress notes as a source of useful patient information. This is because a patient’s SST is taken as the most recent SST in any source. Indeed, incorporating past or additional progress notes as they become available will only help strengthen our results.

1) Structured Electronic Medical Record: We represent the EMR data in the XML Metadata Interchange (XML) format. While not a semantic web standard, XML could easily be expressed in terms of semantic web representations. The choice of XML is to be able to handle the variety of different standards that proliferate within the healthcare domain, while still ensuring a flexible patient centric data model. In the sample EMR below the attribute is_tobacco_user holds the SST value of the patient, which is a member of the set SST = (Yes, Passive, Quit, Not Asked, NULL, Never) according to the healthcare provider.

```
<social_hx id=...>
 <patid> ... </patid>
 <contact_date>dd-MON-yy</contact_date>
 <is_tobacco_user> ... </is_tobacco_user>
 <tobacco_pack_per_day> ... </tobacco_pack_per_day>
 <tobacco_used_years> ... </tobacco_used_years>
 <tobacco_comment> ... </tobacco_comment>
 <smoking_quit_date> ... </smoking_quit_date>
 ...
</social_hx>
```

Dealing with such limited categories is tricky. It requires resolving ambiguous SST type indicators. For example, a progress note with the following type indicators: [patient is a non-smoker], [pt does not smoke], [patient is not a current smoker] is not straightforward to categorize. These type indicators obscure whether the patient has never smoked at all, or the patient in fact was a smoker who quit. We therefore approach the comparison between the structured EMR SSTs and the SSTs generated from our text analytics with the goal of providing supporting evidence and bringing more information to structured EMR, rather than to judge their correctness. We leave it to clinicians to resolve conflicting cases and concentrate more on finding such potential conflicts.

2) Unstructured Clinical Text: Unlike EMRs, SSTs from unstructured clinical text must be extracted. To accommodate ambiguities such as those discussed in the previous subsection we created an additional SST called NonSmoker not present in the EMR. And while this updated set is still not consistent with other sources, we intend to investigate the inclusion of broader and more specific categories in future.

To train our text analytics, we randomly selected a small silver standard dataset (SilverSet) of 100 patients from the 1000 sample. We use the term SilverSet since the creation of a GoldStandard using a larger sample is potentially very costly and time consuming, and thus not appealing to healthcare providers. We then manually identified SSTs for each patient for each progress note visit. Bear in mind that a single patient has a separate timestamped progress note, per visit as shown in the snippet.

```
patid | ... | contactDate | Content ....
PID1 | ... | 13MAR08 | pt claims constant headache going ....
PID1 | ... | 17MAR08 | unable to sleep, temp still abnormal ....
PID1 | ... | 01SEP10 | Counseled to quit smoking ....
```

Each patient visit is distinguished by a ContactDate, and may also potentially contain multiple type indicators in the narrative. Hence, the creation of longitudinal patient views requires capturing each type indicator within each progress note visit, and associating it with a unique SST. For the 1000 patient sample for example, there were 2109 visits and 173 visits across the 100-patient SilverSet. This averages about two visits per patient. While this number is small, the amount of detail captured in progress notes surpasses that of the structured dataset due to the narrative nature of progress notes. Another contributor is the limited range (2010-2011) from which most progress notes were recorded, compared with the EMR. Concrete examples of progress notes are available in the MTSamples dataset available online.

B. Query Pattern Specification & Semantic Type Extraction

We use SystemT and AQL for our pattern-based extraction of type indicators from the unstructured progress notes.

1) SystemT and AQL: is a declarative information extraction (IE) system that is based on an algebraic framework [16], [17]. It borrows heavily from classical database ideas to overcome limitations associated with typical grammar based systems. SystemT uses a document at a time execution model, in that the rules are applied one document at a time. A declarative language called AQL is used to specify the rules in SystemT, and these rules are compiled into algebraic expressions that are executed over a corpus of unstructured text. Syntactically, AQL is similar to SQL. AQL contains a “create view” construct, which specifies AQL rules. The creation of the AQL rules typically involves a knowledge of a small representative document set together with AQL extraction primitives and text-specific predicates which are subsequently used to build a collection of higher level views. To create an AQL view, the first task is to create a dictionary, such as the quitSmokingPhrases dictionary below:

```java
create dictionary quitSmokingPhrases as
{
    'tobacco use: quit', 'smoking: quit',
    'quit smoking', 'stopped smoking',
    'past smoking', 'former smoker'
};;
```

1The Medical Transcription Samples dataset - http://mtsamples.com/)
A view can then be extracted from this dictionary as follows:

```sql
create view QuitSmoking as
  extract dictionaries 'quitSmokingPhrases'
  on D.text as match
from Document D;
```

Additional views can be created as extensions of the existing dictionary (or view) with rules to identify additional patterns that occur frequently in the text. For example, the following pattern will extract the date of a QuitEventPattern, if an EventDate date occurs within 2 tokens of the QuitEvent.

```sql
create QuitEventPattern1 as
  extract pattern
  (<Q.match>) <Token>{0,2} (<E.match>)
return group 0 as quitsmoker
from QuitSmoking Q, EventDate E
```

Likewise, the following view will find any text in which the term “social history” is mentioned within 5 tokens to the right of a specific quit smoking phrase.

```sql
create view socialHistQuitSmoking as
  extract pattern ('social history')
  <Token>{0, 5}
  ('quit smoking|smoked|no longer smoking')
return group 0 as quitsmoker
from Document D;
```

A collective QuitSmoker SST can then be built as a view which is the union of all previously constructed type indicator views:

```sql
create view QuitSmoker as
  (extract dictionaries 'quitSmokingPhrases'
     and ...
     on D.text as quitsmoker
     from Document D, consolidate on quitsmoker)
union all
  (select S.quitsmoker
     from socialHistQuitSmoking S)
  ...
  (select S.quitsmoker
     from miscQuitSmoking S)
```

To execute AQL queries on the 1000 sample set and the SilverSet, we used the Java version of IBMs Unstructured Information Management Architecture (UIMA) [18]. The UIMA output for our analytics is a set of all type indicators occurring in a progress note visit together with its SST classification. In the next subsection we discuss document level SST classification.

C. Ambiguity Resolution and Patient Classification

Given numerous smoker references, individual progress notes may contain several type indicators that map to several SSTs for the same progress note visit. For example, the type indicators [Tobacco use: Never] and [the pt does not smoke] map to NeverSmoker and NonSmoker respectively when they appear in the same note. What is interesting is that semantically this scenarios makes sense, since a patient who does not smoke, may belong the more decisive class of having never smoked. To address this programmatically however, additional precedence rules must be applied to express the strength of indicator (SOI) for the various type indicators. For example, a NeverSmoker indicator ought to be stronger than a NonSmoker indicator. We therefore assigned the following SOIs strengths for SSTs (NeverSmoker=5, NonSmoker=2, PassiveSmoker=1, QuitSmoker=1, YesSmoker=0, InconclusiveSmoker=0).

To resolve tiebreakers, we rely on the number of occurrences of each type indicator in the progress note visit. The overall rule for disambiguation is therefore as follows: “Let the SST of a progress note be the SST with the strongest indicator and greatest number of type indicators.” Programmatically:

```sql
foreach <currSST,currSSTEvidence> in SSTEvidenceMap
  if (currSSTEvidenceCnt >= sstEvidenceCnt)
    then
      SST = currSST
      sstEvidenceCnt = currSSTEvidenceCnt
  end
```

While this rule will fail when an equal number of type indicators exist for equally scored SSTs, we found its performance to be reasonable.

D. Event Dating

An important secondary issue when performing SST classification is also identifying the date of an event. This is important for studying the exposure and possibly treatment of cardiovascular disease. For example, if PatientC quit smoking 4 years ago and PatientD quit 5 months now it could be useful to know when the quit smoking event occurred for both patients. To detect such event dates, we specified rules for identifying various dates. Building on date indicators such as [year], [month], [day], [week] and date determiners such as [ago], [prior], [previous], [since], [last], [now]. We first created an EventPeriod view:

```sql
create view EventPeriod as
  select CombineSpans(I.match,D.match)
  as match
  from DateIndicator I,DateDeterminer D
  where
    FollowsTok(I.match,D.match,0,1);
```

From this view we then built the more complex numbered and worded event period patterns such as [5 years ago], [two years now].

```sql
create view NumberedEventDate as
  select CombineSpans(N.match,P.match)
  as match
  from Number N, EventPeriod P
  where
    FollowsTok(N.match,P.match,0,1);
```

An EventDate for example could be the union of the various event date patterns including views that capture DayMonthYear, MonthYear etc.

```sql
create view EventDate as
  (select N.* from NumberedEventDate N)
union all
  (select W.* from WordedEventDate W);
```
The overall DatedQuitEvent could then be composed by associating a QuitEvent occurring in close proximity to an EventDate as shown in the following view.

```sql
create view DatedQuitEvent as
extract pattern
  (<Q.match>) <Token>{0,2} (<E.match>)
return group 0 as datedquitevent
from QuitEvent Q, EventDate E
```

We extracted DatedQuitEvent patterns from the 1000-sample dataset for visualizing temporal fluctuations in SSTs. The remaining task is to normalize event dates such as [5 years ago], [two years now] etc, to standard date formats, such as mm/dd/yyyy, relative to the ContactDate of a progress note. In the next section we discuss the evaluation of our pattern-based approach for type indicator extraction and SST classification.

### IV. EVALUATION

Table I shows how the SSTs from the manually curated SilverSet of 100 patients compare with SSTs from the structured EMRs and the SSTs from our text analytics. Note that we acknowledge that using the SilverSet to also train the text analytics to recognize patterns across the corpus may lead to over-fitting. However, since type indicators and SST classifications are evidence-based, the opportunity exists for refining the AQL rules that drive the text analytics as more patterns are discovered across the data. We discuss comparisons for each SST in the following subsections.

1) **YesSmoker**: Table I, Row 1 shows that the text analytic found a total of 18 YesSmoker patients, one more than the 17 smokers according to the SilverSet. Of these 18 patients, all 17 true positives from the SilverSet were included, together with only 1 false positive. The one false positive was labeled QuitSmoker in the SilverSet. The analytic misclassified the type indicator [Smoke: smoked 13 pack yrs] as a YesSmoker patient, hence there were more YesSmoker type indicators (i.e. [Smoke: smoked 13 pack yrs], [6 pack year], [1 pack per day]) in the progress note, than QuitSmoker type indicators (i.e. [Stopped smoking]). Given the strength of indicator (SOI) for YesSmoker is greater than QuitSmoker, the false positive was produced.

On the other hand, the structured EMR showed agreement on 14 out of the 18 YesSmoker patients from the analytic. In first instance of disparity, the analytic tagged the patient as YesSmoker, because of numerous YesSmoker type indicators for the patient’s parents, who recently quit. The correct SST should be PassiveSmoker. In the second case, the patient quit smoking, but the EMR was not subsequently updated and so its SST was NULL. The analytic however captured this change through the type indicator [Tobacco Use Yes] from the progress note.

This second case is a concrete example of how progress notes that are updated more frequently than structured EMRs, can contain more accurate information.

In the third case, the analytic again misclassified the type indicator [Smoke: smoked 13 pack yrs] as a YesSmoker while the EMR showed the patient had indeed quit. In the fourth case the text analytic labeled the patient as YesSmoker based on the type indicator [encoraged to quit smoking] (note the misspelling exists in the progress note). The structured EMR had no data and was hence InconclusiveSmoker.

These examples make a concrete example of how progress notes that are updated more frequently than structured EMRs and SST classifications from the unstructured clinical notes could assist in resolution adding value to the structured EMRs in a healthcare system.

2) **PassiveSmoker**: Table I, Row 2 shows that the analytic found 3 PassiveSmoker patients out of 4 from the SilverSet. The one false positive was classified incorrectly by the analytic, as a NonSmoker. The type indicator [Denies fever. Social History: Tobacco] was incorrectly mapped to NonSmoker because ‘Denies’ and ‘Tobacco’ occur in close proximity. In spite of numerous PassiveSmoker type indicators including [Tobacco Use: Passive] and [father smokes], since NonSmoker has a greater SOI than PassiveSmoker. On the other hand, the structured record showed agreement on each SST by the analytic. Note that the structured EMR indicated 2 additional PassiveSmokers. In the first instance, the text analytic correctly labeled the patient as a YesSmoker, based on the type indicator [smoke marijuana]. We consider marijuana users are YesSmoker was because our healthcare provide conforms to a wider definition of smoking, not in the least because marijuana smokers frequently deal in tobacco. The second instance was the misclassified case, in which the type indicator [Denies fever. Social History: Tobacco] was incorrectly mapped to NonSmoker for the same reason as before.

These results establish that we are able to sufficiently capture passive smoking events and even identify discrepancies where a smoking status may not have been captured adequately in the EMR.

<table>
<thead>
<tr>
<th>SST</th>
<th>EMR</th>
<th>SilverSet</th>
<th>Text Analytic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>14</td>
<td>17</td>
<td>18 17 1 0</td>
</tr>
<tr>
<td>Passive</td>
<td>5</td>
<td>4</td>
<td>3 3 0 1</td>
</tr>
<tr>
<td>Quit</td>
<td>32</td>
<td>19</td>
<td>16 16 0 3</td>
</tr>
<tr>
<td>Non-smoker</td>
<td>0</td>
<td>36</td>
<td>35 31 4 5</td>
</tr>
<tr>
<td>Never</td>
<td>45</td>
<td>14</td>
<td>14 12 2 2</td>
</tr>
<tr>
<td>Inconclusive</td>
<td>4</td>
<td>10</td>
<td>14 8 6 2</td>
</tr>
</tbody>
</table>

#### Table I

SilverSet SST Classification Comparison
3) **QuitSmoker:** Table I, Row 3 shows the analytic found 16 of the 19 QuitSmoker patients from the SilverSet. Hence there were no false positives and only 3 false negatives. One false negative was tagged as InconclusiveSmoker by the analytic because of the type indicator [Counseled to quit smoking? n/a]. The second case was tagged as NeverSmoker due to the following type indicators [FORMER SMOKER], [Quit>12 MO. AGO.], [Tobacco Use: Never]. Since NeverSmoker type indicators have higher precedence than QuitSmoker, our analytic misclassified the patient. The last patient was tagged YesSmoker, as the analytic again misclassified [Smoke: smoked 13 pack yrs].

The disparity between the analytic and the structured EMR SSTs was significant. There were 32 QuitSmoker patients according to the EMR, but only 15 as per the analytic. Of the remaining 17 in the EMR, 12 were tagged as NonSmoker, 3 as InconclusiveSmokers and 2 as YesSmoker.

Of the 12 NonSmoker, it was the analytic that was in fact was correct 10 out of the 12 times drawing from type indicators: [smoking denied], [current smoker no], [social history: smoker? no], [denies smoking], [smoker: no], [smoker- no], [non-smoker], [current smoker no] etc. Indeed, this is not a surprising result. A patient that quits smoking does become a NonSmoker. What is striking is that in once instance one patient [quit over 40 years ago]. Evidence-based analytics that enable Identification of such temporal event is significant for estimating future risk of cardiovascular diseases.

For the 3 InconclusiveSmoker cases the analytic could not resolve the following type indicators to any SST: [Counseled to quit smoking? n/a], [quit in 1978], [Social History: smoking history were reviewed]. Finally, for 1 of the 2 YesSmoker cases the patient had indeed quit smoking while in the other the patient had not.

These cases establish that while our results matched the structured QuitSmoker status in more than half the cases, we were consistently able to identify important details from the progress notes to assist with disparity in SST classification in the various datasets.

4) **NonSmoker:** Table I, Row 5 shows that the analytic found 31 of 36 NonSmoker patients from the SilverSet, along with 4 false positives. Of the four false positives, two InconclusiveSmoker patients were misclassified by the analytic, as no rule existed for the type indicator [Non-smoker/No answer given]. For the other case, the analytic misclassified because the patient was a non smoker who had never smoked. Hence the patient was classified as NeverSmoker due to precedence among the type indicators: [Tobacco Use: Never], [Smoker: no]. There was also 5 false negatives of which 4 were InconclusiveSmoker and 1 was NeverSmoker. The type indicators [Tobacco Use: Never], [Smoker: no] again swayed towards NeverSmoker.

For the InconclusiveSmoker case, the ambiguous type indicators were [no smoking],[Exposure to smokers?: no] and [Counseled to quit smoking? N/a]. Since the type indicator [no smoking] could be an instruction to a NonSmoker, or an admonition to a YesSmoker we could not classify it. Since the EMR has no NonSmoker SSTs so there was no comparison to be made.

5) **NeverSmoker:** Table I, Row 5 shows that the analytic found 14 NeverSmoker patients of which two were false positives compared with the SilverSet. Of the two false positives, one was a NonSmoker patient, while in the other case, the analytic classified the patient as NeverSmoker based on the precedence of NeverSmoker among the type indicators [Tobacco Use: Never], [FORMER SMOKER]. Of the two false negatives, one was InconclusiveSmoker, because no rule existed to capture the type indicator Smoking: never. The other patient was classified as NonSmoker, because this SST was the most recent from the progress notes.

The EMR indicated however, that there were 45 NeverSmoker patients. 14 were identified correctly by the analytic, while 22 were labeled as NonSmoker, 8 as InconclusiveSmoker and 1 as QuitSmoker. The QuitSmoker was a misclassification by the EMR not the analytic. The patient had [Quit smoking in 1989] according to the progress note. The 8 InconclusiveSmoker cases has the following type indicators: [Smoking history was reviewed], [Smoking: never], [some exposure to cigarette smoke], [generic warning: always greater for smokers that non smokers and increases with age], [No smoking], [Counseled to quit smoking: N/a], all of which require some human judgement to disambiguate. Of the 22 cases classified as non smokers the analytic found the following type indicators: [3-Smoking denied], [7-Non-smoker], [2-patient does not smoke cigarettes], [2-Smoking: does not smoke], [3-Smoking: No], [3-Smoker: no], [1-pt not a smoker] and [1-Non-smoker/No answer given].

These results clearly indicate that a number of patients were classified as having never smoked, when in fact there was no evidence to suggest this from the text. Instead, many patient are in fact non-smokers. This requires investigation as to whether such patients should be labeled QuitSmoker or PassiveSmoker.

6) **InconclusiveSmoker:** The limitations with the inconclusive case have been covered in the other examples, and mainly involve refining and adding more rules and improving the precedence measures for the SOI scheme. We omit a thorough discussion on this due to space limitations.

Overall, the text analytics achieved 87% precision in SST classification after using the SilverSet for both training and testing. However, what is significant is that all SST classifications contain supporting evidence, which were
crucial when comparing SSTs to the EMR. This is consistent with our ambition, which is not to judge correctness of the EMR SST, but to provide evidence to alert clinicians to potential inconsistencies in their data. To further assess the feasibility of this position, we applied the pattern-based text analytics to the 1000-patient sample.

Table II shows in column 1, that the structured EMR in this larger dataset contained 149 YesSmoker, 37 PassiveSmoker, 225 QuitSmoker, no NonSmoker, 548 NeverSmoker, and 41 InconclusiveSmoker patients. The text analytic (trained on the SilverSet), showed in Row 3, that among the 225 QuitSmoker patients the analytic found 83 that indeed quit, but importantly 29 were still YesSmoker, while 4 were passive. While these 29 instances were not manually verified, given the precision of the text analytic, these are interesting results for healthcare providers. It demonstrates that we were able to identify potential additional at-risk individuals than purely through the unstructured text analysis. Further, the text analytic showed that among the 548 NeverSmoker patients, 4 patients were labeled as PassiveSmoker, 5 were QuitSmoker and 31 were YesSmoker. Again this number is significant to healthcare providers, as it identified potential at-risk patients. Finally, the text analytics found that among the 37 PassiveSmoker patients, 7 were actually indicated as YesSmoker. Similarly, 9 patients were identified as YesSmoker that were labeled Inconclusive by the text analytic. Collectively, all of these individuals may have been exposed to smoking either directly or indirectly, and may therefore be more susceptible to related diseases than currently anticipated.

This result is promising as it validates our approach of complementing structured EMR data with information extracted from unstructured clinical text.

### V. Future Work

One limitation of our work is that the small size of the SilverSet. This limitation in size does not guarantee coverage in patterns representative of the corpus and hence may results in low recall in type indicator extraction. However since type indicators must be manually identified to create a GoldStandard, creating a reasonable GoldStandard using larger datasets is potentially very costly and time consuming, which may be unappealing to healthcare providers. We believe that the smaller subset will help surface some intricate details found only in free-text medical records, and enable us to account for misalignment between the SilverSet and the EMR as an initial step. Hence, we begin with a small sample and intend to build to progressively larger subsets.

An important secondary outcome of this work is to position smoker trends from healthcare providers relative to national smoker averages. Given the accuracy of the existing analytic, it will now be possible to classify the several-million patients served by the healthcare provider into the current categories, and thereby guide clinicians to focus attention on potentially interesting patient records using evidence-based analysis.

Finally, this work fits into a larger goal of computing per-patient risk scores based on established medical guidelines such as the Framingham Heart Study. Smoking is a significant factor in risk prediction for a variety of diseases, and a more accurate measure of an individual’s smoking habits will not only support appropriate patient care and counseling, but also allow epidemiologists to come up with accurate models for cost analysis, project specialty needs and geographic requirements in the future.

### VI. Conclusion

At-risk smoking behavior and/or exposure can give rise to a variety of health complications, not limited to cancer, stroke and a variety of cardiovascular diseases. These conditions have significant bearing not only on patient health and quality of life, but also adversely impact rising healthcare costs. In this study we applied text analytics through a pattern-based and evidence-driven method for smoker type classification, to surface semantically-informed aspects related to smoking habits of individuals. This approach provides the data elements that allow augmenting structured EMR data with evidence from unstructured clinical texts. It also provides a mechanism for improving the longitudinal classification of smoker-types and performing temporal observations of smoking behavior using both structured and unstructured data. Granted that the module on Event Dating using unstructured text is still developing, it is not difficult to project the possible benefits of such analytics upon maturation.

### VII. Acknowledgement

We would like to thank Karen Brannon, Laura Chiticariu, Dan Gruhl, Steve Welch and Neal Lewis for technical guidance and the provision of supplementary material for this project.

### References


