Original Article

Health Informatics Journal

Health Informatics Journal
2020, Vol. 26(4) 2906–2914

DOI: 10.1177/1460458220951719
journals.sagepub.com/home/jhi
SAGE

Identifying relevant information in medical conversations to summarize a clinician-patient encounter

Juan C Quiroz
Liliana Laranjo
Ahmet Baki Kocaballi
Australian Institute of Health Innovation, Macquarie University, Australia

Agustina Briatore
Health Information Systems Office, Ministry of Health, Argentina

Shlomo Berkovsky
Dana Rezazadegan
Enrico Coiera
Australian Institute of Health Innovation, Macquarie University, Australia

Abstract

To inform the development of automated summarization of clinical conversations, this study sought to estimate the proportion of doctor-patient communication in general practice (GP) consultations used for generating a consultation summary. Two researchers with a medical degree read the transcripts of 44 GP consultations and highlighted the phrases to be used for generating a summary of the consultation. For all consultations, less than 20% of all words in the transcripts were needed for inclusion in the summary. On average, 9.1% of all words in the transcripts, 26.6% of all medical terms, and 27.3% of all speaker turns were highlighted. The results indicate that communication content used for generating a consultation summary makes up a small portion of GP consultations, and automated summarization solutions—such as digital scribes—must focus on identifying the 20% relevant information for automatically generating consultation summaries.

Corresponding author:
Juan C Quiroz, Centre for Big Data Research in Health, Level 2, AGSM Building UNSW, Sydney, 2052, Australia.
Email: juan.quiroz@unsw.edu.au

Keywords
Automatic summarization, clinical conversations, digital scribe, GP consultation, Pareto principle, natural language processing

Introduction

Clinical documentation is one of the main factors driving clinician burnout. Recent studies have found that physicians in the United States spend almost half of their time on electronic health records (EHRs) and clerical activities[1–3] and less than one-third on face-to-face time with patients.[3] The consequences of EHR use include decreased professional satisfaction, time-consuming data entry, and disruption of face-to-face patient care.[4] For these reasons, there is a strong need for solutions to reduce the documentation burden on clinicians.

Medical scribes are one solution, as they can increase physician efficiency, satisfaction, and the number of patients cared for.[5] Automated documentation tools such as digital scribes aim to provide a cost-effective and scalable alternative to medical scribes.[6–9] A digital scribe is a system that records a conversation between a clinician and a patient and generates a summary of the conversation, similar to the function performed by human medical scribes. Advances in digital scribe development hinge on solving several technical challenges,[9] including identifying and summarizing salient information in medical consultations.[10]

General Practitioners (GPs) lead consultations with patients through a series of questions to understand the problem and arrive at a diagnosis. This process is not linear due to the characteristics of naturally occurring human conversations and the inherent complexity of GP consultations.[11,12] Not all of the content of a GP-patient conversation needs to be part of the summary entered into the EHR. Summaries of GP consultations usually follow the generally accepted structure Subjective-Objective-Assessment-Plan (SOAP).[12] As such, the summarization algorithm of the digital scribe must analyze all the content of the GP-patient conversation and determine what is relevant for the summary.

Digital scribe research has focused on applying machine learning techniques to improve the performance of components that make up a digital scribe, such as speech recognition and summarization, rather than building end-to-end systems.[9] A recent study described a proof-of-concept digital scribe, with evaluation limited to eight doctor-patient conversations.[13] Machine learning research related to digital scribes includes clinical speech recognition,[14,15] extraction of clinical information from transcripts of medical conversations,[16–19] and summarization of medical conversations to generate medical notes.[20] While existing work has focused on machine learning to advance digital scribe research, no work to date has explored the relationship between what is exchanged between a doctor and a patient during a consultation and relevancy to the documentation of the encounter.

The summarization component of a digital scribe[9] may employ extractive summarization (identifying important words or sentences and stringing them together to form a summary) or abstractive summarization (identifying important words or sentences and rewriting them to form a summary) to generate a summary of the doctor-patient conversation.[10,21] To assist in the design of digital scribes and summarization algorithms, this exploratory study sought to determine what proportion of doctor-patient communication in GP consultations is used by GPs for generating a summary of the consultation. Our main contribution is to empirically demonstrate that only a small proportion of conversation during a GP consultation is relevant for a consultation summary.

GP:	Do you smoke cigarettes?	GP:	Now, what would you like to talk about today?
Patient:	No.	Patient:	It's just my ears. I noticed that today I was in a lecture and - it's been happening occasionally - I find that I get a pulsing sense in my ear and I have had ear infections before so I kind of - and it's happening right now.
GP:	No, do you drink alcohol?		
Patient:	No.		
GP:	Any significant family medical history, so parents, grandparents?	GP:	In your left ear?
Patient:	Yeah, undiagnosed mental health issues.	Patient:	Both ears. I get pain in my ears. Sometimes I will wake up in the middle of the night and I've just got excruciating pain in my ears and I…
GP:	Okay...		
Patient:	That's about it.		
GP:	… for both mum and dad?		
Patient:	Yeah.		

Figure 1. Excerpts from two GP-patient conversations and the text highlighted by a human coder as relevant for generating a summary of the consultation.

Methods

Dataset

Data collection involved audio-recording doctor-patient conversations of 44 GP consultations at a hospital in Sydney, Australia.[11] Physicians and patients were recruited using a convenience sampling strategy. Inclusion criteria for physicians required them to be a primary care doctor and use an EHR for documentation purposes. Patients needed to be at least 18 years of age and have English language competency. Researchers obtained informed consent from all participants. The Macquarie University Ethics Committee approved the study.

The 44 doctor-patient conversations were transcribed verbatim. The transcripts dataset comprised 96,096 word occurrences and 5129 unique words. 28 out of the 44 consultations (63.6%) were with returning patients and 16 (36.4%) with new patients. 29 patients (65.9% (29/44)) were female and no patient appeared in more than one transcript.

Coding task

Two researchers with an MD degree and at least 1 year of clinical experience coded the transcripts independently. One researcher coded all 44 transcripts. To validate the reliability of the coding, the second researcher coded five transcripts (5/44, 11.4%) to determine inter-rater reliability using Cohen's kappa coefficient, with the two coders having agreement if they highlighted a word or phrase within the same speaker turn. We used the overlap coefficient as a similarity metric for the text highlighted by the coders.[22]

Each coder was given the same instructions: to highlight words or phrases in the transcripts they would use for creating the documentation of the consultation (Figure 1), such as the content needed to create the SOAP section of the EHR; each consultation should be documented in a useful and efficient summary using the words highlighted (researchers were not tasked with writing the summary); repeating highlights were allowed (i.e. highlighting "headache" on different parts of the conversation), to bring the context back to the word or phrase being highlighted. Coders were

Table 1. Evaluation metrics for highlighted content. All transcripts had less than 20% of their words marked as relevant for generating a summary of the consultation (<35% without stop words). For medical terms, 75% of the transcripts had ⩽31% highlighted. For speaker turns, 75% of the transcripts had <40% of their speaker turns with a word or phrase highlighted.

	Mean	95 CI	Min	IQR	Max
Words	9.11	(8.05, 10.17)	2.58	5.97–11.72	18.91
Words without stop words	15.98	(14.17, 17.79)	4.86	10.87–20.19	33.25
Medical terms	26.57	(23.86, 29.28)	12.64	18.94–31.0	55.56
Speaker turns	27.32	(23.15, 31.50)	4.75	16.67–39.02	54.12

CI: confidence interval; IQR: interquartile range.

blinded regarding our hypothesis. To maximize use of all coded transcripts, our results make use of the transcripts coded by the first coder (all 44 transcripts) and the second coder (5 transcripts out of the same set of 44), for a total of 49 transcripts.

Evaluation metrics

Each evaluation metric was calculated per transcript, with results averaged across all the transcripts coded by the two coders (49 transcripts). We used the following quantitative metrics to assess the amount and type of content highlighted: (1) word count percentage, (2) word count percentage without stop words (the most common words in a language, commonly removed during natural language processing), (3) medical term count percentage (words or phrases identified with MetaMap[23]), (4) speaker turn percentage, and (5) part-of-speech percentage (labels indicative of the category of a word depending on grammar).

For word count percentage, we divided the number of highlighted words by the total number of words in each transcript. For word count percentage without stop words, we ignored stop words when calculating the number of highlighted words and the total number of words (default stop words from the spaCy Python library—https://spacy.io). However, we counted the negations "no", "nor", and "not", as a consultation summary needs to record negatives.[24]

For medical term count percentage, we divided the number of highlighted medical terms by the total number of medical terms (identified with MetaMap Lite[23]). Medical phrases identified by MetaMap were counted as one word (e.g. "heart attack" was counted as one term). Identification of medical terms was limited to the MetaMap semantic groups of "Anatomy", "Disorders", "Physiology", and "Procedures".

For speaker turn percentage, we divided the number of speaker turns that had at least one word highlighted by the total number of speaker turns in each transcript. For part-of-speech counts, we used the spaCy Python library (model "en_core_web_sm") to identify the part-of-speech tag for each word. The parts-of-speech tags used were nouns, verbs, pronouns, adjectives, adposition, particle, determiner, coordinating conjunction, adverb, subordinating conjunction, interjection, auxiliary, and numerical tokens.

Results

The kappa statistic was 0.52 (moderate agreement) for the coding of the five transcripts.[25] The mean overlap coefficient for the tokens highlighted by the two coders for the five transcripts was 0.71 (standard deviation (SD) 0.05). Table 1 compares the evaluation metrics for the highlighted

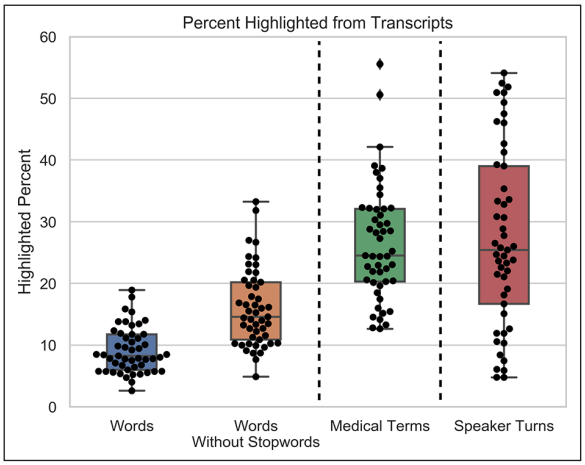

Figure 2. Boxplot and swarm plot of the highlighted percent from each GP-patient consultation transcript relevant for generating a consultation summary. The plots are for each of the quantitative metrics used to assess the amount and type of highlighted content: (1) word count percentage, (2) word count percentage without stop words, (3) medical term count percentage, and (4) speaker turn percentage. All transcripts had less than 20% of their words marked as relevant for generating a summary of the consultation (<35% without stop words).

content and Figure 2 illustrates a boxplot and swarm plot of the distribution of each metric. The highlighted percentage of words in all the transcripts was less than 20% when using all words and less than 35% when removing stop words. On average, 9.11% of all words in the transcripts, 15.98% of all words without stop words, 26.57% of all medical terms, and 27.32% of all speaker turns were highlighted. The distribution of highlighted medical terms per transcript was right-skewed and indicative of duplicates, with a mean skew of 2.74 (95% CI, 2.48, 2.99; SD 0.84). Figure 3 illustrates the percent highlighted for every part-of-speech. The highest percentages were for numerical tokens (25.7%), nouns (15.7%), and adjectives (14.2%). The percentages for all other parts-of-speech were below 10%.

Discussion

Main findings

Our study is the first to show that out of everything said between a GP and a patient during a consultation, only a small percentage is relevant for the consultation summary (less than 20% of the total words). The number of highlighted words from the transcripts being less than 20% suggests that the distribution of content relevant for documentation purposes may follow the Pareto principle (80/20 rule).[26] Analysis of a larger sample is needed to validate this hypothesis.

The moderate agreement and the 0.71 mean overlap coefficient indicate that the coders highlighted similar words, but not necessarily in the same location of the conversation. We note that our emphasis was not on showing that the two coders would highlight the same words or the same speaker turns. Instead, our goal was to show that regardless of how coders highlighted the transcripts, the highlighted content was a small portion of the entire consultation. This is shown in our results, with all metrics calculated using the highlighted transcripts of both coders (49 transcripts).

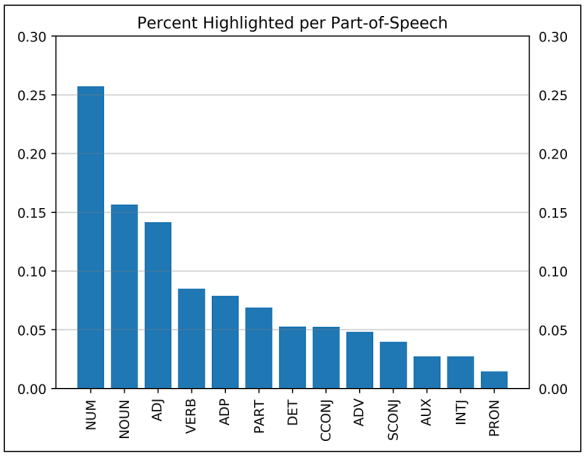

Figure 3. Percent of content from a GP-patient consultation relevant for documentation purposes for every part-of-speech. Numerical tokens (NUM, 25.7%), nouns (NOUN, 15.7%), and adjectives (ADJ, 14.2%) made up the highest percentages. The other parts-of-speech were below 10%: verbs (VERB), adposition (ADP), particle (PART), determiner (DET), coordinating conjunction (CCONJ), adverb (ADV), subordinating conjunction (SCONJ), auxiliary (AUX), interjection (INTJ), and pronoun (PRON).

When removing stop words, at most 30% of the total words were needed for the summary. However, stop words may contain important information for making meaning of the GP-patient interaction, such as negations ("no", "not") and words used by patients to describe their conditions (e.g. "above", "across", "after", "before", "side", "serious", "sometimes"). As such, the percent of highlighted words may have been lower than 30% if the stop words had been filtered to exclude all words potentially useful in a GP-patient conversation.

Medical terms were highlighted at a higher rate than words, but the majority of the transcripts had less than 35% medical terms highlighted. The right skewness of the highlighted medical terms suggests they may be power law distributed, with a few medical terms appearing with high frequency (many duplicates) and a long tail of less frequent medical terms (with few or no duplicates). Given this finding, medical term keyword spotting may not necessarily capture the information needed for a summary and may be misleading. The duplicates do increase the likelihood that a medical term missed in one mention may still be detected in a different mention. Future work is needed to determine if highlighted medical terms are power law distributed and how this can be exploited to tailor summarization methods.

Highlighted speaker turn percentages had the widest distribution spread of all metrics. This suggests that depending on the type of consultation, the content relevant for documentation may be in a few speaker turns or spread over the conversation across a larger number of speaker turns. If the frequency of highlighted speaker turns concentrate on certain areas, this may enable targeted summarization based on location.

When analyzing the parts-of-speech that were highlighted, numbers (25.7%), nouns (15.7%), and adjectives (14.2%) made up the biggest percentages. This may prove useful for digital scribe development by knowing that numerical tokens, nouns, and adjectives are more relevant than other parts-of-speech.

Our results give insight into the quantity and type of content that extractive summarization and information extraction may need to target. When it comes to individual words (the basic unit used in most natural language processing tasks) about 20% or less of the conversation should be captured, with the rest being potentially redundant or not informative enough to warrant inclusion in the

summary. This practically means that a digital scribe—whose goal is to generate a summary of the conversation to remove the documentation burden on the GP[9]—should be capturing about 20% of the conversation to generate an extractive summary. Exploration of patterns associated with highlighted words, medical terms, and speaker turns could also guide the design of rules for information extraction. Finally, if a portion of the conversation can be discarded before applying a summarization or information extraction model, then the performance of the models could be improved.

Limitations

The results of this study are preliminary and exploratory. The transcripts were coded by only two clinicians (5/44 used for inter-rater reliability). While we test inter-rater reliability, our primary goal was to show that different coders highlight only a small proportion of the consultation transcripts (less than 20%), regardless of how they choose to highlight the transcripts and the level of highlighted agreement. Aside from the instructions given to the coders, there was no gold standard and no discussion between the coders. As such, a different set of coders may have highlighted different content. This can be remedied in future studies by having better training of the coders and a discussion between coders on a test set of transcripts before the coding task.

Our analysis was limited to transcripts of consultations by seven GPs working in a single GP clinic in Sydney, Australia. This sample may not be representative of GP consultations in other clinics, countries, and different styles of eliciting information from patients. Languages other than English with a different grammar structure may also result in a different proportion of words from the consultations being included in the summary. This paper focuses on quantitative analysis. As such, future work should address qualitative analysis of the highlighted text.

Conclusion

This study of GP-patient consultations suggests that the communication content used for generating a summary of the consultation makes up a small portion of the entire consultation, with word counts being less than 20% (potentially following the Pareto principle). Further work is needed to generate additional evidence for these observations, including larger samples of (1) transcripts from a wider pool of physicians and (2) transcript coders. Knowing that the information from a GP consultation used to generate a summary is 20% or less can guide future digital scribe and machine learning algorithmic development by focusing on identifying the 20% of information relevant for documentation purposes.

Author's note

Juan C Quiroz is also affiliated with the Centre for Big Data Research in Health, UNSW, Australia. Liliana Laranjo is also affiliated with Westmead Applied Research Centre, The University of Sydney, Australia. Ahmet Baki Kocaballi is also affiliated with the School of Computer Science, University of Technology Sydney, Australia. Dana Rezazadegan is also affiliated with the Department of Computer Science and Software Engineering, Swinburne University of Technology, Australia.

Author contributions

J.C.Q. conceived the paper. J.C.Q., L.L., A.B.K., D.R., and E.C. designed the study. A.B. coded all transcripts. J.C.Q. did the data analysis and wrote the initial draft. All authors participated in critical review and writing of the final text. All authors approved the final draft.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by the National Health and Medical Research Council (NHMRC) grant APP1134919 (Centre for Research Excellence in Digital Health).

ORCID iD

Juan C Quiroz https://orcid.org/0000-0003-0241-5376

References

1. Arndt BG, Beasley JW, Watkinson MD, et al. Tethered to the EHR: Primary care physician workload assessment using EHR event log data and time-motion observations. *Ann Fam Med* 2017; 15(5): 419–426.
2. Overhage JM and McCallie D Jr. Physician time spent using the electronic health record during outpatient encounters: a descriptive study. *Ann Intern Med* 2020; 172(3): 169–174.
3. Sinsky C, Colligan L, Li L, et al. Allocation of physician time in ambulatory practice: a time and motion study in 4 specialties. *Ann Intern Med* 2016; 165(11): 753.
4. Friedberg MW, Chen PG, Van Busum KR, et al. Factors affecting physician professional satisfaction and their implications for patient care, health systems, and health policy. *Rand Health Q* 2014; 3(4): 1.
5. Shultz CG and Holmstrom HL. The use of medical scribes in health care settings: a systematic review and future directions. *J Am Board Fam Med* 2015; 28(3): 371–381.
6. Lin SY, Shanafelt TD and Asch SM. Reimagining clinical documentation with artificial intelligence. *Mayo Clin Proc* 2018; 93(5): 563–565.
7. Finley G, Edwards E, Robinson A, et al. An automated medical scribe for documenting clinical encounters. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, New Orleans, Louisiana, June 2018, pp. 11–15. Association for Computational Linguistics.
8. Coiera E, Kocaballi B, Halamka J, et al. The digital scribe. *npj Digit Med* 2018; 1(1): 58.
9. Quiroz JC, Laranjo L, Kocaballi AB, et al. Challenges of developing a digital scribe to reduce clinical documentation burden. *npj Digit Med* 2019; 2(1): 114.
10. Nenkova A and McKeown K. A survey of text summarization techniques. In: Aggarwal CC and Zhai C (eds) *Mining Text Data*. Boston, MA: Springer US, 2012, pp. 43–76.
11. Kocaballi AB, Coiera E, Tong HL, et al. A network model of activities in primary care consultations. *J Am Med Inform Assn* 2019; 26(10): 1074–1082.
12. Weed LL. *Medical records, medical education, and patient care: the problem-oriented record as a basic tool*. Cleveland: Press of Case Western Reserve University, 1969.
13. Molenaar S, Maas L, Burriel V, et al. Medical dialogue summarization for automated reporting in healthcare. *Adv Inform Sys Eng Workshops* 2020; 382: 76–88.
14. Flemotomos N, Georgiou P and Narayanan S. Linguistically aided speaker diarization using speaker role information. In: *Odyssey 2020 the speaker and language recognition workshop*, Tokyo, Japan, 1 November 2020, pp. 117–124. ISCA.
15. Shafey LE, Soltau H and Shafran I. Joint speech recognition and speaker diarization via sequence transduction. In: *Interspeech* 2019, 2019, pp. 396–400. International Speech Communication Association.
16. Jeblee S, Khan Khattak F, Crampton N, et al. Extracting relevant information from physician-patient dialogues for automated clinical note taking. In: *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, Hong Kong, November 2019, pp. 65–74. Association for Computational Linguistics.

17. Rajkomar A, Kannan A, Chen K, et al. Automatically Charting symptoms from patient-physician conversations using machine learning. *JAMA Intern Med* 2019; 179(6): 836–838.
18. Selvaraj SP and Konam S. Medication regimen extraction from medical conversations. In: *International Workshop on Health Intelligence (W3PHIAI) of the 34th AAAI conference on artificial intelligence*, 2020, New York, USA, 2 January 2020, http://arxiv.org/abs/1912.04961 (accessed 24 June 2020).
19. Shafran I, Du N, Tran L, et al. The medical scribe: corpus development and model performance analyses. *arXiv:2003.11531 [cs]*, http://arxiv.org/abs/2003.11531 (accessed 22 June 2020).
20. Enarvi S, Amoia M, Teba MD-A, et al. Generating medical reports from patient-doctor conversations using sequence-to-sequence models. In: *Proceedings of the first workshop on natural language processing for medical conversations*, July 2020, pp. 22–30. Association for Computational Linguistics, https://www.aclweb.org/anthology/2020.nlpmc-1.4 (accessed 23 June 2020).
21. Gambhir M and Gupta V. Recent automatic text summarization techniques: a survey. *Artif Intell Rev* 2017; 47(1): 1–66.
22. Granovetter MS. The strength of weak ties. *Am J Sociol* 1973; 78(6): 1360–1380.
23. Demner-Fushman D, Rogers WJ and Aronson AR. MetaMap Lite: an evaluation of a new Java implementation of MetaMap. *J Am Med Inform Assoc* 2017; 24(4): 841–844.
24. Chapman WW, Bridewell W, Hanbury P, et al. A simple algorithm for identifying negated findings and diseases in discharge summaries. *J Biomed Inform* 2001; 34(5): 301–310.
25. McHugh ML. Interrater reliability: the kappa statistic. *Biochem Med* 2012; 22(3): 276–282.
26. Newman MEJ. Power laws, Pareto distributions and Zipf's law. *Contemp Phys* 2005; 46(5): 323–351.