Preliminary Validation of a Method to Measure Information Value in Clinical Documentation

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

Efthimis Efthimiadis, PhD  
University of Washington  
eftimiis@uw.edu

Ryan Laundry, BS  
University of Washington  
rjlaundr@uw.edu

Stephen Thielke, MD  
Dept. Veterans Affairs  
stephen.thielke@va.gov

Charlene Weir, PhD  
Dept. Veterans Affairs  
charlene.weir@hsc.utah.edu

Peter Embi, MD  
University of Cincinnati  
peter.embi@uc.edu

Ashley Hedeen, MD  
Dept. Veterans Affairs  
ashley.hedeen@va.gov

Abstract

Well-established electronic medical record (EMR) systems accumulate large volumes of documentation (notes, reports and summaries) of highly variable purpose, scope and quality. To assist treatment decisions, caregivers review electronic charts, but do so in a constrained display space. In this study we investigate whether tf-idf, a computable relevance statistic, exhibits a valid correspondence with user judgments of EMR document quality and usefulness.

110 administrators, nurses, and practitioners rated progress notes on a user-perceived EMR document quality scale. Quality ratings were compared to tf-idf computed for each document evaluated. Over the entire group, several quality dimensions significantly correlated with tf-idf, but practitioner ratings were most strongly correlated. Enhancing tf-idf via stemming and stopword preprocessing produced stronger correlations than the baseline implementation.

Tf-idf relevance shows promise as a valid proxy of practitioner-perceived document quality, with the potential to aid developing a relevance filter that allows EMR users to review documents more efficiently.

1. Introduction

Electronic medical record (EMR) systems are heralded as a cost-effective solution for promoting efficient, high quality health care [1]. EMR systems are increasingly being adopted by health care institutions, and most EMRs support clinical documentation by direct entry of narrative textual information. The advantages of on-line patient care documentation over paper are indisputable. Rapid access to physician and nursing notes, summaries and reports by staff saves time; computer notes are legible, and neither accession nor recording of information requires access to the physical chart [2]. The United States Department of Veterans Affairs Veterans Health Administration (VA) successfully implemented a comprehensive EMR in 1996 that included support for patient care documentation [3]. One consequence of success has been impressive growth in the volume of patient care documentation. Typical active VA patients have records containing hundreds to thousands of clinical notes. In the VA system, the vast majority of notes are entered manually. Document creation is assisted by electronic templates that guide formatting and content, but document entry is accomplished in a hurry: typing and grammar errors are the norm. Templates also furnish boilerplate and insert stored patient data, such as labs, medications and problem lists into documents. Currently the VA stores between one and two billion documents for approximately 8 million patients.

The pressure to document in health care is intense, and this has been accelerated by EMR adoption. In addition to supporting the mnemonic function of recording findings, history and treatment plans, clinical documentation is required for reimbursement and supports a variety of quality management functions. In the VA EMR, for example, data-driven reminders to perform additional documentation or tests must be responded to before a document can be saved. Other bits and pieces – notes of telephone contacts, messages between providers, copied e-mails, and notations by the pharmacy that a patient has picked up a prescription – find their way into the body of documentation with ease. The resulting thicket of documentation can be daunting.

To better understand the VA documentation system, the investigative team interviewed users of EMR documentation. Many insights were obtained, but one of the most prominent themes, common to
practitioners, nurses and administrators alike, was the difficulty of finding a desired piece of information among many patient care documents. Universally, difficulty finding information was seen as an impediment to efficiency[4].

In addition to understanding benefits and barriers in the EMR documentation system, our team sought to discover a statistical approach using Information Retrieval methods to automatically identify the most useful and informative documents in a patient chart. To this end, we collected a 3.7 million-document corpus of clinical notes belonging to all patients treated at the Puget Sound VA Health Care System over a 3 year period (2003-2006). This corpus was indexed and processed using the Lucene text retrieval system [5]. The Lucene system was programmed to compute a relevance measure, tf·idf, summed over all terms occurring in each document.

To bridge the qualitative and quantitative elements of our investigation we developed a web-based simulated EMR system to present test documents to users. Items in the instrument were grounded on quality concepts extracted from the user interviews [6] (table 1). Some of these quality items shown (e.g., #2 and #7, and #1 and #4) show overlap, but all were used in the analysis. Further analysis of the response factor structure will permit refinement and shortening of the questionnaire.

2. Methods

2.1. Questionnaire development

With approval from the Institutional Review Board, we conducted 14 scripted focus groups at four VA sites, interviewing 129 VA practitioners, nurses and administrative personnel in 2007 and 2008. Transcripts of the interviews were analyzed by the investigative team and 10 core themes related to user-perceived document quality were identified and expressed as a semantic differential. Respondents used a seven point scale to indicate where each note’s quality fell between the two extremes of the semantic differential.

2.2. Test documents

A patient determined to be typical of the VA population on the basis of age, gender and diagnoses and having two hospitalizations separated by outpatient care was identified in the corpus population. This patient had 90 documents spanning the care episodes. 12 documents, representing a range of content and usefulness were selected for the test. (To study a related hypothesis that documents cluttered with boilerplate and inserted data would be judged less favorably, each original source document was also transformed to a long or short form, by insertion or suppression of boilerplate and inserted data. Thus, 36 short, long and unchanged test documents derived from

<table>
<thead>
<tr>
<th></th>
<th>This note doesn’t at all tell me what’s going on with the patient</th>
<th>vs</th>
<th>This note fully tells me what’s going on with the patient</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>It’s very difficult to skim to important information in this note</td>
<td>vs</td>
<td>It’s very easy to skim to important information in this note</td>
</tr>
<tr>
<td>2</td>
<td>It’s very difficult to distinguish the author’s text from template text</td>
<td>vs</td>
<td>It’s very easy to distinguish the author’s text from template text</td>
</tr>
<tr>
<td>3</td>
<td>This note doesn’t at all help me anticipate the needs of the patient</td>
<td>vs</td>
<td>This note fully helps me anticipate the needs of the patient</td>
</tr>
<tr>
<td>4</td>
<td>I always skip over this sort of note</td>
<td>vs</td>
<td>I always read this sort of note</td>
</tr>
<tr>
<td>5</td>
<td>I can’t at all follow what the author was really thinking in this note</td>
<td>vs</td>
<td>I can fully follow what the author was really thinking in this note</td>
</tr>
<tr>
<td>6</td>
<td>I have to wade through this note to get what’s important to me</td>
<td>vs</td>
<td>I don’t have to wade through this note to get what’s important to me</td>
</tr>
<tr>
<td>7</td>
<td>This note is incomplete for this type of note</td>
<td>vs</td>
<td>This note is complete for this type of note</td>
</tr>
<tr>
<td>8</td>
<td>I don’t at all trust the information in this note</td>
<td>vs</td>
<td>I fully trust the information in this note</td>
</tr>
<tr>
<td>9</td>
<td>This note is not at all consistent with the overall clinical picture</td>
<td>vs</td>
<td>This note is fully consistent with the overall clinical picture</td>
</tr>
</tbody>
</table>

The test documents were manually stripped of identifiers, names and date-stamps and the de-identification was reviewed by a facility privacy officer prior to deploying the instrument.
2.3. Test design and presentation

Documents were presented to subjects via a previously described web-based system which closely resembled the VA EMR [6]. 110 subjects (41 administrators, 37 nurses and 32 practitioners) from the Salt Lake City and Puget Sound VA medical centers (and distinct from the original focus group participants) responded. (To support the related examination of the influence of document shortening or lengthening, a Latin Square design assured that test subjects viewed equal numbers of short, standard and long versions of the 12 different source documents and that each pattern was presented with equal frequency within each work role group and site.) Research subjects were recruited by e-mail containing a link to the test web page. Documents were presented with the EMR simulator and responses on the ten items in Table 1 were captured for analysis.

2.4. Indexing and corpus preparation

3.7 million EMR text documents, representing all documents produced over a three year period at VA Puget Sound were downloaded to an MS-SQL 2008 database, and exported as xml for analysis. The Lucene system was configured to index this corpus, and to compute term and inverse document frequencies as described next.

2.5. Using tf·idf to weight documents

Tfidf (term frequency X inverse document frequency) is a robust relevance statistic. Empirically derived, it has been shown to capture the “distinctness” of a document in a corpus and is valuable for relevance ranking of documents retrieved by a query [7]. Tf conveys the “importance” of a term in a document, and is expressed as the count (f) of term k occurring in document i containing t distinct terms, expressed as a proportion.

\[ t_i k = \frac{f_{ik}}{f_{ik}} \]

Idf, the inverse document frequency, represents how often documents containing that term occur in the corpus. Because more “interesting” terms occur less often, idf reflects the “importance” of a term in the entire document corpus. In a corpus of N documents, wherein term k occurs in n of the documents, idf of term k is expressed as the logarithm:

\[ idf_k = \log \frac{N}{n_k} \]

Terms that occur very rarely in a corpus occur in correspondingly few documents, so the inverse document frequency associated with a rare term is large. The product of tf and idf, tf·idf, combines document and corpus “importance” of a term in a single scalar value.

A general expression for the weight d of term k in document i in a vector space model is:

\[ d_{ik} = \frac{(\log(f_{ik})+1) \cdot \log(N/n_k)}{\sqrt{\sum_{k=1}^{t} [(\log(f_{ik})+1.0) \cdot \log(N/n_k)]^2}} \]

where N represents the total number of documents in the corpus, n_k represents the number of documents in which term k occurs and f_{ik} represents the number of occurrences of term k in document i. The vector space model is used to compare the similarity of terms in a query with terms in a document to be retrieved.

Without knowledge of what information an EMR user seeks during a record review, it is not possible to formulate a specific query to assist information retrieval. However, it is possible to use tf·idf scoring to compute a proxy of general document relevance by summing individual tf·idf contributions for all terms in the document. Documents scoring higher on tf·idf summed over all terms contain more occurrences of infrequently encountered terms, and are hypothesized to be more “distinctive”, “different” or “interesting” than documents with a low summed tf·idf score.

Tfidf was computed for each of the short, long and standard versions of the 12 source documents presented. The subsequent analysis compared these specific tf·idf scores with test subjects’ corresponding quality responses. (Note: the length transformation exerted a very minor effect on tf·idf scores of transformed documents. We have reported elsewhere [8] that altering document length did not significantly influence quality judgments made by test subjects).

2.6. Basic approach and refinements

Our initial approach utilized the Lucene “white space” analyzer to separate and index all tokens delimited by punctuation or blank spaces. Lucene’s default list was used to eliminate 18 common “stop” words (“the”, “is”, “and”, etc). Tf·idf was computed as described above for each document, adjusting for document length by dividing the summed tf·idf score by the sum of individual tf’s encountered.
Scored documents were sampled and reviewed manually by a physician. The initial reviews consisted of examining high-scoring outliers and comparing the tf·idf scores of documents belonging to individual patients to assess the ability of tf·idf to discriminate “more interesting” from “less interesting” documents. It was observed that frequently repeating tokens, especially date values and words like “patient” found in boilerplate sections tended to boost tf·idf scores without adding to document interest. In an iterative process, the effect of additional preprocessing and algorithm “tuning” was assessed by inspection. Statistics from the corpus were used to inform choices.

Based on manual review of case documents additional preprocessing and algorithm selection choices were selected: elimination of additional “stop” words to reduce noise, aggressive tokenization to filter out confusing punctuation commonly found in EMR documents, filtering of all numeric characters and stemming to reduce inflected word forms. Excessive repetition of terms in documents was damped by using $\sqrt{f_{ik}}$ rather than $f_{ik}$ when summing document tf·idf weights. For this analysis we compared results using Lucene’s basic “white space” analyzer and 18 common stop words with an enhanced analyzer which implemented the following parameters:

- Conversion of all characters to lower case,
- Conversion of all punctuation to white space (e.g., “tf·idf.” becomes “tf idf”
- Filtering the 300 most-frequent words in the corpus (e.g., “patient”, “veteran”, “male”)
- Filtering the 100 most-frequent words that repeated five or more times in any document,
- Filtering numeric characters and alphanumeric compounds (e.g., “A1c”),
- Filtering date expressions (e.g. “9/1/05”, “September 1, 2005”)
- Use of $\sqrt{f_{ik}}$ instead of $f_{ik}$ as the multiplier for summing tf·idf

Baseline and enhanced tf·idf were computed for the long, short and standard versions of the 12 test documents, resulting in two sets of 36 tf·idf scores. Each study subject evaluated 12 of 36 possible test documents (4 standard, 4 short and 4 long versions of the 12 original source documents).

### 2.6. Statistical analysis

Downloaded response data and corresponding baseline and enhanced tf·idf scores were analyzed using the Stata v.9 statistics package. Spearman’s rank order and Pearson correlations and a factor analysis were performed.

### 3. Results

Figure 1 shows the Spearman rank order correlation rho ($\rho$) of the baseline and enhanced tf·idf versus the ten quality items assessed over the entire group of subjects. In this view, both versions of tf·idf were positively but weakly correlated with quality judgments. Most of the correlations were statistically significant, but their weakness suggests that the enhanced tf.idf has limited practical value as a general proxy of document quality.

![Figure 1. Overall: correlation of baseline tf·idf and Quality Judgments](image-url)
**Figure 2. Subgroup rank-order correlation of baseline tf-idf and Quality Judgments**

Figure 2 displays correlations of the baseline tf-idf and quality estimates, by user group. Uniformly, correlations between practitioner’s responses and baseline tf.idf were stronger than those of nurses and administrators. All of the practitioner correlations, while weak, were statistically significant and notably stronger than those observed in the nurse and administrator groups, suggesting a correspondence between baseline tf.idf and practitioner document quality judgments.

Figure 3 displays correlations of the enhanced tf.idf and quality estimates broken down by user group. For practitioners, the enhanced tf.idf showed stronger correlations, all statistically significant. The Spearman rank order coefficient ρ exceeded 0.4 on item 4, “this note helps me anticipate the needs of the patient”, and item 6, “I can follow what the author was really thinking in this note”, and exceeded 0.3 on five other items. Because Spearman’s rho is a rank order correlation, a Pearson product-moment correlation coefficient was computed as a check. For practitioners the Pearson regression coefficient was 0.44 for item 4 and 0.48 for item 6. Correlations of this strength are considered moderately strong in the social and behavioral sciences [9].

**Figure 3. Subgroup rank-order correlation of enhanced tf-idf and Quality Judgments**
Table 2. Factor analysis of quality dimensions using subject responses

<table>
<thead>
<tr>
<th>Item</th>
<th>Factor 1</th>
<th>Factor 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. This note fully tells me what’s going on with the patient</td>
<td>0.7662</td>
<td>0.4164</td>
</tr>
<tr>
<td>2. It’s very easy to skim to important information in this note</td>
<td>0.4472</td>
<td>0.7660</td>
</tr>
<tr>
<td>3. It’s very easy to distinguish the author’s text from template text</td>
<td>0.3608</td>
<td>0.5900</td>
</tr>
<tr>
<td>4. This note fully helps me anticipate the needs of the patient</td>
<td>0.7568</td>
<td>0.4534</td>
</tr>
<tr>
<td>5. I always read this sort of note</td>
<td>0.2311</td>
<td>0.4235</td>
</tr>
<tr>
<td>6. I can fully follow what the author was really thinking in this note</td>
<td>0.6938</td>
<td>0.5381</td>
</tr>
<tr>
<td>7. I don’t have to wade through this note to get what’s important to me</td>
<td>0.4093</td>
<td>0.7658</td>
</tr>
<tr>
<td>8. This note is complete (for this type of note)</td>
<td>0.7799</td>
<td>0.2331</td>
</tr>
<tr>
<td>9. I fully trust the information in this note</td>
<td>0.7204</td>
<td>0.3710</td>
</tr>
<tr>
<td>10. This note is fully consistent with the overall clinical picture</td>
<td>0.7979</td>
<td>0.3315</td>
</tr>
</tbody>
</table>

The jagged pattern of correlation with tf.idf along the quality dimensions suggested that the items may be loading on different underlying factors. An exploratory factor analysis conducted using all subjects’ responses yielded two main factors, as shown in table 2. At face value, Factor 1 appears to relate to the “informativeness” of a note, and Factor 2 appears related to “readability”.

Table 3 displays how these factors correlated with enhanced tf.idf, overall and by subgroup. As seen with the results above, it appears that practitioners’ responses are driving trends seen in the overall group. Practitioner responses were similar at both sites.

Table 3. Correlation of tf.idf and factor scores

<table>
<thead>
<tr>
<th></th>
<th>Factor 1</th>
<th>Factor 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>ρ = 0.1980; p &lt; 0.0001</td>
<td>ρ = 0.2108; p &lt; 0.0001</td>
</tr>
<tr>
<td>Practitioners</td>
<td>ρ = 0.3959; p &lt; 0.0001</td>
<td>ρ = 0.4324; p &lt; 0.0001</td>
</tr>
<tr>
<td>Practitioners – Puget Sound</td>
<td>ρ = 0.4063; p &lt; 0.0001</td>
<td>ρ = 0.4129; p &lt; 0.0001</td>
</tr>
<tr>
<td>Practitioners – Salt Lake</td>
<td>ρ = 0.3938; p &lt; 0.0001</td>
<td>ρ = 0.4729; p &lt; 0.0001</td>
</tr>
<tr>
<td>Nurses</td>
<td>ρ = 0.1309; p &lt; 0.0081</td>
<td>ρ = 0.1250; p &lt; 0.0115</td>
</tr>
<tr>
<td>Administrators</td>
<td>ρ = 0.0983; p &lt; 0.0393</td>
<td>ρ = 0.1082; p &lt; 0.0233</td>
</tr>
</tbody>
</table>

4. Discussion

Tf.idf appears to be a useful proxy for practitioners’ perceptions of document quality. This finding has implications for design of EMR document user interfaces used by practitioners. Incorporating computed relevance measures to filter document displays may be of particular benefit for this user group.

Enhancing tf.idf by preprocessing to reduce term noise improves the correlation of tf.idf and practitioner quality judgments.

Tf.idf does not appear to be as robust a proxy of nurses’ or administrators’ quality judgments. While the data do not offer an explanation of why this is so, it is possible that differences exist in the reading tasks and goals of different users. Another possible explanation is that because a physician guided refinement of the tf.idf enhancement strategy, a bias toward a practitioner perspective was introduced.

The small size of the test document sample limits the generalizability of our observations, but these preliminary results appear promising and warrant further research. Areas of investigation suggested by the findings include:

- Expanded use of other document features (document titles, author role, and associated clinical data such as orders and the patient problem list)
• Investigation of customized relevance algorithms for nurses and administrators
• Incorporation of language modeling in quality determinations
• Investigating the impact of document displays incorporating relevance filtering on decision making performance.

5. 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.

6. References