Voice-Dictated versus Typed-in Clinician Notes: Linguistic Properties and the Potential Implications on Natural Language Processing

Kai Zheng, PhD,[1, 2] Qiaozhu Mei, PhD,[2, 3] Lei Yang, MS, ME,[2] Frank J. Manion, MS,[4] Ulysses J. Balis, MD,[5] David A. Hanauer, MD, MS[4, 6]
[1] School of Public Health Department of Health Management and Policy; [2] School of Information; [3] Department of Electrical Engineering and Computer Science; [4] Comprehensive Cancer Center; [5] Department of Pathology; [6] Department of Pediatrics. University of Michigan, Ann Arbor, MI

Abstract

In this study, we comparatively examined the linguistic properties of narrative clinician notes created through voice dictation versus those directly entered by clinicians via a computer keyboard. Intuitively, the nature of voice-dictated notes would resemble that of natural language, while typed-in notes may demonstrate distinctive language features for reasons such as intensive usage of acronyms. The study analyses were based on an empirical dataset retrieved from our institutional electronic health records system. The dataset contains 30,000 voice-dictated notes and 30,000 notes that were entered manually; both were encounter notes generated in ambulatory care settings. The results suggest that between the narrative clinician notes created via these two different methods, there exists a considerable amount of lexical and distributional differences. Such differences could have a significant impact on the performance of natural language processing tools, necessitating these two different types of documents being differentially treated.

Introduction

Despite the increasing popularity of structured data entry, narrative clinician notes continue to be pervasively used in healthcare practice.[1, 2] How to make computational use and reuse of such unstructured clinical documents is an enduring topic in health informatics research,[2–4] which has stimulated significant recent advances with medical natural language processing (NLP) as well as with novel data entry user interfaces such as structured narratives.[5–7]

A majority of narrative clinician notes stored in electronic health records (EHR) are either created through voice dictation—typically via speech recognition software or professional transcription services—or entered directly by clinicians using a computer keyboard. Intuitively, the nature of dictated notes would resemble that of natural language, because they are results of human speech and have often been preliminarily verified by computer software or human transcribers during the transcription process. While for notes that were typed in, they may demonstrate distinctive language features due to intensive usage of acronyms and abbreviations, symbolic representations, incomplete sentence structures, and unique characteristics originating from local documentation templates, in addition to frequent occurrences of spelling errors and nonstandard use of medical terms.

To automatically extract structured concepts from narrative clinician notes, post-processing using NLP is often needed.[2, 3] The performance of NLP tools is critically contingent upon the quality and the nature of the input data.[3, 8] The distinct linguistic properties of voice-dictated versus typed-in notes could therefore have a significant impact on how well NLP tools may perform, necessitating that these two different types of documents be differentially treated. To the best of our knowledge, no prior research has investigated such differences for the specific goal of better informing the development of context-appropriate post-processing strategies.

In this study, we comparatively examined the linguistic properties of narrative clinician notes created via these two different sources. The empirical dataset, consisting of 30,000 voice-dictated notes and 30,000

notes manually typed, was retrieved from our institutional EHR system. These notes were generated as part of the enterprise clinicians' routine patient care activities. Based on the obtained data, we conducted a comparative statistical corpus analysis to delineate the differences between these two different narrative classes. The objectives were to (1) verify the existence of such differences and quantify their magnitudes and (2) develop a preliminary understanding of how such differences might impact the performance of NLP tools in automated post-processing.

Background

Using electronic systems to acquire, store, and manage patient care data provides great promise to enhance the data's reuse value for computational purposes such as quality improvements, computerized decision-support, and clinical, translational, and health services research.[9, 10] Even though codified data entered through structured forms are highly desirable, structured data entry often imposes an escalated efficiency burden on clinician users and it is not adequate to accommodate all types of documentation needs due to the lack of flexibility and expressiveness.[1, 2] Hence, unstructured, free-text narratives will continue to constitute a vital source of patient care data in the foreseeable future, and a hybrid model combining both structured documentation and post-processing of unstructured clinician notes has been recommended as a more viable approach.[2]

The quality of outputs of post-processing, mainly through NLP tools, is particularly sensitive to the text characteristics of the input data.[3, 8] However, while there exists a significant body of literature evaluating the performance of NLP applied to different genres of documents (e.g., radiology reports, discharge summaries, and progress notes),[3] most studies assumed that all documents of the same genre would demonstrate similar language features, ignoring the fact that the nature of the documents of the same genre could dramatically deviate based on the source via which they were created. In this study, we are interested in testing if this assumption holds among narrative clinician notes created through voice dictation versus those that were manually typed in.

While we are not aware of any prior attempts specifically examining this problem, studies have been conducted in related areas such as assessing the readability of online health content generated from various sources (e.g., WebMD articles versus patient education booklets distributed in medical facilities), and evaluating structural or semantic differences between Open Access articles and papers published in traditional venues[11–13] and between bodies of full text journal articles and article abstracts.[14] This paper was informed by these previous studies while focusing on patient care documents generated in clinical settings.

We expected to observe a considerable level of deviation between the linguistic properties of these two different types. Further, we hypothesized that typed-in notes, due to the reasons mentioned earlier, are more likely to be NLP-resistant and thus require special treatments in post-processing. In the next section, we describe the narrative documents contained in the empirical dataset and the various methods and measures that we used to describe their linguistic properties. Then, we present the study results and their potential impacts on the performance of NLP in post-processing.

Methods

A. The Empirical Dataset

The empirical dataset was retrieved from the institutional EHR used at the University of Michigan Health System (UMHS), a 930-bed quaternary academic medical center that has over 44,000 inpatient admissions and 1.7 million ambulatory visits annually. The system is used in all patient care areas, including the emergency department, inpatient services, and ambulatory clinics and health centers. For simplicity, in this study, we only used narrative clinician notes generated in outpatient settings belonging to deceased patients.

All documents that we analyzed are of the same genre: encounter notes that physicians composed to describe an outpatient encounter or to communicate with other clinicians regarding patient conditions. We chose this document genre because: (1) encounter notes convey rich information that spans across a variety of data elements key to patient care; and (2) working with this common document genre warranted that a large number of documents would be available for the study.

When this study was conducted, a total of 133,296 voice-dictated notes and 31,665 typed-in notes stored in the EHR system met our inclusion criteria (i.e., encounter notes generated in ambulatory care settings belonging to deceased patients). The 60,000 documents included, evenly split between the two documents types, were randomly picked from this pool. Note that at UMHS, a majority of voice-dictated notes are transcribed by professional services either in-house or offshore. Speech recognition software has been tested but has not been in widespread use currently.

To protect patient confidentiality, all notes were de-identified using the MITRE Identification Scrubber Toolkit prior to analysis, which in essence substituted patient identifying information with pseudo tokens (e.g., a fake name, a random phone number, and a made-up street address).[15, 16] While previous research has shown that this resynthesis process could make the data less "realistic," and under certain circumstances may undermine the performance of NLP in subsequent processing,[17] we believe that it would have minimal effects on the results of this study and the effects would affect both document types similarly.

The Medical School Institutional Review Board at the University of Michigan reviewed and approved the research protocol of this study.

A. Analytic Methods and Measures

1) Construction of a comprehensive medical dictionary

First, we constructed a comprehensive dictionary to help delineate the lexical characteristics of the narrative encounter notes. This dictionary is based on an in-house developed vocabulary underlying the spellcheck function of the EHR system, which was in turn built from multiple open-source dictionaries of the English language and commonly used medical terminologies (e.g., GNU Aspell[18] and OpenMedSpel[19]). In addition, we included in the dictionary all concepts and concept names encompassed in the 2010AB release of the UMLS Metathesaurus, which contains millions of entries collected from 158 respective source vocabularies including ICD, SNOMED CT®, LOINC, and MeSH.[20] The dictionary, referred to as UMLS+, should have covered most of the terms that could be used by clinicians in their clinical documentation.

To analyze the documents collected, we applied multiple analytical methods to derive a variety of descriptors of linguistic properties, presented in the remaining parts of this section. Note that all punctuations and symbolic representations had been removed from the data before the study analyses were conducted.

1) Surface metrics

We computed several basic surface metrics to contrast the overall text characteristics of dictated notes versus those that were typed in:

- **Average length:** Average number of words/terms per note.
- **Vocabulary size:** Total number of distinct words/terms across all notes.
- **Vocabulary covered by UMLS+:** Among the unique words/terms that had appeared in the notes, the percent that could not be found in the UMLS+ dictionary.

- **Average fraction of uncovered text:** The amount of text containing words/terms not covered by UMLS+; per note average is reported.

2) Analysis of acronym usage

Because non-standard use of acronyms could be particularly detrimental to NLP performance in post-processing, we specifically compared the acronym usage among the notes that were voice-dictated and those that were manually typed in.

The first step involved the identification of strings in all capital letters. Not all these strings are acronyms, however: many of them could be capitalized forms of regular words. We used a general English dictionary, GNU Aspell,[18] to filter such instances out. It should be noted that many medical acronyms could be identical to the uppercase spelling of regular English words. In this study, we did not distinguish between them.

The remaining capital letter strings, labeled as "candidate acronyms," were then compared to the UMLS+ dictionary. Those candidate acronyms that could not be found in UMLS+ were thereby identified as potential "questionable" acronyms, which could be misspells or non-standard use of medical terms.

3) Analysis of linguistic proprieties

Many NLP techniques utilize statistical methods to learn language models from textual data. In this study, we employed several commonly used measures to describe such statistical properties of the narrative clinician notes collected:

- **Term frequency (TF) and inverted document frequency (IDF):** TF and IDF jointly measure the importance of a term to a document by calculating its occurring frequency in the document (TF) offset by the overall occurring frequency across all documents (IDF). A higher TF score suggests that a word/term is critical to the document, whereas a higher IDF score means a term is more discriminative of the document in the collection. Syntactical and functional words usually have higher TF and lower IDF, as compared to technical concepts.
- **Power-law exponent:** The value of the scaling exponent for a power-law tail of a term frequency distribution, estimated using the maximum likelihood method based on the empirical data. A higher value suggests a more skewed distribution (long tails), which may be associated with increased complexity in post-processing.
- **Burstiness:** The burstiness measure indicates the likelihood of the same word reoccurring within the same document.[21, 22] In the context of this study, a higher burstiness score generally means that the words/terms tend to contain more content, which may potentially facilitate automated post-processing using NLP tools.

4) Entropy analysis

Entropy analysis is a particularly powerful tool for quantifying the expected information value of a distribution; in our scenario, the probabilistic distributions of words/terms among the narrative clinical documents contained in the two collections, respectively.[23] Entropy analysis has been widely applied to construct indicators of the perplexity of language usage,[24] the difficulty of information retrieval, as well as the ambiguity of a linguistic context.

In this study, we conducted a comparative entropy analysis of the two document types by computing mean entropy residuals. Residual entropy is defined as the difference between the entropy computed based on the language model estimated from a document, and that computed as if all words were uniformly distributed in the document. Lower entropy residuals suggest that the language model of the document is closer to a uniform distribution, which is often associated with higher levels of difficulty in

encoding information from the document. The entropy analysis results hence can be used as a predictor of the potential performance of NLP.

Results

The corpus statistics describing the overall characteristics of the two collections are shown table 1.

Table 1. Corpus statistics

Document type	Average length (num. of words per note)	Vocabulary size	Vocabulary covered by UMLS+	Average fraction of text not covered by UMLS+
Voice-dictated	590	128,394	41.1%	1.9%
Typed-in	378	125,848	37.7%	3.1%

As shown in table 1, average length of the voice-dictated collection (average number of words contained in a note) is about 50% longer than that of the notes manually entered. Despite shorter average length, the vocabulary size of the typed-in notes is at the same level as that of the voice-dictated notes. Further, in both collections, more than half of the words/terms could not be found in UMLS+. The amount of text containing these words/terms is relatively small, though: 1.9% and 3.1% respectively for the two document types studied.

A possible source contributing to this low dictionary coverage is the use of acronyms, which is extremely common in clinical documentation and, intuitively, would occur more often among the notes that were manually entered. Table 2 presents the results of the acronym usage analysis.

Table 2. Analysis of acronym usage

Document type	Strings in all capital letters		Candidate acronyms*		"Questionable" acronyms†	
	Unique instances	Occurrences per note	Unique instances	Occurrences per note	Unique instances	Occurrences per note
Voice-dictated	4,960	6.45	3,472	2.02	1,487	0.12
Typed-in	7,837	12.32	4,191	4.78	1,713	0.35

* Not covered by GNU Aspell.
† Not covered by UMLS+.

As shown in table 2, the voice-dictated and typed-in notes both contain a significant amount of acronyms, and not surprisingly, acronyms appeared more often among the notes that clinicians typed in directly via a computer keyboard. Further, with both document types combined, 41.9% of the "candidate acronyms" could not be found in UMLS+. While it may not always be the case, such acronyms could be misspells or non-standard use of medical terms. It is very interesting to note that a higher percent of the acronyms contained in the voice-dictated notes is not covered by the dictionary (42.8%), as compared to the notes manually entered (40.9%).

Table 3. Linguistic properties: Mean measures

Measure	Document type		*p*-value
	Voice-dictated	Typed-in	
Term frequency (TF)	**138.04**	90.19	< 0.001

Inverse document frequency (IDF)	2.39	**2.66**	< 0.001
Burstiness	**1.86**	1.59	< 0.001
Power-law exponent	1.57	1.59	*Does not apply*
Residual of document entropy	0.56	**0.35**	< 0.001

Table 3 shows the statistical measures delineating the linguistic properties of the two collections. On average, the TF measure of the typed-in notes is much smaller than that of the voice-dictated notes (90.19 vs. 138.04), indicating that information about particular terms is much sparser. Mean IDF of the notes that were entered manually is also significantly larger, suggesting that notes of this document type contain fewer syntactical or functional words, and therefore may be more difficult to process using NLP.

The burstiness measure derived based on the voice-dictated notes is significantly higher that that of the typed-in notes. This result indicates that the voice-dictated notes contain more content, which could facilitate automated post-processing using NLP. Further, the occurrences of words/terms in both collections follow a power-law distribution. The higher power-law exponent of the notes manually typed in (1.59) suggests a more skewed term frequency distribution.

Furthermore, the mean entropy residual of the typed-in note is much lower than that of a voice-dictated note. This means that the entropy of narrative clinician notes manually entered through a computer keyboard is closer to the maximum possible level. Such notes can be much more difficult to process using NLP tools.

Table 4. Linguistic properties: Variance measures (variance divided by mean)

Measure	Document type	
	Dictated	Typed-in
Length	0.56	**0.77**
Unique words/terms	0.41	**0.59**
Fraction of text not covered by UMLS+	0.015	**0.028**
Inverse document frequency (IDF)	0.12	**0.16**

Table 4 reports additional variance-based measures (variance divided by mean). The level of heterogeneity among the typed-in notes is consistently higher than that of the voice-dictated notes, across all measures computed. This result suggests that narrative notes manually entered via a computer keyboard are in general "noisier," and therefore requires NLP tools with higher levels of robustness and adaptiveness.

Discussion

As the analysis results show, the voice-dictated outpatient encounter notes collected from our institutional EHR system demonstrate distinct lexical and distributional properties as compared to the notes entered manually. Such differences could have a significant impact on the performance of NLP in post-processing. While this study did not directly evaluate the magnitude of this impact, the results obtained by contrasting the linguistic properties of the two different document types allude to what the potential effects might be, because the performance of NLP tools is highly sensitive to such properties.[3, 8, 21, 25]

First, it is generally more difficult to effectively parse language features (e.g., part-of-speech tagging, chunking, syntactic parsing, entity extraction) from corpuses with a larger size vocabulary, particularly when the vocabulary may not be adequately covered by available dictionaries. Second, more intensive usage of acronyms in medical documents may be associated with higher levels of ambiguity; for example, PANDAS may refer to the animal as well as "autoimmune disorder," and BD may mean "behavioral disorder" or "blood draw," among many other possibilities. Such ambiguity could increase complexity in post-processing. Third, low term frequency in documents, more skewed term frequency distributions, and higher levels of perplexity may increase the likelihood of yielding inaccurate language models, undermining the performance of NLP.

In this study, we found that despite the shorter average length, the vocabulary size of the encounter notes manually entered is comparable to that of the notes transcribed from voice-dictation. The typed-in notes are also more heterogeneous indicated by the variance-based measures. Narrative clinician notes of this type are therefore "noisier," and may be more difficult to process using NLP. The high level of acronym usage among the notes that were manually entered may be a contributing factor, which may also introduce unique post-processing challenges in its own right.

Additionally, information of particular nuggets contained in the typed-in notes is much sparser as signaled by low TF and high IDF scores. The significantly lower burstiness value and higher power-law exponent further suggest that the typed-in notes contain less content and have a much more skewed term frequency distribution—the resulting longer tail could make it more difficult to estimate language models from these documents. The language model of typed-in notes is also closer to a uniform distribution, as indicated by the significantly lower entropy residuals. These results collectively suggest that the nature of the clinician notes of the same document genre (e.g., encounter notes) could be dramatically different based on the source via which they were created. Therefore, the assumption that medical documents of the same genre would demonstrate similar language features does not seem to hold. Further, the results of the analyses confirmed our study hypothesis: typed-in notes are more likely to be NLP-resistant, and thus require special treatments in post-processing.

It is worth noting that for both document types, more than half of words/terms contained in their respective vocabularies are not covered by UMLS+, a rather exhaustive set combing multiple dictionaries of English words, medical terms, and formalized biomedical nomenclatures. This finding indicates that ontology-based post-processing techniques may encounter extreme difficulties when applied to process such narrative notes, especially those that were manually entered by clinicians via a computer keyboard.

This study has several limitations. First, we only studied one type of clinical documentation: encounter notes generated in ambulatory care settings. We are very interested in further testing if the results of this study may also apply to other genres of clinical documents, such as progress notes, pathology reports, and referral letters. Second, to simplify the study, we used several compromised approaches. For example, we did not distinguish between uppercase English words and medical synonyms that may have identical spelling. Further, we did not take into account the amount of time it took to complete documentation using these two different input methods, which may have direct implications on the quality of data resulted. Methodological enhancements are needed to improve the precision of the statistical measures used. Third, it is possible that those clinicians who chose to document via a computer keyboard might have distinct documentation styles comparing to those who chose to dictate; hence, the linguistic differences revealed through this study may originate in individual differences rather than the method of data entry. In addition, our institution works with multiple transcription services. The differences among them may also play a role. We could not test these hypotheses within the scope of this study, however. Finally, we did not directly evaluate how NLP tools may perform on the same types of documents created via different methods, which will be valuable future expansions to validate our study findings.

Conclusion

We comparatively examined the linguistic properties of 30,000 voice-dictated clinician notes versus 30,000 that were manually entered. The documents included in the study are encounter notes generated in ambulatory care settings. We found that typed-in encounter notes are generally "noisier" than voice-dictated notes, and information of particular nuggets contained in them is much sparser. With additional statistical measures we computed, including burstiness, power-law exponent, and residuals of document entropy, we concluded that the language quality of typed-in notes is poorer overall, which may raise unique challenges in post-processing. This study, through quantifying the differences between these two document types, may improve our understanding of the distinct language features originating from the source via which narrative clinician notes were created. The results may therefore provide useful insights into the development of context-appropriate post-processing strategies.

Acknowledgments

This project was supported by Grant HHSN276201000032C received from the National Library of Medicine, and in part by Grant # UL1RR024986 received from the National Center for Research Resources (NCRR), a component of the National Institutes of Health (NIH) and NIH Roadmap for Medical Research.

References

1. **Cannon J,** Lucci S. Transcription and EHRs. Benefits of a blended approach. *J AHIMA* 2010;**81**:36–40.
2. **Rosenbloom ST,** Denny JC, Xu H, *et al*. Data from clinical notes: A perspective on the tension between structure and flexible documentation. *J Am Med Inform Assoc* 2011;**18**:181–6.
3. **Stanfill MH,** Williams M, Fenton SH, *et al*. A systematic literature review of automated clinical coding and classification systems. *J Am Med Inform Assoc* 2010;**17**:646–51.
4. **Zheng K,** Mei Q, Hanauer DA. Collaborative search in electronic health records. *J Am Med Inform Assoc* 2011. (in press)
5. **Friedman C,** Shagina L, Lussier Y, *et al*. Automated encoding of clinical documents based on natural language processing. *J Am Med Inform Assoc* 2004;**11**:392–402.
6. **Los RK,** van Ginneken AM, van der Lei J. OpenSDE: A strategy for expressive and flexible structured data entry. *Int J Med Inform* 2005;**74**:481–90.
7. **Johnson SB,** Bakken S, Dine D, *et al*. An electronic health record based on structured narrative. *J Am Med Inform Assoc* 2008;**15**:54–64.
8. **Zeng-Treitler Q,** Goryachev S, Weiss S, *et al*. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: Evaluation of a natural language processing system. *BMC Med Inform Decis Mak* 2006;**6**:30.
9. **Hersh WR.** Adding value to the electronic health record through secondary use of data for quality assurance, research, and surveillance. *Am J Manag Care* 2007;**13**:277–8.
10. **Etheredge LM.** A rapid-learning health system. Health Aff (Millwood). 2007;**26**:w107–8.
11. **Leroy G,** Eryilmaz E, Laroya BT. Health information text characteristics. *AMIA Annu Symp Proc* 2006:479–83.
12. **Zeng-Treitler Q,** Kim H, Goryachev S, *et al*. Text characteristics of clinical reports and their implications for the readability of personal health records. *Stud Health Technol Inform* 2007;**129**:1117–21.
13. **Verspoor K,** Cohen KB, Hunter L. The textual characteristics of traditional and Open Access scientific journals are similar. *BMC Bioinformatics* 2009;**10**:183.
14. **Cohen KB**, Johnson HL, Verspoor K, et al. The structural and content aspects of abstracts versus bodies of full text journal articles are different. BMC Bioinformatics. 2010;**11**:492.

15. **Wellner B,** Huyck M, Mardis S, et al. Rapidly retargetable approaches to de-identification in medical records. J Am Med Inform Assoc 2007;**14**:564–73.
16. **Aberdeen J,** Bayer S, Yeniterzi R, *et al*. The MITRE Identification Scrubber Toolkit: design, training, and assessment. Int J Med Inform 2010;**79**:849–59.
17. **Yeniterzi R,** Aberdeen J, Bayer S, *et al*. Effects of personal identifier resynthesis on clinical text de-identification. J Am Med Inform Assoc 2010;**17**:159–68.
18. **Atkinson K.** GNU Aspell. http://aspell.net/ (accessed 10 Mar, 2011).
19. **e-MedTools.** *OpenMedSpel - Opensource Medical Spelling*. http://www.e-medtools.com/openmedspel.html (accessed 10 Mar, 2011).
20. **U.S. National Library of Medicine.** *Unified Medical Language System® (UMLS®).* http://www.nlm.nih.gov/research/umls/ (accessed 5 Dec, 2010).
21. **Church KW,** Gale WA. Poisson mixtures. Nat Lang Eng 1995;**1**:163–90.
22. **Clinchant S,** Gaussier E. Information-based models for ad hoc IR. In: *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval (SIGIR'10)*. New York, NY, USA: ACM Press. 2010;234–41.
23. **Shannon CE.** A mathematical theory of communication. Bell Syst Tech J 1948;**27**:379–423, 623–56.
24. **Cole R,** ed. *Survey of the State of the Art in Human Language Technology*. New York, NY, USA: Cambridge University Press. 1997.
25. **Banko M,** Brill E. Scaling to very very large corpora for natural language disambiguation. In: *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics (ACL '01)*. Stroudsburg, PA, USA: Association for Computational Linguistics. 2001;26–33.