

# Evaluating Measures of Redundancy in Clinical Texts

Rui Zhang, MS<sup>1</sup>, Serguei Pakhomov, PhD<sup>1,2</sup>, Bridget T. McInnes, PhD<sup>2</sup>,  
Genevieve B. Melton, MD, MA<sup>1,3</sup>

<sup>1</sup>Institute for Health Informatics; <sup>2</sup>College of Pharmacy; and <sup>3</sup>Department of Surgery  
University of Minnesota, Minneapolis, MN

## Abstract

*Although information redundancy has been reported as an important problem for clinicians when using electronic health records and clinical reports, measuring redundancy in clinical text has not been extensively investigated. We evaluated several automated techniques to quantify the redundancy in clinical documents using an expert-derived reference standard consisting of outpatient clinical documents. The technique that resulted in the best correlation (82%) with human ratings consisted a modified dynamic programming alignment algorithm over a sliding window augmented with a) lexical normalization and b) stopword removal. When this method was applied to the overall outpatient record, we found that overall information redundancy in clinical notes increased over time and that mean document redundancy scores for individual patient documents appear to have cyclical patterns corresponding to clinical events. These results show that outpatient documents have large amounts of redundant information and that development of effective redundancy measures warrants additional investigation.*

## Introduction

Widespread implementation of electronic health record (EHR) systems has resulted not only in fundamental changes in clinical workflow but also rapid proliferation of electronic clinical texts. While EHR system adoption is intended to promote better quality, decrease costs, and increase efficiency in healthcare, there are some secondary effects of EHR use, which may not always be desirable<sup>1,2</sup>. With respect to clinical notes, it has been observed that electronic clinical note creation may be slow<sup>3</sup> and prone to inserting redundant information (including “copy and paste”)<sup>4</sup> from previous notes. While having information readily available in EHR systems is helpful, excessive redundant information can lead to a cognitive burden, information overload, and difficulties in effectively distilling relevant information for effective decision-making at the point of care<sup>5</sup>.

There are few reports on efforts aimed at quantifying redundancy in clinical text<sup>6</sup> indicating that redundancy metrics for clinical texts have not yet been well developed. Further work in this area is motivated by the fact that automated redundancy metrics may play an important role in designing interfaces to clinical texts in both research and production EHR systems. We believe the prevalence of redundant information in electronic documents may hinder review of these documents, an important process performed by clinicians with patient care. Computational methods that are able to identify redundant text and point out new relevant information may decrease the cognitive load of a clinician that is viewing the document. This will be a potential application in next generation EHR systems, which may save clinician time, improve clinician satisfaction and possibly improve healthcare efficiency.

The objective of this study was to explore several possible approaches for measuring redundancy in clinical text, particularly between notes in a single patient record. To this end, we developed an expert-derived reference standard of redundancy, several redundancy metrics with modification of classic dynamic programming alignment techniques, and enhanced these metrics using both statistical and knowledge-based tools.

## Background

### ***Semantic Similarity, Patient Similarity, and Relationship to Information Redundancy***

Similarity is a fundamental concept that is essential to automated information integration, case-based similarity, inference, and information retrieval tasks<sup>7,8</sup>. With respect to assessing similarity at a patient or case level, effective metrics quantify how similar two patients are to one another, based upon the question at hand (such as overall similarity or similarity from a diagnostic standpoint)<sup>9,10</sup>. These comparisons have been conceptualized as measures of similarity based upon complex *sets* of concepts representing each case and has been classically described as a commutative or symmetrical measure (i.e., Similarity of (A,B) = Similarity of (B,A)). However, in the context of measuring information (semantic) redundancy in EHR documents that are created sequentially over time, asymmetrical measures may be required as information is continually added and/or repeated in more recent notes compared to older documents.

In contrast to case-based similarity, concept-level semantic similarity metrics quantify the closeness in meaning between two concepts (versus two groups of concepts such as in the example of case-based similarity)<sup>11,12</sup>. Semantic similarity has been studied extensively both in general language and in biomedicine. Automated measures of both

patient similarity and semantic similarity can be generally classified into knowledge-based approaches and knowledge-free approaches. Knowledge-free approaches rely upon statistical measures such as term frequency and co-occurrence data. Because of the complexity of the medical domain, rich use of synonymy and related concepts, the performance of knowledge-free approaches may not be optimal. Knowledge-based approaches utilize additional information, such as ontological information, definitional data, or domain information to enhance these methods. In the context of automated measurement of information redundancy, measures of semantic similarity may be useful to perform semantic normalization between pieces of text that are being compared to determine the degree of redundancy. For example, theoretically, it may be useful to treat orthographically different but semantically synonymous or highly similar terms as equivalent (e.g., *heart* vs. *cardiac*).

In contrast to similarity, information redundancy (at the semantic level) between two items has been studied less. Information redundancy is conceptually a measure of the degree of identical and/or redundant information in an item of interest (subject item) contained within another item (target item). For example, when comparing subject item A to target item B, the redundancy of information within subject item A contained in target item B is conceptually the information contained in both A and B, normalized by the information in item A --  $|A \text{ and } B|/|A|$ . Redundancy therefore depends upon which item is subject and target and is not commutative.

### ***Global and Local Sequence Alignment***

Alignment techniques can be generally separated into two categories: global alignment and local alignment. In biomedicine, both of these types of algorithms were initially applied for quantifying similarity of two genetic sequences in order to discover or speculate whether sequences might be evolutionary or functionary related. Global alignment identifies the overall similarity of the entire length of a sequence compared to another sequence, and thus is most suitable when the two sequences have a significant degree of similarity throughout and are of similar length. In contrast, local alignment detects similarity of smaller regions within long sequences. This type of alignment is most suitable when comparing substantially different sequences, which possibly differ significantly in length, and contain only short regions of similarity. Typically, there is no difference between global and local alignment when sequences are sufficiently similar. While local alignment allows for the measurement of overlap or similarity over short sequences, it does not provide an aggregate measure of local similarities throughout one sequence compared to another. In contrast, global alignment assumes a single full alignment of two sequences of interest<sup>13</sup>.

### ***Assessing redundancy of clinical text with alignment***

Wrenn et al.<sup>6</sup> recently reported on redundancy in inpatient notes using global alignment to compare documents for 100 inpatient hospital stays. They measured the amount of text duplicated based on global alignment from previous notes and found that signout notes and progress notes had an average of 78% and 54% of information duplicated from previous documents. Duplicative information increased with the length of the hospital stay. This study established that redundancy in inpatient notes is common but did not validate the use of global alignment as a measure of redundancy or take into account conceptual alignment (i.e., semantic similarity), lexical variation, or non-content words (i.e., stop words).

A limitation of the global alignment approach in the context of clinical reports becomes apparent in situations when note sections may appear out of “normal” sequence. For example, if the same two sections in several clinical notes are in a different order but are otherwise highly redundant, the global alignment approach would be unsuitable and would grossly underestimate the degree of redundancy between the notes. In contrast, local alignment techniques alone would not be suitable as these measures would provide a measure of similarity over a short sequence but no aggregate measure over the entire note of information similarity. As an initial approach to this problem, we decided to explore an alignment method using a short, contained window of focus for each alignment and then aggregating the scores of these windows over the entire text of interest as detailed in the “Methods” section.

## **Methods**

### ***Study Setting and Data Preparation***

One hundred and seventy-eight complete outpatient clinical records from University of Minnesota Medical Center, Fairview Health Services in patients with angina, diabetes, or congestive heart failure followed by the Pharmaceutical Care Department for optimal medication management were used for this study. Each complete outpatient record contained all clinical notes including office visits, allied health nursing notes, telephone encounters, and results during a one-year period from December 2008 to November 2009. These notes were originally created in the Epic electronic health record system and extracted in text format. Inpatient notes from any of the Fairview Health Services hospitals were excluded for the purposes of redundancy measurement for this analysis. We assumed that there was no redundant information in the first document. Intra-document redundancy

and semantic alignment were not considered in this study. Outpatient notes were organized chronologically for each patient as detailed in the “Text pre-processing” section of this paper and utilized for this analysis. University of Minnesota institutional review board approval was obtained and informed consent waived for this minimal risk study.

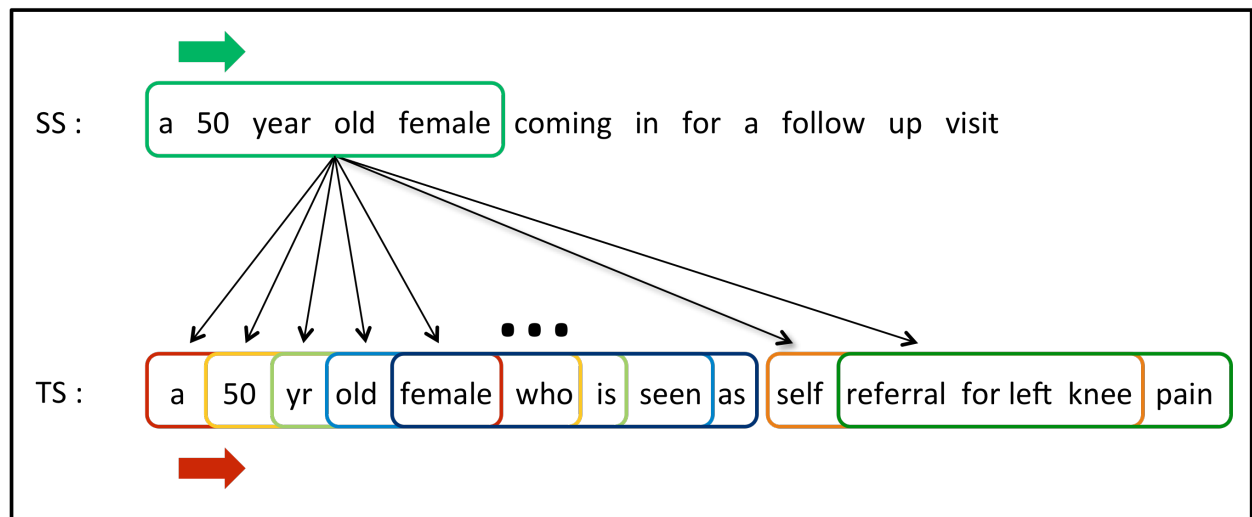
A. Sentence Pair

SS	a 50 year old female coming in for a follow up visit.
TS	a 50 yr old female who is seen as self referral for left knee pain.

B. Prior Alignment

SS	a 50 + year old female + + + + + coming in for + + a follow up visit.
TS	a 50 yr - old female who is seen as self referral - - for left pain.

C. Window Sliding



D. Progressive Modification of Sentences

Subject Sentence (SS)	
Baseline	a 50 year old female coming in for a follow up visit
CSW	50 year old female coming follow up visit
TSW	50 year old female coming follow up
BSW	50 year old female coming follow up
LVG_BSW	50 year old female come follow up
SIM_BSW	50 year old female come follow up
SIM_LVG_BSW	50 year old female come follow up
Target Sentence (TS)	
Baseline	a 50 yr old female who is seen as self referral for left knee pain
CSW	50 yr old female who seen as self referral left knee pain
TSW	50 yr old female who seen as self referral left knee pain
BSW	50 yr old female who seen as self referral left knee pain
LVG_BSW	50 yr old female who see as self referral left knee pain
SIM_BSW	50 year old female who see as self referral left knee pain
SIM_LVG_BSW	50 year old female who see as self referral left knee pain

Figure 1. Schematic of automated redundancy measures between a subject sentence (SS) and a target sentence (TS). A. Sentence pair; B. Prior alignment<sup>6</sup> including matches, additions (+) and subtractions (-); C. First frame of SS align with all frames (differed by color, omit few frames in the middle) of TS; D. Sentences are modified by various measures.

### ***Automated Redundancy Measures***

As a baseline metric, alignment between two texts was performed using a modification of the Needleman-Wunsch algorithm, a dynamic programming technique commonly used in the bioinformatics field to align protein or nucleotide sequences. We modified this algorithm to the constraints of clinical notes, as described, with text pre-processing and a sentence/statement alignment process at the word level (as opposed to a character level). We present an overview of this measurement's processing in Figure 1.

#### *Text pre-processing*

Because of the potential issues of underestimating redundancy with aligning an entire text globally, we chose instead to chunk and process small overlapping portions of text (frames) sequentially within a particular note and then to derive an overall metric for the note based on these measures for individual frames. This approach allows for the exact sequence of words within a note to become minimally important and the content to remain most important. After each complete patient record was split into individual clinical notes, each note was further separated into small chunks of text at a sentence/statement level using a rule-based sentence splitter. Due to the nature of clinical discourse in EPIC notes, not all sentences are well-formed (e.g., "family history of heart disease" may appear as the only text on a line or as part of an enumeration). We refer to the incomplete sentences as "statements" but treat them as sentences for the purposes of this study.

#### *Baseline Sentence/Statement Redundancy Measurement Using Alignment*

The content of one text of interest, the *subject text*, was compared to another text of interest, the *target text*. At the sentence or statement level, each pair of subject and target sentences (SS and TS respectively) was aligned at the word level. We modified the Needleman-Wunsch algorithm to align a window in the SS to the TS in an iterative fashion. The alignment steps were as follows, 1) each SS was split into overlapping frames of five consecutive words; 2) each frame of the SS was then aligned with all frames of the TS by advancing a sliding window one word at a time (Figure 1C); 3) the maximum alignment score for a pair of frames was defined as the number of matched words for each frame of the SS with penalties of word addition or deletion; 4) the window positioned over the SS text was then advanced by one word and aligned as described in the previous step; 5) the final alignment score for the SS text as compared to the TS text was calculated by averaging all maximum TS scores for each SS frame; 6) scores were then normalized by the window size and used to build up a baseline redundancy matrix between pairs of SS vs. TS texts (score range from 0 to 1). Baseline measures were then used to perform stratified random sampling to create the reference standard sentence pair set.

#### *Reference Standard*

Two hundred and fifty sentence pairs selected based on the baseline redundancy scores were chosen from five patient records using stratified random sampling. The sampling consisted of splitting the pairs of sentences in each record into quintiles and then selecting five random sentence pairs with scores in each of the quintiles. Two physicians were asked to judge the redundancy of information on a scale from 1 to 5, with "0.5" scores allowed (i.e. 2.5). The physicians were asked to base their assessments on how much information contained in the SS text they were able to also find within the TS text. They were also asked to compare the information content of the texts as opposed to just comparing the words. The highest score of 5 indicates that all the information in the SS was contained in the TS, while lowest score of 1 indicates that none of the information in the SS was contained in the TS. After calculating agreement, physician scores were averaged to form our reference standard to validate the automated scoring methods.

Inter-rater reliability between the two experts was calculated using Cronbach's Alpha<sup>14</sup>. Inter-rater agreement with Cronbach's Alphas was 0.91. The correlation between the expert ratings was also high (0.871 in Table 1). Expert ratings were averaged to create the reference standard for evaluating all our methods. Both 1) expert evaluations were correlated to one another as a measure of optimal upper-bound performance and 2) automated redundancy measures were correlated to the reference standard using the non-parametric Spearman's rank correlation coefficient<sup>15</sup>.

#### *Implementing enhancements to baseline redundancy measure*

We implemented each technique using a window size of 5 words. The choice of this window size was motivated by prior work showing that the average length of medical terms found in outpatient clinical notes is between 4 and 5 words<sup>16</sup>. In addition to the baseline redundancy metric, which aligned unaltered raw text (Baseline), we experimented with several modifications of our baseline redundancy measure:

- Removal of classic stop words<sup>17</sup> (CSW). This method was based on removing stop words (e.g., “the”, ”a”, ”for”, ”it”, ”this”, etc.) that are generally removed by text indexing and retrieval systems.
- Removal of stop words defined by Term Frequency–Inverse Document Frequency (TFIDF) using optimal thresholds of the TFIDF distribution based upon the entire note corpus (TSW). TFIDF is another method used in standard text indexing and retrieval systems to remove or deemphasize words that occur frequently in many documents and thus are less likely to be useful for ranking documents by their relevance to a query.
- A combination of CSW and TSW, with removal of both classic stop words and stop words defined by optimal TFIDF thresholds (BSW).
- Removal of both stop word types (BSW) and lexical normalization to effectively treat lexically different forms of the same term as equivalent when aligning text using Lexical Variant Generation (LVG)<sup>18</sup> (LVG\_BSW).
- Removal of both stop words (BSW) and treating terms with high semantic similarity as equivalent when aligning text using the Unified Medical Language System (UMLS) and path-based UMLS::Similarity measures<sup>12</sup> (SIM\_BSW). For this, we used a cut-off score of 0.8 from the UMLS::Similarity measure to identify synonymous or near-synonymous terms (e.g., “above” – “upper”, “advice” – “guidance”, etc.).
- Removal of both stop word types (BSW), aligning text using UMLS::Similarity, and lexical normalization using LVG (SIM\_LVG\_BSW).

As an additional baseline, we aligned text as described previously by Wrenn et al.<sup>6</sup> using the Levenshtein edit-distance algorithm at a word-level without window movements (“Prior”). The window size was also examined as a factor and was varied to increments of 4, 5, 8, and 10. Figure 1 shows two example sentences and illustrates how each approach was implemented.

### ***Outpatient record redundancy***

For each patient, all notes were arranged chronologically and each note of interest (subject note) was compared with all the previous notes (target notes). At a document level, windowing was not allowed to cross sentence boundaries so as not to penalize for not preserving information across sentences. The score for each frame was defined as the maximum score with the automated method, comparing the target frame text with the text from all previous notes using LVG\_BSW method. Using this technique, a set of frame scores and their distribution were created for each note. A mean score was assigned for each note by averaging all frame scores. Based on these mean scores, the redundancy between documents was derived first as descriptive statistics of scores over all the documents. A physician (GM) examined three patient records and recorded the purpose of each visit and any noteworthy clinical events. These observations were then overlaid graphically with the mean redundancy scores of documents chronologically.

Last, average document redundancy scores for patients over time were calculated to detect temporal redundancy trends. Redundancy scores for each patient document were normalized to account for different numbers of notes in each record. Normalization was performed by pooling redundancy scores for each patient into even quartiles chronologically over the entire time period, so that the first 25% contained the earliest notes and fourth 25% had the most recent notes. Using the approximately 900 clinical note corpus from 178 patients, each data point with standard error bars illustrates the average redundancy of more than 200 clinical notes. We then included a smoothed curve along with the original data points to visualize the trend in redundancy scores across patient notes with time.

## **Results**

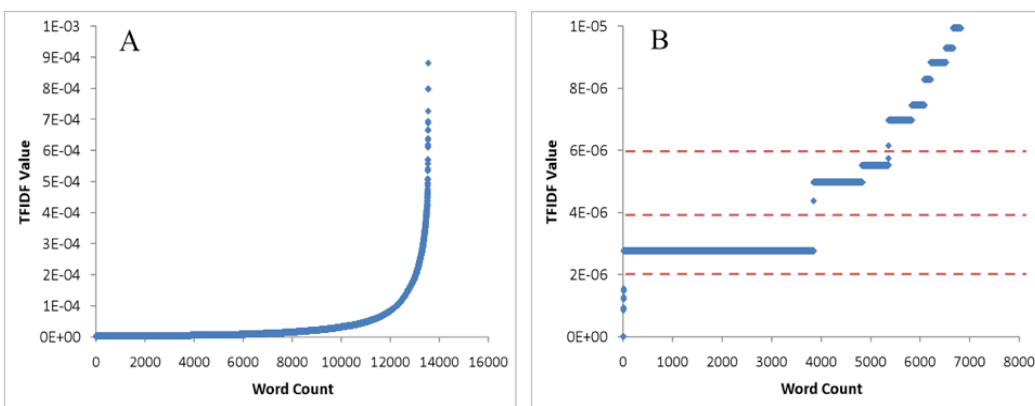
### ***Automated Redundancy Measures***

All measures and their correlation with the reference standard are listed in the Table 1. We performed TF-IDF scoring to experiment with different thresholds for stop-word removal. As illustrated in Figure 2, there were several potential stop-word cutoffs. We tested the first 3 cutoffs of 2E-6, 4E-6, and 6E-6, and found that a cutoff of 2E-6 provided the highest correlation (0.780, 0.777, and 0.778 respectively). Table 1 summarizes the correlations with the reference standard for each of the methods, including using LVG\_BSW with different window sizes. Comparison between various methods for calculating redundancy scores (Table 1) on our reference standard showed that removing stop words with both the classic stop word list and the optimized TFIDF scoring yields higher correlations with human redundancy judgments than using global alignment or the baseline local alignment. Adding lexical normalization further improves correlation albeit by a small amount; in contrast, semantic normalization with UMLS::Similarity path-based measure using a threshold of 0.8 does not either improve the correlation with our reference standard, nor does it make the correlation worse.

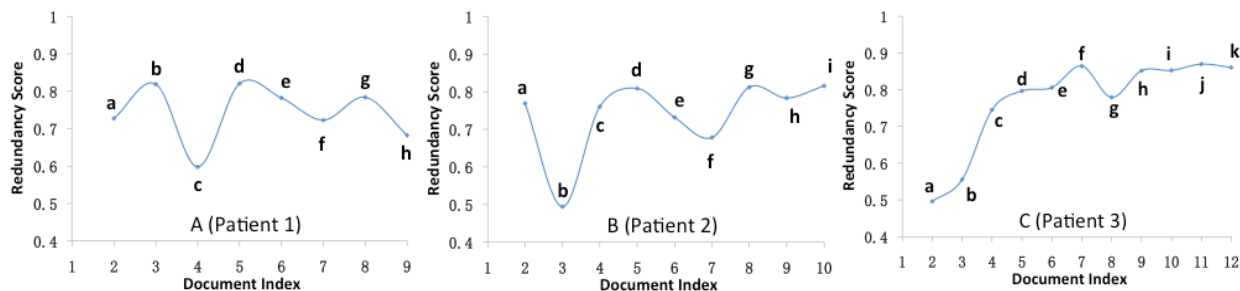
**Table 1.** Correlation of methods compared to reference standard.

\*Experts = correlation of ratings between two raters. Prior = Prior method<sup>6</sup>; Baseline = unaltered raw text.

Method	Spearman Coefficient
Experts*	0.871
Prior (global alignment)	0.759
<u>Methods of redundancy</u>	
Baseline (window 5)	0.781
CSW (window 5)	0.785
TSW (window 5)	0.780
BSW (window 5)	0.814
SIM_BSW (window 5)	0.816
LVG_BSW (window 5)	0.824
SIM_LVG_BSW (window 5)	0.823
<u>Varying window size</u>	
LVG_BSW (window 4)	0.834
LVG_BSW (window 5)	0.824
LVG_BSW (window 8)	0.803
LVG_BSW (window 10)	0.801



**Figure 2.** A. TFIDF value distribution of the whole corpus; B. Magnified view of TFIDF distribution showing three TFIDF cutoff values, which are marked as red dashed lines.

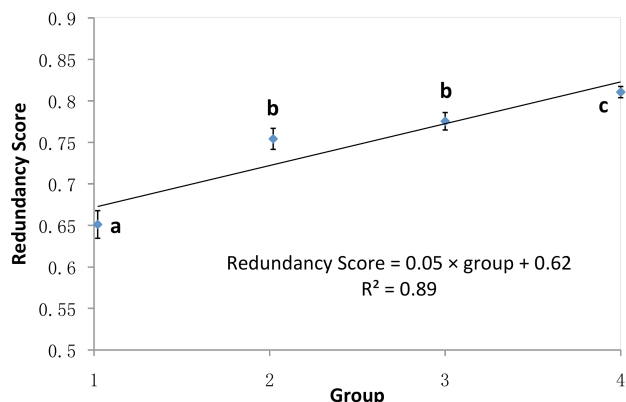


**Figure 3.** Patterns of redundancy scores in outpatient documents with documents indexed in a chronological order.

**A) Patient 1:** a) health maintenance visit; b) health maintenance visit, minimal upper respiratory tract symptoms; c) motor vehicle accident (MVA) with multiple musculoskeletal complaints, headache; d) follow-up of MVA symptoms; e) pre-operative general assessment for minor surgery; f) care following emergency department for congestive heart failure (CHF) exacerbation; g) health maintenance visit; h) visit for total body itchy rash, diagnosed with scabies.

**B) Patient 2:** a, c, & d) health maintenance visit; b) change in insurance and change in medication (short note); e) new upper respiratory tract infection (URI); f) urinary tract infection & fever; g) ongoing URI symptoms; h & i) diabetes-focused health maintenance visit.

**C) Patient 3:** a) right lower extremity (RLE) ankle tender and red (short note); b) recurrent RLE cellulitis and rash; c) follow-up of RLE symptoms; d, e, f, g, h, i, j, & k) health maintenance visit and ongoing RLE symptoms.



**Figure 4.** Redundancy scores (mean  $\pm$  standard error) of document quartiles. A linear regression line (with function and  $R^2$ ) between redundancy scores and group number was drawn to visualize a trend. Means with different letters were significantly different ( $p < 0.05$ ).

### **Outpatient record redundancy**

The mean, standard deviation, maximum and minimum of redundancy scores for all the patient documents in our corpus were 0.74, 0.14, 0.96 and 0 respectively. Several different patterns of redundancy scores were observed when examining individual patient records (Figure 3). Three outpatient records with at least 8 notes were examined with visit purpose and clinical events notes. These events and the mean document redundancy scores were plotted (Figure 3). We observed the presence of cycles in redundancy scores at the individual patient record level, which appeared to correlate with clinical events in most cases.

Figure 4 shows the means and standard errors of the redundancy scores pooled into quartiles (groups 1-4) over all clinical notes in all available patient records with 4 or more notes (because of the split into quartiles). While redundancy at the individual patient record level appears to be cyclical (Figure 3), overall redundancy scores across all patient records temporally have a clear upward trend (Figure 4). The redundancy scores in the fourth quartile (most recent) were significantly higher than all earlier quartiles. The scores in the 3rd quartile were also higher than in the 2nd quartile but this was not statistically different. There also exists a good linear relationship ( $R^2 = 0.89$ ) between the number of groups and corresponding redundancy scores.

### **Discussion**

This study focuses on an understudied and increasingly important problem in clinical documentation, information redundancy. In this exploratory study, we successfully developed an expert-based reference standard and compared our standard to automated measures, including a previous measure based on global alignment, a baseline measure using alignment over short word sequences and enhancements to this measure using a combination of knowledge-based and statistical corpus-based knowledge-free approaches. With respect to the overall level of redundancy with time in clinical text, our results are consistent with previous reports<sup>6</sup> and confirm the finding that information redundancy in clinical notes (in this case outpatient documents) is significant. Our results indicate that content words (as opposed to standard and statistically-based stopwords) are most important to be considered as features for redundancy identification. However, lexical normalization and semantic similarity may also be promising techniques for follow-up studies. Furthermore, the sliding window technique with aggregation performs significantly better than global alignment for assessing clinical text redundancy.

When using different automated methods to quantify redundancy, we achieved good correlation with our expert-derived reference standard. The use of a combination of standard stopword removal and TFIDF threshold stopword removal was optimal over either single type of stopword removal alone. In addition, both lexical normalization and semantic similarity enhanced these measures, although by a small amount. While lexical normalization results were slightly better in this study, it is possible that further enhancements to semantic similarity measures, including more effective mapping to named-entities with text chunking or shallow parsing, would help to identify multi-word concepts (i.e., diabetes mellitus) and improve performance. Other potential enhancements include application of abbreviation and acronym disambiguation as these are common in clinical text. Also, we only utilized the path measure from the semantic similarity package with a single cut-off, and did not look at other semantic similarity and relatedness measures that our group has developed, including second-order vector-based measures<sup>12</sup>. In addition, we did not consider the idea of intra-document redundancy, a problem poorly characterized to date. This would likely be, however, a significantly more computationally difficult problem since this technique would require alignment

not only with previous documents but also with all previous text within the current document, and therefore a dynamic statistical language model within the document.

Our investigation of the effect of various window sizes indicated that the performance of our alignment approach was improved with the decrease in window size as seen in Table 1. For window size of eight and ten, the performance worsened. This could be due to the average length of selected sentence pairs (13 words) approaching this larger window size. Although a smaller window size of four resulted in slightly higher correlation with human judgments of redundancy, smaller window sizes may also result in generating spurious alignments between portions of medical terms rather than entire terms. This may be an issue when the measure is applied to a large set of clinical documents rather than individual sentences. In addition, computational efficiency decreases with decreasing window sizes resulting in more text frames to be compared. The window size of five therefore represents a tradeoff between accuracy and efficiency, as well as meaningfulness of generating alignments that capture most of the content of a medical term. While somewhat inefficient to have the sliding window align over the entire text, we anticipate that this is a tractable method, as we would envision applying these metrics to text a single time and storing this information as part of a display feature in a graphical user interface for electronic text.

We observed several patterns of redundancy scores with different patient documents. Figure 3 shows that most patients demonstrated cyclical patterns in the mean redundancy scores for documents of a given patient. To investigate if redundancy scores could detect redundant and new information, three patient records were reviewed and changes in redundancy scores generally correlated with clinical events, such as a motor vehicle accident, loss of insurance and a medication change, or a new visit to the emergency room for a congestive heart failure exacerbation. We observed that a document with redundant information had a high redundancy score on the peak of the graph and a document with a significant event had a lower redundancy score resulting in a trough on the curve (i.e., Figure 3A(c), 3B(b) and 3B(f)). These findings indicate another potentially interesting and beneficial use of automated approaches for identifying redundant and new information. These approaches may be used to identify salient or unusual events in the patient's history and thus may aid the clinician in quickly constructing the "background" for the current visit.

There was also an observed trend towards an increase in overall redundancy of information with time in the clinical records, which was reported previously by Wrenn et al.<sup>6</sup>. In practical terms, the methods we developed for assessing redundancy of information in clinical notes could potentially be used to automatically identify non-redundant information and present notes to the clinicians at the point of care in an electronic health record system in a more easily digestible manner. Furthermore, our method may be useful for quantifying redundancy during different periods in patient care history and testing for associations with adverse events and other patient outcomes such as hospital admissions, morbidity and mortality.

With respect to limitations and next steps, we plan to confirm and validate our findings on other document sets and to correlate findings of redundancy to cognitive issues that clinicians experience when consuming clinical texts. This study did not utilize a separate development set of documents and, as such, represents pilot data. We plan to further develop additional knowledge-based approaches, as well as enhance our measures through the use of statistical language modeling, including the use of *N*-grams, instead of purely deterministic measurements for redundancy. We also plan to validate these results at a document level to see if our findings at a statement level generalize with an expert-derived reference standard. In addition, future studies beyond this pilot will not be confined to more uniform outpatient documents and will examine all patient documents to understand the effect of document types and clinical sublanguages. Ultimately, these measures have the potential to help provide clinicians with automated visual cues or techniques to help with cognitive processing of electronic clinical documents, as well as to be applied to other secondary uses of clinical notes.

### **Acknowledgements**

This research was supported by the American Surgical Association Foundation Fellowship (GM), the University of Minnesota Institute for Health Informatics Seed Grant (GM & SP), and by the National Library of Medicine (#R01 LM009623-01) (SP). We would like to thank Fairview Health Services for ongoing support of this research.

### **References**

1. Harrington L, Kennerly D, Johnson C. Safety issues related to the electronic medical record (EMR): synthesis of the literature from the last decade, 2000-2009. *J Healthc Manag* 2011;56(1):31-43.
2. Koppel R, Kreda D. Healthcare IT usability and suitability for clinical needs: challenges of design, workflow, and contractual relations. *Stud Health Technol Inform* 2010;157:7-14.



3. Ammenwerth E, Spötl H. The time needed for clinical documentation versus direct patient care. A work-sampling analysis of physicians' activities. *Methods Inf Med* 2009;48(1):84-91.
4. O'Donnell H, Kaushal R, Barrn Y, Callahan M, Adelman R, Siegler E. Physicians' attitudes towards copy and pasting in electronic note writing. *J Gen Intern Med* 2009;24(1):63-8.
5. Stead W and Lin H. *Computational Technology for Effective Health Care: Immediate Steps and Strategic Directions*. Washington (DC): National Academies Press (US); 2009.
6. Wrenn J, Stein D, Bakken S, Stetson P. Quantifying clinical narrative redundancy in an electronic health record. *J Am Med Inform Assoc* 2010;17(1):49-53.
7. Errami M, Wren JD, Hicks JM, Garner HR. eTBLAST: a web server to identify expert reviewers, appropriate journals and similar publications. *Nucleic Acids Res* 2007 Jul;35(Web Server issue):W12-5.
8. Lin J, Wilbur WJ. PubMed related articles: a probabilistic topic-based model for content similarity. *BMC Bioinformatics* 2007 Oct 30;8:423.
9. Cao H, Melton G, Markatou M, Hripcsak G. Use abstracted patient-specific features to assist an information-theoretic measurement to assess similarity between medical cases. *J Biomed Inform* 2008;41(6):882-8.
10. Melton G, Parsons S, Morrison F, Rothschild A, Markatou M, Hripcsak G. Inter-patient distance metrics using SNOMED CT defining relationships. *J Biomed Inform* 2006;39(6):697-705.
11. Pakhomov S, McInnes B, Adam T, Liu Y, Pedersen T, Melton G. Semantic Similarity and Relatedness between Clinical Terms: An Experimental Study. *AMIA Annu Symp Proc* 2010;2010:572-6.
12. McInnes B, Pedersen T, Pakhomov SVS. UMLS-Interface and UMLS-Similarity : open source software for measuring paths and semantic similarity. *AMIA Annu Symp Proc* 2009;2009:431-5.
13. Mount D. *Bioinformatics sequence and genome analysis*. New York : Cold Spring Harbor Laboratory 2001.
14. Cronbach LJ. Coefficient alpha and the internal structure of tests. *Psychometrika* 1951;16:297-334.
15. Gibbons JD. *Nonparametric methods for quantitative analysis*. 3rd ed. Ohio: American Sciences Press; 1997.
16. Savova GK, Harris M, Johnson T, Pakhomov SV, Chute CG. A data-driven approach for extracting "the most specific term" for ontology development. *AMIA Annu Symp Proc* 2003;579-83.
17. <http://www.textfixer.com/resources/common-english-words.txt>
18. McCray AT, Srinivasan S, Browne AC. Lexical methods for managing variation in biomedical terminologies. *Proc Annu Symp Compt Appl Med Care* 1994:235-9.