Voice recognition software can be used for scientific articles

Hans-Christian Pommergaard, Chenxi Huang, Jacob Burcharth & Jacob Rosenberg

ORIGINAL ARTICLE

Department of Surgery, Herlev Hospital

Dan Med J 2015;62(2):A5012

ABSTRACT

INTRODUCTION: Dictation of scientific articles has been recognised as an efficient method for producing high-quality, first article drafts. However, standardised transcription service by a secretary may not be available for all researchers and voice recognition software (VRS) may therefore be an alternative. The purpose of this study was to evaluate the out-of-the-box accuracy of VRS.

METHODS: Eleven young researchers without dictation experience dictated the first draft of their own scientific article after thorough preparation according to a pre-defined schedule. The dictate transcribed by VRS was compared with the same dictate transcribed by an experienced research secretary, and the effect of adding words to the vocabulary of the VRS was investigated. The number of errors per hundred words was used as outcome. Furthermore, three experienced researchers assessed the subjective readability using a Likert scale (0-10). Dragon Nuance Premium version 12.5 was used as VRS.

RESULTS: The median number of errors per hundred words was 18 (range 8.5-24.3), which improved when 15,000 words were added to the vocabulary. Subjective readability assessment showed that the texts were understandable with a median score of five (range 3-9), which was improved with the addition of 5,000 words.

CONCLUSION: The out-of-the-box performance of VRS was acceptable and improved after additional words were added. Further studies are needed to investigate the effect of additional software accuracy training.

FUNDING: not relevant.

TRIAL REGISTRATION: not relevant.

Use of dictation for the writing of scientific articles is an effective method and may reduce the risk of writer's block. Using this method helps researchers create manuscripts of an initial high quality with a suitable language complexity, even when the method is used by inexperienced writers [1]. With thorough preparation, researchers may gain increased confidence in their writing skills owing to a positive writing experience, in some cases characterised by a feeling of flow [2]. However, a considerable amount of time is needed for preparation to achieve the full overview of the contents of the article that allows the author to dictate the first draft. Using the mind-to-paper method, which includes thorough preparation through the production of a detailed outline, dictating the manuscript in 4-5 hours is feasible, even for researchers with no or limited experience [1].

When using dictation for scientific writing, an efficient and reliable transcription method is needed. However, many researchers may not have access to transcription assistance from a secretary, and therefore voice recognition software (VRS) may be a feasible option. Physicians' use of VRS for dictation has been investigated in a variety of clinical scenarios [3-10]. However, VRS has not been evaluated for dictation of scientific articles.

The purpose of this study was to evaluate the out-of-the-box accuracy of a VRS for dictation of scientific articles. Moreover, we evaluated the effect of adding additional topic-specific words to the vocabulary of the software.

METHODS

This study evaluated the first manuscript drafts of articles dictated by 11 young researchers (medical students, PhD students and research nurses) with no previous dictation experience. All manuscripts were for original English language articles investigating health science topics. The dictation was based on a detailed manuscript outline prepared during a 1-month period prior to the dictation [1]. The manuscript outline was prepared in cooperation with the researcher's supervisor to ensure an optimal preparation. None of the researchers had started writing the manuscript or used a premade study protocol for dictation. Thorough preparation gives the researcher full overview of the contents of the article, which is necessary to dictate the first draft. This method has previously been shown to be effective and to produce manuscripts of a high initial quality [1]. For dictation, the researchers used a dictation device such as an iPhone (Apple, California, USA) with a dictation application allowing them to record an audio file. Next, the file was transcribed by the VRS. As VRS, we used Dragon Nuance Premium version 12.5 (Nuance Communications Inc., Burlington, MA, USA). To evaluate the accuracy of VRS and the readability of the transcribed manuscripts, the quality of transcripts was compared with that of an experienced research secretary which we used as gold standard. The secretary in question has 45 years of experience transcribing scientific

FIGURE 1

Number of errors (medians and interquartile ranges) per 100 words using voice recognition software – total and different subtypes.
a) Overall comparison (Friedman's test)
b) Individual comparison (Wilcoxon signed-rank test).
c) p-value could not be calculated, due to no difference in the data.

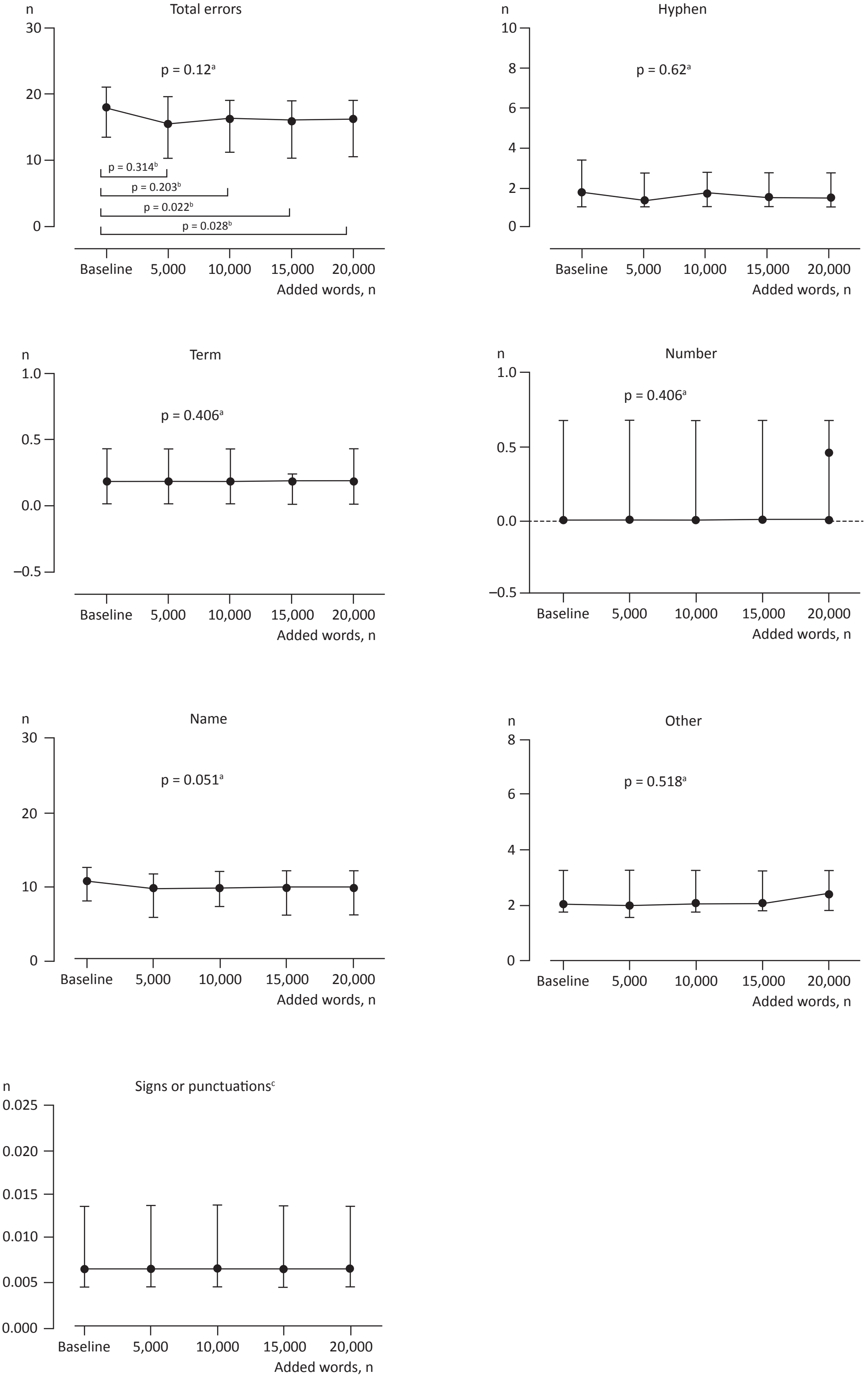
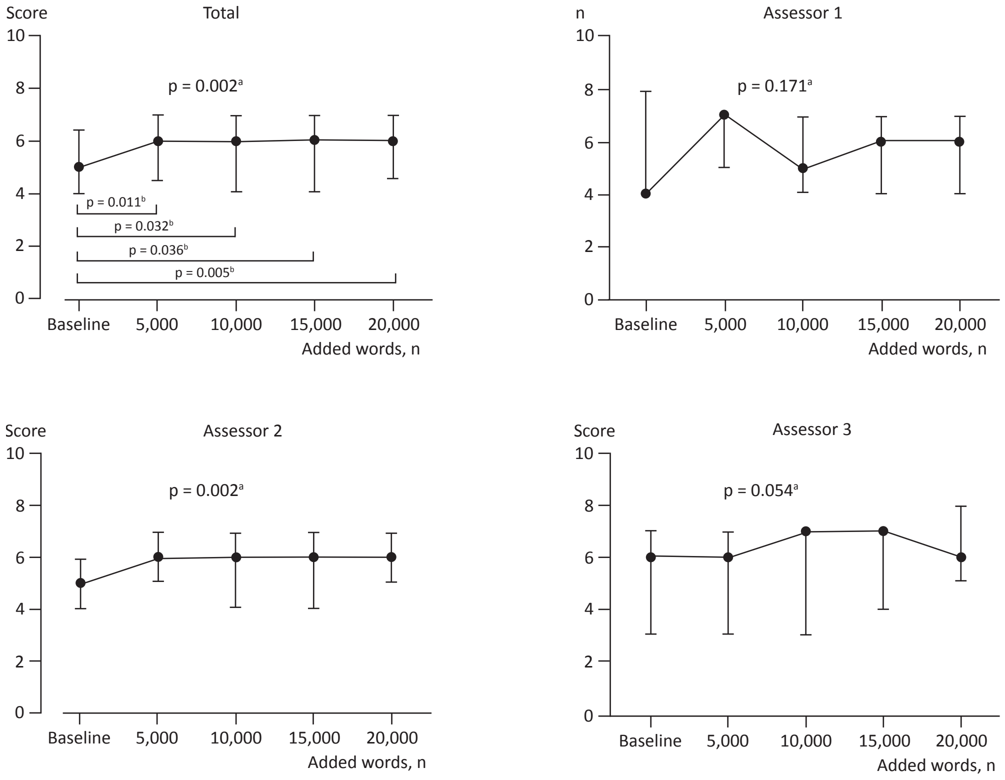


FIGURE 2



Qualitative score (medians and interquartile ranges) of transcripts using voice recognition software – pooled data and data for each assessor.
a) Overall comparison (Friedman's test).
b) Individual comparison (Wilcoxon signed-rank test).

articles. Thus, in our experience, the transcripts done by the research secretary are almost without errors, and are therefore considered optimal transcripts.

Apart from the mandatory set-up of a user profile in the VRS, no additional training was performed. Thus, this study compared the out-of-the-box accuracy of the software with the accuracy of an experienced secretary. The following VRS settings were used: age group 22-54 years, United Kingdom as region, accent as standard, source as handheld or smartphone with recording application, and speech model as BestMatch V. The introductory sections of the articles were transcribed using both transcription modalities and the transcripts were then compared.

To test the accuracy of adding additional topic-specific words to the software vocabulary, five consecutive transcripts using the VRS were carried out with 0 (baseline), 5,000, 10,000, 15,000, and 20,000 words scanned by the software using the "learn from specific document"-wizard. These words were extracted from relevant topic-specific articles provided by the researchers. Typically, the introduction sections of these articles were used. All five transcripts were compared with the gold standard to evaluate the effect of number of added words on the quality of the VRS transcript.

The accuracy of the VRS was measured by the total number of errors per 100 words compared with the gold standard. These errors were subdivided into the following:

- Hyphen or two words, e.g. world wide versus world-wide
- Medical terms, e.g. hysterectomy versus hysterical
- Numbers, e.g. two versus 2
- Name, e.g. Jones versus Joe has
- Signs or punctuation marks: Signs, e.g. "–/." versus hyphen/period. Punctuations, e.g. "," instead of "."
- Other.

Three experienced researchers performed a subjective qualitative assessment of the transcripts. This was done using a readability score (Likert scale 0-10) for measuring the degree of ability to understand the meaning of the text. The following three intervals considering the degree of meaning were created: 0-3: limited (difficult to grasp), 4-6: moderate (the essential part of the text understood with some uncertain details), 7-10: high (easy to understand without any disturbing errors). The assessors were blinded for author, title and number of added words to the VRS vocabulary; and the assessment order of transcripts was randomized. Thus, the assessor did not know the number of added words.

Using Q-Q-plots, histograms and the Kolmogorov-Smirnov test, we found that the number of errors per 100 words and the quality assessment by Likert scale were not normally distributed. Given the dependent nature of data and the repeated measures design, Friedman's test was used to evaluate the overall effect of added words, and significant values would indicate an increase or a decrease with an increasing number of added words. The Wilcoxon signed-rank test was used to compare the individual number of added words versus the baseline, but only when $p < 0.05$ in the Friedman analysis of the parameter in question. This was done to investigate the minimum number of words necessary to obtain an effect. The median number of errors and the median quality score as a result of number of added words are shown in **Figure 1** and **Figure 2**, respectively. Interclass correlation coefficiency (ICC) was used to evaluate the degree of inter-observer variation between the three assessors (absolute agreement, two-way mixed model). ICC can be arbitrarily divided into the following categories: poor (< 0.00), slight (0.00-0.20), fair (0.21-0.40), moderate (0.41-0.60), substantial (0.61-0.80), and almost perfect (0.81-1.00) [11]. SPSS version 20 (SPSS Inc., Chicago, Illinois, USA) was used, and a p-value < 0.05 was considered statistically significant. Informed consent from the participants was obtained prior to the study. The study was exempt from ethical committee approval according to Danish law since it was not considered biomedical research, and no additional approvals were needed.

Trial registration: not relevant.

RESULTS

All eleven researchers completed dictation of a full article in one day despite having no previous experience with dictation. The median number of words for the analysed introduction section was 289 (range 148-548 words).

The quantitative analysis showed that the median number of errors per hundred words was 18 (range 8.5-24.3 words) at baseline. This number of errors was reduced as words were added to the vocabulary. Individual comparisons pointed at 15,000 words as the minimum number of words needed (Figure 1). The majority of the errors were in the "name"-category followed by "other" and "hyphen"-category, whereas only few errors were found in the remaining categories (Figure 1).

For the qualitative analysis, there was moderate agreement between the observers for all levels of added words (**Table 1**). The texts generally had acceptable meaning with a median Likert score of 5 (3-9), which improved after adding words. Here, individual comparisons pointed to 5,000 words as the minimum number of words needed (Figure 2).

TABLE 1

Interclass correlation coefficiency (ICC) for the three assessors at each level.

	ICC value[a]
Baseline	0.492
5,000 words	0.496
10,000 words	0.558
15,000 words	0.577
20,000 words	0.507

a) 1 = perfect agreement; 0 = no agreement at all.

DISCUSSION

The out-of-the-box quality of scientific articles dictated using Dragon Nuance Premium VRS had a median accuracy of 18 errors per 100 words compared with an experienced secretary whose transcription was used as gold standard. Adding of topic-specific words to the vocabulary improved the accuracy by reducing the number of errors and improving the quality rating.

Other studies have compared the use of VRS with secretaries for clinical transcriptions. Experience from an outpatient clinic found that the use of VRS reduced the turnaround time (time from dictation to written note) [4]. The use of VRS for radiology reports reduced the turnaround time, was more cost-effective and reduced the number of spelling errors [7]. In contrast, the use of VRS for pathology reports was less accurate and more time-consuming for the physicians than use of standard secretary service [3]. VRS for clinical notes in an emergency department was nearly as accurate as secretary service, more cost-effective and had a shorter turnaround time [8]. However, a randomised controlled trial in a psychiatry and an endocrinology department comparing software and standard transcription showed no increase in productivity [6].

A study comparing Dragon Nuance VRS with two other software products found that the out-of-the-box accuracy was lower for Dragon Nuance than for the other 2 products [12]. However, the study was published in 2000, and the software has been upgraded repeatedly since then. It is therefore not possible to draw valid conclusions from that study regarding the software available today.

This study is the first to evaluate the use of VRS for dictation of scientific articles. A limitation may be that the scale used for the qualitative assessment was not validated. The out-of-the-box accuracy of the evaluated VRS was not optimal. However, most of the errors were minor, did not disturb the overall meaning, and could easily be corrected afterwards. Nevertheless, most errors were misspelling of names, which may cause the readers of the text to misunderstand its meaning. The author, who has overview of all cited names, can easily correct these errors.

In our research department, we use the VRS as a standard tool to produce transcriptions of scientific articles of a high initial accuracy. However, this may require a considerable amount of training with the VRS. In our experience, this VRS program for transcription of dictated scientific articles may become completely error-free through additional training of the program. This was not tested in the present study and may be the subject for a future study. A study evaluating 1 month's use and training of VRS in a clinical department found that this was insufficient to achieve complete accuracy [5]. However, this was for clinical use and not for dictation of scientific articles, and it was unclear how the training was performed in the study. For dictation of scientific articles, VRS may be a feasible option since a secretary service is not available in many research departments. Moreover, this study evaluated the use of VRS in non-native English researchers dictating in English language. Thus, the transcript may have been more accurate if the authors had been native English speakers. Lastly, using non-validated scales, a statistically significant difference as found in the present study may not necessarily translate into a practically important difference. However, statistical testing was used as a tool to quantify the effect of the number of added words.

CONCLUSION

We found that transcription of dictated scientific articles using Dragon Nuance Premium voice recognition software had an acceptable out-of-the-box performance. Adding 15,000 and 5,000 words to the vocabulary reduced the number of errors and improved subjective assessment. Future studies may investigate the effect of additional voice training for accuracy improvement.

CORRESPONDENCE: *Hans-Christian Pommergaard*, Kirurgisk Gastroenterologisk Afdeling D, Herlev Hospital, Herlev Ringvej 75, 2730 Herlev, Denmark. E-mail: hcpommergaard@gmail.com

ACCEPTED: 10 December 2014

CONFLICTS OF INTEREST: Disclosure forms provided by the authors are available with the full text of this article at www.danmedj.dk

LITERATURE

1. Rosenberg J, Burcharth J, Pommergaard HC et al. Mind-to-paper is an effective method for scientific writing. Dan Med J 2013;60(3):A4593.
2. Spanager L, Danielsen AK, Pommergaard HC et al. A feeling of flow: exploring junior scientists' experiences with dictation of scientific articles. BMC Med Educ 2013;13:106.
3. Al-Aynati MM, Chorneyko KA. Comparison of voice-automated transcription and human transcription in generating pathology reports. Arch Pathol Lab Med 2003;127:721-5.
4. Borowitz SM. Computer-based speech recognition as an alternative to medical transcription. J Am Med Inform Assoc 2001;8:101-2.
5. Issenman RM, Jaffer IH. Use of voice recognition software in an outpatient pediatric specialty practice. Pediatrics 2004;114:e290-3.
6. Mohr DN, Turner DW, Pond GR et al. Speech recognition as a transcription aid: a randomized comparison with standard transcription. J Am Med Inform Assoc 2003;10:85-93.
7. Ramaswamy MR, Chaljub G, Esch O et al. Continuous speech recognition in MR imaging reporting: advantages, disadvantages, and impact. AJR Am J Roentgenol 2000;174:617-22.
8. Zick RG, Olsen J. Voice recognition software versus a traditional transcription service for physician charting in the ED. Am J Emerg Med 2001; 19:295-8.
9. Zheng K, Mei Q, Yang L et al. Voice-dictated versus typed-in clinician notes: linguistic properties and the potential implications on natural language processing. AMIA Annu Symp Proc 2011;2011:1630-8.
10. Chang CA, Strahan R, Jolley D. Non-clinical errors using voice recognition dictation software for radiology reports: a retrospective audit. J Digit Imaging 2011;24:724-8.
11. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159-74.
12. Devine EG, Gaehde SA, Curtis AC. Comparative evaluation of three continuous speech recognition software packages in the generation of medical reports. J Am Med Inform Assoc 2000;7:462-8.