The Relationship Between Structural Characteristics of 2010 Challenge Documents and Ratings of Document Quality

Shuying Shen, MStat[1,2,3], Brett R South, MS[1,2,3], Jorie Butler, PhD[1,2], Robyn Barrus, MS[1], Charlene Weir, PhD[1,2,3]

[1]VA Salt Lake City Health Care System, [2]Departments of Internal Medicine [3]Biomedical Informatics, University of Utah, Salt Lake City, Utah

Abstract

Quality of clinical narratives has direct impact on the perceived usefulness of these documents. With the advent of electronic documentation, the quality of clinical documents has been debated. Electronic documentation is supported by features to enhance efficiency, including copy/paste, templates, multi-level headings, and inserted objects. The impact of these features on perceived document quality has been difficult to assess in real settings as compared to simulations. This study used electronic notes from the 2010 i2b2/VA Challenge to explore the impact of text characteristics on general perception of document quality. We administered a validated instrument to assess document quality, focusing on two dimensions, informativeness and readability. Text characteristics were collected from both subjective ratings and quantitative summary. The results suggested common clinical elements such as templates, headings and inserted objects had strong positive association with document quality. Understanding of such relationship may prove useful in future EHR design and informatics research.

Introduction

Implementation of electronic health records (EHRs) is occurring at an increasing rate across the country. A key component of EHRs is computerized provider documentation (CPD). Narrative notes are at the heart of the medical record, providing the patient's story, the clinicians' perspective and the big picture view of the current clinical situation and plan of care. The clinical narrative supports a shared situation model of the patient for team members and regulatory agencies use the narrative to determine compliance. Clinical texts are also a rich source of information that can be exracted using Natural Language Processing (NLP) techniques as well as other automated processes. Understanding the factors associated with document quality may be key to understanding why and what care has been provided, the medical decision-making process, and clinical communication. These factors also have potential impact for information extraction and classification tasks done through either manual review or automated tools. The purpose of this paper is to report an exploratory analysis of the association between some of the structural and informational characteristics of computerized documentation and ratings of document quality.

Document Quality

Prior work assessing document quality has identified several key components to the cognitive dimensions of document quality. One analysis of provider interviews identified five factors influencing satisfaction with clinical documentation work: Availability, Time Efficiency, Expressivity, Structure and Quality[1]. Another study identified four document dimensions associated with the perception of quality: Well-formed, Comprehensible, Accurate and Compact[2]. In addition, Hammond et al[3] developed and validated an instrument based on the results of focus groups in the VA and found two higher order factors (readability and informativeness) with five distinct themes: 1) communication and coordination, 2) control and forced functions, 3) information availability and reasoning support; 4) workflow alteration and disruption and 5) confidence and trust.

None of these prior works has linked the overarching dimensions of quality to specific structural aspects of notes. This study explores those issues using the 2010 i2b2/VA Challenge[4] documents and a validated scale for assessing quality.

Objectives

A central premise of this study is that document quality impacts the usefulness of clinical narratives and is driven by structural factors such as templated data entry fields, use of headings, the presence of inserted objects, and use of de-identification methods such as replacement of Protective Health Information (PHI) with a tag, or realistic surrogate.

For computerized clinical documentation, document quality becomes complex as a function of structural issues. When clinicians use copy and paste features, they tend to create notes that are longer with less internal cohesion. Sometimes it is difficult to discriminate one note from another when they only change a few words day to day. The use of inserted objects may result in pieces of information that are really collected at different times. When this information is presented as a whole, it is not necessary salient. In addition, object insertion is not fine tuned enough, so that sometimes whole lab results pages are copied, making it difficult to discern which lab results were the most important. Pre-compiled templates make it easy to enter in information, but harder to extract higher level reasoning, especially when it results in page after page of notes with no findings checked. Prior qualitative work has found users of VA clinical notes to have significant problems with comprehension, readability, understanding the intentions of the author and treatment goals[5]. Similar problems might exist in other settings. We hypothesized that perceived difficulty with some of these features, in combination with note length and information density, would be associated with both structural and informational quality of document quality. Because we wanted to extend some of our conclusions to issues of annotation and chart review, we include an evaluation of de-identification and re-identification outputs. Document quality could be affected by the way in which de-identification is conducted and how PHI is redacted or replaced with surrogate information.

Methods

In this section, we describe the data sources, study subjects and procedures including variables used and measurements.

Design

The study design is a cross-section correlational study on randomly selected documents using retrospective manual review. Two clinicians and two non-clinicians were recruited, whom all had prior annotation and chart review experience. They were asked to read and rate the documents using a document quality assessment instrument, and make judgments about the degree that templating, inserted objects and de-identification manipulations posed challenges.

Procedures

The raters received a training session, where the task and each of the questions were explained. They were encouraged but not required to comment on low ratings and provide justifications. Each document was reviewed by a random pairing of two raters.

Variables

Document Quality. An instrument composed of five questions taken from prior work validating a longer instrument were used to assess document quality[3]. The first two represent assessment of readability (ease to read, ease to skim). The third, forth and fifth question represent assessment of .informativeness (complete content, consistent information, and clear meaning). Items for the document quality instrument were extracted from a validated scale developed by Hammond and Weir initially based on 9 focus groups in the VA and further validated using a sample of 93 physicians. Exploratory factor analysis confirmed the basic structure of two factors (readability and informativeness). The items used for this study were the highest loading items in the original study on each factor.

Difficulty/Helpfulness. The raters judged the perceived helpfulness of individual text characteristics, including, templates, headings, inserted objects and PHI replacement. A Likert scale of 1 to 7 was used for the responses, where higher score indicates higher preference (Appendix 1).

Objective Measures. For objective measurement of text characteristics, we utilized the following: a) Length of document as defined by the number of words extracted for each document and b) Information density, or how dense a document is with clinical information, calculated as (number of annotated clinical concepts) / (number of words).

Data

The Department of Veterans Affairs (VA) Consortium for Healthcare Informatics Research partnered with the informatics for Integrating Biology and the Bedside (i2b2) team to generate the reference standard for the 2010 i2b2/VA challenge. A total of 826 documents obtained from three healthcare institutions were annotated and

released to the community for this challenge. Documents were de-identified and released only after appropriate data use agreements were signed by annotators and all participants of challenge teams. These included 230 discharge summaries from Partners Healthcare, 196 discharge summaries from Beth Israel Deaconess Medical Center (BIDMC), and 200 progress notes, and 200 discharge summaries from University of Pittsburgh Medical Center (UPMC). Redacted PHI elements for BIDMC, and Partners data sources were resynthesized with realistic surrogates. For this analysis we randomly sampled 246 documents from the 2010 i2b2/VA corpus.

Analysis

We reported descriptive statistics for document quality scale and structural factor ratings, as well as number of words and information density. Cohen's kappa was used to assess the subjectivity of the instrument. Cronbach's Alpha[7] was calculated for the readability and informativeness scores. To adjust for clustering within each rater, a "within subjects" correlation coefficient was estimated between the document quality scores and the individual text characteristics by removing the variations between raters[8]. A mixed effect regression was used to model relationship between readability and the text characteristics, again accounting for lack of independence among repeated measurements from same rater. Informativeness score was modeled the same way.

Raters' comments and reasons behind low ratings were extracted and closely examined. The authors (SS and BS) group the comments into categories corresponding to each of the questions. Through several group discussions, common themes were identified and named..

Results

We first present the results from the assessment rating in this section. The reviewers' comments are then summarized and discussed.

Descriptive

Overall, 246 documents were reviewed, with two random raters assigned to each document. Forty-six were Partners Healthcare discharge summaries, 50 were Beth Israel discharge summaries, 101 were University of Pittsburgh progress notes, and 49 were University of Pittsburgh discharge summaries. The descriptive statistics for ratings of each question, as well as text characteristics were reported in Table 1. For all nine questions, the minimum response was 1 and maximum response was 7, reflecting a wide spread of perceived document quality and preference on text elements. Average document length was 762 words, while the shortest document having only 93 words and longest document having 2972 words. Information density ranged between 0.011 to 0.32, with a mean of 0.1 (1 clinical concept per 10 words). Inter-rater reliability was calculated using Cohen's kappa and the reviewers were in fair agreement (kappa = 0.3453, $p < 0.0001$).

Table 1. Descriptive statistics of ratings and text characteristics.

	Mean	**SD**	**Min**	**Max**
Subjective measures				
Readability	9.97	3.38	2	14
Easy to Skim	4.63	1.86	1	7
Easy to Read	5.30	1.81	1	7
Informativeness	17.00	4.00	3	21
Complete Content	5.50	1.67	1	7
Consistent Information	6.14	1.24	1	7
Clear Meaning	5.39	1.74	1	7
Text Characteristics				
Templated Information	4.50	1.79	1	7

Headings	4.41	2.22	1	7
Inserted Objects	4.59	1.85	1	7
Re-identification	4.08	2.20	1	7
Objective Measures				
Word Count	770	522	93	2972
Information Density	0.1	0.04	0.011	0.32

Creation of Scales

The Readability Scale was composed of the first two questions (ease to read, ease to skim) and had a Cronbach's alpha of 0.83. The Informativeness Scale consisted of the sum of the ratings from the next three questions (complete content, consistent information, and clear meaning) and had an acceptable Cronbach's alpha of 0.85.

Pearson Correlations

Adjusted Pearson correlation was estimated between these two scores and each document characteristic and clinical element's perceived helpfulness. As seen from Table 2, the helpfulness of each clinical element was positively correlated with both dimensions of document quality. The length of document and information density were negatively correlated with Readability, but had no significant association with Informativeness of document.

Table 2. Adjusted Pearson correlations between document quality scores and individual text characteristics.

	Readability Scale		Informativeness Scale	
Templated Information	0.64	p<0.0001	0.57	p<0.0001
Headings	0.73	p<0.0001	0.55	p<0.0001
Inserted Objects	0.67	p<0.0001	0.53	p<0.0001
Re-identification	0.27	p<0.0001	0.17	p<0.0001
Word Count	-0.29	p<0.0001	0.04	0.3951
Information Density	-0.23	p<0.0001	-0.02	0.6904

Regression Analyses

After estimating the adjusted correlation, we ran mixed models regressions on Readability and Informativeness with the text characteristics and perceived helpfulness as independent variables, while adjusting for annotator clustering effect. The coefficients and associated p-values are reported in Table 3. Consistent with the univariate analysis, the degree to which templates, headings and inserted objects were helpful was associated with both dimensions of document quality. When re-identification was bothersome, it impacted only the readability of document (p=0.003), but not the informational content of the document (p = 0.703). On the other hand, documents with dense information resulted in negative impact on readability (p = 0.003), while quality of information was not affected by how dense or sparse the information was (p = 0.782).

Table 3. Regression of Text Characteristics on document quality scores**.**

	Readability Scale		Informativeness Scale	
Helpful Templated Information	0.34	p<0.0001	0.71	p<0.0001
Helpful Headings	0.66	p<0.0001	0.47	p<0.0001
Helpful Inserted Objects	0.50	p<0.0001	0.39	0.001

Helpful Re-identification	0.13	0.003	-0.03	0.703
Word Count	-0.0013	p<0.0001	0.0007	0.01
Information Density	-6.93	0.003	0.96	0.782

Qualitative Review of Comments

Overall, reviewers generated 34 comments. Some of the comments were based on overall quality of the document, and some were targeted on specific phrases and objects that they identified to be problematic. After the authors iteratively review these comments, they were organized them into the following three general categories:

Table 4. Themes, examples and quotations from reviewers' comments.

Theme	Example	Quotation
Unstructured	Missing headings or templates.	"There were no headings, but it would've been very helpful"
Disorganized	Inserted lists were in wrong order.	"The lists were there and detailed, but it's hard to read and doesn't seem to be in a great order."
Distractive PHI surrogate	PHI tags were distractive.	"The De-id process for this was very hard to read in a continuous flow. Very distracting."
	Non-English PHI surrogates were difficult to read.	"phonetically difficult surrogates really slow me down." "i find the Icelandic-looking De-ID substittions to be jarring and difficult to read."
	Clinical eponyms were incorrectly identified and replaced as PHI.	"The De-id was a little confusing in places."

Discussion

Overall, reviewers rated the quality of documents high, with most items above the medium point on the scale of 3.5. The one exception was the de-identification ratings, suggesting that de-identification was distracting. The qualitative comments supported that impression. In addition, there were differences between institutions in mean quality and helpfulness ratings. The overall pattern of responses indicated that headings and templates were considered helpful. Headings were less helpful (or missing) in UPMC progress notes. Document length and information density were both negatively correlated with Readability, but there was no significant association of either with Informativeness of the document. The regression analysis supported the general perspective that Informativeness and Readability are impacted by different variables and attributes of a document. Re-identification and information density both negatively impacted Readability, but not Informativeness. Evidence from studies of how clinicians read notes suggest that clinicians learn quickly on how to retrieve the "gist" of a note. As a result, Informativeness would likely be impacted by other than structural variables. A recent study by some of the authors of this paper supported this view. Three forms of a note were created to include the same information but different forms (one lean and informative, one the usual VA style and the third, an overly verbose note). Physicians rated the quality of the notes overall. No difference in informativeness was found, despite differences in Readability[3].

Reviewers generally preferred having structured documentation, which they believed provided a logical sequence for quickly skimming and understanding the information. When inserted objects were out of order (potentially due to copying and pasting from other notes), reviewers found them to be bothersome and difficult to comprehend. De-identification and re-identification posed three unique problems. First, PHI tags such as ***PHI type*** were distractive and threw reviewers off. Second, non-English PHI surrogates such as "Patient MAULPLACKAGNELEEB , INACHELLE" were difficult to read and slowed-down reviewers. They preferred surrogates that resembled English words. Third, Clinical eponyms such as "Alzeimer's Disease" could be incorrectly identified as PHI and thus replaced by "Jane's Disease." Reviewers were confused by such replacements.

We used a mixed methods approach for this study to identify and characterize quantitative and qualitative factors related to structural characteristics of documents used for the 2010 i2b2/VA Challenge. Our quantitative results bear out some of the constructs and themes identified from our more in depth qualitative analysis of the 2010 i2b2/VA Annotation task[9]. Two themes of note include the difficulty of managing uncertainty and a second theme the readability and its affects on the annotation process itself.

Implications for NLP

There are several obvious implications when NLP methods are used for information extraction and classification tasks. First, NLP systems were traditionally designed for narrative text. While human begins could discern the meaning of a table or list of checked item relatively easily, a significant amount of customization would be required to retrain an NLP system for such patterns. There have been advances on methods to identify section headers and sub-headings in clinical documents[10]. Challenges remain less tackled when it comes to information extraction from templated text in the form of tables, check-box and questionnaires, which are frequently seen in semi-structured CPD. Less complex NLP task such as negation could become more difficult when concepts are entered using pre-compiled templates like a page-long check list, essentially turning the negation detection algorithm into a discourse analysis.

The impact of Informativeness on NLP may be more nuanced. Informativeness may be impacted by the actual content of the note including internal consistency, errors, copy and paste and actual clinical documentation errors. We did not examine those domains in this study, but the actual efficiency of NLP would likely be impacted by unreliability and inaccuracies of documentation. Future work could examine this question.

Implications for design of clinical notes

The findings of this work have some implications for the design of electronic notes. The use of structured headings and formatted text clearly makes it easier for clinicians to skim the text rapidly looking for the required information. Templates appear to have both a positive and negative impact depending on how well they are used. When organized well, they make it easier to find information, but they could also make it confusing. Future work could examine the human factors considerations for template design (both for input and readability perspectives).

Conclusion

With the recent and widespread Implementation of electronic health records (EHRs) it is important to return to a discussion of the quality of clinical documentation and its potential effects on application of NLP in a real world setting. The 2010 i2b2/VA challenge task represents one available document corpus that researchers can draw upon for training and evaluation data. Structural variables were shown to be significantly related to ratings of document quality and the results were congruent with the qualitative reports.

ACKNOWLEDGEMENTS

Prior to conducting this study appropriate IRB approvals were received. This study was supported using resources and facilities at the VA Salt Lake City Health Care System, the VA Consortium for Healthcare Informatics Research (CHIR), VA HSR HIR 08-374. We also wish to acknowledge the efforts of reviewers responsible for the document quality ratings discussed in this paper.

References

1. Rosenbloom ST, Crow AN, Blackford JU, Johnson KB. Cognitive factors influencing perceptions of clinical documentation tools. J Biomed Inform. 2007 Apr;40(2):106-13.
2. Stetson PD, Morrison FP, Bakken S, Johnson SB. Preliminary development of the physician documentation quality instrument. J Am Med Inform Assoc. 2008 Jul-Aug;15(4):534-41.
3. Hammond KW, Efthimiadis EN, Weir CR, Embi PJ, Thielke SM, Laundry RM, et al. Initial Steps toward Validating and Measuring the Quality of Computerized Provider Documentation. AMIA Annu Symp Proc. 2010;2010:271-5.
4. Uzuner O, South BR, Shen S, DuVall SL. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. Journal of the American Medical Informatics Association : JAMIA. [Research Support, N.I.H., Extramural Research Support, U.S. Gov't, Non-P.H.S.]. 2011 Sep-Oct;18(5):552-6.
5. Weir CR, Nebeker JR. Critical issues in an electronic documentation system. AMIA Annu Symp Proc. 2007:786-90.
6. Uzuner O, Luo Y, Szolovits P. Evaluating the state-of-the-art in automatic de-identification. J Am Med Inform Assoc2007. p. 550-63.
7. Allen M, Yan W. Introduction to Measurement Theory. Long Grove, IL: Waveland Press; 2002.
7. Bland JM, Altman DG. Calculating correlation coefficients with repeated observations: Part 1--Correlation within subjects. Bmj. 1995 Feb 18;310(6977):446.
8. Gueorguieva R, Krystal JH. Move over ANOVA: progress in analyzing repeated-measures data and its reflection in papers published in the Archives of General Psychiatry. Arch Gen Psychiatry. [Research Support, U.S. Gov't, Non-P.H.S. Research Support, U.S. Gov't, P.H.S. Review]. 2004 Mar;61(3):310-7.
9. South BR, Shen S, Barrus R, DuVall SL, Uzuner O, Weir C. Qualitative analysis of workflow modifications used to generate the reference standard for the 2010 i2b2/VA challenge. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium. [Research Support, N.I.H., Extramural
Research Support, U.S. Gov't, Non-P.H.S.]. 2011;2011:1243-51.
10. Denny JC, Spickard A, 3rd, Johnson KB, Peterson NB, Peterson JF, Miller RA. Evaluation of a method to identify and categorize section headers in clinical documents. Journal of the American Medical Informatics Association : JAMIA. [Evaluation Studies Research Support, N.I.H., Extramural]. 2009 Nov-Dec;16(6):806-15.

Appendix 1 Document Quality and Text Characteristic Assessment

Instructions: Answer the following questions according to the document you just read. Annotate the appropriate number that corresponds to your answer.

DOCUMENT QUALITY QUESTIONS

a. In terms of reading the note, the text was:

Difficult to Skim	a.1	a.2	a.3	a.4	a.5	a.6	a.7	Easy to Skim

b. I would describe the note as:

Difficult to Read	b.1	b.2	b.3	b.4	b.5	b.6	b.7	Easy to Read

c. The expected content in the note was:

Incomplete	c.1	c.2	c.3	c.4	c.5	c.6	c.7	Complete

d. The information in the note was:

Inconsistent	d.1	d.2	d.3	d.4	d.5	d.6	d.7	Consistent

e. In terms of identifying the AUTHOR'S MEANING the note was:

Very Unclear	e.1	e.2	e.3	e.4	e.5	e.6	e.7	Very Clear

Please annotate the degree that you found the following to be a problem or to be helpful:

f. Templated Information

A Significant Problem	f.1	f.2	f.3	f.4	f.5	f.6	f.7	Very Helpful

g. Headings

A Significant Problem	g.1	g.2	g.3	g.4	g.5	g.6	g.7	Very Helpful

h. Inserted Objects (medication lists, problem lists, etc)

A Significant Problem	h.1	h.2	h.3	h.4	h.5	h.6	h.7	Very Helpful

i. Use of De-identification/Re-identification (replacements, versus **PHI type**)

A Significant Problem	i.1	i.2	i.3	i.4	i.5	i.6	i.7	Very Helpful