Capturing patient information at nursing shift changes: methodological evaluation of speech recognition and information extraction

RECEIVED 10 April 2014
REVISED 21 August 2014
ACCEPTED 2 October 2014
PUBLISHED ONLINE FIRST 21 October 2014

Hanna Suominen[1], Maree Johnson[2], Liyuan Zhou[3], Paula Sanchez[4], Raul Sirel[5], Jim Basilakis[6], Leif Hanlen[7], Dominique Estival[8], Linda Dawson[9], Barbara Kelly[10]

ABSTRACT

Objective We study the use of speech recognition and information extraction to generate drafts of Australian nursing-handover documents.

Methods Speech recognition correctness and clinicians' preferences were evaluated using 15 recorder–microphone combinations, six documents, three speakers, Dragon Medical 11, and five survey/interview participants. Information extraction correctness evaluation used 260 documents, six-class classification for each word, two annotators, and the CRF++ conditional random field toolkit.

Results A noise-cancelling lapel-microphone with a digital voice recorder gave the best correctness (79%). This microphone was also the most preferred option by all but one participant. Although the participants liked the small size of this recorder, their preference was for tablets that can also be used for document proofing and sign-off, among other tasks. Accented speech was harder to recognize than native language and a male speaker was detected better than a female speaker. Information extraction was excellent in filtering out irrelevant text (85% F1) and identifying text relevant to two classes (87% and 70% F1). Similarly to the annotators' disagreements, there was confusion between the remaining three classes, which explains the modest 62% macro-averaged F1.

Discussion We present evidence for the feasibility of speech recognition and information extraction to support clinicians' in entering text and unlock its content for computerized decision-making and surveillance in healthcare.

Conclusions The benefits of this automation include storing all information; making the drafts available and accessible almost instantly to everyone with authorized access; and avoiding information loss, delays, and misinterpretations inherent to using a ward clerk or transcription services.

Key words: computer systems evaluation, information extraction, nursing records, patient handoff, speech recognition software

RESEARCH AND APPLICATIONS

OBJECTIVE

Fluent channels, communication, contact, and links to pertinent people, that is, *flow of information*,[1] is important in any information-intensive organization but critical in healthcare services. However, failures in the flow are common and lead to adverse events that could have been prevented. For example, in Australian healthcare, these failures are a major contributing factor in over two-thirds of sentinel events in hospitals and associated with over one-tenth of preventable adverse events.[2–5] This includes delays in diagnosis or treatment; administration of wrong treatments or medications; and missed or duplicated tests.

The failures are tangible in *clinical handover* or *handoff* when a clinician is transferring professional responsibility and accountability, for example at shift change.[3–6] In *nursing handover* the outgoing nurse verbally presents critical patient information to the oncoming nurse(s). This often occurs at the point of care or patient bedside. Then follows another separate process where nurses document similar information delivered at verbal handover into the patient's record.[7] Regardless of the verbal part being accurate and comprehensive, anything from two-thirds to all of this information is lost after 3–5 shift changes if handover notes are not documented or they are taken by hand.[8,9] In contrast, effective handover documentation improves care continuity

Correspondence to Adjunct Professor Hanna Suominen, Machine Learning Research Group, NICTA, College of Engineering and Computer Science, The Australian National University, Faculty of Health, University of Canberra, and Department of Information Technology, University of Turku, NICTA, Locked Bag 8001, Canberra ACT 2601, Australia; E-mail: hanna.suominen@nicta.com.au

For numbered affiliations see end of article.

and reduces errors.[4,10,11] Consequently, related guidelines and standards for clinicians and managers exist both internationally and nationally.[12,13]

In this paper, we study the use of *speech recognition* and *information extraction* methodologies to generate drafts of nursing handover suitable for the nursing notes within the patient record automatically (glossary of underlined key terms in online supplementary appendix). A speech recognition system transcribes verbal information into written text. By identifying relevant snippets of this text for each slot of a handover form, an information extraction system fills out the form. This pre-filled form is given to a clinician to proofread and sign-off. The paper presents empirical findings using a real-life clinical dataset with a focus on nurses' shift changes.

BACKGROUND

Result correctness

Speech recognition achieves 90–99% of the words being correctly detected with only 30–60 min of tailoring to a given clinician. This correctness is supported by studies using 12 US-English male physicians' speakers on two medical progress notes, one assessment summary, and one discharge summary;[14] two US-English physicians' speakers on 47 emergency department charts;[15] and speakers of seven Canadian-English pathologists and one foreign-accented researcher on 206 surgical pathology reports.[16] Speech recognition gives 6.7 erroneous words per clinical report; this rate is 0.4 for human transcribers.[16] When comparing speech recognition systems, IBM ViaVoice 98 General Medicine provides the best average correctness (91–93%), but this rate is only slightly worse for L&H Voice Xpress for Medicine 1.2 General Medicine (85–87%) and Dragon Medical 3.0 (85–86%).[14] A study using eight Danish speakers and about 3600 anesthesia comments demonstrates that the correctness is over 77%, even in the presence of background noise and interruptions.[17]

Information extraction as a method for filling out clinical forms has been addressed in over 170 studies between 1995 and 2008, with the best correctness exceeding the 90% F1.[18] The most common tasks include code extraction (eg, for assigned diagnosis codes or performed pathology examinations); de-identification and other report processing to support record research (eg, hiding or replacing patient names, dates of birth, and other privacy-sensitive words); record enrichment or structuring, especially to support computerized decision-making and surveillance in healthcare (eg, filling out forms related to incidences of cold, fever, and seasonal influenza in a given geographic region); and clinical terminology management (eg, creating a minimum dataset for emergency care, nursing, or other clinical specialty). This work mostly focuses on discharge, echocardiogram, pathology, and radiology reports.

Time released from documentation for other tasks

Each healthcare event must be documented in clinical records by law and this takes a lot of clinicians' time away from other duties. To illustrate the large number of events to document, in the OECD (Organisation for Economic Co-operation and Development) countries, seven physician consultations take place on average per capita per year and the annual hospital discharge rate is over 16 000 per 100 000 population.[19] Clinicians type approximately 40% of electronic clinical records as text; the remaining 60% is either manually- or automatically-entered structured information.[20] As an example of the overwhelming amount of manually documented information, the average number of structured items that clinicians enter in intensive care alone is over 1500 items per patient-day, and the amount of textual notes for a patient can be over 60 pages.[21,22]

Free-form text as an entry type is essential to release clinicians' time for other tasks. With electronic clinical records, allowing free-text entry at the point of care, clinicians use typically a few minutes per patient on documentation, but fully structured or centralized solutions can increase this to nearly 60% of their working time.[23–25] Further time saving can be gained by using speech recognition to support clinicians in entering text. At two US emergency departments, speech recognition and manual transcription have the report turnover times of 4 min and 40 min, respectively, with approximately 4 min proofing time in both cases.[15] Similarly in three US military medical teaching facilities, speech recognition with proofing by hand results in 4.7 times faster turnover than manual transcription.[26] This efficiency improvement is 2.1 at a Finnish radiology department and 3.0 in over 40 US radiology practices.[27,28] In a longitudinal study on 5011 US surgical pathology reports, the turnaround was 1.3 times faster after speech recognition was used for over 35 months.[29]

Importance of text structuring

Structuring the record content previously documented as free-form text makes finding and using relevant information easier, while also making information available for computerized decision-making and surveillance in healthcare.[30] However, accomplishing this structuring by hand takes more clinicians' time than entering free-form text, and using structured information without the option of visiting the original unstructured text can lead to a significant information loss as well as differences and errors in the coding.[23–25,31,32] This results from limiting the freedom and expressive power of free-form text, which contains valuable, interpretative information on patients' status and clinicians' decision-making, and provides stronger support to individualized care than structured electronic clinical records alone.[21,33–35] Nurses are keen to use speech recognition and information extraction systems if they enhance patient safety and reduce adverse patient outcomes.[36]

In this paper, we study the combined use of speech recognition and information extraction to generate drafts of structured handover documents. The potential benefits of this automation are threefold.

First, it stores all information. The approach covers the whole workflow from the recording of verbal handover through the speech-recognized transcription and automatically filled-out form to the proofed and signed-off record. In this way, clinicians never lose the context nor the changes, and the change-history can be used to improve the result correctness.

Figure 1: An example scenario. Permitted re-print.[41]

Bed eight, Michael Tian. Forty-eight years under Dr Greenborough. He came in with headache and vertigo. He's got a history of headache, tinnitus, Bell's Palsy to the left side of his face. That's where his headache has been for the last three years. He's also got photophobia. His GCS is 15 pupils equal and reactive. He's just came back from a brain MRI at Park Central. He's ambulant and self-caring but he's a little bit unsteady at times. OBS are stable. He is for carotid doppler, he was supposed to have this morning at 950 but that pushed it back to 1050, 1030, sorry, because they were late. Then the team were here and they said it's cutting it too close to his MRI so he needs another carotid doppler appointment. Other than that he's fine.

Second, it makes the record drafts available and accessible almost instantly to everyone with an authorized access to a particular patient's documents. The transcription for a minute of verbal handover (approximately 160 words) is available 20 s after finishing the handover if processing is performed in real time with the speed of 120 words per minute, while the waiting time for handwriting/typing is at least 3.5 min longer.[37–40] Additional issues with paper-based records include accessibility to only a couple of people in one location at a time together with their poorer readability and unavailability for automated re-use. Automated structuring through information extraction is systematic and almost instantaneous. The downside of only recording the verbal information is the inherent inability to automatically search information from these records if transcriptions are not available.

Third, an alternative approach of having a ward clerk to take the notes is even more prone to errors than the handover clinicians documenting the information themselves. In this approach, the clinician to whom the patient is handed over verbally summarizes the handover information to the clerk who then writes the handover document, by hand or typing. If we extrapolate from the information-loss rate for not taking notes,[8,9] during one shift more than 13% of the information gets lost.

MATERIALS AND METHODS

We evaluated speech recognition correctness and clinicians' preferences for microphones and recorders in simulated clinical settings, where one nurse is presenting a patient at a shift change.[41] The evaluations used six handover scenarios across the specialties of aged care ($n=1$, 366 words), dementia ($n=1$, 106 words), neurological ($n=1$, 144 words), and medical ($n=3$, $189+136+120=445$ words) as well as the baseline of the *Preamble to the Australian Constitution* (figure 1). Derived from existing clinical handover data, these fictitious and de-identified scenarios reflected a range of realistic handover situations.[42]

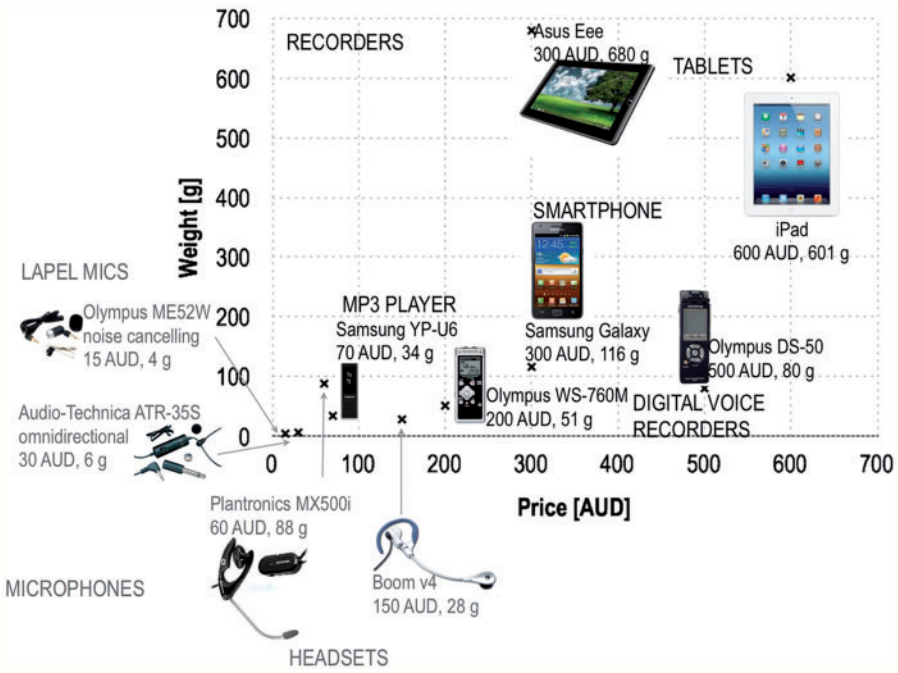
We used the *Dragon Medical Speech Recognition System, V.11* for Australian English with the vocabularies of general medicine, medicine, and nursing. We recorded three speakers (ie, two Australian-English native speakers (male physician and female nursing professor) and one Australian Spanish-accented female nurse) in a studio. We played these recordings using professional-level speakers and recorded this sound across all 15 feasible recorder–microphone combinations of two lapel microphones, two headsets, an MP3 player, two digital voice recorders, a smartphone, and two tablets (figure 2). We chose to tailor the speech recognition system minimally, requiring less than 5 min of prepared text reading by each speaker. As a measure of correctness, we compared the original text read by the three speakers with the Dragon outputs by analyzing *correctly recognized*, *substituted*, *deleted*, and *inserted words*, based on the *edit distance* between the compared texts. For this computation, we used the *SCLITE scoring tool of the Speech Recognition Scoring Toolkit, V.2.4.0.10* with capitalization as a non-distinguishing feature and punctuation removed. To illustrate, when generating *you are* for *your*, we observe a substitution *you—your* and an insertion *are* and when generating *your* for *you* are, this substitution is *your—you* whilst the word *are* got deleted.

We studied preferences via an 18-item pre-survey, 11-item post-survey, and one-to-one post-interview at a virtual clinical ward. Four registered nurses (two male and two female) and one female nursing scientist participated in the study. They had, on average, 28 years' clinical experience. The surveys addressed initial perceptions of speech recognition and perceived usability of the microphones and recorders (eg, power, recording, file upload, and other typical functions).

In information extraction, we used the total of 260 de-identified handover reports related to nursing shift-changes at medical and surgical wards in hospitals based in Sydney, Australia. The reports were based on human transcriptions (without *eh, um*, and other 'backchannels' but containing labels *[unclear]* and *[inaudible]* for incomprehensible parts) of verbal handovers that were collected by audio taping during either the morning or afternoon shift.

The form used to structure these text reports was based on the *ICCCO model* with *five categories* (table 1): *identification* (Iccco), *clinical history/presentation* (iCcco), *clinical status* (icCco), *care plan* (iccCo), *outcomes of care and reminders* (icccO),[42,43] supplemented with the *sixth category* (*not applicable*, NA) for irrelevant text. We chose this model because it has previously been shown to capture and present the information transferred during 81 handover cases related to nursing shift-

Figure 2: The compared microphones and recorders. 1 AUD = 0.91 USD (March 23, 2014). In the soundproof professional studio, we used a professional-level recorder (EDIROL UA-25 24 bit/96 kHz USB Audio Capturer) and microphone (Audio-Technica AT892cW-TH MicroSet Omnidirectional Condenser Headset). We played these recordings using professional-level speakers (EDIROL MA-15D Digital Stereo Micro Monitor Speakers) in a quiet meeting room.



changes within medical–surgical units in 10 major teaching hospitals in Sydney, Australia in 2012.[42,43] This model change resulted in improved nurse satisfaction, reduced falls, and clinical management errors. In this study, the five categories were used as headings of the handover form to be filled out with relevant information.

The final annotation (ie, the fourth round) (table 2) by one expert annotator (ie, PS) used as an information extraction gold standard was formed after three rounds of prior annotations by two expert annotators working independently of each other. The annotators (ie, PS and MJ) were supervised by an information extraction expert (HS) and had a weekly or fortnightly meeting to discuss this annotation process, its guideline, and specific problems. The first, second, third, and fourth rounds used a randomly selected subset of 10 reports out of the 260 reports, 50 reports out of the remaining 250 reports, 50 reports out of the remaining 200 reports, and the remaining 150 reports, respectively. Between each round, the supervisor measured the *inter-annotator agreement*, as defined by the *Knowtator 1.9* Beta *2 plugin for Protégé 3.3.1*, analyzed disagreements, and provided stylistic examples of common disagreements to the annotators.[44] After this feedback, the annotators revised the guideline and performed the next annotation round. This was continued until there was minimal disagreement between the five categories (tables 3 and 4). The remaining disagreements related to annotating many short, scattered snippets of text or fewer long, continuous snippets (ie, stylistic difference). We considered the more thorough but laborious alternative of forming the gold standard based on both annotators' independent assessments of the 150 reports followed by their discussion to solve the disagreements, but after achieving the third round results with the stylistic difference rather than category disagreements, we were confident to continue with only one annotator.

For automated form-filling, we used the *CRF++ conditional random field toolkit*.[45] This method solved the information extraction task by assigning precisely one of the six categories to each word in the report based on patterns learnt by observing words and the respective human-annotated topics in the gold standard, as well as the enriched *feature representation* of the words and their context (see table 5 for feature definitions and examples). These features were used to enrich the original words with additional information that helps to solve the form-filling task. For example, knowing that *can* is a modal verb and *is* is in the present tense in *Patient is conscious and can speak already*, and *can* is a common noun and *have* is in the past tense in *The patient had a can of lemonade but is not feeling better*, helped to disambiguate between the categories of *clinical status* and *clinical history/presentation.* We experimented with numerous features and chose the best 13 out of the 20 feature types to form our *best system compilation*.

Table 1: The ICCCO model and our final annotation guideline

	Definition	Examples		Final annotation guideline
Basic rules	Code separate phrases or words within the text. Strictly follow the class description when selecting the word or phases to be coded. No overlapping or double coding, the theme that corresponds more to the meaning of the context is to be selected A snippet approach is to be taken allowing phases to be shaped into a short succinct nursing note without all the comments and extra unnecessary language: *She was admitted yesterday with a duodenal ulcer.*[iCccо] Brevity of the snippet should be considered. A comma can be included in the snippet when the information is all related to the theme: *came in with shortness of breath, history of asthma and COPD, epilepsy and PE*[Cccо] Do not include full stops in the selection: *she had a fever of unknown origin*[iCccо]. *She has a history of MS*[iCccо]. Do not include unnecessary pronouns/words if the meaning is not altered: *She may be confused—it's 14 or 15*[icCcо], *depends.*			
Iccco: Identification	Identify the patient and introduce the oncoming staff to the patient	Under Dr [Name] In bed [No.] Okay we have this patient in room [No.] So [No.] Mrs [Name]	Allergic to penicillin, codeine, aspirin and panadeine. He's Hep C positive. He's VRE. He was actually TB. He is a voluntary patient.	Patient name, bed no., age, DOB, admitting doctor or team together with other identifying information early in the nursing handover, clinical risk alert, manual handling or mobility alert.
iCcco: Clinical history/ presentation	Describe relevant clinical history and the current patient problems	Renal colic, abdo pain, dehydration and vomiting diagnosis of chronic schizophrenia, recurrent UTI decreased mobility, decrease loss of consciousness Type 2 respiratory failure He's a type 2 diabetic She's a multip, day five, a caesar She was hit by a moving vehicle history of anxiety and depression he's an ex-smoker with a history of Parkinson's, COPD, AF.	She's presented, ?drug induced Came in with fractured C1 and C2 post fall admitted with increased confusion Ultrasound, MRI, C/T, X-Ray, ECG, ECT, laparotomy repair of inguinal hernia, appendicectomy Whipple's procedure, Doppler colonoscopy, right total knee replacement	Clinical history, reason for admission, diagnosis (incl. presenting, old, and all new diagnoses), procedures undertaken, presenting signs and symptoms.
icCco: Clinical status (a.k.a. signs and symptoms)	Current observations Outstanding/abnormal results Current status derived from observations & assessment of activities of daily living/ descriptions of the patient's well-being by following the 'between the flags' criteria (stable/deteriorating)	She's been fine, up & mobilizing she remains much the same Nil thoughts of self-harm on the unit He's settled on the ward No management problem with her Other than that he's fine She is not feeling good today a voluntary p. on care level 3 he is on leave She is under the Mental Health Act 19 He's a bit febrile his last BSL was 18.6 she was a bit tachycardic post-extubation	He's got low urine output for 3h She was hypotensive this morning last blood sugar was 9.2 at 12 Obs are stable postural BP monitoring her sinus rhythm haemodynamically stable foetal hearts are really good her circulation is good normotensive she's lost weight, 82 kilos this morning Mental health behavioral obs: no delusional beliefs expressed she does talk much to everyone not physically aggressive agitated and irritable pacing up and down	Current & general status (ie, stable, unchanged), observations, results, abnormal results, signs/ symptoms, assessment, mental status, and when vital signs are altered by treatment (eg, *did have some times of bradycardia today in the 40 beats per minute. Dr [Cachewala] was here at the time. He said that was due to his Carvedilol*)

(Continued)

Suominen H, et al. J Am Med Inform Assoc 2015;22:e48–e66. doi:10.1136/amiajnl-2014-002868, Research and Applications

Table 1: Continued

	Definition	Examples		Final annotation guideline
icCco: Clinical status (a.k.a. signs and symptoms)	Current observations Outstanding/abnormal results Current status derived from observations & assessment of activities of daily living/ descriptions of the patient's well-being by following the ?between the flags' criteria (stable/ deteriorating)	She's been fine, up & mobilizing she remains much the same Nil thoughts of self-harm on the unit He's settled on the ward No management problem with her Other than that he's fine She is not feeling good today a voluntary p. on care level 3 he is on leave She is under the Mental Health Act 19 He's a bit febrile his last BSL was 18.6 she was a bit tachycardic post-extubation	He's got low urine output for 3h She was hypotensive this morning last blood sugar was 9.2 at 12 Obs are stable postural BP monitoring her sinus rhythm haemodynamically stable foetal hearts are really good her circulation is good normotensive she's lost weight, 82 kilos this morning Mental health behavioral obs: no delusional beliefs expressed she does talk much to everyone not physically aggressive agitated and irritable pacing up and down	Current & general status (ie, stable, unchanged), observations, results, abnormal results, signs/ symptoms, assessment, mental status, and when vital signs are altered by treatment (eg, *did have some times of bradycardia today in the 40 beats per minute. Dr [Cachewala] was here at the time. He said that was due to his Carvedilol*)
iccCo: Care plan	In detail the care delivered, required or to be provided such as Input/output/diet (incl. checking all IV devices) Risk management (eg, Waterlow scale; Falls risk; high risk medications) Wound status & care Skin condition Management of activities of daily living Appointment & investigations	She's been treated with just QID voltaren waiting for aboriginal liaison worker & social worker review nursing care level 3 had day leave today attended ADL's she has a dental appointment tomorrow he's on a full diet he can ambulate with a rollator frame I think they would like to do a family conference she's toileting herself with encouragement and prompting he's on a daily weight which started this morning she needs TLC he is on a food chart he's for a plaster change on Monday she does have a bedpan if she needs it he can get up out of bed and sit in the chair The dressing is intact the wound has been debrided	haven't got a result on the wound swab taken took the dressing off her wound today and put some steri-strips on it I've put an opsite on there because it was oozing a bit he's got a little skin tear on his arm she's had a l of normal saline he's been continent of urine using the bottle he voided around 200mls 1.5 l of fluid restriction Hartmann's running at 10 NG feed running at 60 he's been vomiting, had a few issues with keeping down some fluids and food 80 ml flushes four hourly still got an IDC in and draining well Doctor ceased his fluids She's a medium risk of pressure area, pressure sores His falling risk is high he's ambulant and self-caring but he's a little bit unsteady at times His skin is very dry he does have some reddened areas on his bottom	ADLs, in/output, diet, medical or team reviews (incl. medical officers/teams' names) & instructions, medications, nursing care & tasks attended, oxygen therapy, pathways & instructions, pending reviews, pending tasks & procedures, wounds & dressings, supporting devices, monitoring, *mobilizing well in ward*, *tolerating diet*, *bowels opened*, etc. Time & date of arrival to the ward, moving within the ward Family & visitors, patient wishes/requests & interventions, compliance, family/other's involvement in the patient's care/plans: *son has taken her walke* and *Bed 11's wife wants to know, is he going home* Information on social status & place of residence: *lives with son*, *she is from a retirement village* When not clear whether discharge is occurring information: *going over to the private at two o'clock*

(Continued)

Suominen H, et al. J Am Med Inform Assoc 2015;22:e48–e66. doi:10.1136/amiajnl-2014-002868, Research and Applications

Suominen H, et al. J Am Med Inform Assoc 2015;22:e48–e66. doi:10.1136/amiajnl-2014-002868, Research and Applications

Table 1: Continued

	Definition	Examples		Final annotation guideline
iccc0: Outcomes of care and reminders	Occurred outcomes Expected outcomes Goals that are expected in the next shift The patient's response to care given Tasks to be completed Discharge/transfer plans	She had some buscopan with minimal effect she did eat a little bit, her BSL went up to 6.9, so I gave the insulin he looks a lot better she's tolerating diet and fluid Her IDC's out. She voided post removal when I put the icepack on she said the pain was less using her PCA because she said it is throbbing pain now and then with limit setting she responds Today she's settled. Usually she demands and is quite agitated and irritable she walked to the bathroom, she done really well she's not had any diarrhea today	If his Doppler to his legs was normal then he can go home She is going to be discharged to her own house on Friday with support from [Name the team has rung [Name] Nursing Home and they are quite happy to get him back Social services are in place regarding respite care of placement he'll discharge himself She's for discharge when she is mobilizing She's going home. Everything's organized. She's got her discharge paperwork he's hoping to go home tomorrow he may be discharged on Friday depending on further assessment her expected date is the 23rd to go I was told he can go to the ward maybe this afternoon NFR for MET	Outcome of care/intervention, resuscitation status, medical and/or altered review criteria (ie, NFR, for/not for mets, for/not for clinical review call) end of life issues Any information regarding discharge/transfer of the patient incl. the process Follow up & other care/placement post discharge Result(s) related to an intervention(s) & outcome incl. the initial result: *HB yesterday was 67. Had 2 units of packed cells yesterday. Come up to 87. So having another unit just now. Not feeling dizzy, light-headed* Administration of medication in relation to pain/pain score. *had pain eight out of 10 this morning. He had some Endone and it's come down to about five out of 10* Information that is clearly specified as related to the patient's discharge home/to another facility incl. other place of residency following discharge: *planning for ACAT assessment for long term care*
NA	All other information that is not related to any category and does not add meaning or new information such as: ▶ Nurse responsibilities and allocations ▶ Interactions and comments during nursing handover ▶ Clarifications by nurses ▶ Speaker (*female*, *facilitator*, *male*, *he*, *she*, etc.) Redundant information: repeated words/phrases are not to be coded more than once within the same case.			

Table 2: Final annotation used as a gold standard in information extraction

Snippets in total	
Iccco	602
iCcco	330
icCco	407
iccCo	1238
icccO	94
Number of snippets in a report	
Min	3
Max	51
Average	17.81
SD	7.78
Number of words	
Iccco	min 1, max 9, average 2.12, SD 0.86
iCcco	min 1, max 25, average 4.14, SD 3.57
icCco	min 1, max 40, average 4.73, SD 4.87
iccCo	min 1, max 48, average 6.22, SD 5.12
icccO	min 2, max 102, average 15.21, SD 14.63
Number of reports addressing	
0 categories	0
1 category	1
2 categories	3
3 categories	20
4 categories	75
5 categories	51
Number of words (unique lemmas)	
In total in 150 reports	39 808 (2637)
Min	862 (229)
Max	14 (13)
Average	256.38 (106.57)
SD	172.65 (45.08)

In evaluation, we used *cross-validation* with training set sizes of 30, 60, 90, 120, and 149 reports and the performance evaluation measures of precision, recall, and F1.[46] We evaluated performance both separately in every category and over all categories, as implemented in the *connleval.pl script* (http://www.cnts.ua.ac.be/conll2000/chunking/output.html). When evaluating the latter performance, we used *macro-averaging* because all the five categories are likely to be present in every report and we want to perform well in them all. This was done with the five ICCCO categories only, because NA is the dominating category in our data.

To assess learning performance in relation to the task difficulty, we computed two automated baselines. Our *random* baseline chose one of the six categories randomly for each word. Our *majority* baseline labeled all words as relevant to the most frequent category in our gold standard (ie, *iccCo*).

RESULTS

Speech recognition

The noise-cancelling lapel microphone (15 AUD) in combination with the 200 AUD digital voice recorder gave the best correctness (max 78.7%). In comparison, the correctness percentages for the studio recorder and microphone were 78.5, 64.3, and 52.7 for the native male (MN), native female (FN), and accented female (FA) speaker, respectively. These percentages were more modest than the best rates in clinical speech recognition (ie, 90–99%),[14–16] but explained by our speaker-tailoring being 6–12 times shorter.

Lapel microphones were also preferred over headsets by four out of five participants, and the participants also liked the small size of the digital voice recorders. However, their preference was for tablets that allow them to proof and sign off the resulting handover form, among other reading and writing tasks. Digital voice recorders gave the best correctness (42.8–78.7%), followed by tablets (17.9–74.9%), the MP3 player (13.1–58.8%), and smartphone (13.2–49.5%). Accessory microphones, in particular the noise-cancelling one, improved the recorders' correctness.

The preamble text was always easier for speech recognition than clinical language, and dementia was the easiest clinical specialty (70.0–89.5% vs 58.2–84.0%). The order of the rest of the specialties varied between the speakers.

The best, second best, and worst speech recognition vocabulary, using the best-performing device and the six handover scenarios, were nursing, medicine, and general medicine (figure 3). This held for all three speakers. The difference in the percentage of correctly recognized words between the best and the worst vocabulary was smaller for male than female speakers (MN: 1.4, FN: 7.6, FA: 5.2).

With the best vocabulary, the average percentage of correctly recognized words across the six patient cases was 70.5 at its best (figure 3). Speech recognition for a male speaker gave better results than for a female speaker, and for a native speaker than for an accented speaker. The best vocabulary had the average (median) percentage of correctly recognized words of 65.6 (64.4), with the sample variance of 19.4 across all speakers and patient cases. On average, substitutions were over twice more common than deletions and over nine times more common than insertions. If considering the differences between the patient cases with this vocabulary, the average

Table 3: Inter-annotator analysis across the three prior annotation rounds and five categories

Category	Round	Inter annotator agreement [%]	No. of matches	No. of non-matches
Iccco	10 records in Nov 2012	44.44	30	24
	50 records in Jan 2013	33.01	138	68
	50 records in Feb 2013	75.63	29	90
iCcco	10 records in Nov 2012	46.00	27	23
	50 records in Jan 2013	58.82	67	75
	50 records in Feb 2013	50.77	64	66
icCco	10 records in Nov 2012	28.95	27	11
	50 records in Jan 2013	40.74	80	50
	50 records in Feb 2013	46.01	88	75
iccCo	10 records in Nov 2012	13.38	136	21
	50 records in Jan 2013	37.07	533	314
	50 records in Feb 2013	35.69	501	278
icccO	10 records in Nov 2012	48.15	14	13
	50 records in Jan 2013	53.12	30	34
	50 records in Feb 2013	50.00	29	29

Table 4: Agreements (ie, diagonal elements) and disagreements (off-diagonal elements) between the two annotators across the three prior annotation rounds (ie, 1st (2nd) [3rd]) and five categories

Annotator: category	PS: Iccco	PS: iCcco	PS: icCco	PS: iccCo	PS: icccO	Agreement %
MJ: Iccco	20 (64) [90]	3 (2) [0]	0 (1) [0]	1 (1) [0]		83 (94) [100]
MJ: iCcco	3 (2) [0]	20 (72) [64]		0 (1) [2]		87 (96) [97]
MJ: icCco	0 (1) [0]		8 (48) [72]	3 (6) [3]		73 (87) [96]
MJ: iccCo	1 (1) [0]	0 (1) [2]	3 (6) [3]	12 (304) [272]	5 (2) [1]	57 (97) [98]
MJ: icccO				5 (2) [1]	8 (32) [28]	62 (94) [97]
Agreement %	83 (94) [100]	87 (96) [97]	73 (87) [96]	57 (97) [98]	62 (94) [97]	

Empty cells refer to cases with no disagreements.

percentage of correctly recognized words across the three speakers varied between 57.8 and 68.9, with an average and variance of 64.5 and 14.0, respectively.

Information extraction

The best system compilation was excellent in identifying text relevant to Iccco and iCcco (precision, recall, and F1 percentages of 90.6 vs 76.9 (figure 4A), 83.7 vs 65.0 (figure 4B), and 87.0 vs 70.5 (figure 4C) for the former vs latter category, respectively) and also in distinguishing text relevant to the handover form from irrelevant text (80.4% precision, 89.6% recall, 84.7% F1) (figure 4). This correctness for Iccco and NA was close to the state-of-the-art in information extraction (ie, 90% F1),[18] but with other categories, the results were more modest. Similarly to human annotators, our information extraction system produced confusions in text relevant to the three remaining categories (ie, iccCo, icCco, and icccO) which explain the smaller macro-averaged performance over the five ICCCO categories (ie, 66.5% precision, 57.2% recall, and 61.5% F1). In comparison, the macro-averaged performance for the random (majority) baseline was as low as 10.8% (20.2%) F1 over the ICCCO categories and 26.0% (0.0%) F1 in NA.

As expected, having more data for training contributed to the system performance (figure 4). However, obtaining this

Suominen H, et al. J Am Med Inform Assoc 2015;22:e48–e66. doi:10.1136/amiajnl-2014-002868, Research and Applications

Table 5: Experimented features and template

Compilation	Feature type	Software
The best system	1 *Original word*	
	2 *Lemma* (base form, eg, *patient* for *patients* or *extract* for *extracted*)	Stanford CoreNLP
	3 *Named entity recognition* (NER for person, location, organization, other proper name, date, time, money, and number., eg, *number* for *5*)	Stanford CoreNLP
	4 *Part of speech* (POS, eg, noun, verb, adjective, etc.)	Stanford CoreNLP
	5 *Parse tree*, including the ancestors from the root to the node (eg, *root—noun phrase—common noun* for *5* in *In bed 5, we have. . .*)	Stanford CoreNLP
	6 *Basic dependents (*eg, *5* that refers to the bed ID for *bed* in *In bed 5, we have. . .*)	Stanford CoreNLP
	7 *Basic governors* (eg, preposition *in* and subject *we* for *have* in *In bed 5, we have. . .*)	Stanford CoreNLP
	8 *Phrase context* (eg, *In bed 5* for *bed* in *In bed 5*)	MetaMap 2011
	9 *Top 5 candidates* (eg, *BP* may refer to *Bachelor of Pharmacy*, *bedpan*, *before present*, *birthplace*, or *blood pressure*, among others)	MetaMap 2011
	10 *Top mapping* (eg, *pneumonia* is a type of *respiratory tract infection*)	MetaMap 2011
	11 *Location percentage* (within the [0, 10%], (10%, 20%], (20%, 30%], . . ., (90%, 100%] of the report) on ten-point scale	NICTA
	12 *Medication score* (1 if the word belongs to words annotated as a medicine in an independent set of 100 synthetic handover reports after removing tailored generic stopwords [i.e., the following 14 words: *2*, *4*, *a*, *all*, *and*, *as*, *chest*, *effect*, *for, from*, *i*, *in*, *no*, *of*, *on*, *or*, *other*, *pain*, *the*, *to*, *two*, *used*, *very*, *with*], 0 otherwise).[46]	NICTA
	13 *Systematized Nomenclature of Medicine—Clinical Terms—Australian Release identifiers* (SNOMED-CT-AU IDs, 0 if none found): first, the aforementioned MetaMap was used to extract all Unified Medical Language System (UMLS) concept IDs that are present in the SNOMED-CT subset of UMLS and second, the UMLS concept IDs were transformed to the respective SNOMED-CT IDs through a lookup table derived from UMLS by using the UMLS Installation and Customization Program called MetamorphoSys	SNOMED-CT-AU, MetamorphoSys, and NICTA

(Continued)

Table 5: Continued

Compilation	Feature type	Software
Other experimented features	14 Normalized *term frequency* (ie, number times a given term occurs in a handover report/maximum of this term frequency over all terms in this report)	NICTA
	15 SNOMED-*CT-AU abstraction IDs* (0 if none found): The SNOMED CT-AU concepts found in records were generalized by using, the SNOMED CT-AU reference sets, that is, for each concept, a check was performed to identify all the reference sets the concept was present. This allowed assigning a pseudo semantic class (or abstraction) for each concept. This pseudo semantic class, defined as the list of reference sets the concept was present, was then used as word-level feature (0 if word is not part of SNOMED CT-AU concept or is not present in any of the reference sets).	SNOMED-CT-AU and NICTA
	16 *Australian Medicines Terminology* (AMT) *IDs* (0 if none found): first, a dictionary-based matcher was used in order to identify medicines (both the trade names and their generic versions [eg, *oxycodone* for both *Roxicodone* and *Oxycontin*]) in the texts; second, the medicines were transformed to their generic versions by using the trade products and medicinal products columns in AMT; and third, in order to capture the more general semantic properties carried by the medicinal products, SNOMED CT-AU was used to abstract various medicines to their common parent concept (eg, *opiate* or *CNS (central nervous system) drug* for both *oxycodone* and *codeine*). AMT contained 5066 trade products and 1821 medicinal products. Since there was no existing mapping between medicinal products in AMT and their respective products in SNOMED-CT-AU, string matching was used to find the SNOMED-CT-AU IDs corresponding to the medicinal product IDs in AMT. Such approach allowed to map 842 (46.2%) of total 1821 medicinal products in AMT to their respective concepts in SNOMED CT-AU. Although a substance-level mapping exists between AMT and SNOMED-CT-AU, such approach was discarded because of multiple substances corresponding to a single medicinal product and further complications in transforming concepts to their common abstractions (the structure of SNOMED-CT allows multiple paths between concept nodes).	AMT Sep 2013, Ontoserver, and NICTA
	17 *Corpus check*: since some of the Australian trade products have homonymous names (eg, *Pain*, *Ice*, and *Peg*), a binary feature was introduced to determine if the medicine name is present in the fiction subset of the Brown University Standard Corpus of Present-Day American English.	AMT Sep 2013, Brown Corpus, and NICTA
	18 *Expression type*, *position*, and *temporality*: our hypothesis was that temporal expressions are not distributed homogeneously in the report. For example, iCcco (clinical history/ presentation) should (in theory) contain temporal expressions referencing the past events, whereas iccCo (care plan) or icccO (outcomes of care and reminders) should rather contain expressions referencing future events. From the Med-TTK output, three separate word-level features were created: the first identified the type of the expression (if it's a date, duration, etc.; 0 for words not part of temporal expression), the second identified the relative position of the reference on the timeline (before, after, etc.; 0 for words not part of temporal expression), and the third was a binary feature determining if the word is part of a temporal expression (0 if not).	Med-TKK,[45] medical modification of The Temporal Awareness and Reasoning Systems for Question Interpretation (TARSQI) Toolkit (TTK), and NICTA

(Continued)

Suominen H, et al. J Am Med Inform Assoc 2015;22:e48–e66. doi:10.1136/amiajnl-2014-002868, Research and Applications

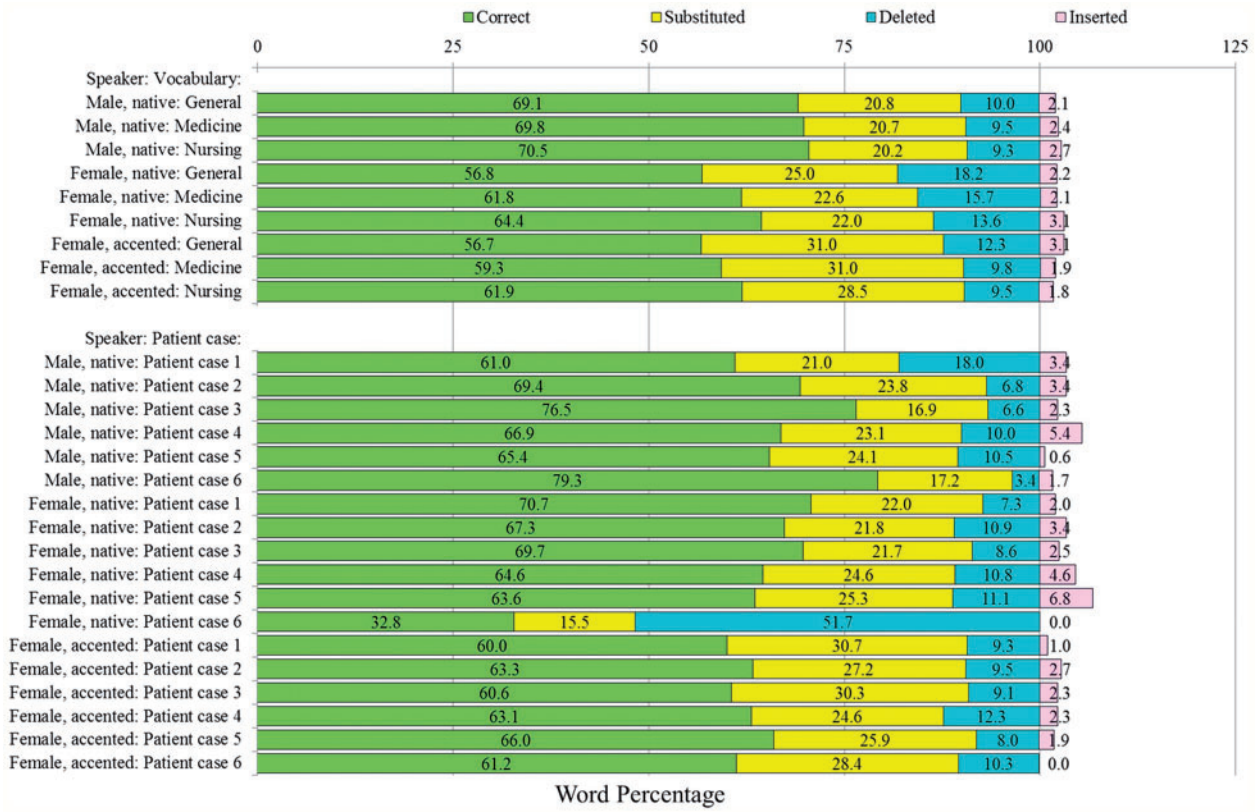
Table 5: Continued

Compilation	Feature type	Software
Other experimented features	19 *Sentiment carrying* and *polarity:* Our hypothesis was that, similarly to temporal expressions, the sentiment carrying words are not distributed homogeneously in the reports. For example, lccco (identification of the patient) should not contain such words at all, whereas icccO (outcomes of care and reminders) should contain a great deal of sentiment—patient getting better, condition improved. The approach used exact string matching to map the transcripts to the list of positive and negative opinion words or sentiment words. As this approach was not tailored for clinical language and word sense disambiguation was not used in this task, common ambiguous stopwords like *patient* and *like* were removed from the list based on our initial error analysis. Additionally, the reports contained words *[unclear]* and *[inaudible]* for identifying incomprehensible portions of text in transcription which we removed. The approach resulted in two separate word-level features: the first identified if the word is carrying any sentiment (1 for yes and 0 for no) and the second identified the sentiment polarity (positive, negative, or 0).	Positive and negative opinion words or sentiment words for English (appr. 6800 words with their respective polarities) supplemented with *discharge* and *discharged* (both were marked with positive polarity) at NICTA
	20 *Rule-based post-processing* containing seven rules for changing the label predicted by CRF++ based on analyzing the labels of previous and following lemmas and their POS tags. The rules were used to change the labels of interjections, modals, personal pronouns, adverbs, and existential theres to *NA*, because these tokens have no useful meaning in current context and thus should be discarded. Additionally, the labels of determiners were changed to *NA*, if the previous and next tokens were labeled as *NA*. Similarly, the labels of conjunctions were changed to match the label of the next token, if both the previous and next token were labeled the same.	NICTA

In the CRF++ template, which controls the usage of features in training and testing, we defined in the unigram part that we use all features of the current location alone, previous location alone, and next location alone; the pairwise correlations of the previous–current location over all features; the pairwise correlations of the current–next location over all features; and the correlation/combination of all features in the current location. In the binary part of the template, we set up the use of only one bigram template, which combines the predicted category for the previous location and the features of the current location as a feature. To download the cited software, please visit http://nlp.stanford.edu/software/corenlp.shtml for Stanford CoreNLP, http://metamap.nlm.nih.gov/ for MetaMap, http://www.nehta.gov.au/our-work/clinical-terminology/snomed-clinical-terms for SNOMED-CT-AU, http://www.ncbi.nlm.nih.gov/books/NBK9683/ for MetamorhoSys, http://www.nehta.gov.au/our-work/clinical-terminology/australian-medicines-terminology for AMT, http://ontoserver.csiro.au:8080/ for Ontoserver, http://khnt.hit.uib.no/icame/manuals/brown/index.htm for Brown Corpus, and http://www.cs.uic.edu/~liub/FBS/opinion-lexicon-English.rar for Positive and negative opinion words or sentiment words for English. For NICTA software, please email hanna.suominen@nicta.com.au.

Suominen H, et al. J Am Med Inform Assoc 2015;22:e48–e66. doi:10.1136/amiajnl-2014-002868, Research and Applications

Figure 3: Speech recognition correctness of the best performing device: the top bar chart details this performance on the six patient cases for the three speakers and three vocabularies over the six patient cases. The bottom bar chart addresses the differences between the six patient cases and three speakers using the best performing vocabulary (ie, nursing). With this nursing vocabulary, the average percentages of correctly recognized, substituted, deleted, and inserted words across the six patient cases were 70.5 (64.4) [61.9], 20.2 (22.0) [28.5], 9.3 (13.6) [9.5], and 2.7 (3.1) [1.8] for MN (FN) [FA]. MN, native male; FN, native female; FA, accented female.



performance gain by annotating more data might not be that straightforward; differences in annotators' opinions need to be explained, but because including the features for the previous and/or next words did not help the automated system (table 5), observing this text in context is unlikely to resolve the human disagreements. In other words, more fundamental alterations of the annotation guidelines are needed.

The best system compilation used 13 feature types and on average, a system using 12 feature types had 1.8% better category-specific F1 (table 6). The system with all 13 feature types had 62.1% macro-averaged F1 over ICCCO (0.9% better without parse tree) and 83.5% F1 for NA (0.8% better without location percentage) in this cross-validation experiment using 120 reports for training. However, because different feature types were advantageous/disadvantageous in different categories, the system using 13 features was the best overall. For example, the aforementioned system without parse tree (without location percentage) had 0.8% worse F1 in NA (1.1% worse macro-averaged F1 over ICCCO) than the best system compilation. For Iccco, the disadvantageous feature types, from the most to least disadvantageous, were top 5 candidates, medication score, and basic governors. For iCcco (icCco) [iccCo] {NA}, they were parse tree, top mapping, NER, medication score, basic governors, and basic dependents (location percentage, medication score, parse tree, top mapping, basic governors, and basic dependents) [SNOMED-CT-AU IDs and medication score] {location percentage, top mapping, basic governors, top 5 candidates, and SNOMED-CT-AU IDs}. For icccO, none of the feature types was disadvantageous.

Suprisingly, as this analysis illustrates, basic grammatical analysis (ie, lemmatization, part of speech tagging, and phrase context) contributed most to information extraction performance, and more advanced grammatical analysis (eg, parsing, dependent and governor identification, named-entity recognition, or sentiment/temporality analyses) or the use of clinical terminologies was not advantageous and even could be disadvantageous. Different feature types benefit information extraction from different types of records, and consequently an extension of this study could use the patient type (eg, cardiovascular vs neurological vs renal vs respiratory patients)[47] or

Figure 4: Precision (4a), recall (4b), and F1 (4c) percentages and learning curves for different cross-validation (CV) settings (ie, training set sizes of 30, 60, 90, 120, and 149 (ie, leave-one-out [LOO] CV) reports with mutually exclusive folds that in combination cover all data). NA refers to the category for irrelevant text. The horizontal direction of the histograms reflects the contribution of having more data for training and the vertical direction the effects of the coupled measured of precision and recall in F1 (see the glossary in online supplementary appendix).

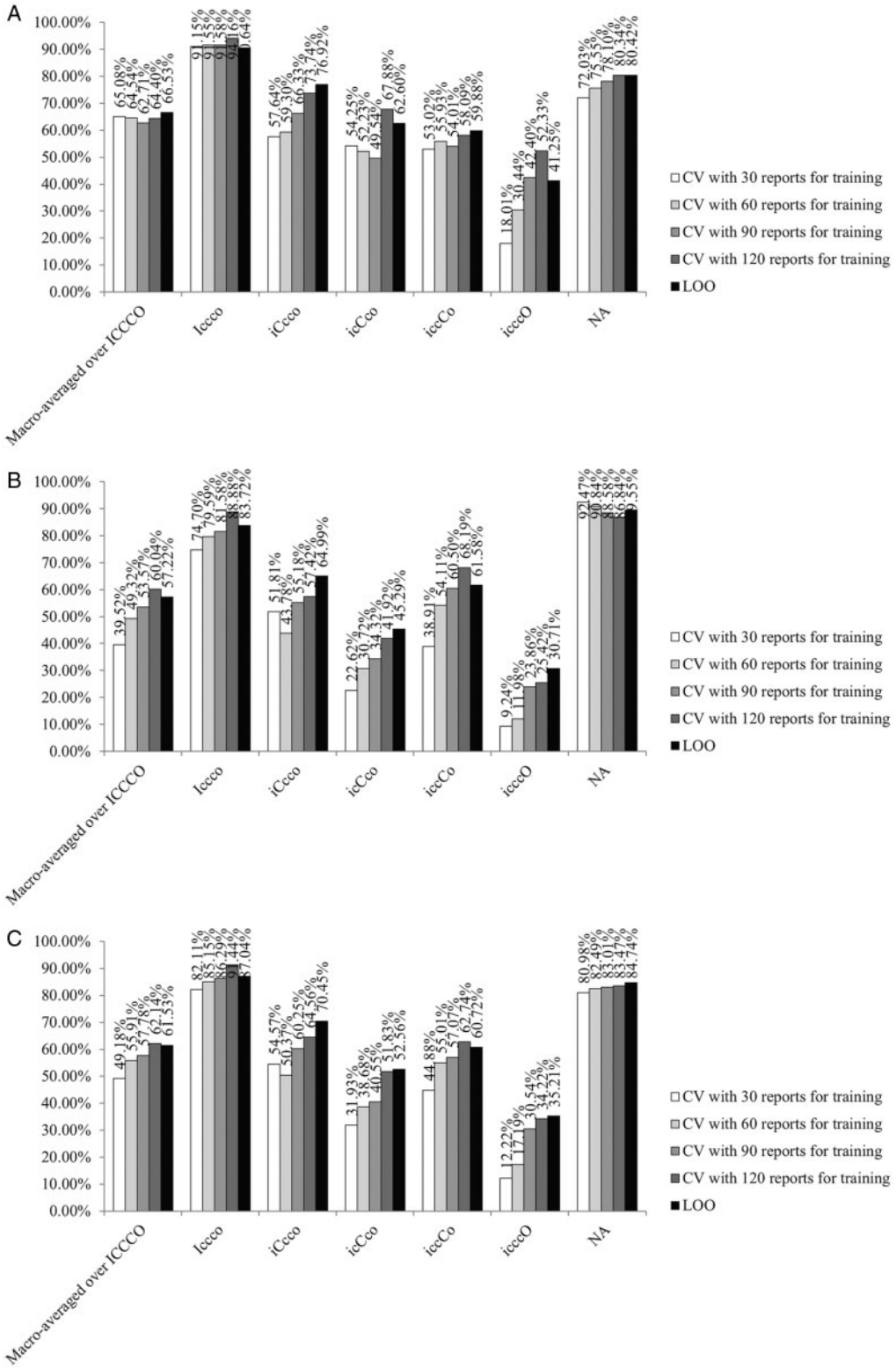


Table 6: Contribution of each feature to the best system compilation using cross-validation with 120 reports for training

Removed feature\category	Macro-averaged over ICCCO	Iccco	iCcco	icCco	iccCo	icccO	NA
Difference in precision [%]							
Word	−1.09	−0.57	−1.77	−3.16	−1.39	−4.49	−0.07
Lemma	−0.93	0.00	0.31	−2.58	−0.93	−4.36	−0.38
NER	−0.91	−1.31	6.69	−2.26	−0.82	−7.14	−0.50
POS	−1.07	−1.98	−3.12	−7.11	−1.22	−9.09	−1.15
Parse tree	−0.43	−2.60	4.85	3.63	−1.73	−7.14	−0.13
Basic dependents	0.35	−0.76	0.72	0.51	−0.78	0.00	−0.19
Basic governors	−1.98	−0.92	1.51	0.33	−1.89	−14.64	0.34
Phrase context	−2.46	−0.34	0.05	−0.55	−3.60	−9.40	−0.65
Top 5 candidates	−0.48	0.00	−2.07	0.73	−1.14	−9.09	−0.26
Top mapping	0.06	−1.15	5.22	−0.15	−0.24	0.00	0.63
Location percentage	0.35	−0.79	**−20.47**	**7.36**	0.32	−4.36	−0.24
Medication score	0.11	−0.57	4.63	1.00	0.20	−4.36	−0.08
SNOMED-CT-AU IDs	−0.34	−0.75	0.73	−3.90	−0.42	−12.69	0.18
Difference in recall [%]							
Word	−1.11	0.00	−4.94	−0.66	−1.42	0.00	−0.01
Lemma	−0.85	0.00	−4.94	−3.81	−0.93	−2.12	−0.37
NER	−1.52	−0.94	−0.10	−2.53	−0.73	−2.12	−0.12
POS	−2.54	−0.38	−4.11	−6.31	−2.44	−2.10	0.43
Parse tree	2.18	0.51	**4.72**	3.95	1.36	−5.82	−1.50
Basic dependents	0.02	−0.27	0.28	0.45	−0.68	0.00	0.00
Basic governors	−2.60	0.93	−0.04	0.60	−4.55	−1.99	0.31
Phrase context	−2.21	−0.53	−2.33	0.00	−3.38	0.00	−0.57
Top 5 candidates	−1.74	0.69	−1.81	−4.51	−1.96	−6.06	0.68
Top mapping	0.08	0.95	1.64	1.40	0.22	0.00	0.30
Location percentage	−2.35	−0.31	**−16.10**	3.20	−2.01	−2.12	1.97
Medication score	−0.11	0.69	0.49	1.07	−0.22	−2.12	−0.18
SNOMED-CT-AU IDs	−0.31	−0.50	−5.54	0.14	0.85	−7.68	−0.03
Difference in F1 [%]							
Word	−1.10	−0.27	−3.86	−1.43	−1.42	−1.02	−0.05
Lemma	−0.88	0.00	−3.14	−3.70	−0.94	−2.85	−0.38
NER	−1.23	−1.11	2.38	−2.61	−0.79	−3.47	−0.33
POS	−1.87	−1.14	−3.80	−6.93	−1.75	−3.92	−0.44
Parse tree	0.94	−0.98	**4.85**	4.06	−0.48	−6.88	−0.78

(Continued)

Table 6: Continued

Removed feature\category	Macro-averaged over ICCCO	Iccco	iCcco	icCco	iccCo	icccO	NA
Basic dependents	0.17	−0.50	0.46	0.49	−0.75	0.00	−0.11
Basic governors	−2.31	0.05	0.55	0.55	−3.05	−5.33	0.32
Phrase context	−2.33	−0.44	−1.48	−0.16	−3.54	−2.29	−0.62
Top 5 candidates	−1.16	0.37	−1.93	−3.41	−1.50	−7.48	0.16
Top mapping	0.07	−0.05	3.02	1.01	−0.05	0.00	0.47
Location percentage	−1.13	−0.54	**−18.02**	4.58	−0.68	−2.85	0.76
Medication score	0.00	0.09	2.04	1.11	0.02	−2.85	−0.13
SNOMED-CT-AU IDs	−0.32	−0.61	−3.40	−1.07	0.10	−9.71	0.08

Negative values indicate that removing a given feature decreases the performance—the larger the absolute value the more this feature contributes to the performance of the best compilation. Positive values indicate that a given feature does not contribute to the performance—the larger the value the more harmful the feature. NA refers to the category for irrelevant text. Minimal and maximal values are in **bold**.

POS, part of speech.

record type (eg, discharge vs echocardiogram vs pathology reports) in cross-validation to study which feature types are the most advantageous for different record/patient types. In addition, this extension could use nursing-specific terminologies only or consider other clinical language processing tools, though some comparative evaluations give evidence for the superiority of the *MetaMap* tool we chose and, based on our *PubMed* search (http://www.ncbi.nlm.nih.gov/pubmed), this is the most common choice in clinical language processing (ie, 12, 1, and 74 hits for *KnowledgeMap*, *Mgrep*, and *MetaMap*, respectively).[48–57] Typically the use of clinical terminologies is advantageous in similar language processing tasks. For example, in tasks related to searching and summarizing biomedical papers, using MetaMap improved the retrieval correctness by 14% and decreased the lexicon by more than 83%.[58,59]

DISCUSSION

Inspired by the collection of data as a by-product of care and the use of this information to design even safer care delivery systems,[60] we have addressed data collection as an intrinsic part of care—not an administrative 'extra'. While information flow during handover is driven by clinical excellence and safety, documentation of data is often secondary. This drives both information loss (eg, time delays and summarization) and inefficiencies (ie, double handling). In some cases, an administrator manually enters a summary of nursing handover into the hospital information system, after the handover occurs. If technology is to support efficiency in care, then capture of data (as well as analysis) is a key point of interest. In this sense, speech recognition is simply capturing data at the source. By capturing data 'closer to the source' we also greatly expand the scope of analysis: a transcribed handover is far richer than a typed nursing note summary. Systems that do not consider the capture of data risk exacerbating existing data inefficiencies.

Delivery of care is central to nursing practice: a nurse dictating the previous shift handover to a clerk is not delivering care—no matter how much that dictation improves care delivery for future shifts. Primary communication between nurses is often team-based and verbal—written notes are secondary. While speech recognition is in its infancy, it has the potential to support the preferred models of nursing communication and augment verbal team based communication. If speech recognition can be realized, there is great scope for nurses to have their clinical discussions automatically incorporated into hospital information systems: essentially having the data systems working for the clinician, rather than the converse. We believe there is substantial scope for developing technological support for communication that does not presuppose access to a keyboard.

Structuring the content of these free-form text records makes finding and using relevant information easier while also making information available for computerized decision-making and surveillance in healthcare. Information extraction enables generating these structured record drafts automatically, systematically, and almost instantaneously for clinicians to sign off. When applying information extraction to speech-recognized text, errors are, however, likely to multiply, but in in our previous study,[47] we have given empirical evidence for correcting such errors by using the phonetic similarity in the context of nursing handover.

CONCLUSION

This paper presented evidence for the feasibility of speech recognition and information extraction to support clinicians in entering textual information and unlock its content for computerized decision-making and surveillance in healthcare. A noise-cancelling lapel-microphone, digital voice recorder, and nursing vocabulary gave the best correctness in speech recognition. Information extraction was excellent in filtering out irrelevant text and identifying text relevant to two categories.

Similarly to the disagreements between human annotators, there was confusion surrounding the remaining three categories in information extraction. The benefits of this automation included storing all information; making the drafts available and accessible almost instantly to everyone with an authorized access to a particular patient's documents; and avoiding information loss, delays, and misinterpretations inherent to using a ward clerk, transcription services, or other third-party for record keeping.

CONTRIBUTORS

In the speech recognition part of this study, HS, MJ, PS, JB, and DE contributed to the design, data gathering, and experimentation for system correctness; the first four of these authors, together with LD, conducted all work related to evaluating clinicians' preferences. In the information extraction part, HS, MJ, and PS first conceptualized study and developed our annotated dataset. HS, LZ, RS, and LH designed and performed all experiments related to our automated system. HS, supported by PS and RS, wrote the first draft of the manuscript and all authors critically commented and revised it. All authors have read and approved the final version of the paper. This research was conducted while Raul Sirel was a visiting PhD student at NICTA.

FUNDING

NICTA is funded by the Australian Government through the Department of Communications and the Australian Research Council through the ICT Centre of Excellence Program. NICTA is also funded and supported by the Australian Capital Territory, the New South Wales, Queensland and Victorian Governments, the Australian National University, the University of New South Wales, the University of Melbourne, the University of Queensland, the University of Sydney, Griffith University, Queensland University of Technology, Monash University, and other university partners. This work was conducted using the Protégé resource, which is supported by grant GM10331601 from the National Institute of General Medical Sciences of the USA National Institutes of Health.

COMPETING INTERESTS

None.

PATIENT CONSENT

Obtained.

ETHICS APPROVAL

This study received the following ethics approvals: (1) Speech to Text Procedures for Health Communication in Nursing, Phase 2A: Clinical Speech Recognition. University of Western Sydney (UWS) Human Ethics Committee. Approval Number: H9597. (2) Concord, for the Clinical Handover project. Sydney Local Health District Human Research Ethics Committee (CRGH). Approval Number: HREC/11/CRGH/122.

PROVENANCE AND PEER REVIEW

Not commissioned; externally peer reviewed.

SUPPLEMENTARY MATERIAL

Supplementary material is available online at http://jamia.oxfordjournals.org/.

REFERENCES

1. Glaser SR, Zamanou S, Hacker K. Measuring and interpreting organizational culture. *MCQ*. 1987;1:173–98.
2. Windows into Safety and Quality in Health Care. Australian Commission on Safety and Quality in Healthcare 2008. http://goo.gl/wB0XZl (accessed 4 Feb 2014).
3. Implementation Toolkit for Clinical Handover Improvement. Australian Commission on Safety and Quality in Healthcare 2011. http://goo.gl/xH7Ncf (accessed 4 Feb 2014).
4. The OSSIE Guide to Clinical Handover Improvement. Australian Commission on Safety and Quality in Healthcare 2012. http://goo.gl/lvS7dc (accessed 4 Feb 2014).
5. Tran DT, Johnson M. Classifying nursing errors in clinical management within an Australian hospital. *Int Nurs Rev*. 2010;57:454–62.
6. Safe Handover: Safe Patients. Australian Medical Association 2006. http://goo.gl/9U8wjm (accessed 4 Feb 2014).
7. Johnson M, Sanchez P, Suominen H, *et al*. Comparing nursing handover and documentation: forming one set of patient information. *Int Nurs Rev*. 2014;61:73–81.
8. Pothier D, Monteiro P, Mooktiar M, *et al*. Pilot study to show the loss of important data in nursing handover. *Br J Nurs*. 2005;14:1090–3.
9. Matic J, Davidson P, Salamonson Y. Review: bringing patient safety to the forefront through structured computerisation during clinical handover. *J Clin Nurs*. 2011;20:184–9.
10. Poletic EB, Holly C. A systematic review of nurses' intershift handoff reports in acute care hospitals. *JBI Library Syst Rev*. 2010;8:121–72.
11. Holly C, Poletic EB. A systematic review on the transfer of information during nurse transitions in care. *J Clin Nurs*. 2013;23:17–18.
12. Patient Safety: Organizational Tools (see the sections of Safe Handover: Safe Patients and Handover (Handoff) Tools). The World Health Organization. http://goo.gl/rhmSCK (accessed 4 Feb 2014).
13. National Safety and Quality Health Service Standards. Australian Commission on Safety and Quality in Health Care 2011–2012. http://goo.gl/rSLHg0 (accessed 4 Feb 2014).
14. 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.
15. 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.
16. 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.
17. Alapetite A. Impact of noise and other factors on speech recognition in anaesthesia. *Int J Med Inform*. 2008;77:68–77.

18. Meystre SM, Savova GK, Kipper-Schuler KC, *et al*. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearb Med Inform*. 2008:128–44.
19. OECD.StatsExtracts. The Organisation for Economic Co-operation and Development 2009 (i.e., the most recent year that has almost all data available in 2014). http://stats.oecd.org/Index.aspx?DatasetCode=HEALTH_STAT (accessed 4 Feb 2014).
20. Dalianis H, Hassel M, Velupillai S. The Stockholm EPR Corpus—characteristics and some initial findings. *Proceedings of The 14th International Symposium for Health Information Management Research, ISHIMIR-09*; Kalmar, Sweden: 2009.
21. Suominen H, Lundgrén-Laine H, Salanterä S, *et al*. Information flow in intensive care narratives. *Proceedings IEEE International Conference on Bioinformatics and Biomedicine Workshops, BIBM 2009*; Washington DC, USA: IEEE, 2009:325–30.
22. Manor-Shulman O, Beyene J, Frndova H, *et al*. Quantifying the volume of documented clinical information in critical illness. *J Crit Care*. 2008;23:245–50.
23. Poissant L, Pereira J, Tamblyn R, *et al*. The impact of electronic health records on time efficiency on physicians and nurses: a systematic review. *J Am Med Inform Assoc*. 2005;12:505–16.
24. Hakes B, Whittington J. Assessing the impact of an electronic medical record on nurse documentation time. *Comput Inform Nurs*. 2008;26:234–41.
25. Banner L, Olney C. Automated clinical documentation: does it allow nurses more time for patient care? *Comput Inform Nurs*. 2009;27:75–81.
26. Callaway EC, Sweet CF, Siegel E, *et al*. Speech recognition interface to a hospital information system using a self-designed visual basic program: initial experience. *J Digit Imaging*. 2002;15:43–53.
27. Koivikko M, Kauppinen T, Ahovuo J. Improvement of report workflow and productivity using speech recognition—a follow-up study. *J Digit Imaging*. 2008;21:378–82.
28. Langer SG. Impact of speech recognition on radiologist productivity. *J Digit Imaging*. 2002;15:203–9.
29. Singh M, Pal TR. Voice recognition technology implementation in surgical pathology: advantages and limitations. *Arch Pathol Lab Med*. 2011;135:1476–81.
30. Allan J, Aslam J, Belkin N, *et al*. Challenges in information retrieval and language modeling: report of a workshop held at the Center for Intelligent Information Retrieval, University of Massachusetts Amherst, September 2002. *SIGIR Forum*. 2003;37:31–47.
31. Lovis C, Baud RH, Planche P. Power of expression in the electronic patient record: structured data or narrative text? *Int J Med Inform*. 2000;58–59:101–10.
32. Walsh SH. The clinician's perspective on electronic health records and how they can affect patient care. *BMJ*. 2004;328:1184–7.
33. Tange H. How to approach the structuring of the medical record? Towards a model for exible access to free text medical data. *Int J Biomed Comput*. 1996;42:27–34.
34. Tange H, Hasman A, de Vries Robbe P, *et al*. Medical narratives in electronic medical records. *Int J Med Inform*. 1997;46:7–29.
35. Kärkkäinen O, Eriksson K. Evaluation of patient records as part of developing a nursing care classification. *J Clin Nurs*. 2003;12:198–205.
36. Dawson L, Johnson M, Suominen H, *et al*. The usability of speech recognition technologies in clinical handover: a pre-implementation study. *J Med Syst*. 2014;38:1–9.
37. Williams JR. Guidelines for the use of multimedia in instruction. *Proceedings of the Human Factors and Ergonomics Society 42nd Annual Meeting*; 1998:1447–51.
38. Karat CM, Halverson C, Horn D, *et al*. Patterns of entry and correction in large vocabulary continuous speech recognition systems. *Proceedings of the SIGCHI conference on Human Factors in Computing Systems, CHI 1999*; New York, NY, USA: ACM, 1999:568–75.
39. Ayres RU, Martinás K, eds. 120 wpm for very skilled typist. In: On the reappraisal of microeconomics: economic growth and change in a material world. Cheltenham, UK & Northampton, MA, USA: Edward Elgar Publishing, 2005:41.
40. Dragon NaturallySpeaking Speech Recognition: More speed. More accuracy. More features! Spectronics. http://goo.gl/FZqHZj (accessed 4 Feb 2014).
41. Suominen H, Basilakis J, Johnson M, *et al*. Preliminary Evaluation of speech recognition for capturing patient information at nursing shift changes: accuracy in speech to text and user preferences for recorders. In: Suominen H, ed. The proceedings of the fourth international workshop on health document text mining and information analysis, Louhi. Sydney, NSW, Australia: NICTA, 2013.
42. Johnson M, Jeffries D, Nicholls D. Developing a minimum data set for electronic nursing handover. *J Clin Nurs*. 2012;31:331–43.
43. Johnson M, Jeffries D, Nicholls D. Exploring the structure and organization of information within nursing clinical handovers. *Int J Nurs Pract*. 2012;18:462–70.
44. Ogren PV. Knowtator: a Protégé plugin for annotated corpus construction. In: Proceedings of the 2006 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, NAACL 2006. Morristown, NJ, USA: ACL, 2006:273–5.
45. Lafferty J, McCallum A, Pereira F. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: Brodley CE, Pohoreckyj Danyluk A, eds. Proceedings of the 18th International Conference on Machine Learning, ICML 2001. Burlington, MA, USA: Morgan Kaufmann, 2001:282–9.
46. Suominen H, Pahikkala T, Salakoski T. Critical points in assessing learning performance via cross-validation. In: Honkela T, Pöllä M, Paukkeri M-S, Simula O, eds. Proceedings of the 2nd International and Interdisciplinary

Conference on Adaptive Knowledge Representation and Reasoning, AKRR 2008. Porvoo, Finland, 2008:9–22.
47. Suominen H, Ferraro G. Noise in speech-to-text voice: analysis of errors and feasibility of phonetic similarity for their correction. In: Karimi S, Verspoor K, eds. Proceedings of the Australasian Language Technology Association Workshop 2013, ALTA 2013. Brisbane, QLD, Australia: ACL, 2013: 34–42.
48. Reeves RM, Ong FR, Matheny ME, *et al*. Detecting temporal expressions in medical narratives. *Int J Med Inform*. 2013; 82:118–27.
49. NANDA International. Nursing diagnoses—definitions and classifications 2012–2014. 9th edn. West Sussex, UK: John Wiley & Sons, 2011.
50. Bulechek G, Butcher H, Dochterman J. Nursing Interventions Classification (NIC). 5th edn. St. Louis, MO: Mosby Elsevier, 2008.
51. Moorhead S, Johnson M, Maas M, *et al*. Nursing Outcomes Classification (NOC). 4th edn. St. Louis, MO: Mosby Elsevier, 2008.
52. Denny JC, Irani PR, Wehbe FH, *et al*. The KnowledgeMap project: development of a concept-based medical school curriculum database. *AMIA Annu Symp Proc*. 2003;2003: 195–9.
53. Denny JC, Smithers JD, Miller RA, *et al*. "Understanding" medical school curriculum content using knowledgemap. *J Am Med Inform Assoc*. 2003;10:351–62.
54. Meystre SM, Haug PJ. Comparing natural language processing tools to extract medical problems from narrative text. *AMIA Annu Symp Proc*. 2005;2005:525–9.
55. Shah NH, Bhatia N, Jonquet C, *et al*. Comparison of concept recognizers for building the Open Biomedical Annotator. *BMC Bioinformatics*. 2009;10(Suppl 9):S14.
56. Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. *J Am Med Inform Assoc*. 2010;17:229–36.
57. Stewart SA, von Maltzahn ME, Raza Abidi SS. Comparing Metamap to MGrep as a tool for mapping free text to formal medical lexicons. *Proceedings of the 1st International Workshop on Knowledge Extraction and Consolidation from Social Media (KECSM2012)*.
58. Aronson AR, Rindflesch TC. Query expansion using the UMLS Metathesaurus. *Proc AMIA Annu Fall Symp*. 1997:485–9.
59. Hofmann O, Schomburg D. Concept-based annotation of enzyme classes. *Bioinformatics*. 2005;21:2059–66.
60. National Research Council. Patient safety: achieving a new standard for care. Washington, DC: The National Academies Press, 2004.

AUTHOR AFFILIATIONS

[1]Machine Learning Research Group, NICTA, College of Engineering and Computer Science, The Australian National University, Faculty of Health, University of Canberra, and Department of Information Technology, University of Turku, Canberra, Australian Capital Territory, Australia

[2]Research Faculty of Health Sciences, Australian Catholic University, Sydney, New South Wales, Australia

[3]Machine Learning Research Group, NICTA, Canberra, Australian Capital Territory, Australia

[4]Centre for Applied Nursing Research (University of Western Sydney and South Western Sydney Local Health District), Sydney, New South Wales, Australia

[5]Institute of Estonian and General Linguistics, University of Tartu, Tartu, Estonia

[6]School of Computing, Engineering and Mathematics, University of Western Sydney, Sydney, New South Wales, Australia

[7]Machine Learning Research Group, NICTA, College of Engineering and Computer Science, The Australian National University, Faculty of Health, University of Canberra, Canberra, Australian Capital Territory, Australia

[8]The MARCS Institute, University of Western Sydney and Department of Linguistics, University of Sydney, Sydney, New South Wales, Australia

[9]Faculty of Social Sciences, University of Wollongong, Wollongong, New South Wales, Australia

[10]School of Languages and Linguistics, The University of Melbourne, Melbourne, Victoria, Australia