Research and Applications

Efficiency and safety of speech recognition for documentation in the electronic health record

Tobias Hodgson, Farah Magrabi, and Enrico Coiera

Centre for Health Informatics, Australian Institute of Health Innovation, Macquarie University, North Ryde, Australia

Corresponding Author: Tobias Hodgson, Centre for Health Informatics, Australian Institute of Health Innovation, Macquarie University, L6 75 Talavera Rd, North Ryde, NSW 2109, Australia. E-mail: tobias.hodgson@hdr.mq.edu.au.

Phone: +61 2 9850 2496

Received 2 February 2017; Revised 12 June 2017; Accepted 17 June 2017

ABSTRACT

Objective: To compare the efficiency and safety of using speech recognition (SR) assisted clinical documentation within an electronic health record (EHR) system with use of keyboard and mouse (KBM).

Methods: Thirty-five emergency department clinicians undertook randomly allocated clinical documentation tasks using KBM or SR on a commercial EHR system. Tasks were simple or complex, and with or without interruption. Outcome measures included task completion times and observed errors. Errors were classed by their potential for patient harm. Error causes were classified as due to IT system/system integration, user interaction, comprehension, or as typographical. User-related errors could be by omission or commission.

Results: Mean task completion times were 18.11% slower overall when using SR compared to KBM ($P = .001$), 16.95% slower for simple tasks ($P = .050$), and 18.40% slower for complex tasks ($P = .009$). Increased errors were observed with use of SR (KBM 32, SR 138) for both simple (KBM 9, SR 75; $P < 0.001$) and complex (KBM 23, SR 63; $P < 0.001$) tasks. Interruptions did not significantly affect task completion times or error rates for either modality.

Discussion: For clinical documentation, SR was slower and increased the risk of documentation errors, including errors with the potential to cause clinical harm compared to KBM. Some of the observed increase in errors may be due to suboptimal SR to EHR integration and workflow.

Conclusion: Use of SR to drive interactive clinical documentation in the EHR requires careful evaluation. Current generation implementations may require significant development before they are safe and effective. Improving system integration and workflow, as well as SR accuracy and user-focused error correction strategies, may improve SR performance.

Key words: patient safety, electronic health record, speech recognition, documentation, medical errors

BACKGROUND AND SIGNIFICANCE

Electronic health records (EHRs) are rapidly becoming mandatory for clinical documentation worldwide, and can result in improved documentation completeness and accuracy. The use of EHRs is, however, often associated with increased documentation time for clinicians compared to use of paper records. Speech recognition (SR) systems are seen as an alternative input modality that may be faster and more acceptable to clinicians than the use of keyboard and mouse (KBM) alone.

In the clinical dictation setting, such as radiology results reporting, SR can decrease overall document turnaround time compared to transcription services. SR is also considered a viable option for clinicians who cannot touch-type or are untrained in EHR use.

SR is an increasingly common input modality across a range of consumer devices, from smartphones to refrigerators, and has...
become the standard method for documentation in specialty areas such as radiology results reporting. Given the significant time clinicians spend on the ubiquitous task of clinical documentation, it is surprising that the benefits of SR for EHR documentation and navigation have been relatively unexplored. Our recent review of the literature identified only a few studies that compared SR to dictation and transcription for clinical report creation. None explored the use of SR for general input and navigation of the EHR, and very few directly compared SR to the use of KBM, the most common input modality for EHR documentation.³

**OBJECTIVE**

The objective of this study was to compare the impact of using SR on the efficiency and safety of EHR documentation compared to KBM alone, using a common commercial EHR in a controlled experimental setting. The 2 input modalities were compared for efficiency (measured by time to complete tasks) and safety (measured by occurrence of documentation errors). The potential influence of task complexity and interruptions on documentation performance was also examined.

**MATERIALS AND METHODS**

A within-subject experimental study was undertaken with emergency department (ED) physicians, with each assigned 8 standardized clinical documentation tasks using a commercial EHR. Participants navigated the EHR and documented patient information for simulated patients. The order of task completion was allocated randomly, with half of the tasks assigned to KBM and half to SR.

The 8 documentation tasks were representative of those commonly undertaken within an EHR by ED physicians and included patient assignment, patient assessment, diagnosis, orders, and patient discharge. Tasks were chosen in consultation with senior ED clinicians, who did not further participate in the trials. All simulated patients had active records available in the experimental version of the standard ED EHR (see Supplementary Appendix A).

To allow for variation in task complexity, 4 of the 8 tasks were designed to be simple and 4 complex. Complexity was measured by the number of subtasks, with the simple tasks having 2 subtasks and complex tasks having 4. Given that interruptions are commonplace in clinical settings and can contribute to documentation error, 4 of the 8 tasks included a randomly assigned interruption condition.⁴⁻⁶ Interrupts were generated by a popup with a multiple-choice question taken from an Australasian College for Emergency Medicine fellowship exam practice set, and occurred at the same predefined stage of task completion. A similar interrupt was generated for simple and complex tasks in the KBM and SR conditions (see Figure 1).

The clinical software used for the experiment was the Cerner Millennium suite with the FirstNet ED component (v2012.01.30) and Nuance Dragon Medical 360 Network Edition UK (version 2.0, 12.51.200.072) SR software. Both were configured to replicate the operation of the EHR that participants used daily. All user actions, down to individual keystrokes, were automatically logged with recording software. Session EHR screens and audio were also separately recorded with a High-Definition Multimedia Interface capture device (see Supplementary Appendices B and C).

Thirty-five participants volunteered from 3 urban teaching hospitals in Sydney, Australia, from an eligible population of approximately 100 ED clinicians. To be eligible, subjects must have previously completed training in the EHR system, including specific SR training (EHR 4 h, SR 2 h). Clinicians were excluded if they had a pronounced speech impediment or physical disability that might affect system use.

It was estimated that a sample size of 27 clinicians would be sufficient to test for differences in time efficiency and error rates when using a t-test with a significance level of 0.05 and power of 0.95. Calculations were performed using G*Power (v3.1).⁹

Participants self-assessed their prior experience with clinical documentation using the EHR and the 2 input methods through a pre-study questionnaire. A brief exit questionnaire solicited opinions on the technologies used and their implementation within their workplace. The study was approved by the university and participating hospitals’ ethics committees. The trials took place over 2 months, commencing March 2015.

**OUTCOME MEASURES**

The efficiency of each input modality was measured by the time taken to complete a documentation task, including separate measurements for all subtasks. Task completion times excluded time spent during any interruptions. The time required to launch the dictation dialog box was also deducted, as this time was unrelated to either user behavior or SR time but was associated with the way in which the SR system was locally integrated with the EHR.

The safety of documentation performance was assessed by the number of errors observed. Each observed error was assigned a label set.

1. Potential for Patient Harm (PPH): This was an assessment of risk that an error had a major, moderate, or minor impact on patient outcomes, based on the scale within the US Food and Drug Administration 2005 guidance document.¹⁰

2. Error Type: The nature of the error was separated into 3 classes: (A) integration/system: associated with technology (including software, software integration, and hardware); (B) user: associated with user action; and (C) comprehension: related to comprehension (eg, user adds words to or omits words from the prescribed task). Errors could be assigned to > 1 class within this label set.

3. User Error Type: Where user errors occurred, they were assigned 1 of 2 additional labels: (A) omission: errors that occurred when

---

**Figure 1.** Experiment conceptual design.
a participant failed to complete an assigned task, and (B) commission: errors that occurred when participants incorrectly executed an assigned task.

The labels for Error Type were not mutually exclusive, and some errors could have multiple labels assigned. Minor typographical errors, such as missing full stops or incorrect capitalization, were treated as a discrete category, as they had no potential for harm and could not be easily assigned a Type category.

Definitions for these error classes are contained in Supplementary Appendix E. The process for allocating observed errors to class and the inter-rater agreement are presented in Supplementary Appendix D. Inter-rater agreement was calculated using Cohen’s kappa.11 Initial observed agreement was robust across all three categories: (1) 81.48% agreement, $\kappa = 0.694$, $P < .001$, 95% confidence interval [CI], 0.476–0.912. (2) 88.89% agreement, $\kappa = 0.800$, $P < .001$, 95% CI, 0.590–1.010. (3) 85.19% agreement, $\kappa = 0.778$, $P < .001$, 95% CI, 0.584–0.971. All discrepancies between the reviewers were resolved through discussion.

Statistical comparisons were made for efficiency and safety outcome variables on equivalent tasks using both KBM and SR, ie, simple or complex tasks. Aggregate data across all task types were reported, but heterogeneity in task type precluded statistical testing. Since the study data do not follow normal distribution, only nonparametric statistical tests, including Wilcoxon signed rank and Mann-Whitney tests, were undertaken using IBM SPSS Statistics v24.0.0.0. Since the study data do not follow normal distribution, only nonparametric statistical tests, including Wilcoxon signed rank and Mann-Whitney tests, were undertaken using IBM SPSS Statistics v24.0.0.0.89

RESULTS

There were 19 female and 16 male participants, who identified themselves predominantly as senior clinicians (24/35). Most rated their SR skills lower than their EHR skill level, with a mean of 3.9/5 for KBM and 2.7/5 for SR (on a scale from 1 = very poor to 5 = excellent). Many participants had <1 year of SR experience (19/35) while few had <1 year EHR experience (7/35).

Efficiency

Across all 4 experimental conditions (simple task KBM, simple task SR, complex task KBM, and complex task SR), there was no association found between task completion time and clinical role ($P = .100–.588$), skill level ($P = .073–.452$), or period of experience with the EHR or SR ($P = .303–.717$). Interruptions had no effect on task completion times for either input modality across both task types (see Supplementary Appendix F).

Overall, complex tasks took 52.82% longer to complete than simple tasks (mean: complex 185.63 s, simple 121.47 s; $P < .001$; 95% CI, 50.55–75.34). Clinical documentation took significantly longer to complete using SR compared to KBM, with mean time to complete all tasks 18.11% longer when using SR (KBM 140.09, SR 165.46; $P < .001$; 95% CI, 9.87–33.91). This significant difference in mean task completion time held for both simple (KBM 112.38, SR 131.44; $P = .005$; 95% CI, −0.07, 30.50) and complex (KBM 170.48, SR 201.84; $P = .009$; 95% CI, 9.61–47.73) tasks (see Table 1 and Figure 2).

Safety

A total of 170 errors was observed, with a significant difference in the number of errors (excluding typographical errors) via input modality (KBM 32, SR 138) (see Figure 3). This increase in errors observed via SR held for both simple (KBM 9, SR 75; $P < .001$; 95% CI, 1.50–2.50) and complex (KBM 23, SR 63; $P < .001$; 95% CI, 0.50–1.50) tasks. There were many typographical errors observed with both input modalities (KBM 213, SR 252), and no significant difference in typographical error frequency by input modality for either simple (KBM 142, SR 133; $P = .345$; 95% CI, −1.00, 0.50) or complex (KBM 71, SR 119; $P = .600$; 95% CI, −0.50, 0.50) tasks was found.

There was no association found between number of errors observed and participant clinical role, skill, or period of experience with the EHR or SR. The introduction of interruptions made no statistically significant difference in the number of errors for complex tasks and simple tasks using SR. However, there was a significant difference in errors observed for simple tasks using KBM (KBM 9, KBM 7; $P = .18$; 95% CI, 0.00–1.00).

Potential for patient harm

There were significant increases in the occurrence of all classes of PPH errors when using SR across both task types: major PPH simple task (KBM 2, SR 29; $P < .001$; 95% CI, 0.50–1.00), and complex task (KBM 11, SR 21; $P = .005$; 95% CI, 0.00–0.50), moderate PPH simple task (KBM 0, SR 13; $P = .008$; 95% CI, 0.00–0.50), minor PPH simple task (KBM 7, SR 33; $P = .002$; 95% CI, 0.50–1.00), and complex task (KBM 9, SR 34; $P < .001$; 95% CI, 1.50–2.50). The exception was moderate PPH errors during complex tasks, where there was no significant difference (KBM 3, SR 8; $P = .083$; 95% CI, 0.00–0.00) (see Table 2).

Error type

There were significant increases in the occurrence of integration/system, user, and comprehension errors while using SR across both task types: integration/system errors simple task (KBM 0, SR 56; $P < .001$; 95% CI, 1.00–2.00) and complex task (KBM 2, SR 36; $P < .001$; 95% CI, 1.00–2.50), user errors simple task (KBM 5, SR 18; $P = .007$; 95% CI, 0.00–0.50) and complex task (KBM 17, SR 26; $P = .002$; 95% CI, 0.00–0.50), and comprehension errors complex task (KBM 4, SR 17; $P < .001$; 95% CI, 1.00–1.50), with the exception of comprehension errors during simple tasks, where no significant difference was found (KBM 4 SR 5; $P = .317$; 95% CI, 0.00–0.00) (see Table 2).

User error type

There were significant increases in the occurrence of omission errors via SR across both task types: simple task (KBM 2, SR 14; $P = .001$; 95% CI, 0.00–0.50) and complex task (KBM 6, SR 6; $P = .005$; 95% CI, 0.00–0.50). However, there were no significant differences in the occurrence of commission errors via either input modality: simple task (KBM 7, SR 5; $P = .719$; 95% CI, 0.00–0.00) and complex task (KBM 15, SR 21; $P = .201$; 95% CI, 0.00–0.50) (see Table 2).

DISCUSSION

This is the first study, to our knowledge, that compares the impact of input modality on the safety and efficiency of clinical documentation using an EHR in a controlled experimental environment. The study’s results show statistically significant differences in outcomes when using KBM compared to SR, with SR created records taking longer and associated with a higher error rate. While these results do not appear to be influenced by the degree of EHR or SR experience,
It is possible that results might be different with changes in task type, participants, setting, or technology. For the tasks undertaken in this study, creating EHR clinical documentation with the assistance of SR was significantly slower (18.11%) than using KBM alone, and this difference held for both simple (16.95%) and complex (18.40%) documentation tasks.

These results are different from those of a recent observational study that found no difference in charting time when a voice driven EHR at one site was compared to a KBM driven EHR at another. However, the study sample size was too small to determine if this result was statistically significant, and the EHRs at these sites were different and may themselves have accounted for the time difference.

### Table 1. Efficiency summary

<table>
<thead>
<tr>
<th>Task Type</th>
<th>N</th>
<th>Mean task completion time (s)</th>
<th>Max task completion time (s)</th>
<th>Min task completion time (s)</th>
<th>Difference in completion time SR vs KBM (%)</th>
<th>Z</th>
<th>Wilcoxon P-value</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>KBM – simple and complex</td>
<td>65</td>
<td>140.09</td>
<td>257.63</td>
<td>49.38</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SR – simple and complex</td>
<td>60</td>
<td>165.46</td>
<td>327.45</td>
<td>71.67</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total KBM vs total SR</td>
<td>58</td>
<td>112.38</td>
<td>214.99</td>
<td>49.38</td>
<td>118.11</td>
<td>−3.302</td>
<td>0.001</td>
<td>9.87, 33.91</td>
</tr>
<tr>
<td>Simple task KBM</td>
<td>34</td>
<td>131.44</td>
<td>285.91</td>
<td>71.67</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Simple task SR</td>
<td>31</td>
<td>170.48</td>
<td>257.63</td>
<td>110.31</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Complex task KBM</td>
<td>31</td>
<td>170.48</td>
<td>257.63</td>
<td>110.31</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Complex task SR</td>
<td>29</td>
<td>201.84</td>
<td>327.45</td>
<td>138.87</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total KBM vs total SR</td>
<td>58</td>
<td>118.11</td>
<td>214.99</td>
<td>71.67</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tasks with interruptions</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Simple task KBM with interrupt</td>
<td>34</td>
<td>112.22</td>
<td>257.63</td>
<td>49.38</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Simple task SR with interrupt</td>
<td>32</td>
<td>134.41</td>
<td>250.20</td>
<td>67.69</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Complex task KBM with interrupt</td>
<td>31</td>
<td>186.65</td>
<td>355.90</td>
<td>102.94</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Complex task SR with interrupt</td>
<td>29</td>
<td>216.45</td>
<td>345.70</td>
<td>133.13</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 2. Boxplot of task completion time for simple and complex tasks via input modality.

Figure 3. Error framework: overview of the breakdown of errors by class.
any conclusions from this study, and comparison with the controlled experiment reported here is not possible.

Interestingly, task interruptions did not significantly increase task completion times in our study. This is consistent with another controlled experiment, where interruptions did not affect electronic prescribing tasks.13 A previous observational study on the impact of interruptions on ED physicians' task time showed a reduction in task time following interruptions, possibly because physicians hurry to complete tasks after being interrupted when they are under time pressure.14 Our results suggest that documentation tasks might not be easily time compressed, and that interruption penalties might therefore be incurred with other tasks associated with patient care.

In this study, more errors occurred when tasks were completed using SR than KBM alone, and this held for both simple and complex tasks.

The occurrence of serious errors with major PPH was significantly greater using SR (simple task: KBM 2, SR 29; complex task: KBM 11, SR 21) (Table 3). Any increase in these errors that could result in death or serious injury to patients requires further examination.

Far more integration/system errors (such as network transmission delays causing SR commands to either not complete or go to the incorrect location in a chart or the wrong chart) occurred while tasks were being undertaken with SR. These included some highly potential patient harm errors, such as mistyping of words (eg, “follow” instead of “FLU,” “fractured” instead of “ruptured”). It is possible that some of these integration errors could be avoided through a redesign or reconfiguration of inter-system communication methods.

Documentation error rates were high independent of input modality, and if replicated in real-world situations would raise questions about the acceptable baseline level of errors for an EHR system. Recent real-world studies have indeed shown that clinical documentation accuracy can be poor, with one study finding accurate documentation rates of 54.4% via paper and 58.4% with

### Table 2. Error summary

<table>
<thead>
<tr>
<th>Error Type</th>
<th>KBM</th>
<th>SR</th>
<th>Wilcoxon Z</th>
<th>P-value</th>
<th>95% CI</th>
<th>KBM vs KBM Z</th>
<th>P-value</th>
<th>95% CI</th>
<th>SR vs SR Z</th>
<th>P-value</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total errors observed</td>
<td>245</td>
<td>390</td>
<td>-4.370</td>
<td>0.000</td>
<td>1.50, 2.50</td>
<td>-2.373</td>
<td>0.018</td>
<td>0.00, 1.00</td>
<td>-1.115</td>
<td>0.265</td>
<td>-0.50, 1.50</td>
</tr>
<tr>
<td>Non-typographical</td>
<td>32</td>
<td>138</td>
<td>-4.085</td>
<td>0.000</td>
<td>0.50, 1.50</td>
<td>-0.294</td>
<td>0.769</td>
<td>0.00, 0.50</td>
<td>-1.086</td>
<td>0.278</td>
<td>-0.50, 1.00</td>
</tr>
<tr>
<td>Potential patient harm</td>
<td>32</td>
<td>138</td>
<td>-2.799</td>
<td>0.005</td>
<td>0.00, 0.50</td>
<td>-1.425</td>
<td>0.154</td>
<td>0.00, 0.50</td>
<td>-0.317</td>
<td>0.751</td>
<td>-0.50, 0.50</td>
</tr>
<tr>
<td>Major</td>
<td>13</td>
<td>50</td>
<td>-2.636</td>
<td>0.008</td>
<td>0.00, 0.50</td>
<td>-1.732</td>
<td>0.083</td>
<td>0.00, 0.50</td>
<td>-0.474</td>
<td>0.635</td>
<td>0.00, 0.00</td>
</tr>
<tr>
<td>Moderate</td>
<td>3</td>
<td>21</td>
<td>-1.732</td>
<td>0.083</td>
<td>0.00, 0.50</td>
<td>-1.732</td>
<td>0.083</td>
<td>0.00, 0.50</td>
<td>-2.648</td>
<td>0.008</td>
<td>0.00, 0.50</td>
</tr>
<tr>
<td>Minor</td>
<td>16</td>
<td>67</td>
<td>-3.101</td>
<td>0.002</td>
<td>0.50, 1.00</td>
<td>-0.686</td>
<td>0.493</td>
<td>0.00, 0.00</td>
<td>-0.198</td>
<td>0.843</td>
<td>-0.50, 0.50</td>
</tr>
<tr>
<td>Error type</td>
<td>32</td>
<td>158</td>
<td>-4.637</td>
<td>0.000</td>
<td>1.50, 2.50</td>
<td>-1.809</td>
<td>0.070</td>
<td>0.00, 0.50</td>
<td>-2.532</td>
<td>0.011</td>
<td>-1.50, 0.50</td>
</tr>
<tr>
<td>Integration/system</td>
<td>2</td>
<td>92</td>
<td>-4.598</td>
<td>0.000</td>
<td>1.00, 2.00</td>
<td>-1.414</td>
<td>0.157</td>
<td>0.00, 0.00</td>
<td>-1.792</td>
<td>0.073</td>
<td>-1.00, 0.00</td>
</tr>
<tr>
<td>User errors</td>
<td>22</td>
<td>44</td>
<td>-4.460</td>
<td>0.000</td>
<td>1.00, 2.50</td>
<td>-0.577</td>
<td>0.564</td>
<td>0.00, 0.00</td>
<td>-1.008</td>
<td>0.314</td>
<td>-1.00, 0.50</td>
</tr>
<tr>
<td>Omission</td>
<td>8</td>
<td>20</td>
<td>-3.207</td>
<td>0.001</td>
<td>0.50, 0.00</td>
<td>-1.265</td>
<td>0.206</td>
<td>0.00, 0.00</td>
<td>-2.138</td>
<td>0.033</td>
<td>-0.50, 0.00</td>
</tr>
<tr>
<td>Commission</td>
<td>22</td>
<td>26</td>
<td>-2.840</td>
<td>0.005</td>
<td>0.00, 0.50</td>
<td>-0.632</td>
<td>0.527</td>
<td>0.00, 0.00</td>
<td>-2.500</td>
<td>0.012</td>
<td>-0.50, 0.00</td>
</tr>
<tr>
<td>Typographical</td>
<td>213</td>
<td>252</td>
<td>-0.252</td>
<td>0.600</td>
<td>-0.50, 0.50</td>
<td>-4.511</td>
<td>0.000</td>
<td>-3.00, 1.50</td>
<td>0.303</td>
<td>0.303</td>
<td>-0.50, 1.00</td>
</tr>
</tbody>
</table>

Bold values indicate general totals and italic bold corresponds to category totals.
EHRs. Our results thus support calls to improve the quality of clinical documentation independent of modality of input.

There continue to be rapid changes in SR technologies; a recent study reports SR systems that now exceed human performance for conversational speech. One recent study examined the potential for EHR transcription using SR together with natural language processing and found potential time and usability benefits for electronic documentation. These results are in accord with our recent systematic review that identified benefits (and risks) of using SR for clinical documentation.

SR as a use model is typically associated with dictation tasks where users are simply creating large volumes of text. Dictation time benefits may be realized in part because the user interaction and screen navigation tasks associated with electronic documentation are not present, unlike the real-world use cases within this study.

SR systems have carved out a role in specific areas of clinical documentation, such as report dictation in radiology, but there has been little attention paid to the potential risks and benefits of using SR as a primary mode of data entry and system navigation in the EHR. The results of this study, which was designed to replicate real world tasks and used a working clinical system with real clinical users, give some pause for thought. SR appears to reduce efficiency by taking more time than KBM, and also appears to increase the rates of error, both minor and serious.

Some of the errors and inefficiencies appear to be inherent in current-generation SR engines and their fit to the task of clinical documentation within EHRs. EHRs are designed with KBM input in mind, and simply expecting SR to directly replace KBM in such a design may be inappropriate. Designing EHRs that are purpose-built for SR may yield different workflows and design constructs, and consequently different results. It is possible that at least some of these weaknesses with SR can be mitigated with system redesign.

The use of SR to drive interactive clinical documentation in the EHR thus requires careful evaluation, and it cannot be assumed that it is a safe and effective replacement for KBM in this particular setting. Current generation implementations may require significant changes before they are considered safe and effective. Improving system integration and workflow, as well as further developing SR accuracy and enhancing the capacity of user interface design to allow clinicians to detect and correct errors, may improve SR performance.

**LIMITATIONS**

Several factors could hinder the results of this study being generalized to other clinical settings or information systems. The study used a routine and standardized version of a widely used commercial clinical record system for EDs, integrated with a common commercial clinical SR system. However, other EHR and SR systems might differ in their individual performance, and different approaches to integrating the two could also lead to varying results.

SR performance may be affected by extrinsic factors, such as microphone quality, background noise level, and user accent. Equally, the tasks created for this study were intended to be representative of typical clinical documentation work in an ED, but different tasks in other settings could yield different outcomes. For example, dictation of investigation reports at high volume by expert clinicians might yield better time performance and recognition rates, although our previous review did not identify this.

The lack of any impact of interruptions should be generalized only cautiously, given that the impact of interruptions is a complex

<table>
<thead>
<tr>
<th>Table 3. Major potential patient harm errors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Error observed</td>
</tr>
<tr>
<td>Simple task KBM</td>
</tr>
<tr>
<td>Complex task KBM</td>
</tr>
<tr>
<td>Simple task SR</td>
</tr>
<tr>
<td>Complex task SR</td>
</tr>
</tbody>
</table>
phenomenon, and many different variables must interact for inter-
ruptions to have a negative effect on performance.

While the study was not specifically designed to test for the
effects of clinical role, skill, or period of experience on outcome vari-
ables and was underpowered to test for these effects individually, it
was possible to demonstrate a significant absence of correlation
when all SR and KBM tasks were pooled. While there were no statist-
ically significant differences in performance observed based on ex-
perience, skill, or seniority, the fact that participants volunteered for
this study may have introduced a recruitment bias. All participants
had completed training in the use of both the EHR and SR, and it is
quite possible that results would be different, and potentially worse,
if users had no such training.

CONCLUSION
The search for the safest and most efficient method of clinical docu-
mentation remains a work in progress. Emerging technologies such as
mobile devices, touch pads, pen control, head-mounted displays, vir-
tual reality, and wearable technology all have different affordances and
are likely to work better in some contexts and provide varying results
over a wide range of tasks. Given the ubiquitous nature of electronic
records in healthcare, and the substantial cost of documentation in
terms of clinician time and patient safety, the foundational act of inter-
acting with an electronic record requires far closer attention, and sub-
stantially higher research priority, than it currently receives.

ACKNOWLEDGMENTS
David Lyell performed the second reviewer functions for error classification.
Thanks are offered to the staff and management teams at participating hospi-
tals, in particular Dr Richard Paoloni and Dr Marty Sterrett, who both
assisted greatly with study design and trial session coordination.

FUNDING
This work was supported by the National Health and Medical Research
Council Centre for Research Excellence in eHealth (APP1032664).

COMPETING INTERESTS
The authors declare that they have no competing interests.

CONTRIBUTORS
TH, EC, and FM conceived the study and its design. TH conducted
the research, the primary analysis, and the initial drafting of the paper.
EC and FM contributed to the analysis and drafting of the paper, and TH, EC, and FM approved the final manuscript. TH is
the corresponding author.

SUPPLEMENTARY MATERIAL
Supplementary material is available at Journal of the American Medical Informatics Association online.

REFERENCES
implementation in ophthalmology at an academic medical center
2013;111:70
3. Hodgson T, Coiera E. Risks and benefits of speech recognition for clinical
documentation: a systematic review. J Am Med Inform Assoc.
2016;23(e1):e169–e79.
4. Shea S, Hrpcsak G. Accelerating the use of electronic health records in
5. Communications N. Speech Recognition Accelerating the Adoption of
Electronic Health Records (white paper). Nuance Communications;
2012.
6. Wolpin S. Samsung’s new refrigerator is also a giant tablet. 2016. http://
www.techlicious.com/blog/samsung-refrigerator-family-hub-touchscreen/.
7. Coiera EW, Jayasuriya RA, Hardy J, Rannan A, Thorpe MEC. Communi-
cation loads on clinical staff in the emergency department. Med J Australia.
2002;176(9):415–18.
tical power analysis program for the social, behavioral, and biomedical
10. Food U, Administration D. Guidance for the content of premarket submis-
sions for software contained in medical devices. Center Devices Radiol Health.
2005.
11. Cohen J. Weighted kappa: Nominal scale agreement provision for scaled
disagreement or partial credit. Psychol Bull. 1968;70(4):213.
12. dela Cruz JE, Shabosky JC, Albrect M, et al. Typed versus voice recogni-
tion for data entry in electronic health records: emergency physician time
13. Magrabi F, Li SY, Day RO, Coiera E. Errors and electronic prescribing: a
controlled laboratory study to examine task complexity and interruption
examination findings in initial progress notes between paper charts and a
2017;24(1):140–44.
16. Xiong W, Droppo J, Huang X, et al. Achieving Human Parity in Conver-
17. Kaufman DR, Sheehan B, Stetson P, et al. Natural language processing-
enabled and conventional data capture methods for input to electronic health