Quantifying clinical narrative redundancy in an electronic health record

Jesse O Wrenn,[1] Daniel M Stein,[1] Suzanne Bakken,[2] Peter D Stetson[3]

[1]Department of Biomedical Informatics, Columbia University, New York, New York, USA [2]School of Nursing, Columbia University, New York, New York, USA [3]Department of Medicine, Columbia University, New York, New York, USA

Correspondence to
Jesse O Wrenn, Vanderbilt Clinic, 622 West 168th Street, 5th Floor, New York, NY 10032, USA; jesse.wrenn@dbmi.columbia.edu

Work presented at AMIA Annual Symposium 2009.

Received 13 August 2009
Accepted 26 October 2009

ABSTRACT

Objective Although electronic notes have advantages compared to handwritten notes, they take longer to write and promote information redundancy in electronic health records (EHRs). We sought to quantify redundancy in clinical documentation by studying collections of physician notes in an EHR.

Design and methods We implemented a retrospective design to gather all electronic admission, progress, resident signout and discharge summary notes written during 100 randomly selected patient admissions within a 6 month period. We modified and applied a Levenshtein edit-distance algorithm to align and compare the documents written for each of the 100 admissions. We then identified and measured the amount of text duplicated from previous notes. Finally, we manually reviewed the content that was conserved between note types in a subsample of notes.

Measurements We measured the amount of new information in a document, which was calculated as the number of words that did not match with previous documents divided by the length, in words, of the document. Results are reported as the percentage of information in a document that had been duplicated from previously written documents.

Results Signout and progress notes proved to be particularly redundant, with an average of 78% and 54% information duplicated from previous documents respectively. There was also significant information duplication between document types (eg, from an admission note to a progress note).

Conclusion The study established the feasibility of exploring redundancy in the narrative record with a known sequence alignment algorithm used frequently in the field of bioinformatics. The findings provide a foundation for studying the usefulness and risks of redundancy in the EHR.

INTRODUCTION

Widespread implementation of electronic health records (EHRs) is changing how clinical information is documented, stored, and shared. These efforts are likely to proliferate rapidly given the path that has been set by the current national goals for health information technology expansion.[1] We must therefore ensure that current systems and practices are well designed and useful.

While the digitization of health records has been shown to be beneficial by several measures,[2] many EHRs closely resemble the paper charts they replace. One of the desired results of EHR implementation is to increase the quantity and utility of data available to clinicians about each patient. On one hand, the increased availability of this information is useful for informed clinical decision-making. On the other hand, too much information can lead to difficulties in navigation and synthesis.[3] Not all of the information in EHRs is likely to be useful to the clinician.[4] However, little is known about the quantity of redundancy that exists and what type of redundancy is beneficial as opposed to harmful.

Several arguments can be made for benefits of redundancy in the medical record. For example, a well-written problem list may be re-examined and improved upon across visits in multiple note iterations. There may also be a cognitive benefit to the process of rewriting pertinent clinical information as it ensures repeated evaluation by a clinician. Rewriting unchanged information in notes may necessarily highlight that certain information about the patient is still true (*eg*, that they still have a II/VI systolic murmur on exam). Additionally, billing compliance regulations require that notes stand on their own, which may promote duplication of text.

There are, however, risks and disadvantages associated with redundancy. A commonly used database design heuristic is that preservation of data integrity is contingent on the elimination of redundancy.[5] For example, this would argue against manual recopying of medication lists which may become 'out of sync'. Physicians may spend more time writing electronic notes[6] which adversely affects satisfaction.[7] Moreover, time spent documenting detracts from direct patient care, which may negatively impact clinical processes, outcomes, and patient satisfaction.

One of the most discussed examples of the benefits and risks associated with clinical narrative redundancy is the use of copy and paste to duplicate information across notes. The time-saving copy and paste function is a natural outgrowth of electronic documentation, as it often takes longer to write electronic notes.[8] Unfortunately, studies suggest that there are significant risks associated with copy and paste. These risks include the introduction of inconsistencies in the record and error propagation.[8–10]

A first step toward addressing this delicate balance of risks and benefits is to determine how much redundancy there is in electronic clinical documentation, and the associated patterns of information duplication. We have applied known computational methods in a novel manner to measure the redundancy in a sample of electronic clinical narrative notes. Through our analysis we seek to initiate timely and critical dialogue concerning the costs and benefits of redundancy in EHRs, and strategies for future EHR development and implementation.

METHODS

Overview

This is a retrospective, descriptive study of the information contained in narrative clinical documentation. We studied admission, resident signout, progress, and discharge summary notes written by physicians at a teaching internal medicine service at New York-Presbyterian Hospital, an urban academic quaternary medical center. All collected notes were written in our web-based clinical information system (WebCIS) in an unstructured free-text entry tool.[11]

Data collection

We identified a set of notes written in the records of patients admitted for greater than 72 h in duration during a 169-day period (June 20—December 6, 2006). We gathered all electronic admission, progress, resident signout and discharge summary notes written during 100 randomly selected admissions. Multiple clinicians contributed to the set of documents for each admission.

Algorithm refinement

We chose to examine redundancy by aligning each of a set of documents with all others in a set. We aligned documents at a token level using a modified Levenshtein edit distance algorithm,[12] a dynamic programming technique used elsewhere to align genetic sequences, to check spelling, and to detect plagiarism.[13] We refined the algorithm to allow the alignment of a series of documents, as opposed to the comparison of only two documents.

The following steps describe the alignment process: the first two documents of the temporally ordered series were aligned by token. The Levenshtein algorithm considers one sequence a transformation of the other. Two documents align with maximum score if they are identical, requiring no transformation. The algorithm penalizes the transformation score for each addition, subtraction or replacement of units. We increased the replacement penalty to be arbitrarily large, thus only additions and subtractions of tokens were considered during alignment. This generated an alignment array of tokens with additions and subtractions noted by placeholders. When these placeholders for additions and subtractions were replaced by their corresponding tokens from each document, an array was generated representing all tokens that existed in both documents. This array was aligned with the next document, producing new representative arrays that were recursively used until all documents were aligned. A simplified version of this technique is demonstrated in figure 1.

We also added a small change to the algorithm to encourage the alignment of longer sequences of tokens across documents, in order to remove noise from the results. We penalized the transitions between the three states encountered during alignment: match, mismatch-addition, and mismatch-subtraction. Subjective review of alignments before and after this change revealed much more appropriate sequence-matching, with less matching of contextually unrelated tokens.

Data analysis

The data set was organized into series, each representing all of the documents written during an individual patient's admission. Redundancy and its inverse, uniqueness, were calculated for each document after the alignment of the series. The amount of redundant information in a document was calculated as the number of words that aligned with previous documents divided by the length, in words, of the document. The amount of unique or new information in a document was calculated as the number of words that *did not* align with previous documents divided by the length, in words, of the document.

Having identified the amount of uniqueness and redundancy in each document in a single patient's admission, we generated graphs illustrating the uniqueness in series of notes over the course of an admission. To visualize the uniqueness of a document type (eg, signout note) over the course of an average admission, we plotted document instance uniqueness throughout the course of an admission. We then used these admission-specific values to generate an average curve. We accounted for the varied number of notes in different admissions by mapping sequences' uniqueness values to arbitrarily long arrays, and linearly imputing missing values.

To assess redundancy between the document types of interest, we calculated redundancies between the four note types: admission, resident signout, progress and discharge summary notes within a series of documents representing a single patient's admission, We calculated the amount of information carried over from the admission note into the first signout and progress notes. We also calculated the amount of information carried over from the last signout note, the last progress note and the admission note into the discharge summary. Each of these statistics is reported in terms of uniqueness of the latter document.

Finally, with these results, we undertook a simple subjective review of the content that was conserved between note types. We manually reviewed the content of the information copied between 10 examples of each of the following pairs of document types: admission to resident signout, admission to progress, resident signout to discharge summary, progress to discharge summary, and admission to discharge summary.

RESULTS

Our sample included 100 admission notes, 1167 resident signout notes, 303 progress notes, and 100 discharge summaries. The quantity of unique information contained in an average signout note, in terms of tokens, was 22% (SD=17%) with an interquartile range of 11%—25%. Progress notes contained an average of 46% (SD=18%) unique information with an interquartile range of 30%—53%.

Figure 2 illustrates the quantity of unique information contained in an average series of signout notes during the course of an admission. The first signout note of each admission was defined as fully unique and the final signout note was 7.3% unique on average.

Figure 3 illustrates the uniqueness of an average series of progress notes during the course of an admission. The first progress note of each admission was defined as fully unique, and the final progress note was 27.7% unique on average.

When the admission note was compared to the initial signout note across admissions, 30.3% of the first signout note was

Figure 1 Representative array for the alignment of two sequences including matches, additions (+) and subtractions (−).

Sequence Alignment										
Rep. Array:	*62*	*year*	*Old*	*female*	*with*	*h/o*	*ESRD*	*on*	*HD,*	*NSTEMI*
Series I:	62	year	Old	female	with	+	+	+	+	NSTEMI
Series II:	62	year	Old	-	with	h/o	ESRD	On	HD,	NSTEMI

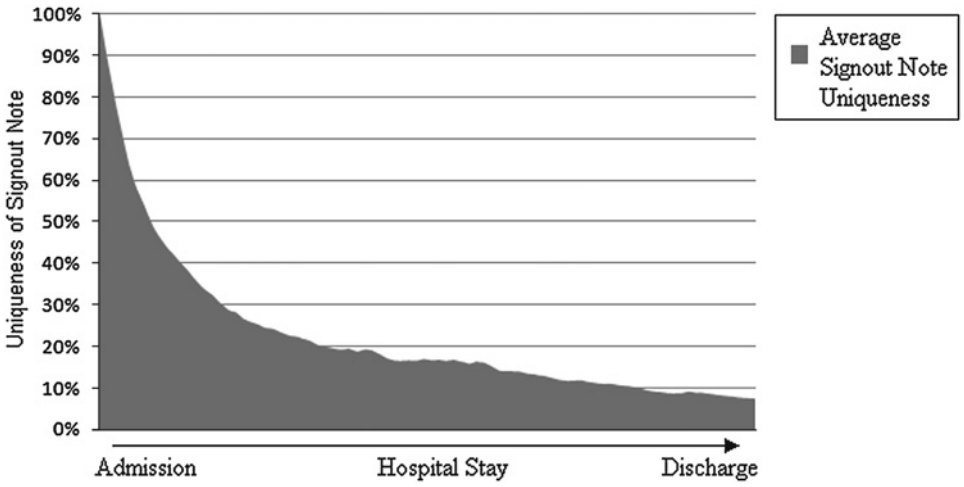

Figure 2 Average signout note uniqueness decreases over the course of an admission.

carried forward from the admission note. Similarly, 29.7% of the first progress note was carried forward from the admission note.

The average discharge summary note was, however, mostly unique with respect to the final signout and progress notes. Only 8.7% of the discharge summary note was carried over from the final signout note, while 7.4% of the discharge summary note was carried over from the final progress note. When we compared the discharge summary note to the admission note, however, we calculated that 30.7% of the discharge summary note was carried forward from the admission note. These results are also displayed in table 1.

Subjective review of the duplicated content revealed that information copied forward from admission notes varied depending on the note type to which the content was copied. Signout notes most frequently inherited medication lists and histories of present illness from the admission note. Progress notes, on the other hand, most frequently inherited assessment and plan sections from the admission note. Less often, medication lists were inherited into the progress note. Finally, discharge summary notes often contained histories of present illness and medication lists that were present in admission notes.

We also manually reviewed the information most frequently copied forward into the discharge summary note from the signout note and the progress note. The discharge summary note frequently contained medication lists inherited from the last signout note. Less often, information found in the discharge summary note's hospital course section appeared to have been inherited from parts of the last signout note. The hospital course section often contained information from the final progress note, and occasionally physical exams and medication lists appeared to have been inherited from the final progress note.

DISCUSSION

Although redundancy in electronic documentation is widely recognized, this is the first paper to describe the use of sequence alignment algorithms to quantify redundancy in clinical narrative documentation. The study established the feasibility of exploring redundancy in the narrative record by using an algorithm typically used in bioinformatics to align genetic sequences and in plagiarism detection. The algorithm aided in the quantification and visualization of redundancy within and between document types and its output afforded us natural visualization of the alignment of a sequence of notes. It revealed instances of copy and paste, as well as copied information that has been edited.

As our study looks at syntactic duplications, it only serves as a proxy for the study of conceptual redundancy. Despite this it provides strong indications that redundancy is prevalent in our data set. Future studies involving a larger set of documents and conceptual alignment facilitated by a combination of natural language processing and manual review may better characterize redundancy with respect to quantity and content.

Signout and progress notes, which are sequential and are generated daily throughout the admission, iteratively evolve. The evolution decelerates, however, as the admission progresses.

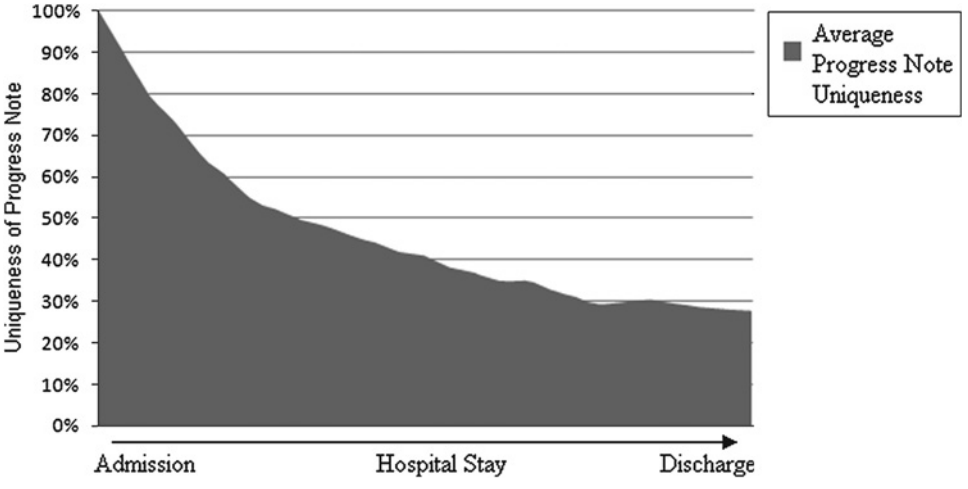

Figure 3 Average progress note uniqueness decreases over the course of an admission.

Table 1 Percentage of information in notes (columns) found to be duplicated from previous notes (rows)

	Signout note	Progress note	Discharge summary
Admission note	30.3%*	29.7%*	30.7%
Signout note	78%‡	§	8.7%†
Progress note	§	54%‡	7.4%†

*First signout note and first progress note.
†Last signout note and last progress note.
‡Compared to all previous notes.
§Information duplication between signout and progress notes was not calculated.

As patients' problems that motivated the admission are diagnosed and treated, less new information is introduced into the clinical narrative.

Despite the overall decrease in amount of unique information in progress notes over the course of an admission, there appears to be a slight increase toward the end of an admission (see figure 3). Although statistical significance cannot be evaluated given our small dataset, a casual review of the aligned notes suggests that this brief rise may correspond with the introduction of new information associated with discharge planning. This leads us to speculate that some clinical events correspond to measurable information 'injections' in the narrative continuum of an admission. A consultation, for example, might result in the introduction of a significant amount of new information. Similarly, adverse events may result in measurable spikes in the amount of unique information introduced into the record.

Although outside the scope of this study, the exploration of such measurable information injections might extend the utility of clinical narrative. Retrospective analysis with this method has the potential to be a valuable research tool, but we may even be able to prospectively surveil narrative documentation for real-time identification of important clinical events, as well as excessive copy and paste.

The subjective review of the content of information that is conserved across document types suggests that there exist circumscribed, persistent modules of information that appear in different notes and evolve in different ways. Not surprisingly, the past medical history, for example, remains relatively constant throughout the admission, but appears in multiple documents. In next-generation EHR implementations, centralization of these different modules of information, and improved capacity to reference or link to existing narrative data could contribute to the reduction of redundancy in clinical notes. Rather than retyping sections of notes one could imagine the EHR facilitating a manual process of daily report generation. A clinician could create pointers to information that is still relevant each day, and add only what is new and clinically relevant. This might relieve the clinician of the burden of retyping or copying, and allow more attention and time for clinical decision-making and patient care.

Our study focused on the processing of free-text narrative, and we therefore did not address the issues of redundancy associated with structured documentation. For example, many EHRs allow clinicians to note only abnormal findings, or, in some instances of nursing, to document by exception only interventions beyond those in the institutional standard of care.[14] The EHR subsequently generates a relatively large document that includes normal/standard of care as well as exceptions. While such an approach may ease redundancy in data entry, the potential impact on comprehensibility has not yet been quantified.

Our findings suggest future directions for investigation. Most importantly, we may need to consider a model for next-generation documentation where billing/compliance information becomes an epiphenomenon of clinical documentation—parallel, auditable, and separate from more salient clinical narrative. Such a model may move us from a debate over 'good' versus 'bad' redundancy to one of how to enable 'smart' redundancy. This would ensure that facts which are valuable for clinical communication are propagated (eg, an abnormal but stable physical exam), and that summary documents (eg, signout notes and discharge summaries) summarize so that they are semantically redundant but concise.

Limitations

Although quantification of information duplication via the direct alignment of words in very 'noisy' text is not optimal, conceptual alignment was out of the scope of this study. The use of lexical alignment as opposed to conceptual alignment may slightly underestimate the amount of information duplication in clinical documents. On the other hand, we believe this is offset by the fact that occasionally words are aligned that are not *contextually* related, which very slightly overestimates the amount of information duplication in clinical documents.

We limited the scope of this study to four note types: admission, signout, progress, and discharge summary, because they are typically generated by multiple clinicians from the same service. We believe, however, that there are numerous other document types to and from which clinicians are likely to duplicate information. A more thorough study might include notes from rehabilitation and social work services, reports, procedure and consultation notes, and notes from previous admissions. We also only studied a small sample of documents from one service of a single academic medical center. We only studied documents written electronically in WebCIS, a system used only at our institution. However, the WebCIS notes we studied were 'free-text' so the findings may be applicable to documents created in systems with similarly unstructured notes. Neither templated, dictated, nor handwritten notes were studied.

CONCLUSIONS

EHRs should be designed to be clinically useful, practical, and efficient. We propose that this and future studies of redundancy in narrative documentation may inform the development of future EHRs with modular, reusable documentation components that reduce the effort required to write clinical documents. The findings of our study support the feasibility of our methods for studying redundancy. It is, of course, up to the medical and informatics communities to debate and determine the proper balance of 'smart' redundancy in the health record, weighing the benefits against the dangers of inefficiency and the possibility of propagation of errors and loss of data integrity.

Funding T15LM007079 (Wrenn, Stein), K22LM8805 (Stetson), R01NR008903 (Bakken). Other Funders: NLM.

Competing interests None.

Ethics approval This study was conducted with the approval of the Columbia University.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1. **United States Statutes at Large**. American Recovery and Reinvestment Act of 2009, PL 111—5, February 7, 2009.
2. **Amarasingham R,** Plantinga L, Diener-West M, *et al*. Clinical information technologies and inpatient outcomes: a multiple hospital study. *Arch Intern Med* 2009;**169**:108—14.
3. **Stead WW,** Lin HS. *Computational technology for effective health care: immediate steps and strategic directions*. Washington, DC: National Academies Press, 2009.

4. **Stetson PD,** Morrison FP, Bakken S, *et al*. Preliminary development of the physician document quality instrument. *J Am Med Inform Assoc* 2008;**15**:534—41.
5. **Codd EF.** *The relational model for database management*. Version 2. Boston, MA: Addison-Wesley Longman Publishing Co, Inc, 2009.
6. **Ammenworth E,** Spötl HP. The time needed for clinical documentation versus direct patient care. A work-sampling analysis of physicians' activities. *Methods Inf Med* 2009;**48**:84—91.
7. **Mechanic D.** Physician discontent: challenges and opportunities. *JAMA* 2003;**290**:941—6.
8. **O`Donnell H,** Kaushal R, Siegler E, *et al*. Physicians attitudes towards copy and pasting in electronic note writing. *AMIA Annu Symp Proc* 2008:1073.
9. **Siegler EL,** Adelman R. Copy and paste: a remediable hazard of electronic health records. *Am J Med* 2009;**122**:495—6.
10. **Hammond KW,** Helbig ST, Benson CC, *et al*. Are electronic medical records trustworthy? Observations on copying, pasting and duplication. *AMIA Annu Symp Proc* 2003;269—73.
11. **Hripcsak G,** Cimino JJ, Sengupta S, *et al*. WebCIS: large scale deployment of a Web-based clinical information system. *Proc AMIA Symp* 1999:804—8.
12. **Levenshtein VI.** Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady* 1966;**10**:707—10.
13. **Su Z,** Ahn B, Eom K, *et al*. Plagiarism detection using the Levenshtein distance and Smith-Waterman algorithm. *ICICIC '08: Proceedings of the 2008 3rd International Conference on Innovative Computing Information and Control* Washington, DC: IEEE Computer Society, 2008:569.
14. **Nicely N.** Charting by exception. *Advance for Health Information Executives* 2006;**10**:10.