Use of Text Search to Effectively Identify Lifetime Prevalence of Suicide Attempts Among Veterans

Kenric W. Hammond  
Department of Veterans Affairs  
kenric.hammond@va.gov

Ryan J. Laundry  
University of Washington  
rilaunadr@uw.edu

T. Michael O’Leary  
University of Washington  
tmoleary@uw.edu

William P. Jones  
University of Washington  
williamj@uw.edu

Abstract
Suicide is an important emerging problem among active duty personnel and veterans of recent conflicts in Iraq and Afghanistan. Suicide rates among veterans treated in the VA System are about 50% higher than in the general population, and the rate of suicide among active duty service personnel has recently exceeded the rate in the general population. Preventing suicides and providing services is a high priority for the Department of Veterans Affairs Health Administration (VA). The VA electronic medical record system stores a very large corpus of clinical notes. Application of search engine technology to a subset of this corpus demonstrated that text search confers an 8- to 10-fold improvement in the ability to identify individuals who have made prior suicide attempts. This technology is highly relevant to health care institutions that have adopted electronic medical record systems.

1. Introduction
Suicide is a significant public health problem. Every year, more than 36,000 US citizens take their own lives and over 1 million make a suicide attempt. Suicide is of special interest to the Department of Veterans Affairs (VA), the agency responsible for treating service related illness of former members of the US military because the rate of suicide in patients treated in the VA well as military veterans in general, exceeds that of the general population. Blow et al. reported a rate of 38 per 100,000 in the years 2000-2007 which is 50% higher than the general population rate [1]. Recently the rate of suicide among active duty military personnel has also begun to exceed that of the general population as well [2]. In most cases, suicide is an “unnecessary tragedy”, potentially preventable in the short term, and amenable to risk reduction strategies consisting of appropriate clinical care in the longer term. Nearly a million veterans will have returned from combat deployment in Iraq and Afghanistan. In 2004 the VA implemented a comprehensive suicide prevention program, including telephone hotlines and facility suicide prevention coordinators as part of its Mental Health Initiative [3].

Although it has been studied for many years, the epidemiology of suicide has been quite difficult to characterize satisfactorily. Death by suicide occurs at a low base rate, around 20 per 100,000 persons per year in the general population. Suicide is a stigmatizing diagnosis, and may be under-reported as a cause of death. Newer statistical approaches and the National Death Index and National Violent Death Reporting System have improved reporting and permit classification of suicides by finer-grained demographic characteristics and diagnoses found on death certificates.

Death statistics provide sparse information about the clinical precursors of suicide. Clinical insights about suicide risk tend to come from the study of patients who present to health care settings with a suicide attempt or from follow up studies of treated patients. Sample sizes tend to be small in these studies, and follow-up over long periods is difficult and incomplete. Nevertheless, clinical research presents a consistent picture. Combined with death statistics, useful generalities about suicide risk have emerged: men complete suicide more often than women, and in the general population, women attempt suicide more often than men. The effect of age is salient: suicidal behavior begins at a low rate in childhood, increases rapidly with adolescence, and then shows a first peak between ages 20 and 35 which then lowers through middle age before climbing again at the end of life. Patient factors that elevate risk include having a diagnosed mental disorder, especially depression, and having made a previous suicide attempt.
Regrettably, clinical characteristics that predict increased suicide risk are extremely common compared to completed suicide. Heuristic risk-rating algorithms capable of identifying significant proportions of those at actual risk have a high false-positive rate. Population “screening” questionnaires have an even higher false positive rate [4].Clinicians assessing patients for suicide risk utilize these indicators in a rough way, but ultimately the decisions about managing the suicidal patient rely on judgment, careful analysis of individual contextual factors, intuition and caution. Relative risk factors summarized from the meta-analysis of Harris et al. are presented in Table 1 [5].

1.1 Identifying suicide risk in electronic medical records

Electronic medical record systems may help improve recognition of patients at risk for suicide. The VA maintains very complete health care information on more than ten million regular patients. Structured data elements (diagnoses, prescriptions, laboratory results, etc.) are supplemented by free-text clinical notes documenting patient encounters. Currently VA stores 2 billion of these notes. In the remainder of this article we will describe how structured and narrative information can be used together to identify the most important (Table 1) risk factor for suicide: history of a previous suicide attempt.

In the US, the Veterans Health Administration is distinctive in that it is a very large and integrated health care system. VA treats a defined population and maintains electronic records for all its patients. Most patients receive the majority of their care from VA. Each VA site maintains a local clinical data base, but transmits its data to a central data warehouse where patient records are merged. Warehoused data is available for business analysis, and serves as the source for a separate data store known as VINCI (Veterans Informatics and Computing Infrastructure) to support medical research [6]. VINCI furnished us data on 9.8 million veterans.

To date, most research and business use of VA data has relied on structured data fields, and the principal analytic approach has been the relational model. These data resources have been valuable for strategic planning, quality improvement and health services research [7]. Structured data, though, represents but a fraction of the total data available: most of the data consist of text unsuitable for direct analysis using relational database techniques. Recognizing the potential value of textual data, the VA research community has begun initiatives to apply natural language processing methods to extract useful information from text repositories [6, 8, 9].

Electronic data systems permit automated surveillance of disease risk factors that can prompt issuance of clinical warnings and recommendations when the electronic chart is opened. This method is widely used in electronic record systems for conditions such as heart disease, diabetes and hypertension [11]. Despite its importance in assigning risk, the suicide attempt history was not consistently recorded as a diagnosis unless the attempt resulted in an episode of

<table>
<thead>
<tr>
<th>Risk Factor</th>
<th>SMR *</th>
<th>95% Confidence Interval</th>
<th>Sampled population size</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Suicide Attempt (poisoning)</td>
<td>40.8</td>
<td>(37.0-44.7)</td>
<td>8,000</td>
<td>1</td>
</tr>
<tr>
<td>Suicide Attempt (any method)</td>
<td>38.4</td>
<td>(34.0-43.1)</td>
<td>2,700</td>
<td>2</td>
</tr>
<tr>
<td>Major Depressive Disorder</td>
<td>20.4</td>
<td>(18.3-22.6)</td>
<td>8,000</td>
<td>3</td>
</tr>
<tr>
<td>Bipolar Disorder</td>
<td>15.0</td>
<td>(12.3-18.4)</td>
<td>3,700</td>
<td>4</td>
</tr>
<tr>
<td>Schizophrenia</td>
<td>8.5</td>
<td>(8.0-9.0)</td>
<td>30,000</td>
<td>5</td>
</tr>
<tr>
<td>Alcoholism</td>
<td>5.9</td>
<td>(5.4-6.3)</td>
<td>45,000</td>
<td>6</td>
</tr>
</tbody>
</table>

*Standardized Mortality Ratio where 1.0 = no elevated risk compared to the general population.**Sampled population after combining studies.
care. More often, prior suicide attempts are documented in clinical notes as historical facts and are not readily available to surveillance algorithms.

2. Motivation

Our goal was to assess whether narrative records in VA data bases may be useful in estimating lifetime prevalence of suicide attempts, thus enhancing the VA’s suicide prevention efforts. This project was undertaken as part of a larger effort to use multiple natural language processing methods to extract epidemiologic information from VA clinical notes. The focus of our team was to evaluate the applicability of text search in this effort.

4. Methods

Data from the VINCI research data warehouse were available for study. Access to this collection with waiver of patient informed consent was authorized by Investigative Review Boards at VA Puget Sound Health Care System, VA Salt Lake Health Care System and the University of Washington. The data consisted of visit diagnoses encoded in the International Classification of Diseases ICD-9 system, encrypted unique patient identifiers, sex, age and clinical documents containing Visit, Title, Text, Date and Site fields. Documents consisted of clinical notes and addenda entered by VA staff. The available data set consisted of 1.2 billion documents and data belonging to 9.74 million patients extracted between September, 2001 and July 2011, stored in a MS-SQL server data base.

To support text search, random subsets of the population were defined and text documents and associated data belonging to 10,000 and 100,000 patients were extracted. These sample sizes were determined by processing and memory capacity. Parent documents were consolidated with their addenda. The 10,000-patient data set contained 1.14 million documents and the 100,000 patient set contained 11.32 million documents.

Lucene [12] indexes were created from the SQL database using the SOLR utility [13] and custom analyzers. These analyzers included a Porter stemmer, a small stop word list, and a filter which excluded numeric characters and date variants. Indexed fields included document and addendum text and metadata: document Title, Visit Diagnosis, Patient ID, and VA Site. Document term vectors were analyzed and stored. The sizes of the 10,000 patient and 100,000 patient indexes were 6.28 and 65.8 gb, respectively. The larger sample represented approximately 1% of the available text corpus. An overview of the system is presented in Figure 1.

A customized user interface (Figure 2) was developed to allow analysts to select a target index and execute single word, phrase, wildcard and adjacency (span) queries for multiple terms. The interface enabled designations of which fields were returned and storage of returned results as an Excel spreadsheet or SQL table. The default Lucene “practical scoring function” for terms \( t \) in query \( q \) for document \( d \),

\[
\text{score} (q,d) = \text{coord}(q,d) * (\text{queryNorm}(q) * \sum \left( \frac{\text{tf}(t \text{ in } d) * \text{idf}(t) ^ 2 * \text{tgetBoost()} * \text{norm}(t,d))}{14} \right)
\]

was used to compute a relevance score based on the vector space model. However, we did not use the relevance score to triage documents in the present analysis. Instead, SQL tables of all returned results were joined to effect inclusion or exclusion of results depending on the intent of the query. The resulting tables, consisting of document and patient identifiers, were joined to patient data stored in the clinical data warehouse.

A manual query strategy was progressively refined. Initial searches for the term “suicide attempt” returned many documents containing this phrase, but only a small fraction contained useful information. Inspection revealed that the phrase “suicide attempt” occurred often in invariant or “boilerplate” sections of documents recording suicide screenings and that most of these screenings were negative. Some screening instruments consisted of checklists where items relating to suicide assessment were followed by X’s in a “Yes” or “No” column. Simple and complex negated forms were also retrieved with the simple query, e.g. “Patient denied any history of making a suicide attempt”. Among the patients who did have positive documentation of a prior suicide attempt we often retrieved multiple documents attesting to the fact, probably reflecting closer follow-up and regular reassessment. The redundancy of positive signals permitted an aggressive filtering strategy.

This initial review informed a multi-step search approach diagrammed in Figure 3. Initially, a very inclusive query was issued, to identify all documents containing the two-word term “suicide attempt” and excluding commonly found boilerplate phrases, simple negations, and notes containing the word “Screening” in their title. Stemming allowed capture of variant expressions such as “suicidal attempts”, “suicide attempted”, etc.

In subsequent steps documents containing negations of the term “attempt”, identified by finding a
Figure 1. Search overview

Figure 2. Search interface and query example
negated term close to this target word, were excluded, using span queries. Results of queries identifying documents containing negated spans such as “no … attempt”, “never … attempt”, “attempt … denied”, etc, were consolidated. Documents in this set that matched those in the first query were excluded.

The remaining documents were grouped by patient and for the 10,000-patient sample, reviewed to obtain an estimate of retrieval precision. Patients possessing at least one true positive document were classified as being positive for a prior suicide attempt.

We also compared the results of text search with two other methods to determine suicide attempt prevalence from structured data. First, we searched administrative records to identify patients with International Classification of Disease (ICD) diagnosis codes that corresponded to a suicide attempt and second, we searched for “Health Factors”. Health Factors are structured data elements that are automatically posted when items in specialized screening documents, including some suicide screenings are selected. We treated presence of a diagnostic code or a Health Factor designating a prior suicide attempt as an unambiguous indication that the patient had made a suicide attempt.

5. Results

The inclusive search of a data set of 10,032 patients and 1.14 million documents retrieved 4103 documents belonging to 1152 patients. Span queries identified 10,141 documents containing basic negations of the term “attempt”. Eliminating the 1893 documents that matched in these two sets yielded a set of 2210 documents belonging to 507 patients. These cases were reviewed by a psychiatrist, and 402 patients were classified as true positives (Precision 0.793). To compare this result with ICD diagnosis and Health Factor data, the patient sample was searched for presence of a suicide attempt diagnosis or a Health Factor asserting a suicide attempt. 58 cases were identified as having a suicide diagnosis, and 165 patients were identified by the Health Factor method. To obtain a “best estimate” prevalence for lifetime suicide attempt, true positive patient IDs from the
reviewed text search, ICD and Health Factor groups were merged, yielding a total of 581 patients.

A similar procedure was repeated on a set of 100,979 patients, but a manual review of results was not conducted. Instead, patient IDs from text search, diagnosis and Health Factors were merged, and the total was decremented by a number representing 20.07% of the patients identified by text search. Lifetime suicide attempt prevalence estimates for the two samples are shown in Table 2. The combined lifetime suicide prevalence rates found were similar to the 4.6% rate established in a carefully designed interview survey of the general population [15]. In contrast, the lifetime suicide attempt prevalence rate obtainable from structured data alone was lower than the rate reported for the general population.

6. Discussion

Adding text search to traditional data base methods identifies more veterans who have attempted suicide than methods relying on searching structured data alone. The most recent study of VA-treated veterans’ suicide rates reported an overall rate of completed suicide among veterans that was 50% higher than that of the general population [1]. It would therefore be unexpected to find a true lifetime prevalence of suicide attempts among treated veterans lower than that found in general population. Despite a very different data collection methodology, our method found a lifetime suicide attempt prevalence similar to that found in the National Co-morbidity Survey [15], and much higher than that found using structured data alone. This suggests that text search results may correct under-reporting of phenomena that are not well captured from administrative and structured data.

Most importantly, we show the potential for text search to readily identify high-risk individuals who have attempted suicide at some point in their lives. Despite a very basic query approach, the method showed 80% precision for case identification in a sample of 10,000 patients.

Thanks to advances in information retrieval and the availability of a high performance open source search system such as Lucene [10], the feasibility of performing rapid and complex full text search is not remote. We were limited in the sample size that could be indexed by available computing resources, but this is likely to be a temporary restriction. Creation of the 100,000-patient index required 3 hours of processing time on a single machine with 4 GB of memory and produced an index 69 gigabytes in size. Query performance on this index was excellent. Scaling up to a text corpus belonging to 10 million patients might require a 100 fold increase in processing time (300 hours) and storage (7 terabytes). Parallel computing solutions are being explored.

Currently, few institutions possess a text corpus that compares in size to that of the VA, but the rapid adoption of computerized health care information systems assures that this will not long be the case. Our most important finding is that text search greatly improves recall of important information from clinical databases.

There are numerous caveats and limitations in the results we present. While the data indicate that determination of lifetime suicide prevalence for a population is feasible and the result is congruent with community sampling [15], the method we used resulted in only 80% precision. Higher precision would be desirable before, for instance, a policy of flagging medical records for increased suicide risk could be adopted. We did not undertake experiments aimed at improving precision, such as utilizing the relevance score to screen out documents, nor did we explore semantic query expansion, language modeling or proximity analysis techniques.

Negation and uninformative boilerplate text were barriers to better precision, and in this investigation we used multiple queries to include and exclude documents. These confounders are likely to be present in other epidemiologic applications of text search, given the nature of electronic health care documentation[16]. The use of techniques such as SpotSigs [17] to designate near-duplicate boilerplate text sections and implementation of linguistically motivated negation detection [18, 19] at indexing time deserve additional exploration.

Another potential criticism is that the search method did not use sophisticated natural language processing techniques such as Lucene [10], the feasibility of...
processing methods, such as ontology-informed concept extraction. This may be more difficult to implement at indexing time because of the very large number of concepts present in clinical documentation. A suitable search strategy may reduce the data load for computationally intensive linguistic processing, making application of these techniques outside the search system more practical. To support this, the system described was designed to produce output suitable for natural language processing pipelines. Another concern relates to the topic chosen: lifetime event prevalence. While this is relevant to the topic of suicide risk assessment, many important topics exhibit temporal dependency, and resolution of temporal sequences from text analysis remains a challenge [20].

On the other hand, access to even very basic search tools could be of immediate benefit to personnel charged with suicide risk assessment, such as clinicians in emergency departments and suicide hotline operators. As hinted in the second paragraph of section 1.1, locally created documents are currently more available to clinicians. A global document index could improve access to useful information captured at other facilities.

The chief advantage of adding text search to the investigative toolkit may be its ability to add value to currently archived medical records, leveraging the efforts of the clinicians who create them. Our experiment illustrates that valuable information is contained in clinical notes, but without techniques to find the information, the value is unrealized. The current use case has the potential for application in the important area of suicide prevention among veterans. Other applications are likely to be identified. The advantage of our approach – integration of search with relational database techniques and eventually with sophisticated natural language processing – is that a single index of the text corpus can serve a variety of needs.

7. Acknowledgement

This work was supported by the United States Department of Veterans Affairs, Health Services Research and Development Service, the VA Puget Sound and Salt Lake Health Care Systems and the University of Washington Information School. 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.

8. References


