Application of a Hybrid Text Mining Approach to the Study of Suicidal Behavior in a Large Population

Kenric W. Hammond
University of Washington
khammond@uw.edu

Ryan J. Laundry
VA Puget Sound Health Care System
ryan.laundry@va.gov

Abstract
To fulfill the promise of electronic health records to support the study of disease in populations, efficient techniques are required to search large clinical corpora. The authors describe a hybrid system that combines a search engine and a natural language feature extraction and classification system to estimate the annual incidence of suicide attempts and demonstrate an association of adverse childhood experiences with suicide attempt risk in a cohort of 250,000 patients. The methodology replicated a previous finding that a positive association between suicide attempt incidence and a history of childhood abuse, neglect or family dysfunction exists, and that the association is stronger when multiple adverse events are reported.

1. Introduction

Great interest exists in applying natural language processing (NLP) techniques to mine clinical records for relevant epidemiologic information. Structured clinical and administrative data: (e.g., lab studies, pharmacy records, visit data and discharge diagnoses) fail to capture many important details about illness and health care, necessitating manual chart review.

The advent of electronic health records has driven recent advances in automated NLP analysis of machine readable documents and has demonstrated the feasibility of using NLP to process free text [1].

To date, much emphasis has been placed on concept extraction, using such tools as Ctales [2], which maps concepts in free text to clinical concepts catalogued in the National Library of Medicine’s Unified Medical Language System (UMLS) terminology resource [3]. Sophisticated techniques using ontologies have shown some success in extracting meaning from carefully edited text corpora such as MEDLINE abstracts [4]. More recently investigators have begun to tackle medical free text and have encountered challenges occasioned by the important differences between carefully edited scholarly text and the more unruly narrative found in EHR documentation.

Promising approaches have been found. The i2b2 [5] and TREC [6] challenges have driven development of effective algorithms of demonstrated value. However, current information extraction approaches are computationally intensive and compared to index-based full text search, slower. This means that in their native form they may not be suitable for direct analysis of very large corpora. One barrier NLP researchers face in resolving this limitation is the difficulty of accessing large volumes of medical text due to privacy protections. As a result, few investigators have been in a position to evaluate the efficiency of their methods for processing very large clinical corpora.

The United States Department of Veterans Affairs Veterans Health Administration (VA) adopted a full scale EHR with universal clinician-entered clinical documentation in 2000 [7]. Presently VA stores approximately 2 billion test documents belonging to approximately 14 million veterans in its research data warehouse. VA’s Health Services Research and Development Service has supported development of NLP analysis of the VA’s clinical text because of the urgent need to unlock important data maintained in its clinical records. Various projects have been funded to address surveillance of veteran populations, such as those who served in recent wars.

Suicide is an area of special concern in the U.S. military and veteran populations, both of whom have a rate of suicide that exceeds that of the general population [8]. Mitigating this elevated risk is a key agency priority.

1.1 Suicide Epidemiology

The study of suicide risk is complex. Suicide is a medical disaster, but occurs at a relatively low frequency. The extensive body of suicide and self-harm risk research has identified many risk factors, among them: diagnosed mental disorders, substance abuse, and a history of prior suicide attempts. However, much of this research has been based on retrospective study of suicide attempters and completers. The rarity of suicidal behavior limits the
The specificity of predicting future suicidal behavior based on relatively common risk factors. A recent survey of literature on screening tools to predict suicide attempts and death by suicide concluded that current evidence is insufficient to support reliance on screening tools based on presence or absence of risk factors in clinical practice [9], and so, the current standard of care for the at-risk patient remains in the domain of a skilled clinical assessment. Despite this limitation, concern about suicide in the veteran population has prompted widespread screening efforts, traces of which were seen in clinical documentation. The motivation for screening is laudable: although its effectiveness has not been demonstrated, inclusion of screening protocols may have increased clinician awareness of suicide risk and prompted discussions with patients leading to inclusion of additional information. Data from the present study are consistent with this notion (see Table 3 and the discussion section).

It is not possible to discuss all the risk factors for suicide, but the two most salient ones are history of a prior suicide attempt (increasing relative risk by a factor of 40), and a diagnosis of depression, which increases relative risk by a factor of 20 [10]. While a diagnosis of depression is readily found in administrative data, history of a prior suicide attempt is not. A prior study of a random cohort of 100,000 veterans utilized text search to identify lifetime history of suicide attempts, and found that compared to administrative data, text search was able to detect eight to ten times as many patients reporting a lifetime history of suicide attempt [11].

In this study the text search method was extended to identify de novo suicide attempts on an annual basis in a larger population of 250,000 veterans treated in the VA who had served in the 1990-1991 Persian Gulf war. An additional goal of the study was to examine the influence on risk of “Adverse Childhood Experiences” (ACE) on the observed annual suicide attempt rate. In 1998, Felitti et al. published data based on a questionnaire survey of 11,000 health plan subscribers that showed that individuals reporting multiple categories of ACE (including physical, sexual and emotional abuse; family dysfunction; and childhood neglect) experienced higher rates of adult physical and behavioral problems including suicide attempts [12]. This observation seemed to be a good use case for text mining, because ACE data are rarely recorded in structured data (such as ICD Diagnoses used for visit billing) but are often recorded in clinical histories obtained by caregivers.

If Felitti’s observations were to hold true in the population studied, one would expect to see an increased suicide attempt rate among veterans with ACE exposure, and one would expect an even higher rate among individuals reporting exposure in multiple ACE categories. Accordingly, the approach taken was to ascertain the annual rate of suicide attempts in the population as a whole, and to compare this to the rates of suicide attempts in the subpopulations exposed to one or more ACE factors. Estimates of suicide attempts and ACE exposure were accomplished via a hybrid strategy incorporating text search and NLP techniques.

2. Methods

2.1 Population Definition

The target population was veterans who had served in the 1990-1991 Persian Gulf war. Available administrative data furnished only a broad designation, “Persian Gulf Era”, which spanned service in the Middle East and Central Asia from 1990 to the present. To narrow the group, we required that members be at least age 16 in 1991 and have an indication based on text search of having served in the 1990-91 Operation Desert Shield/Desert Storm conflict. To do this, records were searched for phrases such as “Desert Storm”, “Gulf War”, “SCUD missile”, “Kuwait border”, etc. This allowed 250,518 veterans treated in VA facilities to be identified. Review of several hundred records thus retrieved permitted estimation that Operation Desert Storm service was identified with 85% accuracy and we used this cohort for the remainder of the analysis. 54,967,366 clinical documents recorded between January 1, 2000 and June 30, 2011 were attributed to this group.

2.2 Search Method

A search tool, Veterans Indexed Search for Analytics, or VISA was used. VISA, developed by the authors, is based on the Lucene search engine [13] and utilizes an index of clinical notes. Tokens in text were stemmed with the Porter stemmer, a small set of stop words was implemented, and date and numeric forms were omitted. The search engine accepts literal text and “Span” queries (tokens occurring within a specified span length, in order, or order-independent). VISA was configured to write the retrieved documents, a 400-character snippet containing the query terms, a relevance score and document identifier to a SQL data base.

Queries were developed and refined by the first author, a VA psychiatrist familiar with the clinical topic. The principal query approach was to use the Lucene “SpanOr” query. SpanOr permits stacking multiple Span queries together to search for anticipated variants of the target. Query strategy was informed by inspecting search results and numbers of documents.
retrieved. To retrieve, for example, affirmative documentation of suicide attempts occurring within a year, Span queries in the form: (Suicide Attempt [today, last week, etc] span-7, any order) were issued\(^1\). A similar approach was used to search for ACE factors. Next, a limited number of negated forms of the target were searched, typically in the form (Suicide Attempt [no, not, denies, etc] span-7, any order). Subsequently, these documents were excluded.

To identify “boilerplate” text inserted by document templates – a common occurrence because of frequent “suicide risk screening” and other templated documentation – snippets retrieved by “affirmative” queries were grouped using a SQL query. Invariant snippets occurring multiple times were thus readily identified as boilerplate. Span queries containing exact, uninformative boilerplate phrases retrieved documents to exclude from analysis. Finally, clinical notes whose titles indicated “Group Therapy” or “Education” were identified for exclusion, because the narrative they contained might not pertain to the patient.

Exclusion was accomplished using a SQL statement to eliminate “affirmative” documents that matched “exclusion” documents. This procedure was accomplished for suicide attempts within a year and for each of the ACE categories.

Table 1 summarizes text search activities for each topic and shows the number of documents remaining in each category after applying filtering and exclusion procedures.

### Table 1 – Summary of Text Search

<table>
<thead>
<tr>
<th>Topic</th>
<th>Affirmative documents</th>
<th>Negated forms</th>
<th>Group/Edu.*</th>
<th>Remaining</th>
</tr>
</thead>
<tbody>
<tr>
<td>Suicide Attempt</td>
<td>32,236</td>
<td>533,220</td>
<td>3.5 M</td>
<td>23,010</td>
</tr>
<tr>
<td>Physical Abuse</td>
<td>130,936</td>
<td>344,060</td>
<td>3.5 M</td>
<td>73,356</td>
</tr>
<tr>
<td>Sexual Abuse</td>
<td>124,457</td>
<td>305,414</td>
<td>3.5 M</td>
<td>61,933</td>
</tr>
<tr>
<td>Emotional Abuse</td>
<td>52,800</td>
<td>205,138</td>
<td>3.5 M</td>
<td>30,774</td>
</tr>
<tr>
<td>Neglect</td>
<td>24,437</td>
<td>470,807</td>
<td>8094</td>
<td>15,643</td>
</tr>
<tr>
<td>Household Dysfunction</td>
<td>20,281</td>
<td>0</td>
<td>3.5 M</td>
<td>15,643</td>
</tr>
</tbody>
</table>

*Group therapy and Education notes

---

1 (Additional information about query construction will be furnished on request)

### 2.3 Classification of Retrieved Documents

The next step of the analysis was performance of linguistic feature extraction and supervised classification of retrieved, filtered affirmative 400-character document snippets as true or false positive using the Automated Retrieval Console (ARC) developed by D’Avolio [14, 15]. The ARC utility integrates the widely used Mayo Clinic Ctakes [2] tooolset for linguistic feature extraction and the MALLET [16] conditional random fields classifier. The overall process is illustrated in Figure 1.

For each topic, training and gold standard snippet sets were randomly sampled from the filtered search results indicated in Table 1. True positive (i.e., precision) rates among the training samples ranged from 49% for the “Neglect” topic to 81% for the “violence against mother” Household Dysfunction subtopic. Each topic was trained on 300-500 snippets judged by the first author, and validated on gold standard reference sets of about 200 snippets. Because the snippets were short, manual adjudication was accomplished rapidly. Precision (the fraction of relevant documents among the documents retrieved), recall (the fraction of relevant documents retrieved of all relevant documents in the training sample) and F-measure (the harmonic mean of precision and recall) for the trainings shown in Table 2 indicate that the ARC procedure resulted in improved precision compared to raw search results.

### Table 2 – ARC Classifier Statistics

<table>
<thead>
<tr>
<th>Topic</th>
<th>Recall</th>
<th>Precision</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Suicide Attempt</td>
<td>0.872</td>
<td>0.797</td>
<td>0.831</td>
</tr>
<tr>
<td>Physical Abuse</td>
<td>0.957</td>
<td>0.859</td>
<td>0.905</td>
</tr>
<tr>
<td>Sexual Abuse</td>
<td>0.950</td>
<td>0.903</td>
<td>0.926</td>
</tr>
<tr>
<td>Emotional Abuse (mean of 3 items)</td>
<td>0.895</td>
<td>0.762</td>
<td>0.822</td>
</tr>
<tr>
<td>Neglect</td>
<td>0.643</td>
<td>0.738</td>
<td>0.683</td>
</tr>
<tr>
<td>Household Dysfunction (mean of 4 sub-items)</td>
<td>0.988</td>
<td>0.838</td>
<td>0.906</td>
</tr>
</tbody>
</table>
2.4 Calculation of Suicide Attempt Rates

The annual rate of suicide attempts was obtained by tallying patients who had at least one positive document in each index year after concept extraction and classification with the ARC tool and dividing by the population size (250,518). Years 2000 and 2011 had only 9 months of data, so the rate was adjusted accordingly to permit comparison with other periods. Rates for each ACE exposure category subgroup, defined as having at least one positive assertion of an exposure, were tallied and divided by the subgroup population size.

3. Results

Table 3 shows the baseline rate of Suicide Attempts by Fiscal Year (years marked by * were partial years) and the rates observed among patients exposed to three categories of Adverse Childhood Experiences (ACEs). It is seen that the rate of recorded suicide attempts appears to increase over time, from 77.2 per 100,000 to 595 per 100,000 in Fiscal Year 2011. While this could reflect an increase in actual suicide attempts in the population, it is also consistent with the possibility that detection of suicide attempts improved over time due to increased vigilance and screening. The annual suicide attempt rates and annual relative risk in the three ACE exposure groups are increased over the baseline rate for each year.

Presence of multiple ACE exposures is associated with further increase in the observed suicide attempt rate (Table 4). Presence of all three major categories of ACE exposure results in additional increase. This pattern corresponds closely to the increase in rate of suicide attempts reported by Felitti for multiple ACE exposures [12].
4. Discussion

This work has a number of implications for approaching large scale clinical data. From a methodological perspective, it demonstrates that combining text search with text feature extraction and classification is practical in a large text corpus. Most NLP work reported to date has focused on much smaller data sets, largely due to the difficulty of accessing very large clinical corpora because of privacy restrictions. Use of a text search engine to generate snippets of text relevant to the query topic confers considerable efficiency and appears to be effective. Both the manual adjudication for training and testing and NLP processing appear to have been facilitated by the use of snippets. Focusing computationally intense NLP processing on filtered text search results saved considerable time. For instance, the VISA search engine was able to retrieve and store 124,457 documents affirming sexual abuse from the 55 M document corpus in 2.5 hours (>6,000 documents per second). After applying exclusions, the 61,933 snippets submitted to ARC required 6.8 hours to complete the classification task (2.6 documents per second). Using ARC to classify the entire corpus would not have been possible. ARC, however, performed feature extraction and classification reasonably well (Table 2), whereas the VISA search engine lacked this capability. The system used was not fully optimized. Performance penalties were seen during transfer of large numbers of snippets to the SQL data base and the one-time task of building an index of 55 million documents posed an additional challenge. Parallelization was employed in the indexing task, but was limited by the computing environment available to the investigators (two virtual machines accessing 16 GB of memory). Scaling the index task to handle the full VA corpus of 1.8 billion documents would definitely require additional computing resources.

From a clinical perspective, the data are consistent with trends reported in Felitti’s ACE study wherein the increased odds of ever having attempted suicide were 1.8-, 3-, 6.6- and 12.2-fold for exposure to one, two, three or four ACE factors, respectively among his questionnaire subjects [12]. Felitti’s 1998 study assessed seven factors: psychological, sexual and physical abuse in a broader Abuse category; and a Household Dysfunction category that included violence toward the mother, an incarcerated household member, a mentally ill household member, or a substance-abusing family member. Later iterations of the ACE scheme added a Neglect category that included emotional and physical neglect subcategories, and added a “divorced or separated parent” subcategory to the Household Dysfunction category.

Our query approach was unable to differentiate the physical versus emotional neglect subcategories because most documents addressing neglect did not differentiate the two concepts and hence we used a single category. In document classification, the Neglect concept performed the least well (Table 2).
The Abuse and Household Dysfunction results presented represent combined subcategories as described above.

From a suicide epidemiology perspective, the increasing annual rates presented in Table 3 raise questions. While it is generally true that suicide rates increase after middle age following a peak in the third decade and thus with time, it is difficult to imagine that the actual rate of suicide attempts increased as shown. More likely, increased awareness of suicide risk among veterans and more vigilant detection, especially since 2007 when a VA suicide prevention program was implemented, account for the apparent increase. Still, the annual relative risk (RR) data in Tables 3 and 4 shows that the association with ACE exposure was fairly constant over time. Inasmuch as ACE exposure is determinable from documents written at any time and by definition is an unchanging patient characteristic, ACE detection should not be time-sensitive. Once a patient’s ACE exposure is established, it does appear to have value in estimating a patient’s relative, but not absolute, suicide risk.

Our experience with using VISA search to identify ACE factors showed us that most of the ACE exposure information retrieved was found in notes created by mental health clinicians. This was unsurprising, considering that a childhood history is a standard part of a mental health assessment, but a concern, because it suggests that adequate assessment for ACE factors is not evenly distributed in the clinical population. Given that the ACE literature has demonstrated that adverse childhood experiences are associated with numerous other health conditions, consideration should be given to including ACE information in standard general medical data bases. Sixty-one publications describing ACE findings are accessible at the United States Centers for Disease Control web site [3].

Another concern is that finding pertinent ACE exposure information in a particular note is not an easy task in many EHR documentation systems. In emergency situations, a search engine may be very helpful in rapidly retrieving information relevant to a clinician’s suicide risk assessment. The success of the hybrid retrieval and classification methods used in this study suggests that preformed queries are feasible and potentially valuable clinically.

From a technical standpoint, this study illustrates, by essentially replicating Felitti’s results, the feasibility of using natural language processing methods to conduct large scale epidemiologic research on a topic which would have been impossible to study by analyzing available structured elements (e.g., diagnostic codes, procedure codes, laboratory data, etc.) in electronic health records. Felitti determined ACE exposure by administering thousands of questionnaires, whereas using NLP techniques allowed us to demonstrate a very similar effect in a large veteran population through direct analysis of text records. Conducting such a study using traditional questionnaire methods would be much more costly and time-consuming. An attractive feature of the approach described is that it permits investigators to readily explore a wide range of topics that are poorly captured in structured data.

This initial study has limitations. The most important is that the data demonstrate an associative, but not a causal relationship between ACE exposure and suicide attempt risk. The second caveat is that all data were based on a population in active treatment and thus likely to differ in important respects from the general veteran population. A third limitation is that the results of search and classification depend on retrieval processes of limited precision. Fourth, it needs to be recalled that that clinical documentation can only provide suggestive, but not definitive data about the true annual incidence of suicide attempts in the population. For example, the process used to assign a reported event to a given year using linguistic cues was approximate and arbitrary. While estimates of the ratio of attempted suicides to completed suicides exist, the best estimates rely on observed events, not assertions in clinical narrative, making it premature to claim, based on our analysis, that we know the actual annual rates of attempted suicides or could extrapolate that to yield the rate of completed suicides in the population studied.

5. Conclusion

Use of a text mining approach that combines search engine and NLP-based text classification capabilities was more efficient, more feasible and probably more accurate than using either method alone when analyzing a very large clinical text corpus. Indexing the corpus permits a coarse-grained search for targeted concepts, whereas NLP based concept extraction and classification permits finer-grained triaging of search results. Furthermore, the ability to replicate Felitti’s important findings in a larger and entirely new population demonstrates that this approach extends the reach of large scale epidemiologic research to topics that cannot be assessed using structured data alone.

Application of our hybrid methodology demonstrates that a robust association between exposure to adverse childhood events and attempted suicide exists in the treated population of veterans who served in the 1990-1991 Persian Gulf war. This finding supports assessment of childhood history when
evaluating suicide risk. In the future, the data set and methodology described will permit study of the association between adverse childhood events and additional medical conditions reported by others.

6. Acknowledgment

The authors would like to thank others who contributed to this work, including T. Michael O’Leary, William Jones, Leonard D’Avolio and Matthew Samore. This work was supported by the United States Department of Veterans Affairs, Health Services Research and Development Service and the VA Puget Sound and Salt Lake Health Care Systems. The views expressed in this article are those of the authors and do not necessarily reflect the position or policy of the Department of Veterans Affairs.

7. References