Evaluating the Efficiency and Safety of Speech Recognition within a Commercial Electronic Health Record System: A Replication Study

Tobias Hodgson[1] Farah Magrabi[1] Enrico Coiera[1]

[1] Centre for Health Informatics, Australian Institute of Health Innovation, Macquarie University, New South Wales, Australia

Appl Clin Inform 2018;9:326–335.

Address for correspondence Tobias Hodgson, BSc, MMgt, MBA, Centre for Health Informatics, Australian Institute of Health Innovation, Macquarie University, L6 75 Talavera Road, North Ryde, NSW 2109, Australia (e-mail: tobias.hodgson@hdr.mq.edu.au).

Abstract

Objective To conduct a replication study to validate previously identified significant risks and inefficiencies associated with the use of speech recognition (SR) for documentation within an electronic health record (EHR) system.

Methods Thirty-five emergency department clinicians undertook randomly allocated clinical documentation tasks using keyboard and mouse (KBM) or SR using a commercial EHR system. The experiment design, setting, and tasks (E2) replicated an earlier study (E1), while technical integration issues that may have led to poorer SR performance were addressed.

Results Complex tasks were significantly slower to complete using SR (16.94%) than KBM (KBM: 191.9 s, SR: 224.4 s; $p = 0.009$; CI, 11.9–48.3), replicating task completion times observed in the earlier experiment. Errors (non-typographical) were significantly higher with SR compared with KBM for both simple (KBM: 3, SR: 84; $p < 0.001$; CI, 1.5–2.5) and complex tasks (KBM: 23, SR: 53; $p = 0.001$; CI, 0.5–1.0), again replicating earlier results (E1: 170, E2: 163; $p = 0.660$; CI, 0.0–0.0). Typographical errors were reduced significantly in the new study (E1: 465, E2: 150; $p < 0.001$; CI, 2.0–3.0).

Discussion The results of this study replicate those reported earlier. The use of SR for clinical documentation within an EHR system appears to be consistently associated with decreased time efficiencies and increased errors. Modifications implemented to optimize SR integration in the EHR seem to have resulted in minor improvements that did not fundamentally change overall results.

Conclusion This replication study adds further evidence for the poor performance of SR-assisted clinical documentation within an EHR. Replication studies remain rare in informatics literature, especially where study results are unexpected or have significant implication; such studies are clearly needed to avoid overdependence on the results of a single study.

Keywords

- electronic health record
- speech recognition
- integration
- medical errors
- patient safety

Background and Significance

Speech recognition (SR) is a relatively mature modality for interaction with information technology and is regularly used in many healthcare settings. When used for dictation tasks such as reporting radiology or pathology results, SR can improve overall process efficiency.[1] When used to interact with an electronic health record (EHR), emerging evidence suggests that SR use is associated with significant patient safety risks and time penalties.[2] Given the well-reported benefits of SR for dictation in general, these results are perhaps surprising and raise concerns for the safety and efficiency of using SR for EHR documentation tasks.

However, as with all research, such results need to be treated cautiously, given the many limitations of research methods. While statistical testing provides a measure of the

received
December 24, 2017
accepted after revision
March 24, 2018

DOI https://doi.org/10.1055/s-0038-1649509.
ISSN 1869-0327.

likelihood that results could have arisen just by chance, it does not provide certainty. Other studies using similar methods might arrive at different results. The replication of existing studies, especially when they produce unexpected results that can have real-world implications, is thus crucial.

In many research fields, there is currently a "crisis of replicability" where the inability to replicate existing studies by either original or subsequent researchers is calling major research results into question. For studies in psychology, the phenomenon has led to a collaborative replication of 100 existing published experiments. This replication effort has resulted in the conclusion that a "large portion of replications produced weaker evidence for the original findings despite using materials provided by the original authors."[3] This "crisis" also extends to the medical sciences with one recent study suggesting that "irreproducible preclinical research exceeds 50%" of all studies.[4]

Replication studies take many forms, depending on the purpose of replication.[5] When the *validity* of a study is in question, high fidelity replications can provide evidence that the results of a specific protocol are correct. At the other end of the spectrum, replications can be undertaken that test *generalizability* by exploring different experimental settings, protocols, and indeed interventions, while still sharing underlying hypotheses with an original study. Typically, replications that test validity will be untaken before moving on to replications that test generalizability.[6]

This article reports on a replication study that tests the validity of the poor efficiency and safety performance reported when using SR for clinical documentation tasks in the EHR. Specifically, our recent controlled study comparing SR to keyboard and mouse (KBM) found significant SR risks arising from more frequent and potentially harmful data entry errors, as well as a significant increase in documentation time.[2] In that study, one identified limitation of the experimental setup was that different results might have arisen with better technical integration between SR and EHR.

Therefore, for this study, a series of modifications were made to optimize both workflow and technical integration of the EHR and SR systems. In all other respects, the original study design was replicated as far as possible including setting, users, and tasks. This type of study, which is known as a *partial replication study*, tests the validity of an earlier study by directly addressing identified experimental limitations that may have impaired a fair comparison between SR and KBM.[5]

Methods

A within-subject experimental study, was undertaken with 35 emergency department (ED) physicians, replicating the methods of a previous experiment (Experiment 1). Each participant was assigned standardized clinical documentation tasks requiring the use of a commercial EHR, to be completed using either KBM or with the assistance of SR. The order of task completion was allocated randomly, with half of the tasks assigned to SR and half to KBM.

Tasks performed during the experiment were representative of clinical documentation duties performed daily by ED clinicians and included patient assignment, patient assessment, viewing vital signs, performing diagnosis, creating orders, and patient discharge. Each participant undertook four tasks, a simple task and a complex task, performed via both input modalities (KBM and SR). Task complexity was measured by the number of subtasks, simple tasks with two and complex tasks with four subtasks (see ►**Supplementary Material, Appendix B** [available in the online version]).

Example Tasks

Simple task:

1. Assign yourself as the patient's provider.
2. Perform an ED assessment on the patient.

Complex task:

1. View patient's vital signs and note latest blood glucose level (BGL).
2. Add a diagnosis for the patient.
3. Add an order for the patient.
4. Create a discharge note for the patient.

The methods used for this experiment (Experiment 2) were the same as those within Experiment 1 (see ►**Supplementary Material, Appendix A**, available in the online version) with the exception of the following modifications[2]:

1. The number of tasks was reduced from eight to four by eliminating cases in which an external interruption occurred (see ►**Supplementary Material, Appendix B**, available in the online version).
2. A pre-trial demographic survey and a post-trial opinion survey were eliminated (see ►**Fig. 1**).
3. The versions of EHR and SR software were updated to the latest versions available. The EHR to Cerner Millennium suite with the FirstNet ED component (v2015.01.11) and Nuance Dragon Medical 360 Network Edition (UK) (v2.4.2) speech recognition.
4. An updated high-definition multimedia interface capture device was utilized to record participant sessions (Elgato Game Capture HD60; (see **Supplementary Material Appendices C** and **D** [available in the online version]; ►**Fig. 1**).

Thirty-five participants volunteered from four 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 hours, SR: 2 hours). Clinicians were excluded if they had a pronounced speech impediment or a 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).

The study was approved by the university and participating hospitals' ethics committees. The trials took place over 2 months, commencing May 2016 (initial study March 2015).

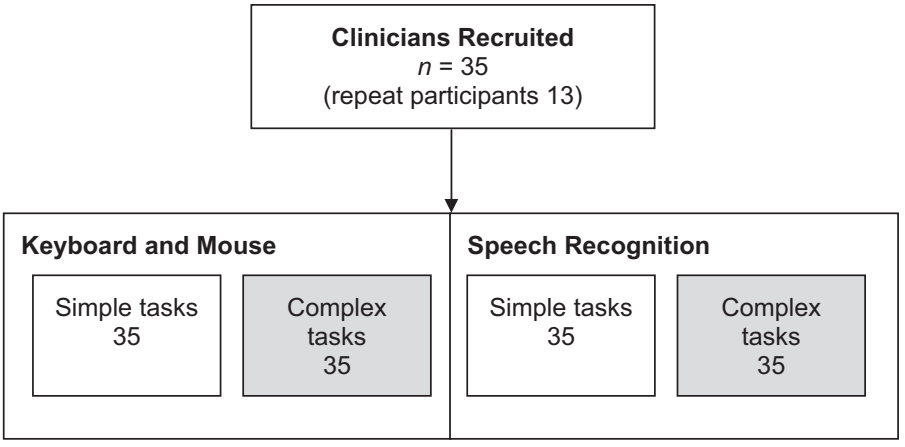

Fig. 1 Experimental conceptual design.

Optimizing the Integration of Speech Recognition into the Electronic Record

One of the limitations of the original study was that some SR errors and time delays could have been attributed to system configuration factors influencing SR performance. As a consequence, it was possible that SR might have performed significantly better had these issues been addressed. Therefore, a review of the technical setup used for Experiment 1 was conducted to identify any factors that might have biased the study results by hampering the performance of SR. The analysis extended from low-level network integration through to user workflows and interaction.

To assist identification of such potential factors, an analysis of all the identified issues and errors in Experiment 1 was conducted to determine if a system issue might have contributed to the error. Additionally, the human–computer interaction framework of activity theory (AT) and the AT checklist by Kaptelinin et al were used to assist in the identification of problems with user-interaction design.[7,8]

A total of 33 issues were identified (see ►**Supplementary Material Appendix E**, available in the online version). Similar issues were grouped into one of seven categories: command reliability, system stability, patient safety, workflow usability, quality of data, typographical issues, or recognition and documentation issues. A series of revisions were made to the EHR and SR systems to address these issues (see ►**Tables 1** and **2**).

Outcome Measures

The efficiency of KBM and SR was measured by the time taken to complete each assigned task, with separate measurements for any subtasks.

Table 1 Summary of fixes implemented prior to Experiment 2

Change name	Description	Expected benefit
1. Command changes	a. SR commands were modified to be able to be run from any chart/section of the EHR b. Delays or wait periods within steps of commands were adjusted to better suit the specifics of this implementation	a. Clinicians will not need to be at the correct chart or location to call a command, the command itself will ensure (or move to) the correct chart/location b. The commands will become far more robust with reduced execution of command sequences
2. Domain change	An alternative, more reliable network domain was used to host the EHR system	The EHR system should be more robust, reducing or removing the occurrences of network related system lag or crashes
3. Integration modifications	System integration between the EHR and SR systems was revised to better facilitate interaction between the local SR and Citrix session. EHR—vSync Citrix integration was utilized	The resolution of numerous technical issues due to the local to Citrix session should be resolved. These include system lag, errors, and stability
4. Software option enabled	Spell check was enabled within all elements of the EHR system	Various typographical errors would be highlighted and/or automatically addressed independent of input modality
5. Revised software	Latest versions of the EHR and SR systems were implemented. Updated software including patches and fixes	Numerous bugs fixed leading to improvements in performance, operation, integration, and system robustness

Abbreviations: EHR, electronic health record; SR, speech recognition.

Table 2 Summary of issues addressed prior to Experiment 2

Desired improvement	Errors and issues to be addressed	Implemented solution(s)
Command reliability	All elements of a command did not complete Navigational command went nowhere or to wrong place/chart	1. Command changes 2. Domain change 3. Integration modifications
System stability	EHR slow—system lag EHR crashed Element(s) of EHR down	2. Domain change 3. Integration modifications
Patient safety	Incorrect patient Incorrect patient—user corrected No BGL entered Incorrect BGL entered Incorrect order collection date entered Incorrect order collection method selected Data entered in incorrect EHR field Section of EHR missed	1. Command changes 5. Revised software
Workflow usability	Clinician closed EHR Incorrect method of EHR menu navigation used	1. Command changes 5. Revised software
Quality of data	Incorrect diagnostic word entered Incorrect trivial word entered Incorrect trivial word entered—user corrected Close chart after task step missed Incorrect unimportant word entered Plural form error "s" Additional word(s) capitalization Missing comma(s) Template brackets not removed	1. Command changes 3. Integration modifications 4. Software option enabled 5. Revised software
Typographical issues	Spelling error(s) Missing full stop(s) Missing word capitalization	3. Integration modifications 4. Software option enabled
Recognition and documentation	Additional unnecessary word(s); e.g., "and" Omitted unnecessary word(s); e.g., "is" Omitted diagnostic word Miss recognition of word(s) by SR Miss recognition of word(s) by SR—user corrected Hyphen error Word mangled Word mangled—user corrected	3. Integration modifications 4. Software option enabled 5. Revised software

Abbreviations: BGL, blood glucose level; EHR, electronic health record; SR, speech recognition.

The safety of documentation performance was assessed by the number of errors observed. Each observed error was assigned labels in three categories (see ►**Supplementary Material Appendix F**, available in the online version).

1. *Potential for patient harm (PPH):* The risk that an error had a major, moderate, or minor impact on patient outcomes based on the scale within the U.S. Department of Health and Human Services Food and Drug Administration 2005 guidance document.[9]
2. *Error type:* The nature of the error was separated into three classes: (a) Integration/System: associated with technology (including software, software integration, and hardware), (b) User: user action–related errors, and (c) Comprehension: errors related to comprehension (e.g., user adds or omits words to the prescribed task). Errors could be assigned to more than one class within this label set.
3. *Use error type:* Where use errors occurred, they were assigned one of two additional labels— (a) Omission: errors occurred when a subject failed to complete an assigned task and (b) Commission: errors occurred when subjects incorrectly executed an assigned task.

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

Statistical comparisons were made for efficiency and safety outcome variables on equivalent tasks using both KBM and SR, and between the outcomes of Experiment 1 (E1) and Experiment 2 (E2). 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 were undertaken, including Wilcoxon's signed-rank test, Mann–Whitney tests, and chi-square tests, using IBM SPSS Statistics (v24.0.0.0) and Minitab

17 (v17.3.1) statistics packages. For statistical tests that ranked paired observations, comparisons were only possible where values for both input modalities were available. In cases where a task had no value for one input modality (such as a missed or incomplete task), the pair was excluded.

Results

The ratio of male to female participants was similar for both rounds of experiments (male E1: 16, E2: 18), (female E1: 19, E2: 17). Thirteen participants were involved in both experiments (7 females and 6 males).

Documentation Efficiency

No difference in mean completion time for simple tasks was observed (KBM: 126.4 s, SR: 126.8 s; $p = 0.701$; CI, 6.7–13.2). Complex tasks, however, were significantly slower when completed with SR when compared with KBM (KBM: 191.9 s, SR: 224.4 s; $p = 0.009$; CI, 11.9–48.3). Complex tasks took significantly longer than simple tasks overall (complex: 185.6 s, simple: 121.5 s; $p < 0.001$; CI, −86.2 to 53.6; see ►Table 3).

Comparing these results with those of our earlier experiment, there were no statistical differences in mean task completion times observed for any of the four individual task types: simple tasks via KBM (E1: 112.4 s, E2: 126.4 s; $p = 0.060$; CI, −26.0 to 0.4), simple task via SR (E1: 131.4 s, E2: 126.8 s; $p = 0.646$; CI, −17.3 to 10.9), complex task via KBM (E1: 170.5 s, E2: 191.9 s; $p = 0.199$; CI, −39.0 to 7.4), and complex task via SR (E1: 201.8 s, E2: 224.4 s; $p = 0.230$; CI, −42.9 to 10.0; see ►Table 3; ►Fig. 2).

Comparing the performance of subjects who participated in both the current and the earlier experiments, no difference in mean task completion times was observed: simple tasks via KBM (E1: 120.4 s, E2: 118.1 s; $p = 0.308$; CI, −20.4 to 11.9), simple task via SR (E1: 129.3 s, E2: 129.5 s; $p = 0.286$; CI, −45.3 to 26.7), complex task via KBM (E1: 179.4 s, E2: 159.7 s; $p = 0.059$; CI, −37.30 to 2.66), and complex task via SR (E1: 205.1 s, E2: 208.4 s; $p = 0.814$; CI, −31.6 to 61.6).

Documentation Safety

Significant differences in the number of errors were observed between KBM and SR for the following classes (►Table 4):

- Major PPH errors with simple task (KBM: 1, SR: 35; $p < 0.001$; CI, 1.0–1.0).
- Minor PPH errors with simple task (KBM: 1, SR: 42; $p < 0.001$; CI, 0.5–1.0), and complex task (KBM: 10, SR: 44; $p < 0.001$; CI, 0.5–1.5).
- Integration/system errors with simple task (KBM: 0, SR: 53; $p < 0.001$; CI, 1.0–1.5), and complex task (KBM: 1, SR: 45; $p < 0.001$; CI, 1.0–1.5).
- Use errors with simple task (KBM: 3, SR: 28; $p < 0.001$; CI, 0.5–1.0).
- Omission errors with simple task (KBM: 1, SR: 21; $p < 0.001$; CI, 0.5–1.0).
- Commission errors with simple task (KBM: 2, SR: 12; $p = 0.008$; CI, 0.0–0.5; see ►Table 4, ►Fig. 3).

Table 3 Summary of efficiency results

Task completion times: Experiment 1						
	Simple task KBM time (s)	Simple task SR time (s)	Complex task KBM time (s)	Complex task SR time (s)	Combined simple and complex KBM	Combined simple and complex SR
Mean	112.38	131.44	170.48	201.84	140.09	165.46
Max	214.99	285.91	257.63	327.45	257.63	327.45
Min	49.38	71.67	110.31	138.87	49.38	71.67
Task completion times: Experiment 2						
	Simple task KBM time (s)	Simple task SR time (s)	Complex task KBM time (s)	Complex task SR time (s)	Combined simple and complex KBM	Combined simple and complex SR
Mean	126.39	126.78	191.89	224.39	159.61	176.29
Max	214.54	175.40	349.84	400.01	349.84	400.01
Min	74.14	67.69	104.38	124.42	74.14	67.69
Experiment 1 vs. Experiment 2						
Efficiency tasks			Mean task completion time comparison (%)	*N*	Mann–Whitney *p*-Value	95% CI
Simple task KBM E2 vs. simple task KBM E1			112.46	34	0.060	−26.04 to 0.44
Simple task SR E2 vs. simple task SR E1			96.46	31	0.646	−17.30 to 10.91
Complex task KBM E2 vs. complex task KBM E1			112.56	31	0.199	−39.03 to 7.39
Complex task SR E2 vs. complex task SR E1			111.17	29	0.230	−42.93 to 9.96

Abbreviations: KBM, keyboard and mouse; SR, speech recognition.

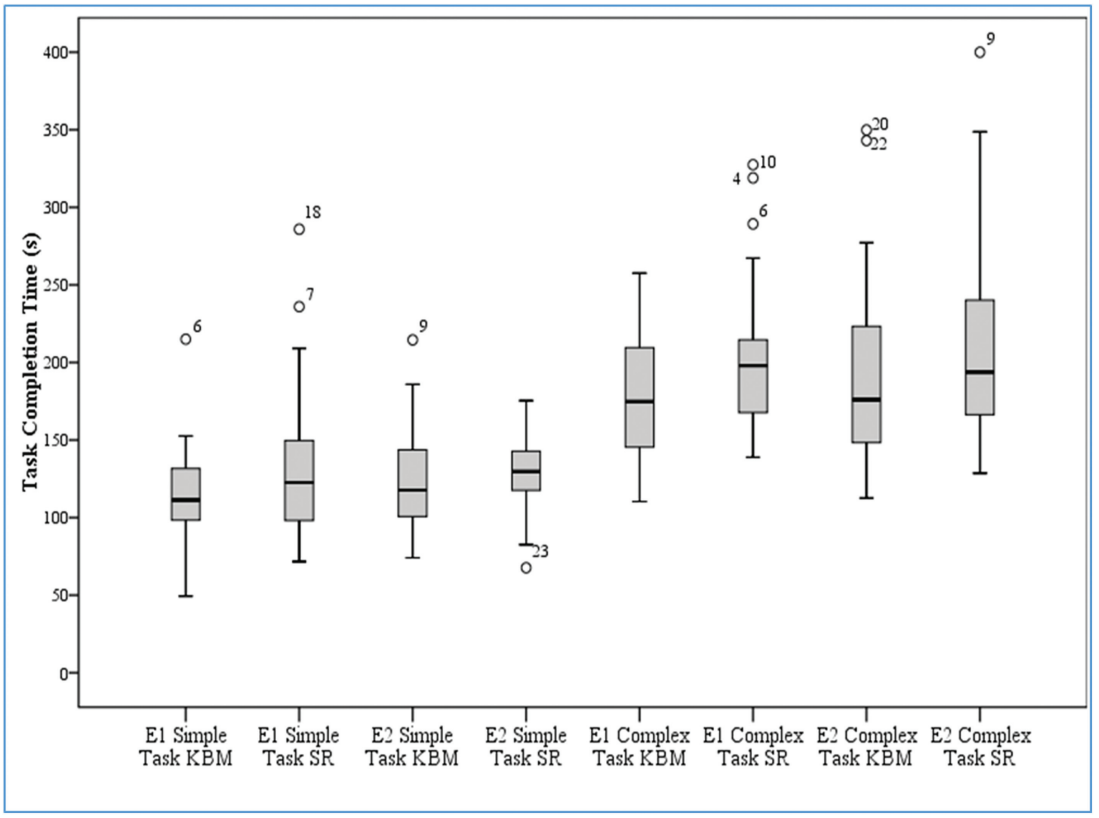

Fig. 2 Boxplot of task completion time for simple and complex tasks via input modality.

Comparing these results with those of our earlier experiment, there were no significant differences in the total number of non-typographical errors observed during the two experiments (E1: 170, E2: 163; $p = 0.660$; CI, 0.0–0.0). However, significant differences in the number of errors observed between the two experiments were found for use errors with complex task with SR (E1: 26, E2: 8; $p = 0.009$; CI, −0.0 to 1.0), comprehension errors with complex task with SR (E1: 17, E2: 0; $p < 0.001$), and omission errors with complex task with SR (E1: 6, E2: 0; $p = 0.010$).

Significant differences in typographical errors were observed between the experiments (E1: 465, E2: 150; $p < 0.001$; CI, 2.0–3.0). These were observed across all typographical error types: simple task with KBM (E1: 142, E2: 57; $p < 0.001$, CI, 2.0–4.0), simple task with SR (E1: 133, E2: 40; $p < 0.001$; CI, 2.0–4.0), complex task with KBM (E1: 71, E2: 29; $p < 0.001$; CI, 1.0–2.0), and complex task with KBM (E1: 119, E2: 24; $p < 0.001$; CI, 2.0–3.0; see ►Table 3).

There was no overall difference in the number of errors observed between experiments for repeat participants (E1: 69, E2: 57; $p = 0.311$; CI, 0.0–1.0). However, significant differences were seen in the number of errors observed between the two experiments in three scenarios: use errors with complex task with SR (E1: 15, E2: 3; $p = 0.006$; CI, −1.5 to 0.5), commission errors with complex task with SR (E1: 11, E2: 3; $p = 0.005$; CI, −1.0 to 0.5), and typographical errors with simple task with SR (E1: 46, E2: 20; $p = 0.010$; CI, −3.5 to 0.5; see ►Supplementary Material Appendix G, available in the online version).

Discussion

The results of this study largely replicate those reported in the original experiment. The use of SR while performing clinical documentation tasks within an EHR system was associated with decreased time efficiencies and increased data entry errors. Errors observed included some with risk of serious patient harm. This replication of results increases confidence that the risks identified with the original experiment are valid and not a statistical abnormality.

A series of modifications were made to the SR and EHR integration to minimize any potential bias in the original experimental setup toward KBM and to optimize the performance of SR. Several improvements were seen in the performance of SR, but these were insufficient to fundamentally change the overall results. While there were no statistical differences in overall error rates despite these technical improvements, the number of error types observed was reduced, with eleven observed in Experiment 1 eliminated: Incorrect patient, Incorrect test/order collection date entered, Data lost during text transfer (no EHR record created), Clinician closed EHR, Incorrect trivial word entered, Incorrect trivial word entered—user corrected, Word mangled (letters repeated or cut off), Word mangled—user corrected (letters repeated or cut off), EHR slow—system lag, Element of EHR down (e.g., vitals), and Missing comma(s).

The low numbers observed for some error types across both experiments makes generalizing lessons about the association between SR and those error types difficult.

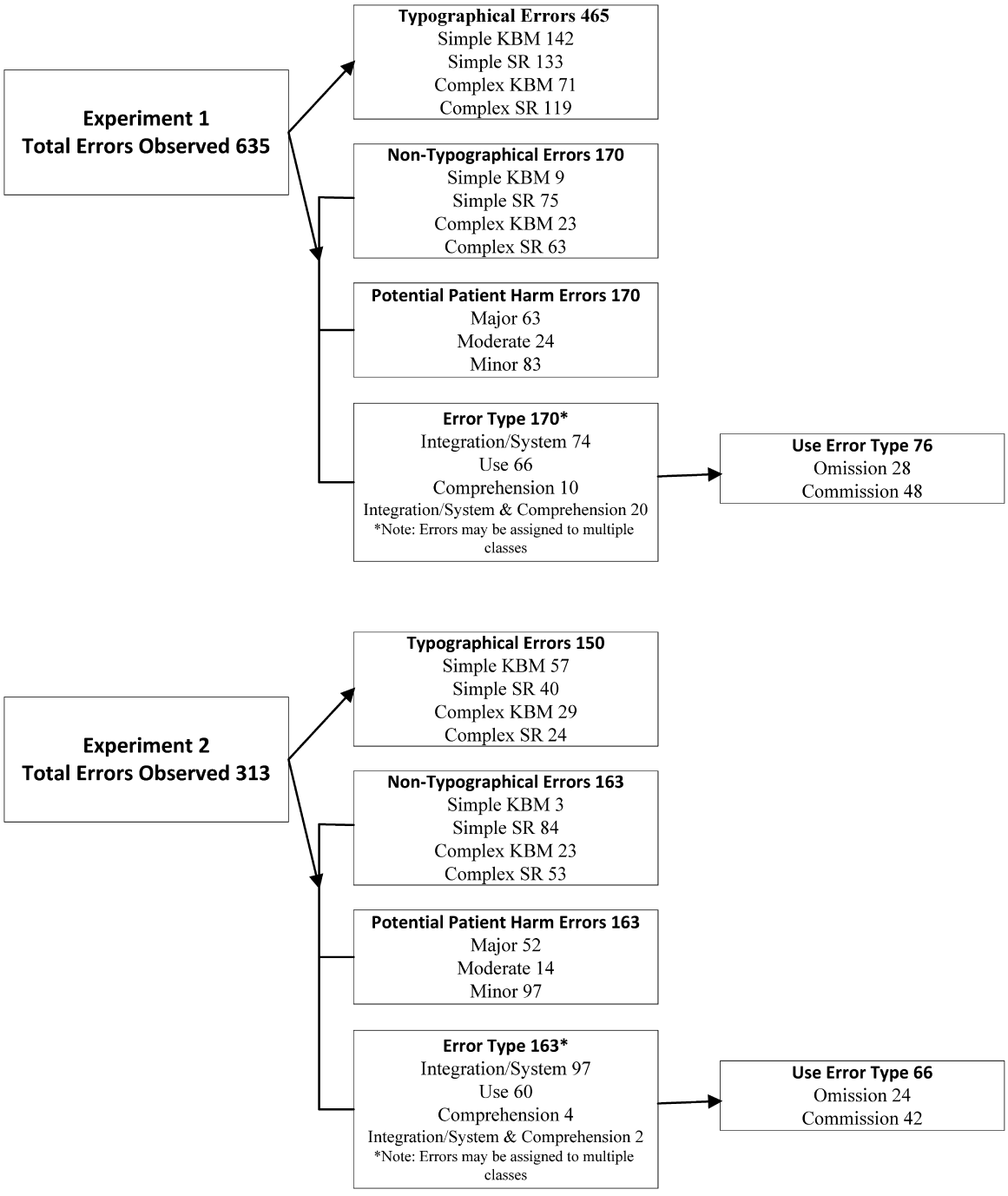

Fig. 3 Error framework. Overview of the breakdown of errors by class.

Rather, the picture is of SR on average continuing to generate higher error rates than KBM, including errors with risk of patient harm. Most of the reduction in volume of errors came from reductions in typographical errors across both modalities, predominately due to improved integration and the enabling of spell check components within the EHR. It seems likely that the specific distribution of different error classes will be highly influenced by the task undertaken, the specific EHR, and SR systems in use. Studies wishing to explore these would require much larger subject or tasks numbers than the present study to yield appropriate statistical power.

The variety of EHR documentation methods and technologies available to clinicians continues to evolve. Options such as SR are, for example, being combined with mobile devices and virtual or augmented reality to create new interaction models that could improve the efficiency of documentation. Decision support systems have the potential to mitigate some of the errors and problems observed within this study.[10–12] Context aware decision support systems could detect information entered in the wrong EHR fields, prompt for additional data when fields deemed appropriate are left blank, or trigger the use of a specialized vocabulary by the

Table 4 Error Summary Table

Experiment 1			Experiment 2			Experiment 1 vs. Experiment 2 (Mann–Whitney)		
							p-Values	
	KBM	SR		KBM	SR		KBM	SR
Total errors observed	**245**	**390**	**Total errors observed**	**112**	**201**		0.6595	
Non-typographical	**32**	**138**	**Non-typographical**	**26**	**137**	**Non-typographical**		
Simple	9	75	Simple	3	84	Simple	0.814	0.897
Complex	23	63	Complex	23	53	Complex	0.921	0.226
Potential patient harm	**32**	**138**	**Potential patient harm**	**26**	**137**	**Potential patient harm**		
Major	**13**	**50**	**Major**	**9**	**43**	**Major**		
Simple	2	29	Simple	1	35	Simple	0.842	0.229
Complex	11	21	Complex	8	8	Complex	0.428	0.310
Moderate	**3**	**21**	**Moderate**	**6**	**8**	**Moderate**		
Simple	0	13	Simple	1	7	Simple	N/A	0.664
Complex	3	8	Complex	5	1	Complex	0.828	0.152
Minor	**16**	**67**	**Minor**	**11**	**86**	**Minor**		
Simple	7	33	Simple	1	42	Simple	0.677	0.916
Complex	9	34	Complex	10	44	Complex	1.000	0.245
Error type	**32**	**158**	**Error type**	**26**	**139**	**Error type**		
Integration/System	**2**	**92**	**Integration/System**	**1**	**98**	**Integration/System**		
Simple	0	56	Simple	0	53	Simple	N/A	0.530
Complex	2	36	Complex	1	45	Complex	0.842	0.224
Use errors	**22**	**44**	**Use errors**	**24**	**36**	**Use errors**		
Simple	5	18	Simple	3	28	Simple	0.828	0.162
Complex	17	26	Complex	21	8	Complex	0.538	**0.009**
Comprehension	**8**	**22**	**Comprehension**	**1**	**5**	**Comprehension**		
Simple	4	5	Simple	0	5	Simple	0.039 (Chi-Square)	1.000
Complex	4	17	Complex	1	0	Complex	0.538	**0.000 (C-S)**
Use error type	**30**	**46**	**Use error type**	**25**	**41**	**Use error type**		
Omission	**8**	**20**	**Omission**	**3**	**21**	**Omission**		
Simple	2	14	Simple	1	21	Simple	1.000	0.190
Complex	6	6	Complex	2	0	Complex	0.414	**0.010 (C-S)**
Commission	**22**	**26**	**Commission**	**22**	**20**	**Commission**		
Simple	7	5	Simple	2	12	Simple	0.668	0.224
Complex	15	21	Complex	20	8	Complex	0.344	0.023
Typographical	**213**	**252**	**Typographical**	**86**	**64**	**Typographical**		
Simple	142	133	Simple	57	40	Simple	**0.000**	**0.000**
Complex	71	119	Complex	29	24	Complex	**0.000**	**0.000**

Abbreviations: KBM, keyboard and mouse; SR, speech recognition.

speech recognition system when specific contexts are recognized, thus improving recognition accuracy. Semantic technologies can monitor the clinical sense of data entries and help identify clinically incorrect data entries.

These results suggest that simply adding SR to an EHR that has been predominately designed for KBM interaction is unlikely to be an efficient or safe choice. While it is known that SR is reported to be effective for dictation, it may be that SR is better suited for entry of longer blocks of text, and is less well suited for system navigation and item selection.[1] Future studies could explore a hybrid approach to the use of SR in EHR systems, with KBM-like interaction being used to

support navigation and interaction with items such as drop-down menus, and SR being available as an option for free text data entry.

The Need for Replication Studies

Replication studies are rare in health informatics. When repeated studies are conducted, they are often in very different settings and achieve different results. Such differences are often ascribed to changes in context. Other explanations for failure to replicate a study include that it was some way statistically underpowered, biased in sample selection, or otherwise methodologically flawed.[13] Separating the influence of context and of experimental design is difficult, but focusing first and foremost on experimental factors to explain failure to replicate is a conservative and sound approach to take.

Given the significant impact that informatics interventions can have on clinical processes and patient outcomes, it is perhaps surprising that replication studies are not a standard feature of informatics research. The ongoing challenges in informatics, with implemented systems not always performing in the way expected, may say as much about the robustness of the informatics evidence base as it does about the influence of contextual or implementation factors.[6]

Limitations

Several factors may affect the likelihood that results from this study can be generalized to other clinical settings or information systems. The study used a routinely 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 may also vary results.

Modifications made to optimize the systems may have inadvertently introduced new problems. When implementing changes to any system, additional issues may be created, and this can be magnified when introducing multiple changes at the one time. The identification of the specific change that leads to a particular issue may be difficult when multiple changes are made simultaneously.

SR performance may be affected by extrinsic factors such as microphone quality, background noise level, or 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 may yield different outcomes. For example, dictation of investigation reports in high volume by expert clinicians might yield better time performance and recognition rates, although our previous review did not identify this.[1]

The reduction in the number of tasks in the second experiment (due to the removal of interruptions) may have affected task completion times or error rates for Experiment 2, perhaps because participants became more effective in system use with a larger number of tasks. Alternatively, the fewer tasks of Experiment 2 may have reduced fatigue compared with the greater number of tasks in Experiment 1.

Conclusion

The results of the replication study provide strong supporting evidence that SR-assisted clinical documentation in an EHR is both slower and produces more documentation errors than KBM alone.

Independent validation utilizing other EHR and SR systems would be a welcome extension to this research. The use of SR as a navigation and data input methodology in EHRs requires caution and continued monitoring and evaluation. More generally, replication studies are to be encouraged in health informatics, to ensure that unusual or highly impactful single studies are not acted upon without careful effort to ensure that their findings are indeed generalizable to other experiments and working settings.

Clinical Relevance Statement

This work is relevant to all clinicians who undertake electronic documentation within an EHR system. It provides insights into efficiency and safety of alternative input modalities and system optimization. Ultimately, this work should assist with identification of the most appropriate input modality for electronic clinical documentation.

Multiple Choice Questions

Which of the following is a valid reason for undertaking replication studies?

a. Increasing the author's number of journal published articles.
b. Discrediting the original study's authors and attacking their reputation.
c. Validation of initial findings, generalizability of results, and real-world application of results.
d. To maintain employment of research teams and utilize existing research funding.

Correct answer: The correct answer is option c, validation of initial findings, generalizability of results, and real-world application of results. Validation of initial findings, generalizability of results, and real-world application of results are all sound reasons for undertaking replication studies. The other options are all unreasonable motives for performing replication studies.

Authors' Contributions

T.H., E.C., and F.M. conceived the study and its design. T.H. conducted the research, the primary analysis, and the initial drafting of the manuscript. E.C. and F.M. contributed to the analysis and drafting of the manuscript, and T. H., E.C., and F.M. approved the final manuscript. T.H. is the corresponding author.

Protection of Human and Animal Subjects

This study is approved by the Sydney Local Health District Local Health District Human Ethics Committee—Concord

Repatriation and General Hospital (CRGH) (LNR/14/CRGH/272).

Funding
This work was supported by the NHMRC Centre for Research Excellence in eHealth (APP1032664).

Conflict of Interest
None.

Acknowledgments
The authors thank the staff and management teams of the participating hospitals who helped to facilitate the study, and also the study participants for volunteering and fitting us into their busy schedules.

References
1 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–e179
2 Hodgson T, Magrabi F, Coiera E. Efficiency and safety of speech recognition for documentation in the electronic health record. J Am Med Inform Assoc 2017;24(06):1127–1133
3 Open Science Collaboration. PSYCHOLOGY. Estimating the reproducibility of psychological science. Science 2015;349(6251):aac4716
4 Freedman LP, Cockburn IM, Simcoe TS. The economics of reproducibility in preclinical research. PLoS Biol 2015;13(06):e1002165
5 Nakagawa S, Parker TH. Replicating research in ecology and evolution: feasibility, incentives, and the cost-benefit conundrum. BMC Biol 2015;13(01):88
6 Coiera E, Ammenwerth E, Georgiou A, Magrabi F. Does health informatics have a replication crisis? J Am Med Inform Assoc 2018. Doi: 10.1093/jamia/ocy028
7 Engeström Y. Activity theory as a framework for analyzing and redesigning work. Ergonomics 2000;43(07):960–974
8 Kaptelinin V, Nardi BA, Macaulay C. Methods & tools: the activity checklist: a tool for representing the “space” of context. Interactions 1999;6(04):27–39
9 Guidance for Industry and Food and Drug Administration Staff. Guidance for the content of premarket submissions for software contained in medical devices. Rockville, MD: Center for Devices and Radiological Health; 2005
10 Castaneda C, Nalley K, Mannion C, et al. Clinical decision support systems for improving diagnostic accuracy and achieving precision medicine. J Clin Bioinforma 2015;5(01):4
11 Demner-Fushman D, Chapman WW, McDonald CJ. What can natural language processing do for clinical decision support? J Biomed Inform 2009;42(05):760–772
12 Jaspers MW, Smeulers M, Vermeulen H, Peute LW. Effects of clinical decision-support systems on practitioner performance and patient outcomes: a synthesis of high-quality systematic review findings. J Am Med Inform Assoc 2011;18(03):327–334
13 Gross MT. The Need for Replication Studies—Is It Really a Done Deal? J Orthop Sports Phys Ther 1997;25(03):161–162