JMIR Medical Informatics

Impact Factor (2022): 3.2
Volume 3 (2015), Issue 2 ISSN 2291-9694 Editor in Chief: Christian Lovis, MD, MPH, FACMI

Contents

Original Papers

Viewpoint

Original Paper

Comprehensive Evaluation of Electronic Medical Record System Use and User Satisfaction at Five Low-Resource Setting Hospitals in Ethiopia

Binyam Tilahun[1,2], MSc, MPH; Fleur Fritz[1], PhD

[1]Institute of Medical Informatics, University of Münster, Münster, Germany

[2]Department of Health Informatics, University of Gondar, Gondar, Ethiopia

Corresponding Author:
Binyam Tilahun, MSc, MPH
Institute of Medical Informatics
University of Münster
Albert-Schweitzer-Campus 1, Gebäude A11
Münster, D-48149
Germany
Phone: 49 (251) 83 55262
Fax: 49 (251) 83 52259
Email: Binyam.Tilahun@uni-muenster.de

Abstract

Background: Electronic medical record (EMR) systems are increasingly being implemented in hospitals of developing countries to improve patient care and clinical service. However, only limited evaluation studies are available concerning the level of adoption and determinant factors of success in those settings.

Objective: The objective of this study was to assess the usage pattern, user satisfaction level, and determinants of health professional's satisfaction towards a comprehensive EMR system implemented in Ethiopia where parallel documentation using the EMR and the paper-based medical records is in practice.

Methods: A quantitative, cross-sectional study design was used to assess the usage pattern, user satisfaction level, and determinant factors of an EMR system implemented in Ethiopia based on the DeLone and McLean model of information system success. Descriptive statistical methods were applied to analyze the data and a binary logistic regression model was used to identify determinant factors.

Results: Health professionals (N=422) from five hospitals were approached and 406 responded to the survey (96.2% response rate). Out of the respondents, 76.1% (309/406) started to use the system immediately after implementation and user training, but only 31.7% (98/309) of the professionals reported using the EMR during the study (after 3 years of implementation). Of the 12 core EMR functions, 3 were never used by most respondents, and they were also unaware of 4 of the core EMR functions. It was found that 61.4% (190/309) of the health professionals reported over all dissatisfaction with the EMR (median=4, interquartile range (IQR)=1) on a 5-level Likert scale. Physicians were more dissatisfied (median=5, IQR=1) when compared to nurses (median=4, IQR=1) and the health management information system (HMIS) staff (median=2, IQR=1). Of all the participants, 64.4% (199/309) believed that the EMR had no positive impact on the quality of care. The participants indicated an agreement with the system and information quality (median=2, IQR=0.5) but strongly disagreed with the service quality (median=5, IQR=1). The logistic regression showed a strong correlation between system use and dissatisfaction (OR 7.99, 95% CI 5.62-9.10) and service quality and satisfaction (OR 8.23, 95% CI 3.23-17.01).

Conclusions: Health professionals' use of the EMR is low and they are generally dissatisfied with the service of the implemented system. The results of this study show that this dissatisfaction is caused mainly and strongly by the poor service quality, the current practice of double documentation (EMR and paper-based), and partial departmental use of the system in the hospitals. Thus, future interventions to improve the current use or future deployment projects should focus on improving the service quality such as power infrastructure, user support, trainings, and more computers in the wards. After service quality improvement, other departments (especially inter-dependent departments) should be motivated and supported to use the EMR to avoid the dependency deadlock.

(JMIR Med Inform 2015;3(2):e22) doi:10.2196/medinform.4106

KEYWORDS

electronic medical record; evaluation; low-resource settings; Ethiopia; DeLone and MacLean model

Introduction

Background

Electronic medical record (EMR) systems are increasingly being implemented in hospitals to achieve the following six aims of improved care: (1) safety, (2) effectiveness, (3) patient centeredness, (4) timeliness, (5) efficiency, and (6) quality [1]. However, many EMR systems which are technically sound for developers and healthcare managers, face resistance from users and may end up in failure [2].

Measuring success of an information system is difficult because success does not have a common explicit definition [3], and is dependent on expectations. The agreed hypothesis to say an information system is successful is when the implemented system is accepted to be used by the end user and the users are satisfied with the system. As a result, a number of researchers suggested that user satisfaction and system use are the primary determinants of user adoption, and therefore are suitable to measure information system success [4-7]. Mazzoleni et al describe health professionals' satisfaction towards EMR system as “essential to the survival” of the system [8]. The different implementation projects which were reported as failed have often been those in which the end users were dissatisfied or the core system functions were not properly used [9].

Even though most health professionals generally perceive that technology can help eliminate the burden of paper-based documentation and the unavailability of patient data in critical situations, they also get easily dissatisfied when an introduced system or support does not meet their expectations [10]. Pare et al [11] assessed the factors in clinical information system implementation success and reported on the necessity of identifying risk factors and involving health professionals starting from the development and pilot phase to avoid failure. Many factors affect the adoption of EMR systems and they vary within the system users, hospital setting, and the type of system in use [12,13].

EMR systems are also increasingly being incorporated into the healthcare organizations of developing countries that do not have well-developed infrastructure and well-trained technical personnel to use and manage the systems. As outlined by Sood [14], the determinant factors that affected the information system success in those settings might be different from factors in developed countries. Hence, rigorous evaluation studies on different health information system implementation projects in those settings are necessary to understand the critical success and failure factors. To date, since only a few reports are available [15], this study was conducted to fill this gap by evaluating the use and user satisfaction of an EMR system implemented in Ethiopia.

Rationale

In the current health sector development plan of Ethiopia, strengthening the health management information system and incorporating a computerized health management information system (HMIS) are priority policy plans to ensure health service quality and equity [16]. As a result, an HMIS documentation package was implemented to standardize the patient documentation and reporting systems in all health facilities of the country [17]. The implementation of this standardized documentation system was mandatory and a prerequisite to implement the EMR system to make sure that the paper workflow is in line with the EMR.

In 2009, the Ministry of Health, with support of the Tulane University Technical Assistance Project in Ethiopia (TUTAPE), started the development and implementation of a comprehensive EMR system for hospitals called SmartCare. The system was deployed in 5 hospitals in Addis Ababa [18] and other hospitals in regional cities. In 2013, the Ministry of Health adapted the system as a national EMR for all hospitals, and planned to scale it up to further hospitals and regions [19].

Even though the system developers claimed it the best EMR, it had not been thoroughly evaluated by a neutral investigator. As explained by Joaquin [20] and Fraser [21], information system projects, especially those which were developed by NGO’s, need thorough independent evaluation prior to scale-up to determine if system expansion is both worthwhile and feasible. Hence, this study conducted by an independent and neutral investigator, filled this gap by identifying the main factors which needed to be addressed before a costly expansion.

Objectives

The main objectives of this study are (1) to assess the current EMR use rate among health professionals, (2) to assess the use level of core EMR functions, (3) to determine the user satisfaction level of health professionals, and (4) to identify determinants factors of user satisfaction towards the EMR system in the study hospitals. The study was conducted in accordance with the Guidelines for Good Evaluation Practices in Health Informatics (GEP-HI) [22] and reported based on the Statement on Reporting of Evaluation Studies in Health Informatics (STARE-HI) [23].

Study Context

Organizational Setting

This evaluation study was conducted in 5 hospitals in Ethiopia. All are government hospitals located within a 15 km radius in Addis Ababa, the capital city of Ethiopia. Of the 5 hospitals, one is an 80-bed children and mothers care specialized hospital with inpatient and outpatient clinics, 2 are 300-bed teaching referral hospitals with different specialized clinics, and the remaining 2 are 200-bed general hospitals with both inpatient and outpatient services. All of the hospitals implemented the HMIS in 2009, and started EMR implementation in 2011 [24].

System Detail and System in Use

SmartCare is a portable, integrated EMR system that is currently used by three African countries (Zambia, Ethiopia, and South Africa), and presumably is the largest EMR system in use in Africa [25]. The system was designed in Africa to be robust in environments with limited infrastructure. The system also offers a touch screen interface to minimize the learning curve.

This comprehensive EMR system has different components (modules) that can be used in the various units of healthcare facilities (Figure 1). The main modules of SmartCare include registration, outpatient department, inpatient (to admit, follow, and discharge patients in wards), tuberculosis, pediatrics, HIV/AIDS (to manage patients in antiretroviral therapy clinics), antenatal care, postpartum, pharmacy, drug stock control, laboratory (to store and send laboratory results to the requesting clinic), eHMIS (to generate monthly, quarterly, and annual reports), and finance. Currently all but the financial module are implemented and used in the hospitals of this study.

Installation of the network, server infrastructure, and the EMR system at all hospital sites was conducted by TUTAPE. After implementation, 5 day-long onsite user training sessions were provided to all health professionals of each hospital. Additionally, TUTAPE computer and network experts are responsible to provide continuous on-call service for technical assistance during system failure.

On average, the SmartCare system has been in use in the 5 hospitals of this study since 2011. In parallel, the paper-based medical record system is also still in use which means that the health professionals are expected to document both on paper and within the EMR system. The plan of the government is to expand the system to the other 127 existing hospitals in the country after the pilot testing. Additionally, the government is training health informatics professionals to support the health management information system and implementation of EMR in the country [26].

Figure 1. Screenshot of the SmartCare EMR system currently implemented in Ethiopian hospitals. The main modules are listed on the left side of the image. The main modules have sub-modules that will be displayed upon clicking. The screenshot shown is displayed when "bed management" is clicked.

Methods

Study Design

A quantitative, cross-sectional study design based on a validated questionnaire was used to assess the use pattern, user satisfaction level, and determinant factors of SmartCare in 5 government hospitals located in Addis Ababa. To better understand the use and challenges of the system, we also assessed the current method of documentation on the EMR server and the fluctuation levels of power access in the study hospitals. The selected hospitals were chosen because the EMR system had been implemented for 3 years. Additionally, as 2 are teaching hospitals and 3 are general hospitals, representative hospital types were included.

Theoretical Background

This study was conducted based on the DeLone and MacLean (D&M) information system success evaluation model [4], a validated and the most commonly used information system success evaluation method among the informatics community [4]. The basic dimensions in this model are system quality, information quality, service quality, system use, user satisfaction, and net benefit.

For this evaluation, we chose 5 factors from the D&M model that are relevant for user satisfaction and use rate evaluation, by excluding net benefit. Instead of net benefit, user background was included as a determinant factor to be tested because many researchers reported it as a determinant factor especially in low literacy working environments [27]. Additionally, we assessed

the level of use of core EMR functions since they have been found to be a main factor of user satisfaction [28,29].

Participants and Sample Size

The participants of this study, health professionals across the 5 study hospitals, were categorized into the following four groups (1) physicians (doctors and health officers), (2) nurses (clinical and midwifery), (3) lab and pharmacists (laboratory and pharmacy professionals), and (4) HMIS (health data entry and management secretaries, information system officers, and data clerks). The sample size of 422 participants was calculated assuming a 95% Cl and 10% non-response rate. All health professionals, who were selected by a simple random sampling technique among their professional category and who also served for >6 months in the hospitals, were approached to complete the questionnaire.

Study Flow

This study began in January, 2014 after obtaining ethical clearance. The first step was to choose data collectors from each hospital and familiarize them with the objective and methodology of the research. Seven data collectors were chosen and trained on how to collect the questionnaire and the level of support they should give to avoid bias. The questionnaires were distributed to the participants by visiting them in their offices, mostly during the afternoon. To motivate participants, we provided one Samsung Galaxy III phone as a reward, by a lottery method, to all of the participants who fully completed the questionnaire. Data collection took place over a one-month period.

Outcome Measure and Evaluation Criteria

The outcome measures and corresponding evaluation criteria are shown in Textbox 1.

Textbox 1. Outcome measures and evaluation criteria.

Measure

- EMR use rate
  - Measured by the proportion of respondents who are currently using the system and server log analysis of current patient data documentation in the EMR system.
- Use rate of core EMR functions
  - Measured by the frequency of use of 12 core functionalities of the implemented EMR system.
- User satisfaction level
  - Evaluated by a median of 5 different user satisfaction measurement items based on a 5-point Likert scale (1-strongly agree to 5-strongly disagree).
- Factors determining user satisfaction
  - Measured by a binary logistic regression analysis of all user characteristic and organization factors.

Data Acquisition and Measurement

A questionnaire was developed based on standardized and previously validated instruments. The questions were divided into three categories. The first category, on the user background, had 15 questions about general socio-demographic data, computer training, and current use of the EMR system. Some of them were adapted from Mahmood et al [12], Lawrence and Low [30], and Igbara and Nachman [31]. The second category was designed to measure the perceived system quality, information quality, service quality, satisfaction, and expectation towards future benefits. To assess system quality, 7 items were used, whereas 10 were used to assess information quality, 9 to assess service quality, 5 to assess user satisfaction questions, and 3 to assess expectations towards future benefit. The items were adapted from Seddon et al [32] and Doll et al [33]. For the service quality, we added additional setting-specific questions to reflect the power interruptions and the computer access challenges faced in the study hospitals. The third category contained 12 questions on core EMR functions which were adapted from Moustafa et al [9] with amendments from EMR officers on the main core functionalities of the system.

A pretest of the questionnaire was conducted in a hospital that was not part of the study in which 5 physicians, 8 nurses, 3 lab/pharmacists, and 5 HMIS staffs participated. Based on the pretest results, 2 questions were amended for wording as they were reported to be unclear from a health professional's perspective. The reliability of the items was evaluated with Cronbach's alpha, and the values were all above .84, indicating satisfactory reliability of the questionnaire.

Data Analysis

Descriptive statistics were performed to describe the characteristics of the participants, EMR use rate, and user satisfaction. Binary logistic regression analysis was used to identify determinant factors of user satisfaction among the study participants.

The selected dependent variable for this study was "user satisfaction". In the questionnaire, participants were asked to rate their satisfactions on a 5-point Likert scale. Median and interquartile ranges (IQRs) with percentages were used. In the cross tabulation of our data, we found that responses of "very satisfied" and "very dissatisfied" were very low. Consequently, for the logistic regression, the 5-item scales were merged into

two groups from “very satisfied and satisfied” and "very dissatisfied and dissatisfied" to “satisfied” “dissatisfied”, respectively. After this dichotomization, the determinant factors were analyzed using binary logistic regression. All analyses were performed using SPSS Software version 22.

Ethical Statement

Ethical approval was granted by the Institutional Review Board (IRB) of the University of Gondar and the Addis Ababa City Administration Health Bureau. Permission for data collection was also obtained from each of the hospitals. The participants were informed about the study, its importance, and confidentiality of the information collected, as well their right to leave the study at any time. Written consent was obtained from participants in a form provided with the questionnaire and the procedure was approved by the IRB.

Results

Socio-Demographic Characteristics

Out of the 422 participants of this study, 96.2% (406/422) completed the questionnaire. Of all the questionnaires, 7 were not completed and 9 were not returned. The mean age of the participants was 34 years (SD 8.5). Of all the participants, 53.4% (217/406) were males, and the majority of the participants were nurses (43.3%, 176/406), followed by physicians (20.4%, 83/406), HMIS staff (18.2%, 74/406), and laboratory and pharmacy staff (18%, 73/406). The participants had a mean work experience of 8.6 years (SD 7.2) in the current hospital. The detailed socio-demographic characteristics of the respondents are shown in Table 1.

Table 1. Frequencies of the socio-demographic characteristics of the study participants (n=406).

Characteristics		Frequency	Relative frequency, %
Age of respondents, years			
	<30	161	39.7
	31-40	136	33.5
	41-50	84	20.7
	>50	25	6.2
Sex			
	Male	217	53.4
	Female	189	46.6
Work experience in current hospital, years			
	<5	166	41.1
	5-15	172	42.6
	<15	66	16.3
Professional category			
	Physicians	83	20.4
	Nurses	176	43.3
	Lab and pharmacists	73	18.0
	HMIS staff	74	18.2
Part-time job			
	Yes	106	26.1
	No	300	73.9

Study Findings

EMR Use Pattern and Related Characteristics

Among the respondents, 76.1% (309/406) started using the system immediately after implementation and user training. In this context, 'use' refers to a complete use of the EMR to document patient information, in addition to the patient card. Among them, the major proportion of users were HMIS staff 20.7% (64/309), followed by nurses 44.0% (136/309), laboratory and pharmacy staff, and physicians 18.4% (57/309). However, during the data collection, only 31.7% (98/309) of the professionals reported to use the system with the majority being HMIS staff (54.0%, 53/98), followed by nurses (33.6%, 33/98), lab and pharmacy staff (33.6%, 33/98), and physicians (7.1%, 7/98). Those who completely stopped using the EMR reported that they were using the computers for other purposes such as browsing the Internet, word processing, while others returned them to the store.

Training, Information Technology Qualification, and EMR Experience

In terms of training, 64.0% (260/406) participated in the EMR user training and 60.6% (246/406) had been previously trained on the HMIS implementation. Almost half of the staff (47.2%, 192/406) responded to having a "reasonable information technology (IT) qualification". The majority (76.1%, 309/406) did not have previous EMR experience, 20.4% (83/406) reported to having individual computer access in the office, and of those, the majority were HMIS staff (55.4%, 46/83). The others shared the computer access with 2 people (11.5%, 47/406), 3 people (16.7%, 68/406), 4 people (17.7%, 72/406) and the majority shared with more than 5 people (57.8%, 235/406), mainly nurses (94.8%, 223/235) (Table 2).

Table 2. Training, information technology (IT) qualification, experience, and current EMR use status of physicians, nurses, laboratory, pharmacy, and HMIS staff in the study participants (n=406).

Characteristics		n (%)NursesLaboratory& PharmacyHMIS			
		Physicians	Nurses	Laboratory & Pharmacy	HMIS
Computer access in hospital					
	Individual	11 (20.4)	12 (9.0)	14 (26.9)	46 (71.9%)
	For 2 practitioners	4 (7.4)	25 (18.7)	11 (21.2)	7 (10.9)
	For 3 practitioners	13 (24.4)	33 (24.6)	14 (26.9)	8 (12.5)
	For 4 practitioners	23 (42.6)	42 (31.3)	5 (9.6)	2 (3.1)
	For >5 practitioners	3 (5.6)	223 (16.4)	8 (15.4)	1 (1.6)
HMIS training					
	Yes	22 (26.5)	131 (74.4)	35 (47.9)	58 (78.4)
	No	61 (73.5)	45 (25.6)	38 (52.1)	16 (21.6)
IT qualification					
	No qualification	18 (21.7)	43 (24.4)	37 (50.7)	6 (8.1)
	Reasonable qualification	58 (69.9)	82 (46.6)	25 (34.2)	27 (36.5)
	Good qualification	7 (8.4)	51 (29.0)	11 (15.1)	41 (55.4)
SmartCare training					
	Yes	28 (33.7)	120 (68.2)	45 (61.6)	67 (90.5)
	No	55 (66.3)	55 (31.3)	28 (38.4)	7 (9.5)
Another EMR experience					
	Yes	35 (42.7)	16 (9.1)	25 (34.2)	20 (27.0)
	No	47 (57.3)	160 (90.9)	48 (65.8)	54 (73.0)
SmartCare use since implementation					
	Yes	57 (68.7)	136 (77.3)	52 (71.2)	64 (86.5)
	No	26 (31.3)	40 (22.7)	21 (28.8)	10 (13.5)
Current SmartCare use					
	Yes	7 (12.1)	33 (24.3)	5 (9.4)	53 (81.5)
	No	51 (87.9)	103 (75.7)	48 (90.6)	12 (18.5)

Usage Pattern From Server Log Analysis

The observations of the patient chart in the registration department of the hospitals showed that on average 184,594 patients per hospital have paper-based records. Out of those, 58.7% (108,450/184,594) were also available in the EMR system database. However, only 4.8% (5244/108,450) of those patients had a documented main diagnosis patient history in the EMR server.

In terms of infrastructure, hospitals on average had 61 computers and one server for the EMR system. Four of the hospitals had one or more IT staff members, however, they were not specifically hired for the EMR system. Rather, they primarily worked for the statistics office and they took the EMR system work as their secondary task. The number of IT staff, computers, and the number of medical records in the hospital paper and server databases is shown in Table 3.

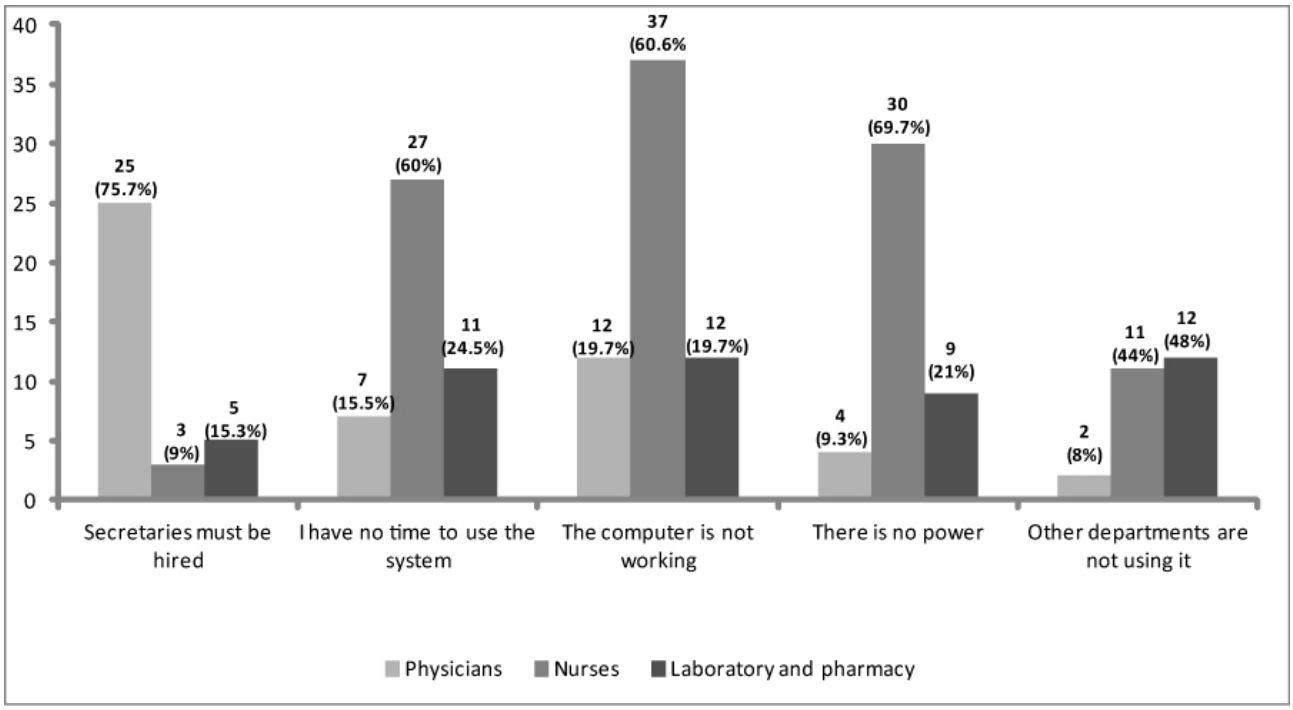
Table 3. Information technology (IT) professionals, reported number of EMR-dedicated computers, and patient records the Addis Ababa study hospitals from January-February, 2014. Most of the information available in the paper-based record system was not registered on the computer. Patient registration in the card room was done by data clerks while the main diagnosis was written by physicians or nursing assistants.

Characteristics	Hospital 1	Hospital 2	Hospital 3	Hospital 4	Hospital 5
Number of IT staffs	1	2	0	1	1
Number of computers for the EMR system	73	61	55	66	51
Number of patients with paper-based records	222,937	179,327	71,985	171,292	277,421
Number of patients registered in the EMR system	199,866	155,967	55,644	95377	35,398
Number of patients who have a main diagnosis in the EMR	7841	4848	4500	5323	3721

Main Reported Reasons of Not Using the System

An open-ended item question was provided to assess the main reasons for not using the system (Figure 2). Among the respondents who reported that they did not have the time to use it, the majority were physicians (75.7%, 25/33). Among the respondents who reported that the main reason for not using the system was that the computers were not working were nurses (60.6%, 37/61), while 69.7% (30/43) reported that power fluctuations in their department were the main obstacle. Of the respondents who reported that "the other departments were not using the system" was the main reason for not using the EMR, 48% (12/25) were laboratory and pharmacy staff.

Figure 2. Main reported reasons for not using the EMR system by physicians, nurses, and laboratory and pharmacy staff (n=197).



Use of the Main EMR Components in the Study Hospitals

Participants were also assessed on the use of the main components of the EMR as shown in Table 4. The most frequently used functionalities of the respondents were "find patients with certain characteristics" (45.6%, 141/309), and "produce patient summary" (45.9%, 142/309). The detailed use of the EMR components by the various categories of health professionals is shown in Table 4.

Table 4. Use of the basic EMR components (n=309).

Component	Physicians, n (%) N=56			Nurses, n (%) N=136			Lab and pharmacists, n (%) N=51			HMIS staff, n (%) N=63		
	U[a]	R[b]	F[c]	U	R	F	U	R	F	U	R	F
Find patient with certain characteristics	5 (8.9)	7 (12.5)	44 (78.5)	19 (13.9)	31 (22.7)	86 (63.2)	9 (17.6)	40 (78.4)	2 (3.9)	19 (30.1)	35 (55.5)	9 (14.2)
Create notes (history and physical exam)	45 (80.3)	6 (10.7)	5 (8.9)	51 (37.5)	65 (47.7)	20 (14.7)	10 (19.6)	4 (7.8)	37 (72.5)	23 (36.5)	33 (52.3)	7 (11.1)
Enter order (lab, radiology)	42 (75.0)	8 (14.2)	6 (10.7)	59 (43.3)	61 (44.8)	16 (11.7)	9 (17.6)	40 (78.4)	2 (3.9)			
Review/obtain lab and radiology results	4 (7.1)	9 (16.0)	43 (76.7)	27 (19.8)	64 (47.0)	45 (33.0)						
Update diagnosis	42 (75.0)	7 (12.5)	7 (12.5)	23 (16.9)	90 (66.1)	23 (16.9)						
Review currently received medications	6 (10.7)	5 (8.9)	45 (80.3)	20 (14.7)	88 (64.7)	28 (20.5)	7 (13.7)	6 (11.7)	38 (74.5)			
Write prescriptions	43 (76.7)	6 (10.7)	7 (12.5)	53 (38.9)	58 (42.6)	25 (18.3)						
Admit a patient	6 (10.7)	6 (10.7)	44 (78.6)	61 (44.8)	47 (34.5)	28 (20.5)						
Refer a patient	9 (16.0)	46 (82.1)	1 (1.9)	24 (17.6)	86 (63.2)	26 (19.11)						
View/schedule appointment for a patient	5 (8.9)	9 (16.0)	42 (75.0)	75 (55.1)	30 (22.0)	31 (22.7)				55 (87.3)	6 (9.5)	2 (3.1)
Communication using SmartCare's communication	10 (17.8)	5 (8.9)	41 (73.2)	59 (43.3)	33 (24.2)	44 (32.3)	8 (15.6)	6 (11.7)	37 (72.5)	58 (92.0)	2 (3.1)	3 (4.7)
Produce patient summary reports	10 (17.8)	5 (8.9)	41 (73.2)	62 (45.5)	25 (18.3)	48 (35.3)				3 (4.7)	7 (11.1)	53 (84.1)

[a]Unaware of the function (U)

[b]Rarely used the function (R)

[c]Frequently used the function (F)

EMR Satisfaction and Expectation for Future Benefit

Among the participants, 64.4% (199/309) responded to be dissatisfied with the use of the implemented EMR system. Of those dissatisfied, 24.6% (49/199) were physicians and 52,7% (105/199) were nurses. The participants responded with a strong disagreement towards the statements “The system helps me finish my task faster” (median=5, IQR=1) and “The system has a positive effect on quality of care” (median=5, IQR=1). Of all the professionals, 67.9% (210/309) preferred the paper-based record to the EMR system. Overall, the median satisfaction level was at the range of “Disagree” (median =4, IQR=1). The overall median responses with IQRs and percentages are shown in Table 5.

Table 5. Median satisfaction level of the study participants (n=309).

Characteristics of EMR Satisfaction	Physicians		Nurses		Laboratory and pharmacy		HMIS	
	n (% DA[a]) n=57	mean (IQR[b])	n (% DA) n=136	mean (IQR)	n (% DA) n=52	mean (IQR)	n (% DA) n=64	mean (IQR)
SmartCare help me to finish my work faster	49 (85.9)	5 (1)	89 (65.4)	4 (2)	46 (88.4)	4 (0)	2 (3.1)	2 (1)
EMR Improves my productivity	47 (82.4)	4 (0)	52 (38.2)	3 (1)	41 (78.8)	4 (0)	1 (1.5)	3 (1)
I prefer the EMR than the paper record	13 (22.8)	4 (1)	50 (36.7)	3 (2)	7 (13.4)	2 (1)	0 (0.0)	2 (2)
System has positive impact on quality of care	48 (84.2)	5 (1)	107 (78.6)	4 (0)	42 (80.7)	5 (1)	2(3.1)	2 (1)
Overall, I am satisfied with the EMR system	49 (85.9)	5 (1)	105 (77.2)	4 (0)	44 (84.6)	5 (0)	2 (3.1)	2 (1)
Category median score (95% CI), median (IQR)	5 (1)		4 (1)		4 (2)		2 (1)	
Over all median score (95% CI), median (IQR)	4.0 (1.0)							

[a]Disagree (DA)

[b]Interquartile range (IQR)

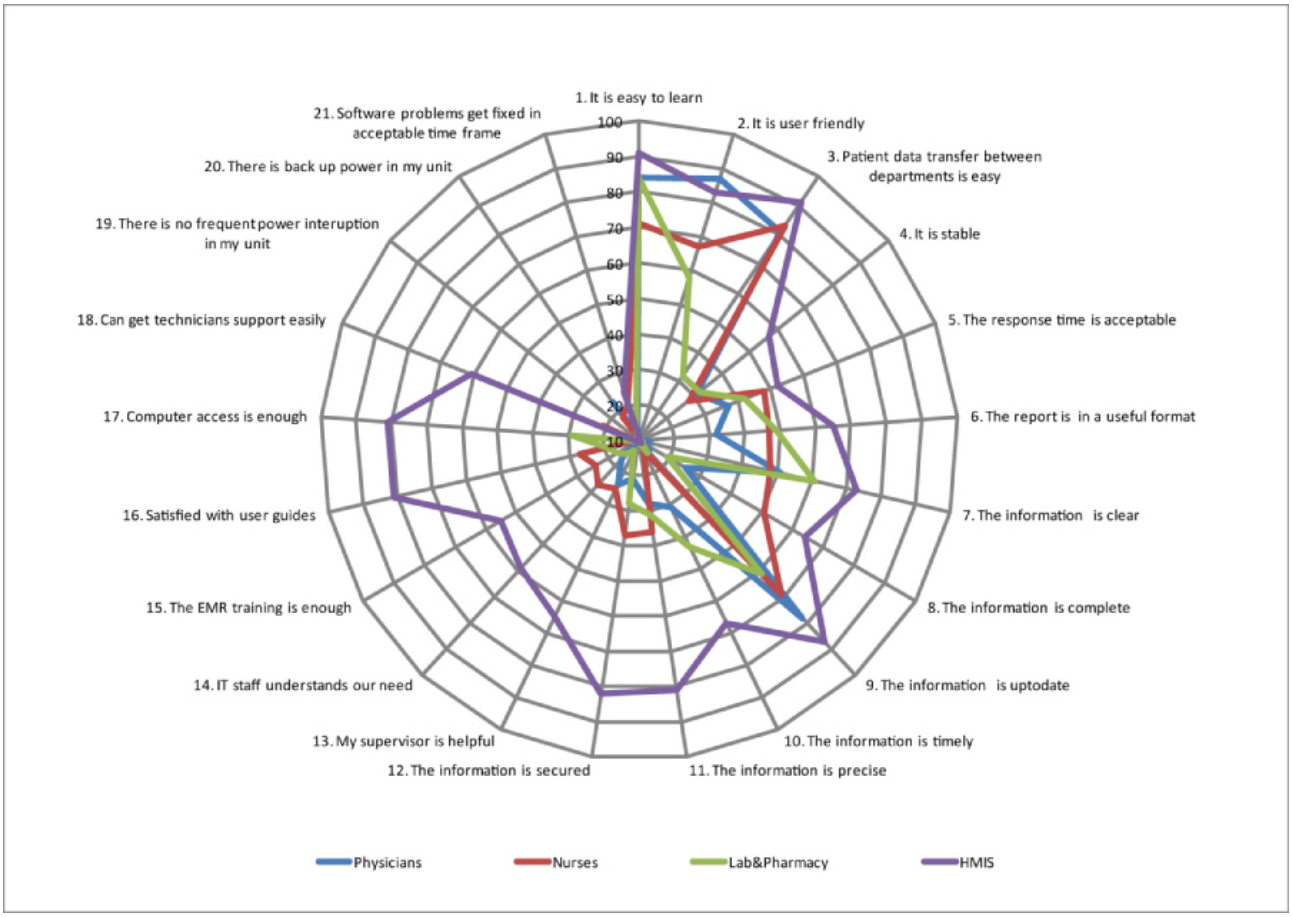
Perceived System, Information, and Service Quality

The respondents indicated an agreement with the statement that the implemented EMR system had an acceptable quality with an overall median score in the range of “agree” (median=2, IQR=0.5). Of the health professionals, 77.6% (240/309) found the system easy to learn, 61.1% (189/309) user friendly, 58.2% (180/309) stable, and 55.9% (173/309) found the response time acceptable. Overall, HMIS staff perceived the system to have more quality when compared to the other professional categories (median=2, IQR=1). All the criteria to measure system quality (1-5), information quality (6-12), and service quality (13-21) with percentages are shown in Figure 3.

The participants of this study also agreed with the statement that the information quality of the implemented system was acceptable with an overall median range of “agree” (median=2, IQR=1.0) with more acceptance within HMIS staff (median=1, IQR=1.0) and less agreement by physicians (median=4, IQR=2.0). Of all the participants, 93.5% (289/309) found the output of the system useful, and 76.6% (237/309) also found the information on the modules sufficient for their clinical practice. Only 42.3% (131/309) reported that they felt secure when using the system.

More dissatisfaction was reported with the service quality with an overall median score in the range of “disagree” (median=4.5, IQR=1.5). Of the respondents, 56.6% (175/309) believed their immediate supervisors were not helpful in using the EMR system, 71.8% (222/309) thought the IT support staff did not understand their needs, and 61.8% (191/309) believed the training given was not adequate. Additionally, 66.9% (207/309) responded that they could not get a computer in the ward during patient treatment, 66.0% (204/309) were unhappy with the computer technicians support, 73.4% (227/309) were also unhappy with the frequent power interruptions, and among them, 58.2% (180/309) responded that their department was not backed up by the standby generator.

Figure 3. Perceived system, information, and service quality of the study (n=309). The numbering on the label is to show in which category the criteria belong (1-5=System quality; 6-12=Information quality; 13-21=System quality). The main axis is the reported percentage. As shown in the figure, HMIS staff give more positive responses than phycisians and nurses .



Expectation Towards Future Benefit

Expectations of the respondents about the benefit of the EMR system for the patient, professionals, and the hospital were also assessed (Table 6). The majority of the respondents who never use the EMR system (91.7%, 89/97) and the current users (53.6%, 52/97) expect that the EMR system will be beneficial. Of the respondents who used to use the system, 45.3% (140/309) reported that the EMR system will be beneficial to the hospital. An independent sample *t* test revealed a statistically significant difference between “those who never use the system” and “those who used to use the system” ($P<.001$), but not significant between “those who used to use the system” and “current users”.

Table 6. Expectations of future benefits of EMR users and non-users (n=406).

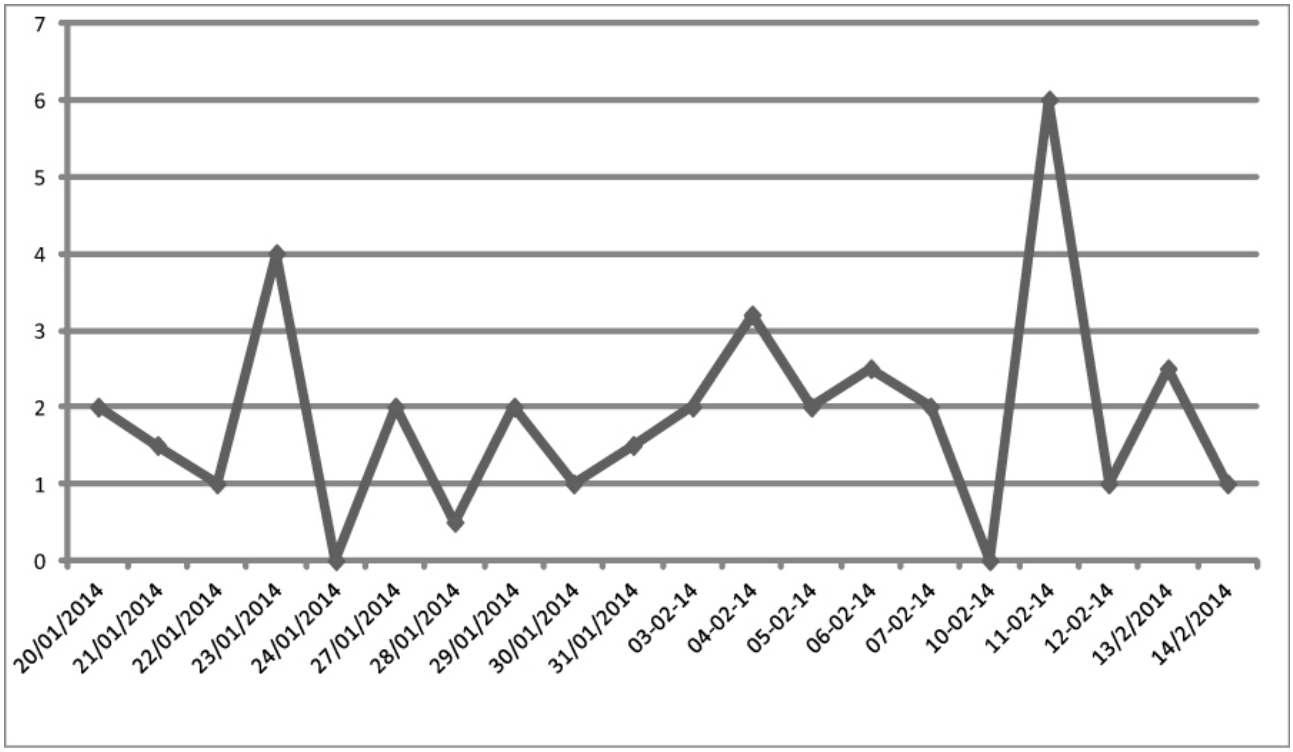
Characteristic	Those who never use EMR, n (%A[a]), n=97	Those who used to use, n (%A), n=309	Current users, n (%A), n=98
I expect EMR to benefit patients in the future	89 (91.7)	140 (45.3)	52 (53.0)
I expect EMR to benefit staff in the future	80 (82.4)	130 (42.0)	46 (46.9)
I expect EMR to benefit the hospital in the future	94 (96.9)	120 (38.9)	50 (51.0)

[a]Agree

Power Interruption Rate in the Study Hospitals

During the study period, the power fluctuation frequency in the study hospitals was monitored for one month. In one of the hospitals, the power supply was too weak to run the computers and the EMR was not functional during the study period. In the other hospitals, the daily hours of the power interruption were recorded, and the average of the 4 hospitals is shown in Figure 4. Accordingly, the mean time the power was interrupted for was 1.7 hours per day (SD 0.3). Of all the hospitals, 3 had a standby generator, but the generators could only reach the emergency and surgical departments and hence could not support the full running of the EMR in all of the departments.

Figure 4. The average daily power interruption rate in the four hospitals of Addis Ababa (January-February, 2014). Fluctuations were measured during work hours (8 hours) and days.



Determinants of EMR User Satisfaction

In the binary logistic regression analysis, the following were found to be significantly associated with EMR satisfaction (Table 7): computer access method in the hospital, IT qualification, EMR use, training, perceived system quality, perceived information quality, and perceived service quality. Respondents who reported to have good IT qualification were 3 times (adjusted OR 3.21, 95% CI 3.05-8.12) more likely to be satisfied with EMR systems when compared with those reported not having IT qualification, and those who had individual computer access were 4 times (adjusted OR 4.10, 95% CI 2.85-21.95) more likely to use the EMR than those who shared the computer with more than 5 people. Respondents who were currently using the system were 8 times more likely to be dissatisfied (adjusted OR 7.99, 95% CI 5.62-9.10), and those who received initial EMR training were 3 times (adjusted OR 3.04, 95% CI 2.05-8.12) more likely to be satisfied with EMR systems. The respondents who perceived the system to be of good quality were 2 times (adjusted OR 2.2 95% CI 1.34-3.09) more likely to be satisfied with the EMR, while those who perceived the information to be of good quality (adjusted OR 1.94, 95% CI 1.12-3.23) and those who perceived service to be of good quality (adjusted OR 8.23, 95% CI 3.23-17.01) were 2 and 8 times more likely to be satisfied, respectively. The result of respondent characteristics and its associated factors are shown in Table 7.

Table 7. Binary logistic regression analysis of factors associated with EMR satisfaction with a 95% CI and a significance level of $P<.05$ (n=309).

		EMR satisfaction, n (%)			
Characteristics		Dissatisfied	Satisfied	OR (95% CI)	AOR[a] (95% CI)
Computer access in hospital					
	Individual	18 (21.6)	65 (78.3)	17.77 (12.62-26.42)	4.10 (2.85-21.95)
	2 practitioners	27 (57.4)	20 (42.5)	11.85 (2.53-55.34)	2.91 (1.94-6.13)
	3 practitioners	51 (75.0)	17 (25.0)	5.33 (1.15-24.64)	1.5 (0.10-2.25)
	4 practitioners	66 (91.6)	6 (8.4)	1.45 (0.27-7.61)	1.11 (0.18-2.37)
	>5 practitioners	32 (94.1)	2 (5.9)	1.0	1.0
IT qualification					
	None	70 (86.4)	11 (13.6)	1.0	1.0
	Reasonable	77 (60.1)	51 (39.9)	4.21 (2.03-8.72)	2.11 (1.58-8.77)
	Good	52 (52.0)	48 (48.0)	5.87 (2.78-12.39)	3.21 (3.05-8.12)
SmartCare training					
	Yes	109 (53.1)	96 (46.9)	8.01(18.56-27.78)	3.04(2.31-7.34)
	No	89 (86.4)	14 (13.6)	1.0	1.0
Current SmartCare use					
	Yes	12 (12.2)	86 (87.8)	11.8 (6.68-26.86)	7.89 (3.62-9.10)
	No	187 (88.6)	24 (11.4)	1.0	1.0
Perceived system quality					
	Good	160 (68.9)	72 (31.1)	3.21 (1.34-4.23)	2.2 (1.34-3.09)
	Not good	23 (29.8)	54 (69.2)	1.0	1.0
Perceived information quality					
	Good	140 (65.1)	75 (34.9)	2.8 (1.23-3.78)	1.94 (1.12-3.23)
	Not good	32 (34.0)	62 (66.0)	1.0	1.0
Perceived system quality					
	Good	73 (90.1)	8 (9.8)	9.34 (4.23-18.34)	8.23 (3.23-17.01)
	Not good	197(86.4)	31(13.6)	1.0	1.0

[a]Adjusted OR

Unexpected Observations

There was one unexpected observation we want to point out. There was no dedicated information communication technology (ICT) support center in all of the hospitals despite the implementation of such an expensive server and network infrastructure. Of all 5 hospitals, 3 did not even have any professional IT technical support. The other 2 hospitals did have professional IT support; however, they were not primarily responsible for the EMR. The country is indeed training health information technicians at both the diploma and masters level to manage such systems but there were no health information technicians hired in all of the hospitals. Additionally, we observed that the technical support from TUTAPE was not sufficient during the study duration. The technical support team usually took 2-3 days to visit the hospital and solve the problem.

Discussion

Principal Findings

The purpose of the study was to assess the use and the determinant factors of user satisfaction in an implemented EMR system through a comprehensive assessment of usage patterns, user satisfaction, and determinant factors which affect the EMR system. This study had four main results.

First, the usage of the system was found to be low. To increase the use of the system, most of the physicians expressed the need to hire secretaries as the nurses expressed a lack of time to input information and the proper maintenance of computers, and the laboratory and pharmacy staff complained about the lack of use of the system by other departments. This result is not actually surprising given that health professionals are expected to do dual documentation both on the computer and on paper which makes them feel that transferring data to the EMR is not their duty. Hence, most of the doctors and the nurses were

complaining about the lack of time and most of them demanded secretaries to be hired so that the secretary can transfer the paper documentation to the EMR. The other aspect of the challenge is the partial use of the system in the hospital departments. Hospital work flows are interconnected, in which the activity of one department affects the other. Therefore, there is a need to implement the system to interdependent departments especially to those that are pillars of the hospital system (eg, laboratory, pharmacy, and radiology).

Second, the result of this study shows that only 2 of the core EMR functions were frequently used, 3 of the functionalities were never used, and participants were also unaware of 4 out of 12 core EMR functions. The participants enrolled in this study were all employed for >6 months. Thus, they were likely to be familiar with the main system functionalities. The low rate of use and awareness might be attributed to the general low adaptation rate of the EMR in the hospital and the training quality.

Third, the user satisfaction of the respondents was also found to be low. The majority of them reported to be dissatisfied with the use of the system. The main reported reason of the dissatisfaction was the service quality in the hospital. This was mainly due to lack of IT support, the shared computer access, and frequent power interruption.

Fourth, the user satisfaction was strongly correlated with service quality and system use. It was also moderately correlated with IT qualification, computer access method, perceived system, and information quality. Given the infrastructural and organizational challenges, such a strong correlation between service quality and use and user satisfaction was expected. However, the level of strength of the relationship was high, which shows that there was a need to improve the service quality and the current way of using the EMR in the hospitals.

Study Strengths and Weaknesses

Totally, 406 professionals (96.2% response rate) from 5 hospitals participated in this study. The response rate was very high when compared to other evaluation studies, which can be attributed to the use of data collectors from each hospital and our encouragement for participation by providing rewards. We addressed different potential system users by including physicians, nurses, laboratory, pharmacy, and HMIS staff.

There are some limitations to this study. Firstly, our data collection period was short (1 month). As pointed out by Meijden [3], system implementation is dynamic and the level of use rate fluctuates over time. Hence, this result may not exactly reflect the current status of the EMR implementation in those hospitals. The second limitation is with respect to the hypothesized determinant factors. This study is only based on the six constructs of the D&M model but there are many other organizational and human factors which affect acceptance and thus implementation success of an EMR system. Future studies can include those additional variables to have a complete picture of EMR success in those settings.

Results in Relation to Other Studies

There are different evaluation studies regarding the use and user satisfaction of health professionals with respect to EMR systems and different adoption rates were reported. For example, Laerum et al [34] reviewed EMR system use in 19 hospitals and reported that system use was low and only 2 out of 7 implemented functions were frequently used, which is in line with our results. Another similar study in Saudi Arabia [9] also reported that the use of different core EMR components were minimal. In that study, 54.9% of the physicians never used ≥1 of the 10 investigated core EMR functions. Mikkelsen et al also assessed the challenges of parallel documentation (EMR and paper-based records) and found that it is a source of dissatisfaction and inconsistency, which is similar to our study [35].

Alharthi et al [36] assessed physicians' satisfaction of an EMR system and reported only 40% of them were satisfied. A similar low satisfaction rate was reported in Malaysia [37], Oman [38], and Kenya [39] which all is similar to our result. However, a recent study by Jia-lin [40] in two big hospitals in China reported a satisfaction rate of 70.7%. This difference might be because of the infrastructural differences in the study setting hospitals. Another study by Palm et al [5] reported that medical secretaries were more satisfied than nurses and physicians and Moody et al [41] and others [35,39,40] reported that nurses were more satisfied than physicians with the use of EMR, which is similar to our result.

Common in most studies and in our study are the factors that affect the success of implementation of an EMR system. Palm et al [5], in his assessment of determinants of user satisfaction of EMR systems, reported that female gender, perceived system quality, usefulness, and service quality are strongly correlated with satisfaction. Similarly, another study by Chatzoglou et al [27] reported that user background, information quality, and service quality directly and positively affect user satisfaction. Consistent with many evaluation studies and models [3,5,12,26,42-44], system quality, information quality, and service quality are determinant factors for user satisfaction. However, in our study we found out that there was a strong correlation between user satisfaction and system quality. This difference might be due to the infrastructural challenge in our study hospitals in which there were frequent power interruptions and no dedicated ICT support centers.

We agree with Meijden [3] that the D&M's conceptual framework does not address different contingent factors for the success of an EMR system. In our study, user IT qualification and computer access methods were found to be significant determining factors but we were not able to accommodate them in the framework. These are also reported as determinant factors in other studies [9,45], but computer access method was a significant determinant factor in our study. We believe that this factor is significant for low-resource settings, given that most clinicians (43% of the physicians and 31% of nurses) shared one computer for ≥4 people. .

Meaning and Generalizability

Even though many hospitals implemented an EMR system in developing countries, very few evaluation studies exist on use,

user satisfaction, and factors affecting it. In this study we attempted to close this gap by assessing use and user satisfaction in low-resource setting hospitals, and we believe that the result will be helpful to health care managers and decision makers as an input for future EMR implementation or expansion projects. The ministry of health of Ethiopia plans to expand the EMR to all other hospitals, and we are hopeful that this study will help them as an input. As outlined above, our result shows that more emphasis must be given to service quality in implementing the EMR system to the other hospitals. Since our study includes both teaching and general hospitals as well different professionals in the hospitals, we believe that our findings are generalizable to other similar setting hospitals in developing countries.

Unanswered and New Questions

The informatics community perceives the D&M model as the best and most validated model to measure the success of an implemented information system [32,43,46-49]. However, as also stated by Meijden [3], the D&M's conceptual framework does not address all information system success factors. In our study results, we were unable to categorize the computer skill and experiences into the model. Hence, a more comprehensive model, which takes into account the different factors in low-resources setting hospitals, is necessary.

The other necessary but unaddressed aspect of the evaluation is the financial feasibility of the system. In the study hospitals, the system was implemented by external NGO-funding that provided a higher budget for EMR implementation than governmental funding. Therefore, we recommend the Ministry of Health to conduct more comprehensive, cost-benefit analyses which include a qualitative evaluation on the system before implementing the system to the other hospitals.

Conclusions

Health professionals' use of the EMR is low and they are generally dissatisfied with the service of the implemented system. The result of this study showed that the dissatisfaction was caused mainly and strongly by the poor service quality, the current practice of double documentation (EMR and paper-based), and partial departmental use of the system in the hospitals. Thus, future interventions to improve the current use or future deployment projects should focus on improving the service quality, such as power infrastructure, user support, trainings, and more computers in the wards. After service quality improvement, other departments (especially inter-dependent departments) may be motivated and supported to use the EMR to avoid the dependency deadlock. Further evaluation studies that include a cost-benefit analysis are recommended.

Acknowledgments

We would like to acknowledge Professor Martin Dugas for his valuable comments on the study design and final manuscript. We acknowledge support by Deutsche Forschungsgemeinschaft and Open Access Publication Fund of University of Muenster for supporting the final publication cost of this paper. We also would like to acknowledge anonymous reviewers, the data collectors, and all participants of this study.

Conflicts of Interest

None declared.

Authors' Contributions

Binyam Tilahun initiated the study, developed the study design, collected and analysed the data and wrote the manuscript. Fleur Fritz contributed to the study design and critically revising all versions of the manuscript. All authors read and approved the final manuscript.

References

1. Otieno GO, Hinako T, Motohiro A, Daisuke K, Keiko N. Measuring effectiveness of electronic medical records systems: towards building a composite index for benchmarking hospitals. Int J Med Inform 2008 Oct;77(10):657-669. [doi: 10.1016/j.ijmedinf.2008.01.002] [Medline: 18313352]
2. Garcia-Smith D, Effken JA. Development and initial evaluation of the Clinical Information Systems Success Model (CISSM). Int J Med Inform 2013 Jun;82(6):539-552. [doi: 10.1016/j.ijmedinf.2013.01.011] [Medline: 23497819]
3. Van Der Meijden MJ, Tange HJ, Troost J, Hasman A. Determinants of success of inpatient clinical information systems: a literature review. J Am Med Inform Assoc 2003;10(3):235-243 [FREE Full text] [doi: 10.1197/jamia.M1094] [Medline: 12626373]
4. DeLone W, MacLean E. The DeLone and McLean Model of Information Systems Success: A Ten-Year Update. J Manag Inf Syst 2003;19(4):9-30.
5. Palm JM, Colombet I, Sicotte C, Degoulet P. Determinants of user satisfaction with a Clinical Information System. AMIA Annu Symp Proc 2006:614-618 [FREE Full text] [Medline: 17238414]
6. Likourezos A, Chalfin DB, Murphy DG, Sommer B, Darcy K, Davidson SJ. Physician and nurse satisfaction with an Electronic Medical Record system. J Emerg Med 2004 Nov;27(4):419-424. [doi: 10.1016/j.jemermed.2004.03.019] [Medline: 15498630]

7. Ives B, Olson M, Baroudi J. Commun ACM. The measurement of user information satisfaction URL: https://archive.nyu.edu/bitstream/2451/14594/1/IS-82-27.pdf [accessed 2015-04-30] [WebCite Cache ID 6YBG4R3w5]
8. Mazzoleni MC, Baiardi P, Giorgi I, Franchi G, Marconi R, Cortesi M. Assessing users' satisfaction through perception of usefulness and ease of use in the daily interaction with a hospital information system. Proc AMIA Annu Fall Symp 1996:752-756 [FREE Full text] [Medline: 8947766]
9. Nour El Din MM. Physicians' use of and attitudes toward electronic medical record system implemented at a teaching hospital in saudi arabia. J Egypt Public Health Assoc 2007;82(5-6):347-364. [Medline: 18706293]
10. Chisolm DJ, Purnell TS, Cohen DM, McAlearney AS. Clinician perceptions of an electronic medical record during the first year of implementaton in emergency services. Pediatr Emerg Care 2010 Feb;26(2):107-110 [FREE Full text] [doi: 10.1097/PEC.0b013e3181ce2f99] [Medline: 20093997]
11. Paré G, Sicotte C, Jaana M, Girouard D. Prioritizing the risk factors influencing the success of clinical information system projects. A Delphi study in Canada. Methods Inf Med 2008;47(3):251-259. [Medline: 18473092]
12. Adam Mahmood M, Burn JM, Gemoets LA, Jacquez C. Variables affecting information technology end-user satisfaction: a meta-analysis of the empirical literature. International Journal of Human-Computer Studies 2000 Apr;52(4):751-771. [doi: 10.1006/ijhc.1999.0353]
13. Msukwa MKB. User Perceptions on Electronic Medical Record System (EMR) in Malawi. 2011. URL: http://www.medcol.mw/commhealth/mph/dissertations/martin%20msukwa_Approved.pdf [accessed 2015-05-02] [WebCite Cache ID 6YDyn85vQ]
14. Sood S, Nwabueze S, QMbarika V, Prakash N, Chatterjee S, Ray P, et al. Electronic Medical Records: A Review Comparing the Challenges in Developed and Developing Countries. 2008 Presented at: Proceedings of the 41st Hawaii International Conference on System Sciences -; 2008; Hawaii Int Conf p. 1-10 URL: http://www.computer.org/csdl/proceedings/hicss/2008/3075/00/30750248.pdf
15. Verbeke F, Ndabaniwe E, Van Bastelaere S, Ly O, Nyssen M. Evaluating the Impact of Hospital Information Systems on the Technical Efficiency of 8 Central African Hospitals Using Data Envelopment Analysis. Journal of Health Informatics in Africa 2013;1(1).
16. Federal Ministry of Health Health Management Information System (HMIS) / MonitoringEvaluation (M & E) Strategic Plan for Ethiopian Health Sector. Addis ababa; 2008 URL: http://phe-ethiopia.org/resadmin/uploads/attachment-58-Health_Managment_Information_System_%28HMIS%29.pdf [accessed 2015-05-02] [WebCite Cache ID 6YDz5ajAs]
17. Hirpa W, Tesfaye H, Nigussie F, Argaw H. Q Heal Bull. Implementation Of An Integrated Health Management Information System And Monitoring And Evaluation System In Ethiopia: Progress And Lessons From Pioneering Regions URL: http://www.who.int/healthmetrics/library/countries/ETH_HIS_LessonsLearned.pdf [accessed 2015-05-02] [WebCite Cache ID 6YDzEF8Ht]
18. Mengesha T. Electronic solutions for Ethiopian health sector. 2011. URL: https://www.theseus.fi/bitstream/handle/10024/36264/Mengesha_Tewodros.pdf?sequence=1 [accessed 2015-05-02] [WebCite Cache ID 6YDzJxymz]
19. Biruk S, Yilma T, Andualem M, Tilahun B. Health Professionals' readiness to implement electronic medical record system at three hospitals in Ethiopia: a cross sectional study. BMC Med Inform Decis Mak 2014;14:115 [FREE Full text] [doi: 10.1186/s12911-014-0115-5] [Medline: 25495757]
20. Nguyen L, Bellucci E, Nguyen LT. Electronic health records implementation: an evaluation of information system impact and contingency factors. Int J Med Inform 2014 Nov;83(11):779-796. [doi: 10.1016/j.ijmedinf.2014.06.011] [Medline: 25085286]
21. Fraser HSF, Blaya J. Implementing medical information systems in developing countries, what works and what doesn't. AMIA Annu Symp Proc 2010;2010:232-236 [FREE Full text] [Medline: 21346975]
22. Nykänen P, Brender J, Talmon J, de Keizer N, Rigby M, Beuscart-Zephir MC, et al. Guideline for good evaluation practice in health informatics (GEP-HI). Int J Med Inform 2011 Dec;80(12):815-827. [doi: 10.1016/j.ijmedinf.2011.08.004] [Medline: 21920809]
23. Talmon J, Ammenwerth E, Brender J, de Keizer N, Nykänen P, Rigby M. STARE-HI--Statement on reporting of evaluation studies in Health Informatics. Int J Med Inform 2009 Jan;78(1):1-9. [doi: 10.1016/j.ijmedinf.2008.09.002] [Medline: 18930696]
24. Vital Wave Consulting. Health Information Systems in Developing Countries URL: http://www.minsa.gob.pe/ogei/conferenciaops/Recursos/43.pdf [accessed 2015-05-02] [WebCite Cache ID 6YDzn7rjd]
25. Mengesha T. Electronic Solutions for Ethiopian Health Sector. 2011. URL: https://www.theseus.fi/bitstream/handle/10024/36264/Mengesha_Tewodros.pdf?sequence=1 [accessed 2015-05-18] [WebCite Cache ID 6YcuSM8xB]
26. Tilahun B, Zeleke A, Fritz F, Zegeye D. New bachelors degree program in health informatics in Ethiopia: curriculum content and development approaches. Stud Health Technol Inform 2014;205:798-802. [Medline: 25160297]
27. Chatzoglou PD, Fragidis LL, Doumpa T, Aggelidis P. Hospital Information System Evaluation. In: Hospital Information System Evaluation.: 10th International Conference on ICT in Health; 2012 Presented at: 10th International conference on Information and Communication technologies for health; Jully 12-14, 2012; Island, Grek p. 12-14.

28. Chen RF, Hsiao JL. An investigation on physicians' acceptance of hospital information systems: a case study. Int J Med Inform 2012 Dec;81(12):810-820. [doi: 10.1016/j.ijmedinf.2012.05.003] [Medline: 22652011]
29. Bokhari RH. The relationship between system usage and user satisfaction: a meta‐analysis. Journal of Ent Info Management 2005 Apr;18(2):211-234. [doi: 10.1108/17410390510579927]
30. Lawrence M, Low G. Exploring Individual User Satisfaction within User-Led Development. MIS Quarterly 1993 Jun;17(2):195-208. [doi: 10.2307/249801]
31. Igbaria M, Nachman SA. Correlates of user satisfaction with end user computing. Information & Management 1990 Sep;19(2):73-82. [doi: 10.1016/0378-7206(90)90017-C]
32. Seddon PB, Kiew M, Agency R. A partial test and development of delone and mclean's model of is success. Australia: Australian Journal of Information systems; 1995:90-109.
33. Doll BWJ, Torkzadeh G. The Measurement of End-User Computing Satisfaction. MIS Quarterly 1988 Jun;12(2):259-274. [doi: 10.2307/248851]
34. Laerum H, Ellingsen G, Faxvaag A. Doctors' use of electronic medical records systems in hospitals: cross sectional survey. BMJ 2001 Dec 8;323(7325):1344-1348 [FREE Full text] [Medline: 11739222]
35. Mikkelsen G, Aasly J. Concordance of information in parallel electronic and paper based patient records. Int J Med Inform 2001 Oct;63(3):123-131. [Medline: 11502428]
36. Alharthi H, Youssef A, Radwan S, Al-Muallim S, Zainab A. Physician satisfaction with electronic medical records in a major Saudi Government hospital. Journal of Taibah University Medical Sciences 2014 Sep;9(3):213-218. [doi: 10.1016/j.jtumed.2014.01.004]
37. Mohd Amin I, Sumarni Hussein S, Isa WARWM. Assessing User Satisfaction of using Hospital Information System (HIS) in Malaysia. 2011 Presented at: International Conference on Social Science and Humanity; 2011; Singapore p. 210-213 URL: http://www.ipedr.com/vol5/no1/45-H00097.pdf
38. Al Farsi M, West DJ. Use of electronic medical records in Oman and physician satisfaction. J Med Syst 2006 Feb;30(1):17-22. [Medline: 16548410]
39. Kipturgo MK, Kivuti-Bitok LW, Karani AK, Muiva MM. Attitudes of nursing staff towards computerisation: a case of two hospitals in Nairobi, Kenya. BMC Med Inform Decis Mak 2014;14:35 [FREE Full text] [doi: 10.1186/1472-6947-14-35] [Medline: 24774008]
40. Jia-lin L, Siru L, Fei L. Physician satisfaction with electronic medical record in a huge hospital (China). Stud Health Technol Inform 2013;192:920. [Medline: 23920694]
41. Moody LE, Slocumb E, Berg B, Jackson D. Electronic health records documentation in nursing: nurses' perceptions, attitudes, and preferences. Comput Inform Nurs 2004 Dec;22(6):337-344. [Medline: 15602303]
42. Bossen C, Jensen LG, Udsen FW. Evaluation of a comprehensive EHR based on the DeLone and McLean model for IS success: approach, results, and success factors. Int J Med Inform 2013 Oct;82(10):940-953. [doi: 10.1016/j.ijmedinf.2013.05.010] [Medline: 23827768]
43. Zaied ANH. An Integrated Success Model for Evaluating Information System in Public Sectors. Journal of Emerging Trends in Computing and Information Sciences 2012;3(6):814-825 [FREE Full text]
44. Cheung CS, Tong EL, Cheung NT, Chan WM, Wang HH, Kwan MW, et al. Factors associated with adoption of the electronic health record system among primary care physicians. JMIR Med Inform 2013;1(1):e1 [FREE Full text] [doi: 10.2196/medinform.2766] [Medline: 25599989]
45. Alquraini H, Alhashem AM, Shah MA, Chowdhury RI. Factors influencing nurses' attitudes towards the use of computerized health information systems in Kuwaiti hospitals. J Adv Nurs 2007 Feb;57(4):375-381. [doi: 10.1111/j.1365-2648.2007.04113.x] [Medline: 17291201]
46. Chin WW, Lee MKO. A proposed model and measurement instrument for the formation of IS satisfaction: the case of end-user computing satisfaction. : Twenty-First International Conference on Information Systems; 2000 Presented at: Proceedings of the twenty first international conference on Information systems; 2000; Atlanta,USA p. 553-563 URL: https://www.researchgate.net/profile/Wynne_Chin/publication/220268832_A_proposed_model_and_measurement_instrument_for_the_formation_of_IS_satisfaction_the_case_of_end-user_computing_satisfaction/links/09e4151492510e5cdb000000.pdf
47. Xiao L, Dasgupta S. Measurement of user satisfaction with web-based information systems: an empirical study. : Eighth Americas Conference on Information Systems; 2002 Presented at: Eighth Americas Conference on Information Systems; 2002; USA p. 1149-1155 URL: http://www.sighci.org/amcis02/CR/Xiao.pdf
48. Petter S, DeLone W, McLean E. Measuring information systems success: models, dimensions, measures, and interrelationships. Eur J Inf Syst 2008 Jun;17(3):236-263. [doi: 10.1057/ejis.2008.15]
49. Perezmira B. PhD Dissertation.: Louisiana State University and Agricultural and Mechanical College; 2010. Validity of Delone and McClean's Model of Information Systems Success at the Web Site Level of Analysis URL: http://etd.lsu.edu/docs/available/etd-04162010-001906/unrestricted/Perez-Mira_diss.pdf [accessed 2015-05-02] [WebCite Cache ID 6YE1BKQrw]

Abbreviations

D&M: DeLone and MacLean
EMR: Electronic medical record
HMIS: Health management information system
ICT: Information communication technology
IQR: Interquartile range
IRB: Institute Review Board
IT: Information technology
NGO: Non governmental organization

Edited by G Eysenbach; submitted 04.12.14; peer-reviewed by M Wong, F Verbeke; comments to author 05.02.15; revised version received 12.02.15; accepted 17.02.15; published 25.05.15.

Please cite as:
Tilahun B, Fritz F
Comprehensive Evaluation of Electronic Medical Record System Use and User Satisfaction at Five Low-Resource Setting Hospitals in Ethiopia
JMIR Med Inform 2015;3(2):e22
URL: http://medinform.jmir.org/2015/2/e22/
doi:10.2196/medinform.4106
PMID:26007237

Viewpoint

Balancing the Interests of Patient Data Protection and Medication Safety Monitoring in a Public-Private Partnership

Nancy A Dreyer[1]; Stella Blackburn[2], MBBS; Valerie Hliva[3]; Shahrul Mt-Isa[4]; Jonathan Richardson[5]; Anna Jamry-Dziurla[6]; Alison Bourke[7]; Rebecca Johnson[8]

[1]Quintiles Real-World & Late-Phase Research, Scientific Affairs, Cambridge, MA, United States
[2]Quintiles Real-World & Late-Phase Research, Reading, United Kingdom
[3]Quintiles Real-World & Late-Phase Research, St. Prex, Switzerland
[4]School of Public Health, Imperial College London, London, United Kingdom
[5]Newcastle University, Newcastle upon Tyne, United Kingdom
[6]Poznan University of Medical Sciences, Poznan, Poland
[7]Cegedim Strategic Data, London, United Kingdom
[8]International Alliance of Patients' Organizations, London, United Kingdom

Corresponding Author:
Nancy A Dreyer
Quintiles Real-World & Late-Phase Research
Scientific Affairs
201 Broadway
5th floor
Cambridge, MA, 02139
United States
Phone: 1 617 715 6810
Fax: 1 617 206 9464
Email: nancy.dreyer@quintiles.com

Abstract

Obtaining data without the intervention of a health care provider represents an opportunity to expand understanding of the safety of medications used in difficult-to-study situations, like the first trimester of pregnancy when women may not present for medical care. While it is widely agreed that personal data, and in particular medical data, needs to be protected from unauthorized use, data protection requirements for population-based studies vary substantially by country. For public-private partnerships, the complexities are enhanced. The objective of this viewpoint paper is to illustrate the challenges related to data protection based on our experiences when performing relatively straightforward direct-to-patient noninterventional research via the Internet or telephone in four European countries. Pregnant women were invited to participate via the Internet or using an automated telephone response system in Denmark, the Netherlands, Poland, and the United Kingdom. Information was sought on medications, other factors that may cause birth defects, and pregnancy outcome. Issues relating to legal controllership of data were most problematic; assuring compliance with data protection requirements took about two years. There were also inconsistencies in the willingness to accept nonwritten informed consent. Nonetheless, enrollment and data collection have been completed, and analysis is in progress. Using direct reporting from consumers to study the safety of medicinal products allows researchers to address a myriad of research questions relating to everyday clinical practice, including treatment heterogeneity in population subgroups not traditionally included in clinical trials, like pregnant women, children, and the elderly. Nonetheless, there are a variety of administrative barriers relating to data protection and informed consent, particularly within the structure of a public-private partnership.

(JMIR Med Inform 2015;3(2):e18) doi:10.2196/medinform.3937

KEYWORDS

pharmacovgilance; pregnancy; Internet; public-private partnerships; data protection; ethics

Introduction

First Do No Harm

The Declaration of Helsinki extends the ancient medical tenet of "*Primum non nocere*", "first do no harm", and provides protection to human subjects of medical research by establishing ethical principles to ensure that medical research can never take precedence over the rights and interests of individual research subjects [1]. While laudable, harm can also occur by over-zealous interpretation of rules and regulations that overcomplicate studies, while adding little, or nothing, to the protection of subjects. The European Union (EU) "EU Data Protection Directive" by the European Commission (EC) (European Directive 95/46 EC) was intended to enable personal data "to flow freely from one Member State to another", while safeguarding the fundamental rights of individuals, yet its implementation into national law has given rise to a myriad of interpretations, making multi-country studies challenging.

Medication Safety in Pregnancy

Consider, as an example, the importance of understanding which medications can be safely used during pregnancy, especially during the first trimester, since some exposures at this time may have teratogenic potential [2,3]. Inclusion of pregnant women in preclinical randomized controlled trials is generally considered unethical due to the unknown risks which may be posed to the developing fetus, and as such, pregnant patients are often excluded unless the medicine is specifically for a pregnancy related condition. Consequently, safety data for pregnancy outcomes must be collected after licensing via noninterventional observational studies, which often utilize pharmacoepidemiologic techniques to analyze large databases, such as electronic health records to look for rare events such as specific congenital anomalies. However, these databases may not contain information about lifestyle and other factors, which may also affect the outcome of pregnancy, or may not contain adequate details concerning concomitant risk factors. These omissions could bias study interpretation. Hence, the development and testing of alternative methods of data collection for pharmacovigilance is important.

Here, we describe the legislative challenges in data ownership and barriers to approval faced by a public-private partnership in conducting an observational study of self-reported maternal medication use and pregnancy outcomes.

Our Experiences

Example of Challenges of Data Ownership and Barriers to Approval

This observational study of direct-to-consumer data collection on various exposures during pregnancy was conducted through a public-private partnership known as the Pharmacoepidemiological Research on Outcomes of Therapeutics by a European ConsorTium (PROTECT) [4], which was coordinated by the European Medicines Agency. The PROTECT project received support from the Innovative Medicines Initiative (IMI) Joint Undertaking, which included financial contribution from the EU's Seventh Framework Programme (FP7/2007-2013) and in-kind contribution from the European Federation of Pharmaceutical Industries and Associations. PROTECT consisted of 35 partners including pharmaceutical companies, academic organizations, national and international regulatory agencies, patient organizations, and other interested parties, and the IMI has now extended this public-private partnership model to address other important public health concerns [5].

While other PROTECT work packages focused on methodological challenges using existing data sources, we explored digital technologies for frequent and timely data collection from consumers for the purposes of determining whether this is a viable alternative as a pharmacovigilance tool.

This study was conducted according to the current best practices for noninterventional drug safety research including full protocol, specification of analytic methods and data to be collected, and a description of the plan for protecting human subjects [6]. Pregnant women were invited to participate via Internet or using an automated telephone response system (Interactive Voice Response System). Information was collected via a secure website from women in Denmark, the Netherlands, Poland, and the United Kingdom (UK) who identified themselves as pregnant, and were recruited through websites, emails, leaflets, television, and social media platforms. Health care professionals were not involved directly in study recruitment or promotion. Data were collected on prescription, nonprescription and herbal medications, recreational drug use, age, ethnicity, and lifestyle factors. Data were treated with strict confidential measures; for example, contact details were key-coded and deleted after the study end, and medical data were stored on a separate, secure server with restricted physical and password access. Local academic centers and a national health system entity served as country study leads, and notified the local ethics committee and data protection agencies. Regulatory and data protection submissions were performed according to the local requirements in the participating countries.

Some Examples of Variations by Country

There was substantial variation in the requirements for ethical review. Table 1 shows the differences in protocol requirements and the length of time needed for ethical and data protection review in each country and by the European Medicines Agency.

Table 1. Country specific protocol differences; ethical and data protection requirements and timing.

Protocol differences	Denmark	Netherlands	Poland	United Kingdom	European Medicines Agency
Country lead	Statens Serum Institute	University of Groningen	Poznan University of Medical Sciences	Newcastle University	N/A[a]
Minimum age (years)	18	18	18	16	N/A[a]
Informed consent	Electronic only, IVRS[b] not acceptable	Both Internet and IVRS[b] possible	Written informed consent required in addition to Internet and IVRS[b] informed consent	Both Internet and IVRS[b] possible	N/A[a]
Consent for individual record linkage	Required for study entry	N/A[a]	N/A[a]	Separate consent requested	N/A[a]
Ethical approval timing	Not required	Waiver (certificate of nonobjection)	1 week	3 weeks	N/A[a]
Time for data protection approval	~3 months	1 day	9 months	2 weeks	3 months opinion, 5 months prior check

[a] N/A=not applicable

[b] IVRS = Interactive Voice Response system

Some Examples of Variations by Country

Denmark did not require ethical review for an observational study. In the Netherlands, a waiver (literally, a certification of nonobjection) was granted since the personal identifiers were securely retained and maintained separately from study analysis files. In Poland and the UK, ethics submission required submitting the study protocol and all study documents (informed consent, questionnaires, etc) and other administrative information.

It is also worth noting the differences between countries in enrollment requirements and informed consent. Although the study was designed to give the choice of participating by phone or Internet to facilitate recruitment of low-income women, one country required all participants to enroll on the Internet before being able to respond by phone, and another required printing and mailing written informed consent in addition to consent by phone or Internet.

Formal notifications were required for data protection. The European Medicines Agency, as required under Article 27 of Regulation (EC) number 45/2001, submitted a notification for prior check with the European Data Protection Supervisor (EDPS) in October 2010. The EDPS opinion was that, since all study partners were involved in the development of the protocol and all could decide on the "means and purposes of the processing of personal data" and review results, all study partners effectively determined the purposes of the collection of the data and were "joint controllers". As a result of this ruling, a formal memorandum was prepared detailing each partner's role and participation in the study, responsibilities to the study and other partners, and to data protection. It took about 14 months to get these agreements in place, since they required agreement from all study partners. After these provisions were in place, the EDPS confirmed that the processing operations would not involve any breach of Regulation (EC) No 45/2001.

In the Netherlands, approval of data protection was granted on the same day the request was submitted, and review was also relatively quick in the UK and in Denmark. However, review by the Polish Data Protection Agency took 9 months and required submission of special items including the characteristics of the Personal Data Administrator, the technical and organizational conditions, and how those conditions would be fulfilled to comply with Polish legislation.

Results

Data collection for this study closed in the first quarter of 2015. Analyses examining the type of information reported by respondents are in progress, including comparisons of self-reported data with that available from electronic medical records and with the Danish National Prescription Registry. Analyses will be completed in 2015.

Discussion

Benefits and Challenges of Direct-to-Consumer Health Research Findings

Using direct reporting from consumers to study the safety of medicinal products allows researchers to address a myriad of research questions relating to everyday clinical practice, including treatment heterogeneity in population subgroups not traditionally included in clinical trials, like pregnant women, children, and the elderly. Internet-based studies such as this may also be useful for studying illicit drug use and other risky behaviors, since there is some evidence suggesting that patients will tell computers things that they might not tell health care professionals [7]. These studies can be supplemented with clinical validation and pharmacy prescription data, but

direct-to-patient data collection may provide additional information about potentially harmful exposures that would not have been recorded elsewhere, and consequently would not be available to researchers. Nonetheless, there are a variety of administrative barriers, including obtaining informed consent for subjects participating by phone or Internet. The variations in informed consent requirements encountered here largely reflect challenges of recruitment without intervention of health care professionals, and are one of many complexities faced by Ethics Committees from use of emerging technologies [8].

Added Complexities of Public-Private Partnerships

There were substantial barriers due to the nature of the funding structure, in addition to the challenges typically encountered in conducting direct-to-patient medical research. Public-private partnerships like this IMI project are becoming more prevalent as desirable funding mechanisms for research on the safety of medications and medical devices used in everyday clinical practice, for example, IMI Get Real [9] in Europe and the Food and Drug Administration's efforts to build a postmarket National Medical Device Safety System in the United States [10]. In fact, at this time, the IMI is Europe's largest public-private initiative aiming to speed up the development of better and safer medicines for patients. With these large efforts come tremendous opportunities, but also substantial additional work relating to partnership governance, including shared liability. In this study, for example, assuring compliance with data protection requirements took about two years, which delayed data collection, reduced the overall time available for study conduct, and required substantial investment of legal and administrative time over and above any traditional research project. Moreover, most countries did not initially recognize the status of joint controller, arguing that only two partners had control of personal data, those who handled data collection and those who conducted study analyses. The concept that all parties to a research study must bear the full legal burden of being joint controllers, which includes accepting responsibility for legal damages regardless of culpability, needs updating. Fortunately, in this case all partners agreed to accept joint controller status, but refusal by one or more partners, or refusal by a country to accept that a person, agency, or institution had this status and/or to refuse a notification, could jeopardize other such collaborations.

The text of the proposed data protection regulation, which was endorsed by the European Parliament at its first reading in March 2014, if adopted into law, will do little to improve the situation [11]. The joint controller status still exists and although a single "competent" supervisory authority of the EU territory of the researcher's main establishment can be requested to certify that the processing of personal data complies with the regulation, amendments to the proposed regulation require cooperation of supervisory authorities from other Member States. At the same time, supervisory authorities in disagreement with decisions are allowed the right of appeal to the European Data Protection Board. Uncertainty remains as to how this "cooperation" mechanism will operate to give much needed consistency. Moreover, the proposed regulation allows for multiple codes of conduct to be developed and approved by the supervisory authority of individual Member States and/or the European Commission, once again opening the door for disharmonized interpretations, now with much higher stakes since fines relating to failure to comply with the regulation can be as high as €100 million or 5% of annual worldwide turnover [11].

Data protection legislation is intended to allow freedom of movement of data, while protecting people from the theoretical harm of disclosure of personal data. This theoretical harm of disclosure of data that could be linked to an individual needs to be balanced against the potential for actual harm that could result from failure to identify safety signals in a timely fashion. Further, issues of data protection which require joint controller status to be shared among multiple parties may discourage participation, and might even drive health research away from regions of most interest to areas with potentially weaker protection of patient privacy and medication use that is quite different [4,12]. The potential financial consequences are considerable for an enterprise and may mean that companies or institutions may be reluctant to join consortia where the negligent actions of one partner could have such huge repercussions on the others, thus weakening the value of the public-private partnership investment.

Acknowledgments

The research leading to these results was conducted as part of the PROTECT consortium [4], which is a public-private partnership coordinated by the European Medicines Agency. The PROTECT project has received support from the IMI Joint Undertaking under Grant Agreement number 115004, resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007-2013) and in-kind contributions from the European Federation of Pharmaceutical Industries and Associations. The views expressed are those of the authors only. In Poland, the scientific work was cofinanced from the 2011-2014 allocation of funds for international projects from the Ministry of Science and Higher Education. Quintiles also provided funding for one senior scientist. The authors wish to thank A Latos-Bieleńska, Department of Medical Genetics, Poznan University of Medical Sciences; J-P Balling, Lundbeck; L Comisky, Genzyme; M Laursen, Statens Serum Institut; Omer Mol, Genzyme; B Patel, Amgen; S Stevens, Institute of Cellular Medicine, Newcastle University; S Thomas, Newcastle University; LTW de Jong-van den Berg, University of Groningen; and AP Zetstra-Van der Woude, University of Groningen.

Conflicts of Interest

None declared.

References

1. Williams JR. The Declaration of Helsinki and public health. Bull World Health Organ 2008 Aug;86(8):650-652 [FREE Full text] [Medline: 18797627]
2. Margulis AV, Mittleman MA, Glynn RJ, Holmes LB, Hernandez-Diaz S. Effects of gestational age at enrollment in pregnancy exposure registries. Pharmacoepidemiology and Drug Safety 2015 (forthcoming). [doi: 10.1002/pds.3731]
3. Mitchell AA. Systematic identification of drugs that cause birth defects--a new opportunity. N Engl J Med 2003 Dec 25;349(26):2556-2559. [doi: 10.1056/NEJMsb031395] [Medline: 14695418]
4. Innovative medicines initiative PROTECT. URL: http://www.imi-protect.eu/partners.shtml [accessed 2015-03-21] [WebCite Cache ID 6XC4RYedV]
5. IMI. ADVANCE accelerated development of vaccine benefit-risk collaboration in Europe URL: http://www.imi.europa.eu/content/advance [accessed 2015-03-21] [WebCite Cache ID 6XC58eJzl]
6. Epstein M, International Society of Pharmacoepidemiology. Guidelines for good pharmacoepidemiology practices (GPP). Pharmacoepidemiol Drug Saf 2005 Aug;14(8):589-595. [doi: 10.1002/pds.1082] [Medline: 15918159]
7. Beauclair R, Meng F, Deprez N, Temmerman M, Welte A, Hens N, et al. Evaluating audio computer assisted self-interviews in urban South African communities: Evidence for good suitability and reduced social desirability bias of a cross-sectional survey on sexual behaviour. BMC Med Res Methodol 2013;13:11 [FREE Full text] [doi: 10.1186/1471-2288-13-11] [Medline: 23368888]
8. Grady C. Enduring and emerging challenges of informed consent. N Engl J Med 2015 Feb 26;372(9):855-862. [doi: 10.1056/NEJMra1411250] [Medline: 25714163]
9. IMI. Get Real URL: http://www.imi-getreal.eu/ [accessed 2015-03-21] [WebCite Cache ID 6XC4k7CYO]
10. Daniel G, McClellan M, Colvin H, Aurora P. The Brookings Institute. Washington DC; 2015. Strengthening patient care: Building an effective national medical device surveillance system URL: http://www.brookings.edu/~/media/research/files/papers/2015/02/23-medical-device-policy-surveillance/med-device-reportweb.pdf [accessed 2015-03-24] [WebCite Cache ID 6XGpLZs4X]
11. European Commission. Proposal for the EU general data protection regulation. 2012. URL: http://ec.europa.eu/justice/data-protection/document/review2012/com_2012_11_en.pdf [accessed 2015-04-08] [WebCite Cache ID 6Xdx8YP28]
12. Dreyer NA, Eloff BC, Kirklin JK, Naftel DC, Toovey S. Registries for evaluating patient outcomes: A user's guide. Rockville, MD: US Agency for Healthcare Research & Quality; 2014 Apr. Public-private partnerships URL: http://www.quintiles.com/library/brochures/ahrq-registries-for-evaluating-outcomes-a-users-guide [accessed 2015-04-08] [WebCite Cache ID 6XdxNaKQW]

Abbreviations

EC: European Commission
EDPS: European Data Protection Supervisor
EU: European Union
IMI: Innovative Medicines Initiative
PROTECT: Pharmacoepidemiological Research on Outcomes of Therapeutics by a European ConsorTium
UK: United Kingdom

Edited by G Eysenbach; submitted 12.10.14; peer-reviewed by A Ahmadvand, D Schmidt; comments to author 19.02.15; revised version received 02.03.15; accepted 16.03.15; published 15.04.15.

Please cite as:
Dreyer NA, Blackburn S, Hliva V, Mt-Isa S, Richardson J, Jamry-Dziurla A, Bourke A, Johnson R
Balancing the Interests of Patient Data Protection and Medication Safety Monitoring in a Public-Private Partnership
JMIR Med Inform 2015;3(2):e18
URL: http://medinform.jmir.org/2015/2/e18/
doi:10.2196/medinform.3937
PMID:25881627

Original Paper

Benchmarking Clinical Speech Recognition and Information Extraction: New Data, Methods, and Evaluations

Hanna Suominen[1,2,3,4], BSc, MSc, PhD; Liyuan Zhou[1,2], MSc; Leif Hanlen[1,2,3], BSc, BEng, PhD; Gabriela Ferraro[1,2], PhD

[1]Canberra Research Laboratory, Machine Learning Research Group, NICTA, Canberra, ACT, Australia

[2]College of Engineering and Computer Science, Australian National University, Canberra, ACT, Australia

[3]Faculty of Health, University of Canberra, Canberra, ACT, Australia

[4]Department of Information Technology, University of Turku, Turku, Finland

Corresponding Author:
Hanna Suominen, BSc, MSc, PhD
Canberra Research Laboratory
Machine Learning Research Group
NICTA
Locked Bag 8001
Canberra, ACT, 2601
Australia
Phone: 61 431913826
Fax: 61 262676220
Email: hanna.suominen@nicta.com.au

Abstract

Background: Over a tenth of preventable adverse events in health care are caused by failures in information flow. These failures are tangible in clinical handover; regardless of good verbal handover, from two-thirds to all of this information is lost after 3-5 shifts if notes are taken by hand, or not at all. Speech recognition and information extraction provide a way to fill out a handover form for clinical proofing and sign-off.

Objective: The objective of the study was to provide a recorded spoken handover, annotated verbatim transcriptions, and evaluations to support research in spoken and written natural language processing for filling out a clinical handover form. This dataset is based on synthetic patient profiles, thereby avoiding ethical and legal restrictions, while maintaining efficacy for research in speech-to-text conversion and information extraction, based on realistic clinical scenarios. We also introduce a Web app to demonstrate the system design and workflow.

Methods: We experiment with Dragon Medical 11.0 for speech recognition and CRF++ for information extraction. To compute features for information extraction, we also apply CoreNLP, MetaMap, and Ontoserver. Our evaluation uses cross-validation techniques to measure processing correctness.

Results: The data provided were a simulation of nursing handover, as recorded using a mobile device, built from simulated patient records and handover scripts, spoken by an Australian registered nurse. Speech recognition recognized 5276 of 7277 words in our 100 test documents correctly. We considered 50 mutually exclusive categories in information extraction and achieved the F1 (ie, the harmonic mean of Precision and Recall) of 0.86 in the category for irrelevant text and the macro-averaged F1 of 0.70 over the remaining 35 nonempty categories of the form in our 101 test documents.

Conclusions: The significance of this study hinges on opening our data, together with the related performance benchmarks and some processing software, to the research and development community for studying clinical documentation and language-processing. The data are used in the CLEFeHealth 2015 evaluation laboratory for a shared task on speech recognition.

(JMIR Med Inform 2015;3(2):e19) doi:10.2196/medinform.4321

KEYWORDS

computer systems evaluation; data collection; information extraction; nursing records; patient handoff; records as topic; speech recognition software

Introduction

Information Flow Failures

Information flow, defined as channels, contact, communication, or links to pertinent people [1], is critical in health care. Failures in information flow lead to preventable adverse events, including delays in diagnosis and intervention, administration of incorrect treatments, and missed or duplicated tests among others [2-4]. In Australian hospitals, these failures are associated with over a tenth of preventable adverse events. Information flow is critical in clinical handover, when a clinician or group of clinicians is transferring professional responsibility and accountability, for example, at shift change [3].

Nursing handover is a form of clinical narrative [5], where the documented (written) material is only a small component of the complete information flow. There are multiple approaches to clinical handover at shift change; however, nursing handover typically occurs with a combination of whole-team in a private area, followed by whole-team in the presence of the patient or carer. Best practice in Australian hospital settings [6,7] recommends verbal handover in the patient's presence, supplemented with written material.

Australian Privacy Laws

The Australian National Health and Medical Research Council [8] places a number of restrictions on the use of Australian clinical data, most notably, avoidance of so-called *deidentified* data. The data referenced as *deidentified* in US publications [9,10] is considered as *reidentifiable* under Australian privacy law [8]. While approaches exist for semiautomatically deidentifying clinical texts [11,12], all such processes (whether automatic or manual) do not meet the stringent privacy requirements of Australian law.

An audio recording of a complete nursing handover requires ethical consenting of the nursing team, patients, visitors, and all other incidental clinical staff. It is difficult to obtain a "natural" recording—that could be provided without restriction on its use—under such conditions. Audio recordings also present significant difficulties in terms of identification of patients [13]. Reidentifiable data [8] must have restricted use, appropriate ethical use, and approval from all data generators (eg, patients, nurses, other clinicians', and visitors at the wards).

Ethical deidentification of the nursing handover for open data is not realistic. The *British Medical Journal* recommends [14] not publishing verbatim responses or transcriptions of clinical discussions. Existing sources of clinical data have limitations such as research-only use [15], nondisclosure of data [16], or limited commercial licenses [17].

In the case of clinical nursing notes and handover, precise data does not exist in an open form. By open we mean without restriction [18], including commercial use. Due to the lack of existing datasets and the difficulty of providing an ethically sound "free" data resource, we have developed a synthetic dataset that closely matches the typical data found in a nursing shift change. Synthetic clinical documents have also been used in other clinical informatics studies. For example in 2013-2014, the *MedNLP* track on medical natural language processing (NLP) used synthetic clinical notes [19].

Free-form text, as an entry type, is essential to release clinicians' time from documentation for other tasks [20-22]. NLP (a.k.a. automated text analysis or text mining) [10,23-28], including speech recognition (SR) and information extraction (IE), provides a way to fill out a handover form for clinical proofing and sign-off (see Multimedia Appendix 1), but this cascaded system evokes significant research challenges.

The development of these techniques is hindered by access to data for research, development, and evaluation [29]. Medical shared tasks by, for example, *NII Testbeds and Community for Information access Research* [19], *Text Retrieval Conference* [30], and *Conference and Labs of the Evaluation Forum* (CLEF) *eHealth* [31] (see this reference also for a review of related shared tasks), have provided deidentified datasets to researchers, who developed new clinical language technologies to improve clinical notes and credit patient outcomes. In 2013, the *Health Design Challenge* had a shared task aiming to make clinical documents more usable by and meaningful to patients, their families, and others who take care of them [32]. This design/visualization task attracted over 230 teams to participate.

By providing an open clinical dataset, that includes verbatim conversations and associated audio recordings, we anticipate a greater impact from the shared computational tasks, and increased development in natural language technologies for clinical text. Consequently, the significance of this study hinges not only on opening our data and some processing software to the research and development community, but also on publishing our performance evaluation results as a benchmark for tracking of performance improvements in time.

Methods

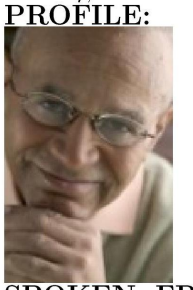
Data Creation

Creation Process

We created a synthetic dataset of 101 handover records (see Multimedia Appendix 1). Each record consisted of a patient profile; spoken, free-form text document; written, free-form text document; and written, structured document (Figure 1 shows this). The creation process included the following five steps: (1) generation of patient profiles; (2) creating a synthetic, but realistic nursing handover dataset, in collaboration with a registered nurse (RN); (3) development of a structured handover form; (4) using this form and the written, free-form text documents to create written, structured documents; and (5) creation of spoken, free-form text documents.

Figure 1. An example record that originates from our dataset.

ID: *14*, **TYPE:** *cardiovascular*
PROFILE:



Name: *Ken Harris*
Age: *71 years*
Admission story: *Ken is suffering from arrhythmia for the first time in his life. He is feeling pretty sick but this does not seem to be too serious.*
In-patient time: *He has been at the ward for three days.*
Familiarity: *Both you and the next nurse have looked after him earlier.*

SPOKEN, FREE-FORM TEXT DOCUMENT:
WAV file (93 words, 48 seconds, 4.25 MB)
The first author typed a transcription only for this spoken document
On a bed three is Ken Harris, 71 years old under Dr Gregor. He came in with arrhythmia. He complained of chest pain this morning and ECG was and was reviewed by the team. He was given some anginine and morphine for the pain and he is still tachycardic and new meds have been ordered in the medchart. Still for pulse checks for one full minute. Still awaiting for echo this afternoon. His blood pressure is just normal though he is scoring MEWS of three for the tachycardia. Otherwise he still for monitoring.

WRITTEN, FREE-FORM TEXT DOCUMENT:
Ken harris, bed three, 71 yrs old under Dr Gregor, came in with arrhythmia. He complained of chest pain this am and ECG was done and was reviewed by the team. He was given some anginine and morphine for the pain. Still tachycardic and new meds have been ordered in the medchart. still for pulse checks for one full minute. Still awaiting echo this afternoon. His BP is just normal though he is scoring MEWS of 3 for the tachycardia. He is still for monitoring.

WRITTEN, STRUCTURED DOCUMENT:

Ken[1] *harris*[2], *bed three*[5], *71 yrs old*[3] *under* *Dr Gregor*[6.1], *came in with* *arrhythmia*[7]. *He*[4] *complained of* *chest pain*[1] *this am and* *ECG*[2] *was done*[1] *and was reviewed by the team*[1]. *He was given some* *anginine*[1] *and* *morphine for the pain*[1] . *Still* *tachycardic*[2] *and* *new meds*[1] *have been ordered in the medchart. still* *for pulse checks for one full minute*[1]. *Still awaiting* *echo*[2] *this afternoon*[3]. *His* *BP is just normal*[2] *though he is* *scoring MEWS of 3 for the tachycardia*[2] *He is* *still for monitoring*[1].

PATIENT INTRODUCTION:
1. GivenNames/Initials: *Ken*
2. LastName: *harris*
3. AgeInYears: *71 yrs old*
4. Gender: *He*
5. CurrentBed: *bed three*
6. UnderDr: 6.1. LastName: *Dr Gregor*
7. AdmissionReason/Diagnosis: *arrhythmia*

MY SHIFT:
1. Status: *chest pain*
2. OtherObservation: *tachycardic; BP is just normal; scoring MEWS of 3 for the tachycardia*

APPOINTMENTS:
1. Status: *was done; was reviewed by the team*
2. Description: *ECG; echo*
3. Time: *this afternoon*

MEDICATION:
1. Medicine: *anginine; morphine for the pain; new meds*

FUTURE CARE:
1. Goal/TaskToBeCompleted/ExpectedOutcome: *for pulse checks for one full minute; still for monitoring*

Generation of Patient Profiles

The patient profile was developed using common user profile generation techniques [33]. The first author of this paper (Adj/Prof in machine learning for communication and health computing) considered an imaginary medical ward in Australia. With an aim for balance in patient types, she created simulated profiles for 101 patients. This included 1 sample patient together with 25 cardiovascular, 25 neurological, 25 renal, and 25 respiratory patients of the ward. These patient types were chosen

because they represent the most common chronic diseases and national health priority areas in Australia [34]. This provided a balanced demographic sample from which various handover scenarios could be created.

Each imaginary profile was given a stock photo from a royalty-free gallery, name, age, admission story, in-patient time, and familiarity to the nurses giving and receiving the handover. All patients were adults, but both young and old people were included. Some patients were recently admitted to the ward, some had been there for some days already, and some were almost ready to be discharged. For some patients, the in-patient time was short and for other patients it was longer. Within the admission story, the reason for admission was always an acute condition, but some patients had also chronic diseases.

Creation of Written, Free-Form Text Documents

The first author created a synthetic, written, free-form text document for the sample profile and supervised a RN in creating these documents for the remaining 100 profiles.

The RN had over twelve years experience from clinical nursing. She spoke Australian English as her second language and was originally from the Philippines. The RN's written consent was obtained for gathering, using, and releasing the spoken and written documents she created. She performed all these creative speaking and writing tasks as a National Information and Communications Technology, Australia (NICTA) employee alone in an office environment.

The RN was guided to imagine herself working in the medical ward and delivering verbal handovers to another nurse at a nursing shift change by the patient's bedside (see Multimedia Appendix 1). The handovers were to be monologues that include all handover information at once rather than discussions.

The RN was asked to write, for each patient profile, a realistic, but fully imaginary text document (ie, TXT file) as if she was talking and using normal wordings. The document length was set to 100-300 words.

Development of a Structured Handover Form

In consultation with Nursing Handover domain experts, the first and third authors developed a handover form (Figure 2 shows this) to be filled out. The form is compatible with existing handover forms, and matches the Australian and international standards/best practice for handover communication [35,36], as well as mimicks the RN's practical experiences from two Australian states/territories.

The form consisted of six headings (ie, *HANDOVER NURSE*, *PATIENT INTRODUCTION*, *MY SHIFT*, *APPOINTMENTS*, *MEDICATION*, and *FUTURE CARE*) with mutually exclusive categories (ie, *Title*, *Given names/initials*, *Last name*, and other subheadings together with subsubheadings like *Year*, *Month*, *Day* under *Date of birth*) for patient information, supplemented with the category of *Not Applicable* (*NA*) for irrelevant information. The number of categories was in total fifty with five, eighteen, eight, twelve, three, and three categories under *HANDOVER NURSE*, *PATIENT INTRODUCTION*, *MY SHIFT*, *APPOINTMENTS*, *MEDICATION*, and *FUTURE CARE*, respectively, and one category for *NA*.

This form structure is also consistent with the five-step nursing process model by the American Nurses Association (ANA): Assessment, Diagnosis, Outcomes/Planning, Implementation, and Evaluation [37].

ANA specifies that information about the first three steps should be documented under the patient's care plan in the patient's record so that nurses and other health care professionals caring for the patient have access to it. The Assessment step refers to a nurse collecting and analyzing patient information, including, physiological data together with psychological, sociocultural, spiritual, economic, and life-style factors. The Diagnosis step refers to his/her clinical judgment about the patient's response to actual or potential health conditions or needs. The Outcomes/Planning step refers to the nurse setting, based on the two previous steps, measurable and achievable short- and long-range goals for this patient. In our form, these three steps were covered under the headings of *PATIENT INTRODUCTION* with own, specific subheadings of *Admission reason/diagnosis* and *Care plan* for the initial Diagnosis and Outcomes/Planning steps.

The Implementation step refers to the implementation of nursing care in accordance with the care plan in order to assure the continuity of care for the patient during hospitalization and in preparation for discharge. Also, this delivered care is to be documented in the patient's record. In our form, it was covered under the headings of *MY SHIFT*, *MEDICATION*, *APPOINTMENTS*, and *FUTURE CARE*.

The Evaluation step refers to the continuous evaluation of the patient's status and the effectiveness of the nursing care and the respective modifications of the (written) care plan. Our form captured this step by considering the longitudinal series of handover documents in time.

Figure 2. Descriptive statistics of text snippets highlighted by the registered nurse in the 101 written, structured documents used as a reference standard in information extraction together with the performance of our best information extraction system. RN: registered nurse; RS: reference standard; IE: information extraction; NA: not applicable; min: minimum; max: maximum.

CATEGORY	LENGTH OF A A HIGHLIGHTED TEXT SNIPPET [WORDS]				NUMBER OF HIGHLIGHTED TEXT SNIPPETS	NUMBER OF POSITIVES IN THE REFERENCE STANDARD	NUMBER OF TRUE POSITIVES BY OUR BEST IE SYSTEM	PERFORMANCE		
	Min	Max	Mean	Standard deviation				Precision [%]	Recall [%]	F1 [%]
A. PATIENT INTRODUCTION						**2,064**	**1,880**	**91.33**	**92.52**	**91.92**
1. Given names/ initials	1	2	1.12	0.33	107	119	114	94.22	93.73	93.98
2. Last name	1	2	1.01	0.10	99	99	96	94.55	94.55	94.55
3. Age in years	1	5	2.49	0.90	100	246	238	93.43	95.60	94.50
4. Gender	1	6	1.05	0.51	95	489	476	93.93	97.82	95.83
5. Current room	2	2	2.00	0.00	27	54	54	100.00	100.00	100.00
6. Current bed	1	2	1.80	0.40	100	180	179	98.02	98.02	98.02
7. Under Dr: Given names/ initials	1	2	1.50	0.50	10	15	7	45.45	45.45	45.45
8. Under Dr: Last name	1	3	1.97	0.31	91	181	171	91.59	91.85	91.72
9. Admission reason/ diagnosis	1	13	3.01	2.30	135	414	321	73.03	81.97	77.24
10. Allergy	2	4	2.80	0.58	5	14	0	0.00	0.00	0.00
11. Chronic condition	1	7	12.41	1.70	29	70	12	28.00	25.71	26.81
12. Disease/ problem history	1	10	3.19	2.73	42	147	40	40.00	23.48	29.59
13. Care plan	6	6	2.69	1.23	13	36	0	0.00	0.00	0.00
B. MY SHIFT						**1,353**	**926**	**73.55**	**73.83**	**73.69**
14. Status	1	12	2.23	1.65	151	483	346	73.50	69.52	71.46
15. Contraption	1	9	4.00	2.63	11	44	0	0.00	0.00	0.00
16. Input/ diet	1	7	3.39	1.59	28	101	53	69.14	58.75	63.52
17. Output/ diuresis/ bowel movement	1	9	4.30	2.66	10	52	20	41.07	40.94	41.00
18. Wounds/ skin	1	11	4.75	2.72	12	55	0	0.00	0.00	0.00
19. Activities of daily living	1	10	4.19	2.11	59	245	152	85.28	76.11	80.43
20. Risk management	2	5	3.33	1.20	3	12	0	0.00	0.00	0.00
21. Other observation	1	24	3.69	2.90	99	361	86	27.83	17.81	21.72
C. APPOINTMENTS						**393**	**109**	**32.10**	**22.24**	**26.28**
22. Status	1	10	4.91	2.54	33	159	23	12.67	9.24	10.69
23. Description	1	8	3.04	1.47	50	157	24	23.73	15.86	19.02
24. Clinician: Given names/ initials	2	2	2.00	0.00	1	2	0	0.00	0.00	0.00
25. Clinician: Last name	2	2	2.00	0.00	1	2	0	0.00	0.00	0.00
26. Date and time: Day	1	4	1.56	0.91	25	40	1	2.38	4.76	3.17
27. Date and time: Time	1	3	1.47	0.60	19	28	7	33.33	30.00	31.58
28. Date and time: City	2	2	2.00	0.00	1	3	0	0.00	0.00	0.00
29. Date and time: Ward	1	2	1.50	0.50	2	2	0	0.00	0.00	0.00
D. MEDICATION						**262**	**159**	**68.19**	**57.49**	**62.38**
30. Medicine	1	5	1.74	0.95	91	157	100	63.47	58.53	60.90
31. Dosage	1	5	2.18	1.52	17	37	6	10.94	8.59	9.63
32. Status	1	6	2.91	1.02	23	68	41	76.67	68.75	72.49
E. FUTURE CARE						**644**	**320**	**57.58**	**52.09**	**54.70**
33. Alert/ warning/ abnormal result	1	8	3.35	1.94	17	59	4	15.38	7.69	10.26
34. Goal/ task to be completed/ expected outcome	1	17	4.75	2.97	104	496	282	49.42	49.79	49.60
35. Discharge/ transfer plan	2	14	6.85	3.37	13	89	15	31.67	22.62	26.39
F. NA						**3,771**	**3,352**	**82.71**	**90.12**	**86.26**
36. NA						3,771	3,481	79.40	92.91	85.62

Creation of Written, Structured Documents

The first author created a model structuring of the sample patient's written, free-form text document with respect to the mutually exclusive categories of the handover form and supervised the RN in creating these written, structured documents for the remaining 100 profiles. The RN proofed and agreed on this sample structuring. The first author installed Protégé 3.1.1 with the Knowtator 1.9 beta [38] on the RN's computer and guided her in using it to structure the documents (see Multimedia Appendix 1).

The RN was reminded that, on one hand, not all documents include information for all form categories and, on the other hand, some documents have relevant information to a given category multiple times (eg, if a given patient was referred to in a document with both a given name Michael and nickname Mike, both these occurrences were to be assigned to the category of *PATIENT INTRODUCTION: Given names/initials*).

The first and second author performed light proofing of these 101 structured documents in total. More precisely, they improved the consistency in including/excluding articles or titles, as well as in marking gender information in each document if it was available.

Creation of Spoken, Free-Form Documents

The first author supervised the RN in creating the spoken, free-form text documents by reading the 100 written free-form text documents out loud as the nurse giving the handover. She was guided to try to speak as naturally as possible, avoid sounding like reading text, and repeat the take until she was satisfied with the outcome (see Multimedia Appendix 1).

The Olympus WS-760M digital recorder and Olympus ME52W noise-canceling lapel-microphone (see Multimedia Appendix 1) that were previously used and shown to produce a superior word correctness in SR [36] captured the RN's voice. The use of the recorder and microphone was practiced before the actual recording and the recording took place in a quiet office environment.

The first author edited each Windows Media Audio (WMA) audio recording to include only one handover document. This included assuring the file beginning and end did not include recordings that the RN was unsatisfied with, file identifiers, or other additional content.

Processing and Evaluation Methods for Speech Recognition

Processing Methods

We used Dragon Medical 11.0 to convert the audio files to written, free-form text documents. This software was initialized with respect to the RN's details of age of 22-54 years and accent of Australian English, and trained to her voice by recording her reading the document of The Final Odyssey (3893 words, 29 minutes 22 seconds, 4 minutes needed) using the aforementioned recorder and microphone. This training, tailoring, or adaptation

to a speaker's voice was left minimal, since it could limit comparability with other studies and might not be feasible for every clinician in practice. To meet the software requirements, the first author converted WMA recordings from stereo to mono tracks and exported them from WMA to WAVeform (WAV) files on Audacity 2.0.3 [39].

We compared the Dragon vocabularies of *general*, *medical*, *nursing*, *cardiology*, *neurology*, and *pulmonary disease*. That is, we used the most general clinical vocabulary of *general*, the vocabulary suitable for a medical ward (ie, *medical*), the vocabulary suitable for nursing handovers (ie, *nursing*), and the vocabularies that were the closest matches with our patient types (ie, *cardiology* for cardiovascular patients, *neurology* for neurological patients, and *pulmonary disease* for respiratory patients).

Evaluation Methods

We applied the SCLITE scoring tool of the SR Scoring Toolkit 2.4.0 [40] in the analysis of the correctly recognized, substituted, inserted, and deleted words. The reference standard (RS) in all comparisons consisted of the original written, free-form text documents by the RN (ie, not transcriptions by hand), where punctuation was removed and capitalization was not considered as a distinguishing feature.

We chose the vocabulary resulting in the best performance in terms of the correctly recognized words (see the Results section) for a more detailed error analysis. The correct, substituted, inserted, and deleted words were defined by the aforementioned SCLITE scoring tool. As the most fundamental concept in this analysis, we measured the phonetic similarity (PS), defined as a perceptual distance between speech sounds [41], between words in the RS and speech-recognized text in order to find sound-alike substitution errors (eg, "four" vs "for" or "doctors signed" vs "dr san") for their correction. In the error analysis, we used the entire dataset and the subset that affects the IE system (ie, "inside" refers to text identified as relevant to the slots of the handover form). The correction could be based on linguistic postprocessing that combines PS with grammatical context [42-44].

We implemented a simple PS measure, which combines the Double Metaphone phonetic encoding algorithm [45,46] on the Apache Commons Metaphone [47] with the unweighted edit distance of the SimMetrics library [48]. We chose this algorithm because it approximates accented English from Slavic, Germanic, French, and Spanish, among others languages, and can be therefore seen as suitable for our accented RN's speech. The encoding algorithm translated each consonant into a limited set of characters where similar sounds are represented by the same character (eg, "b" and "p" both sound like "p"). The unweighted edit distance calculated the similarity between the encoded words or word sequences as the minimum number of substitution, insertion, and deletion operations required to transform an encoded word into another. Because the algorithm is designed to encode a single word at a time, we first encoded each word in a multi-word sequence, then combined the encoded words as a sequence, and finally calculated the edit distance to measure the similarity between the sequences.

Processing and Evaluation Methods for Information Extraction

Processing Methods

We used our expert-annotated dataset to train and evaluate IE systems. We considered this learning problem as a task where each word in text is considered as an entity with features and the goal is to assign it automatically to one or none of the categories. We chose to apply the conditional random field (CRF) [49], a probabilistic model for processing, segmenting, and labeling sequence data. This method solved the IE task by assigning precisely one category to each word of the document(s) based on patterns it has learned by observing words and the RN's expert-annotated categories, as well as the enriched feature representation of the words and their context. We adopted an open-source implementation of CRFs called CRF++ [50].

We generated the features by processing the original records using Stanford CoreNLP (English grammar) by the *Stanford Natural Language Processing Group* [51], MetaMap 2012 by the *US National Library of Medicine* [52], and Ontoserver by the *Australian Commonwealth Scientific and Industrial Research Organisation* [53] (Tables 1-3). Our best system used eight syntactic, three semantic, and twelve statistical feature types. We also experimented with additional feature types, but this did not contribute to the IE system performance.

In the CRF++ template, we defined in the *unigram part* that we use all features of the current location alone; all features of the previous location alone; all features of the next location alone; the pairwise correlations of the previous and current location over all features; the pairwise correlations of the current and next location over all features; and the combination of all features in the current location. In the *binary part*, we combined the predicted category for the previous location and the features of the current location to form a new feature.

Table 1. Experimented syntactic features.

ID	Name	Definition	Example	Software	In our best IE system
1	Word	Word itself	"Patients" or "had"	None	Yes
2	Lemma	Lemma of the word	"patients" or "have"	CoreNLP	Yes
3	NER[a]	NER[a] tag of the word for named entities (ie, person, location, organization, other proper name) and numerical entities (ie, date, time, money, number)	"number" for "5"	CoreNLP	Yes
4	POS[b]	POS[b] tag of the word	"IN" (ie, preposition) for "in", "NN" (ie, common noun as opposed to Proper Name, "PN") for "bed", "CN" (ie, cardinal number) for "5"	CoreNLP	Yes
5	Parse tree	Parse tree of the sentence from the root to the current word	"ROOT-NP-NN" (ie, root-noun phrase-common noun) for "5" in "In bed 5 we have..."	CoreNLP	Yes
6	Basic dependents	Basic dependents of the word	"Cardinal number 5" that refers to the bed ID for "bed" in "In bed 5 we have..."	CoreNLP	Yes
7	Basic governors	Basic governors of the word	Preposition "in" and subject "we" for "have" in "In bed 5 we have..."	CoreNLP	Yes
8	Phrase	Phrase that contains this word	"In bed 5" for "bed" in "In bed 5 we have"...	MetaMap	Yes

[a] NER = named entity recognition

[b] POS = part of speech

Table 2. Experimented semantic features.

ID	Name	Definition	Example	Software	In our best IE system
9	Top 5 candidates	Top 5 candidates retrieved from UMLS[a]	"BP" may refer to, for example, "Bachelor of Pharmacy", "bedpan", "before present", "birthplace", or "blood pressure"	MetaMap	Yes
10	Top mapping	Top UMLS[a] mapping for the concept that is the best match with a given text snippet	"pneumonia" is a type "respiratory tract infection"	MetaMap	Yes
11	Medication score	1 if the word is a full term in ATCL[b]; else 0.5 if it can be found in ATCL[b]; 0 otherwise	1 for "acetylsalicylic acid"	NICTA	Yes

[a] UMLS = Unified Medical Language System

[b] ATCL = Anatomical Therapeutic Chemical List

Table 3. Experimented feature types, statistical features.

ID	Name	Definition	Example	Software	In our best IE system
12	Location	Location of the word on a ten-point scale from the beginning of the document to its end	"1" for the first word and "10" for the last word	NICTA	Yes
13	Normalized term frequency	Number of times a given term occurs in a document divided by the maximum of this term frequency over all terms in the document		NICTA	No
14	Top 5 candidates'	As 9 using SNOMED-CT-AU[a]		Ontoserver	No
15	Top mapping'	As 10 using SNOMED-CT-AU[a]		Ontoserver	No
16	Top 5 candidates''	As 9 using AMT[b]		Ontoserver	No
17	Tom mapping''	As 10 using AMT[b]		Ontoserver	No

[a] SNOMED-CT-AU = Systematized Nomenclature of Medicine - Clinical Terms - Australian Release

[b] AMT = Australian Medicines Terminology

Evaluation Methods

To evaluate the system performance, we used cross-validation (CV) with 100 documents for training and leaving out one for testing (ie, *leave-one-out*, LOO, CV over 101 documents). In addition, to assess the task difficulty and adequacy of the amount of data used for training, we computed system learning curves for training set sizes of 20, 40, 60, and 80 documents (together with the aforementioned training with 100 documents). For this purpose, we chose 21 documents to be used for testing by sampling the entire document set randomly without replacement. Then, we chose the documents to be used for training by sampling the remaining set of documents randomly without replacement. That is, we used all remaining documents for training when the training set size was 80, and otherwise chose a document subset of an appropriate size randomly without replacement. In order to assess the contribution of each feature to the overall system performance, we performed a leave-feature-out experiment on our best system and LOO CV. See, for example, [54] for these evaluation methods.

In these evaluations, we measured the Precision, Recall, and F1 (ie, the harmonic mean of Precision and Recall) as implemented in use in *CoNLL 2000 Shared Task on Chunking* [55]. We evaluated performance both separately in every category and over all categories. When evaluating the latter performance, we used both macro- and micro-averaging over all other categories than *NA*. We also documented the performance in the dominating category of *NA* category-specifically. Because our desire was to perform well in all classes, and not only in the majority classes, the macro-averaged results are to be emphasized over the micro-averaged results.

We also used two baseline systems: (1) the random baseline assigned a class to each word randomly and (2) the majority baseline the most frequent class (ie, *Future goal/Task to be completed/Expected outcome*).

Finally, to assess the stability and robustness of our categorization form, expert annotations, and IE system, we performed an experiment, where our goal was to predict only the highest-level classification task to the heading categories of *HANDOVER NURSE*, *PATIENT INTRODUCTION*, *MY SHIFT*, *APPOINTMENTS*, *MEDICATION*, *FUTURE CARE*, and *NA*. We compared two systems with exactly the same features, template, and LOO CV setting. The first system was trained on subheading and subsubheading level annotations as above and then its predictions were abstracted to the highest level. The second system was trained on these heading-level categories directly.

This experiment tested the null hypothesis of detailed annotations not being helpful for system performance. On the one hand, if we gained evidence to support the alternative hypothesis of detailed annotations being helpful, we would need to divide the more loosely defined and verbose categories (eg, *Care plan* and *Future goal/Task to be completed/Expected outcome*) to subcategories. On the other hand, if we accepted the null hypothesis, we could be satisfied with our form structure and annotations. This division of headings to subheadings would also then be a likely cure for issues we observed in our former study [36] that used a handover form with five high-level headings only.

In any case, even though it was more laborious to annotate free-form text with respect to the fifty categories of our form versus using the seven heading-level categories only, automatically generated structured documents, enabled by these more detailed annotations have many benefits. Namely, they support the documents reuse in computerized decision making and surveillance in health care better than the loosely classified documents.

Results

National Information and Communications Technology, Australia Synthetic Nursing Handover Data, Descriptive Statistics and Validation

The released dataset, called NICTA Synthetic Nursing Handover Data [56], included the following data records: (1) Dragon initialization details for the RN (ie, 1. DOCX for the written, free-form text document that originates from the Dragon software release and is to be used as the RS text and 2. WMA for the spoken, free-form text document by the RN) in the folder *handoverdata/initialisation* of the expanded file *handoverdata.ZIP*; (2) 100 patient profiles (DOCX) created by the first author and the respective 100 written, free-form text documents (TXT) created by the RN together with the sample text by the first author in the folders *handoverdata/ 100profiles* and *handoverdata/101writtenfreetextreports*, respectively; (3) 100 spoken, free-form text documents by the RN (WAV) in the folder *handoverdata/ 100audiofiles*; (4) 100 speech-recognized, written, free-form text documents for each of the six vocabularies (TXT) in the vocabulary-specific subfolders (eg, *Dragon-cardiology*) of the folder *handoverdata/100x6speechrecognised*; and (5) 101 written, structured documents for IE that include the RS text, features used by our best system, and form categories with respect to the RS and our best IE system when using LOO CV and the respective template (TXT, CRF++ format) in the folder *handoverdata/101informationextraction*.

Descriptive statistics of the dataset are given in Tables 4 and 5 and Figure 2.

Data Release

The licensing constraints were set as follows, the license of the spoken, free-form text documents (ie, WMA and WAV files) was set as "Creative Commons - Attribution Alone - Noncommercial - No Derivative Works" [57], for the purposes of testing SR and language processing algorithms in order to allow others to test their computational methods against these files with appropriate acknowledgment. The license of the remaining documents (ie, DOCX and TXT files) was set as "Creative Commons-Attribution Alone" [58] with our intention to allow others to use these text and image files for any purpose with appropriate acknowledgment. In both cases, the acknowledgment requirement is to cite this paper.

All documents were made publicly available on the Internet. They will be used in the CLEFeHealth 2015 evaluation laboratory for a shared task on SR [59].

Data Validation

The technical pipeline (ie, recorded voice, transcription, analysis) has been validated in clinical settings and published [36,60,61]. We have also evaluated the model of the handover [60,61] and systematically reviewed relevant technical literature [62].

Although the data we provided are a simulation of nursing handover, the written text for the handover scenario was based upon 150 live audio recordings of nursing handover in several Sydney-based hospitals [36,60,61]. These recordings were manually transcribed under confidentiality conditions and the results used to inspire new handover scenarios. The audio recordings contained 71/150 examples (47.3%) with a single person speaking, 59/150 (39.3%) with two people speaking, and 20/150 (13.3%) with three people speaking. Based on these recordings, and anecdotal evidence from clinical experts, a single speaker scenario appears to occur in half of the team handovers in the Australian Capital Territory and New South Wales-based hospitals. (Each state, and in some cases each health jurisdiction, in Australia has a slightly different model for handover. Discussions with domain experts suggested similar percentages in all health jurisdictions, but we are not aware of any systematic evidence.) Our clinical advisers noted that English-as-a-second-language is common in nursing handover. Patient voices were present only in 2 of the 150 recordings. The final scenarios, including audio files and transcripts, were presented to Nursing Managers and verified as a reasonable facsimile of true handover scenarios.

Finally, also the technical performance, including the suitability of different vocabularies for SR and features resulting in the best IE system, was similar [36] and in this current study. When using the same SR software with the nursing vocabulary and very similar approach for recording and initialization, the recognition correctness was from 0.62 (accented female) through 0.64 (native female) to 0.71 (native male) in [36]. Now, this correctness was 0.73, as we will learn in the next subsection. Similarly in IE, the F1 was 0.62 in both cases when macro-averaging over the five form-categories. For the irrelevant text, F1 was 0.85 in the former study and 0.86 now. These IE experiments used CRF++ with very similar features, template setting, and form headings.

Table 4. Descriptive statistics of the 100 written, free-form text documents produced by the RN.

Descriptor	Subdescriptor	Patient type Cardiovascular	Patient type Neurological	Patient type Renal	Patient type Respiratory	All
Documents	Number documents	25	25	25	25	100
	Number of words	1795	1545	1818	2119	7277
	Number of unique words	556	500	496	604	1304
	Number of inside words	1140	1006	1086	1305	4547
	Number of unique inside words	447	397	408	483	1106
Number of words in a document	Minimum	19	26	29	31	19
	Maximum	162	106	149	209	209
	Mean	70	60	71	83	71
	SD	37	22	33	39	34
Top 10 words in documents	1st (n)[a]	and (95)	and (64)	and (88)	and (100)	and (347)
	2nd (n)[a]	he (59)	is (60)	is (72)	is (69)	is (256)
	3rd (n)[a]	for (58)	he (54)	he (69)	on (63)	he (243)
	4th (n)[a]	is (55)	she (38)	is (46)	he (61)	in (170)
	5th (n)[a]	the (43)	in (35)	she (46)	with (51)	for (163)
	6th (n)[a]	with (43)	with (34)	the (38)	in (49)	with (162)
	7th (n)[a]	in (40)	on (33)	with (34)	for (43)	she (151)
	8th (n)[a]	to (32)	for (31)	came (32)	she (42)	on (141)
	9th (n)[a]	of (30)	to (29)	for (31)	the (37)	the (138)
	10th (n)[a]	came (27)	came (24)	to (30)	to (33)	to (124)
Top 10 inside words in documents	1st (n)[a]	he (57)	he (52)	he (63)	and (51)	he (220)
	2nd (n)[a]	for (47)	she (35)	she (39)	he (48)	she (139)
	3rd (n)[a]	and (26)	for (25)	and (34)	she(40)	and (131)
	4th (n)[a]	bed (25)	dr (22)	bed (24)	for (27)	for (118)
	5th (n)[a]	she (25)	and (20)	is (24)	dr (25)	dr (88)
	6th (n)[a]	dr (23)	old (20)	to (23)	is (20)	to (84)
	7th (n)[a]	to (22)	bed (19)	old (21)	on (20)	bed (80)
	8th (n)[a]	the (21)	to (19)	yrs (21)	to (20)	is (76)
	9th (n)[a]	her (18)	yrs (17)	all (20)	room (18)	old (72)
	10th (n)[a]	old (18)	her (16)	for (19)	of (16)	all (61)

[a] The notation “word, n” specifies that the word “word” occurred “n” times.

Table 5. Descriptive statistics of the 100 written documents produced by the RN.

Descriptor	Subdescriptor	Sample document	Patient type Cardiovascular	Patient type Neurological	Patient type Renal	Patient type Respiratory	All
Documents	Number of documents	1	25	25	25	25	101
	Number of words	167					8487
	Number of unique lemmas	92					1283
Number of words in a document	Minimum	167	26	37	35	32	26
	Maximum	167	181	170	239	120	239
	Mean	167	80.80	82.24	98.12	71.96	84.10
	SD	0	38.70	35.24	43.46	24.06	38.02
Number of unique lemmas in documents	Minimum	92	22	22	27	27	22
	Maximum	92	99	96	126	79	126
	Mean	92	53.64	54.48	63.84	48.60	55.50
	SD	0	19.83	17.44	21.84	12.80	19.35
Top 10 lemmas in documents	1st (n)[a]	be (15)	be (115)	be (119)	be (126)	be (111)	
	2nd (n)[a]	he (13)	and (95)	he (95)	and (100)	he (68)	
	3rd (n)[a]	and (4)	he (75)	and (88)	he (79)	and (64)	
	4th (n)[a]	to (4)	for (58)	she (63)	on (63)	she (57)	
	5th (n)[a]	a (3)	she (44)	in (46)	she (59)	in (35)	
	6th (n)[a]	headache (3)	the (43)	the (38)	with (51)	with (34)	
	7th (n)[a]	it (3)	with (43)	have (36)	in (49)	on (33)	
	8th (n)[a]	that (3)	in (40)	with (34)	for (43)	for (31)	
	9th (n)[a]	the (3)	to (32)	come (33)	the (37)	to (29)	
	10th (n)[a]	carotid (2)	of (30)	for (31)	to (33)	have (26)	
Number of highlighted text snippets in a document	Minimum						8
	Maximum						33
	Mean						16.15
	SD						5.29

[a] The notation "word, n" specifies that the word "word" occurred "n" times.

Evaluation Outcomes From Speech Recognition

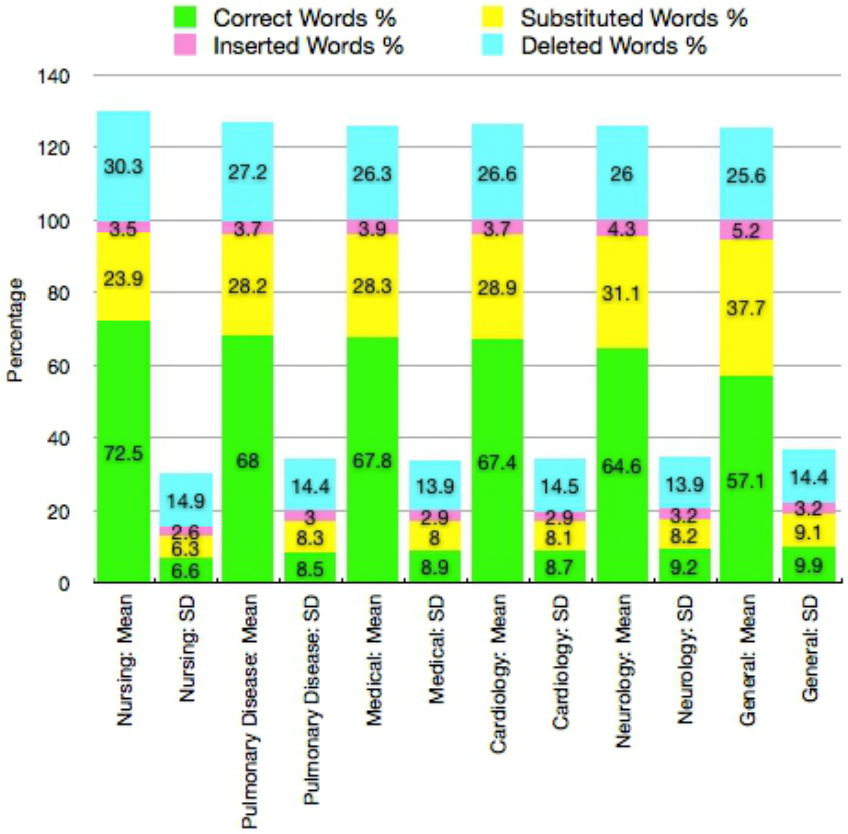
The best vocabulary for SR was *nursing*, resulting in the largest mean (5275/7277 words, ie, 0.725) and smallest SD (0.066) of correctly recognized words over the total of 7277 words (1 hour, 8 minutes, 5 seconds) in our 100 documents (Figure 3 shows this, see Multimedia Appendix 1). This correctness had the minimal, maximal, and median values of 0.547, 0.864, and 0.737 for this vocabulary. The *nursing* vocabulary also gave the largest number of correct words in 74 out of 100 cases. For the 25 cardiovascular patients, the matching vocabulary (ie, *cardiology*) gave more correct words than any other vocabulary only three times. For the 25 neurological patients with the *neurology* vocabulary and 25 respiratory patients with the *pulmonary disease* vocabulary, this number was four and zero, respectively. The number of times when the matching vocabulary gave more correct words than the *nursing* vocabulary was only four, three, and six for the cardiovascular, neurological, and respiratory patients, respectively. The *medical* vocabulary performed very differently from other vocabularies; its word

distribution to correct, substituted, inserted, and deleted words had more inserted and deleted words, but less correct words. An example of speech-recognized text using the *nursing* vocabulary is given in Textbox 1.

When considering the different patient types and the *nursing* vocabulary, the mean of correctly recognized words was 0.733 for the 25 cardiovascular patients, 0.732 for the 25 neurological patients, 0.724 for the 25 renal patients, and 0.713 for the 25 respiratory patients with the respective SDs of 0.073, 0.059, 0.063, and 0.071. That is, SR was slightly easier on cardiovascular patients, on average. Also the minimal and maximal values for the word correctness (ie, 0.619 and 0.864) were the largest for this patient type.

In text relevant to the form, 836 unique errors were present when using the *nursing* vocabulary [63]. Substitutions and insertions were the most common error types. Nearly a fifth of word substitutions sounded exactly the same as the correct word and over a quarter of the substitutions had a PS percentage above 75. Half of the substitutions occurred with words shorter than 4 characters that were obviously harder for SR than longer words. The most common single-word substitutions were “years” versus “yrs” and “in” versus “and” (n≥20). This error type was generally related to proper names (a quarter of errors and some of them sounded exactly the same, eg, “Lane” vs “Laine”, and often were just spelling variants, for example, “Johnson” vs “Johnsson”) and singular versus plural forms (eg, “fibrosis” vs “fibroses”). In conclusion, around a quarter of substitutions were candidates for their correction, and most of these errors were not SR errors, but rather explained by our written documents. The most common insertions included short words (eg, “and”, “is”, “in”, “she”, “are”, “all”, “arm”, “for”, “the”, “he”, “that”, or “a”, *n*≥20), typically when the RN used “aa”, “mm”, “eh”, or other back-channels that were not included in the written free-form text documents. The majority of the insertion and deletion errors corresponded to functional words with little semantic meaning. The most common deletion was “is” (n=20). Almost all remaining errors were caused by the following four types of systematic differences between the written free-form text documents and SR: (1) Australian versus US spelling (eg, “ catheterisation” vs “catheterization”); (2) digits versus letters (eg, “0” vs “zero”); (3) the RN’s use of abbreviations and acronyms in her writing, but complete forms when speaking (eg, “AM” vs “this morning”, “obs” vs “observations”, and “K” vs “potassium”); and (4) RN’s typing mistakes (eg, “ arrythmia” vs “arrhythmia”).

Figure 3. Speech recognition performance with the vocabularies of general, medical, nursing, cardiology, neurology, and pulmonary disease illustrated as a summary over the 100 documents. The notation of the x axis details the mean and SD for each Dragon vocabulary.



Textbox 1. Speech-recognized text corresponding to the example record.

TRANSCRIPTION OF SPOKEN, FREE-FORM TEXT DOCUMENT:

"On a bed three is Ken Harris, 71 years old under Dr Gregor. He came in with arrhythmia. He complained of chest pain this morning and ECG was and was reviewed by the team. He was given some anginine and morphine for the pain and he is still tachycardic and new meds have been ordered in the medchart. Still for pulse checks for one full minute. Still awaiting for echo this afternoon. His blood pressure is just normal though he is scoring MEWS of three for the tachycardia. Otherwise he still for monitoring."

WRITTEN, FREE-FORM TEXT DOCUMENT:

"Ken harris, bed three, 71 yrs old under Dr Gregor, came in with arrhythmia. He complained of chest pain this am and ECG was done and was reviewed by the team. He was given some anginine and morphine for the pain. Still tachycardic and new meds have been ordered in the medchart. still for pulse checks for one full minute. Still awaiting echo this afternoon. His BP is just normal though he is scoring MEWS of 3 for the tachycardia. He is still for monitoring."

WRITTEN, SPEECH-RECOGNIZED, FREE-FORM TEXT DOCUMENT USING THE NURSING VOCABULARY:

"Own now on bed 3 he is then Harry 70 is 71 years old under Dr Greco he came in with arrhythmia he complained of chest pain this morning in ECG was done and reviewed by the team he was given some and leaning in morphine for the pain in she is still tachycardic in new meds have been ordered in the bedtime is still 4 hours checks for one full minute are still waiting for echocardiogram this afternoon he is BP is just normal though he is scarring meals of 3 for the tachycardia larger otherwise he still for more new taurine."

Evaluation Outcomes From Information Extraction

Our best IE system classified 6349 out of the 8487 words correctly with respect to the 36 categories present in the RS (Figure 2). Figure 4 shows an example of an automatically structured document. The system performed excellently in filtering out irrelevant text (ie, *NA* category with 0.794 Precision, 0.929 Recall, and 0.856 F1 or 3481 correct out of 3771). The macro-averaged F1 over the 35 nonempty sub and subsubheading categories of the RS was 0.702 (Precision 0.759 and Recall 0.653). As expected, the larger amount of data for training, the better was the system performance (Figure 5 shows this). The system also performed substantially better in well-defined, compact categories (eg, perfect or nearly perfect Precision, Recall, and F1 in identifying the patient's current room and bed, respectively) than in more abstract and verbose categories (eg, 0.217 and 0.496 F1 in identifying *other observations* for *MY SHIFT* and *goals, tasks to be completed*, and *expected outcomes* for *FUTURE CARE*, respectively).

Most frequent category confusions related to irrelevant words (Figure 6 shows 1057 false positives and 290 false negatives). Other common confusions included differentiating: (1) *APPOINTMENTS, Description*, *APPOINTMENTS, Status*, and *MY SHIFT, Activities of daily living* from *FUTURE CARE, Goal/task to be completed/expected outcome* (n=58, n=29, and n=29); (2) *Disease/problem history* and *Chronic condition* from *Admission reason/diagnosis* under *PATIENT INTRODUCTION* (n=49 and n=22); (3) *Other observation* from *Status* (n=36) and vice versa (n=28) under *MY SHIFT*; and (4) *FUTURE CARE, Goal/task to be completed/expected outcome* from *MY SHIFT, Other observation* (n=35), where the first category is always with respect to the RS and the second refers to our best IE system.

In comparison, the majority baseline achieved overall a very modest performance (macro-averaged Precision, Recall, and F1 of 0.051, 0.091, and 0.065 over the 35 form categories and zero Precision, Recall, and F1 for *NA*). Its Precision, Recall, and F1 in the majority category were 1.00, 0.051, and 0.093. The random baseline was even weaker (macro-averaged Precision, Recall, and F1 of 0.015, 0.026, and 0.019 over the 35 form categories and 0.372, 0.030, and 0.055 for *NA*).

Each system feature contributed to the 36 categories differently (see Multimedia Appendix 1). However, on "average" (μ) over the 36 categories, Lemma was the most influential type (μ = 1.07), followed by Top 5 candidates (μ = 0.69), Part of speech (POS, μ = 0.56), Top mapping (μ = 0.35), and named entity recognition (NER, μ = 0.26). If considering this decrease in the macro-averaged F1 over the 35 form categories, the five types that influenced the most were Location (0.89), Top 5 candidates (0.25), POS (0.24), Basic governors (0.23), and Parse tree (0.16). In filtering out irrelevant words, they were POS, Lemma, Basic dependents, Location, and Top 5 candidates, with the decrease in the F1 of 0.0151, 0.0060, 0.0050, 0.0034, and 0.0023 respectively. This demonstrates that both the syntax and semantics together with the word location in the document is advantageous.

In the highest-level classification task with all but the *MY SHIFT* category present in the RS, the system trained on the highest-level annotations outperformed the system trained on the subheading and subsubheading level annotations (6731 vs 6710 words out of the 8487 words right, Figure 2). The respective category-specific statistics were: the F1 of 0.919 versus 0.918 for *PATIENT INTRODUCTION* (1882 vs 1880 correct out of 2064); the F1 of 0.737 versus 0.712 for *MY SHIFT* (915 vs 926 correct out of 1353); the F1 of 0.263 versus 0.279 for *APPOINTMENTS* (101 vs 109 correct out of 393); the F1 of 0.624 versus 0.650 for *MEDICATION* (153 vs 159 correct out of 262); the F1 of 0.547 versus 0.536 for *FUTURE CARE* (328 vs 320 correct out of 644); and the F1 of 0.863 versus 0.867 for *NA* (3352 vs 3316 correct out of 3771).

Figure 4. Automatically structured text that corresponds to our example document (Figure 1). When compared with the reference standard, added text is shown as bold and removed text is shown in grey. Risk-carrying errors include: (1) "chest pain" moving from "MY SHIFT, Status" to "PATIENT INTRODUCTION, Disease/problem history", (2) not identifying "tachycardic" and "scoring MEWS of 3 for the tachycardia" for "MY SHIFT, Other Observation", (3) not identifying "echo" for "APPOINTMENTS, Description", and (5) not identifying "anginine" and "new meds" for "MEDICATION, Medicine". RN: registered nurse and IE: information extraction.

HIGHLIGHTED TEXT BY THE RN:

Ken[1] *harris*[2], *bed three*[5], *71 yrs old*[3] *under* *Dr Gregor*[6.1], *came in with* *arrhythmia*[7]. *He*[4] *complained of* *chest pain*[1] *this am and* *ECG*[2] *was done*[1] *and* *was reviewed by the team*[1]. *He was given some* *anginine*[1] *and* *morphine for the pain*[1] . *Still* *tachycardic*[2] *and* *new meds*[1] *have been ordered in the medchart. still* *for pulse checks for one full minute*[1]. *Still awaiting* *echo*[2] *this afternoon*[3]. *His* *BP is just normal*[2] *though he is* *scoring MEWS of 3 for the tachycardia*[2]. *He is* *still for monitoring*[1].

WRITTEN, STRUCTURED DOCUMENT BY THE RN VS. OUR BEST IE SYSTEM:

PATIENT INTRODUCTION:
1. GivenNames/Initials: *Ken*
2. LastName: *harris*
3. AgeInYears: *71 yrs old*
4. Gender: *He*
5. CurrentBed: *bed three*
6. UnderDr: 6.1. LastName: *Dr Gregor*
7. AdmissionReason/Diagnosis: *arrhythmia*
8. **Disease/ProblemHistory: *chest pain***

MY SHIFT:
1. Status: *chest pain;* **BP is just normal**
2. OtherObservation: *tachycardic; BP is just normal; scoring MEWS of 3 for the tachycardia;* ***ECG was done***

APPOINTMENTS:
1. Status: *was done; was reviewed by the team*
2. Description: *ECG; echo*
3. Time: *this* *afternoon*

MEDICATION:
1. Medicine: *anginine; morphine for the pain; new meds*

FUTURE CARE:
1. Goal/TaskToBeCompleted/ExpectedOutcome: *for pulse checks for one full minute; still for monitoring*

Figure 5. Learning curves for cross-validation settings that included training set sizes of 20, 40, 60, 80, and 100 (ie, leave-one-out) documents with mutually exclusive folds, which in combination covered all data. CV: cross validation; and LOO: leave one out.

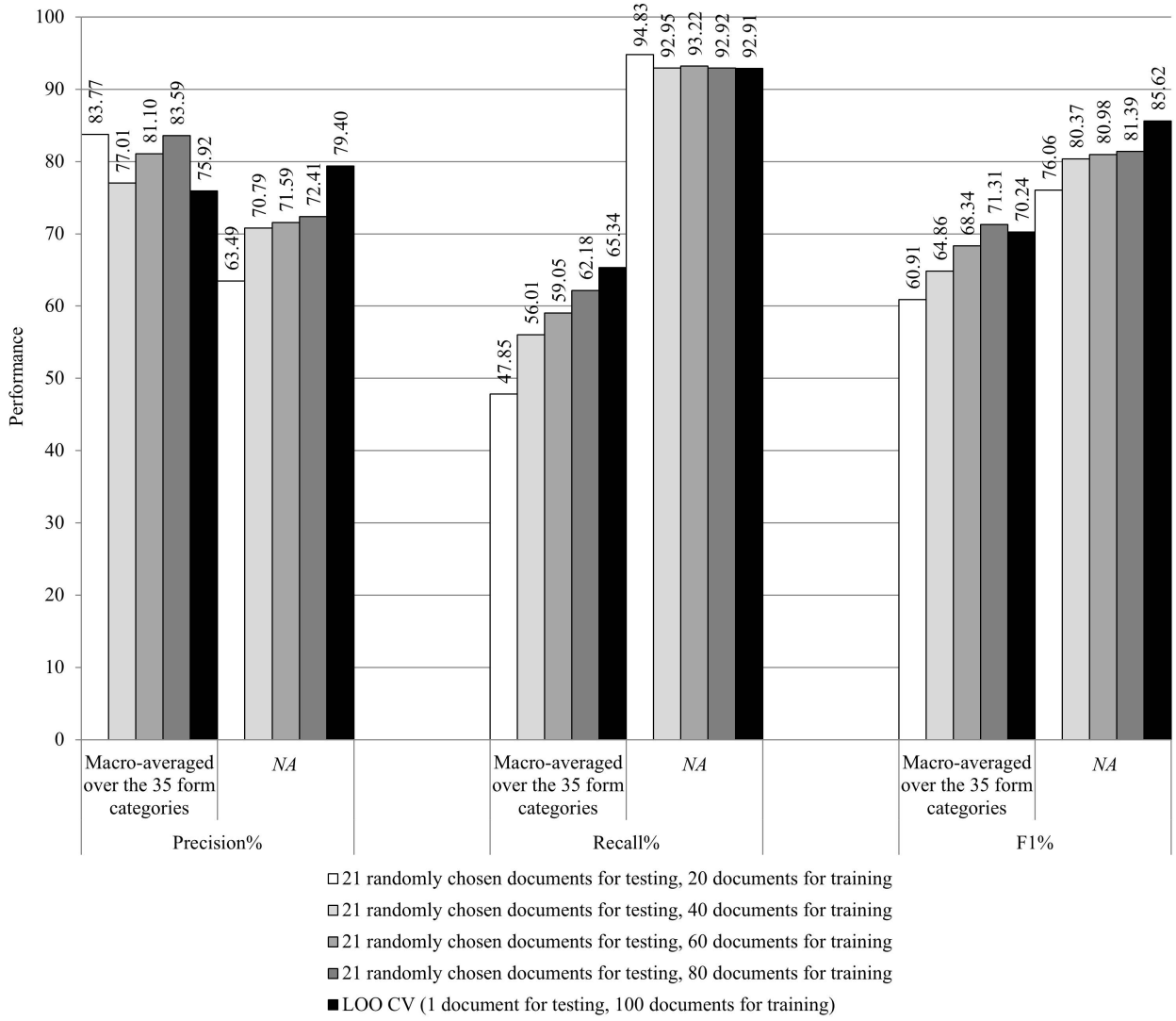


Figure 6. Confusion matrix between the reference standard (rows) and our best information extraction system (columns) in the 36-class multi-class classification task. Zero columns of 10, 15, 18, 20, 24, 25, and 28 have been removed for space constraints. For clarity, diagonal elements have been emphasized, and zero elements have been left empty. The category numbering corresponds to Figure 2. IE: information extraction.

	1	2	3	4	5	6	7	8	9	11	12	13	14	16	17	19	21	22	23	26	27	29	30	31	32	33	34	35	36
1	**114**																												5
2	1	**96**																											2
3		1	**238**			2																							5
4				**476**					1								1										6		5
5					**54**																								
6	1					**179**																							
7	2						**7**	6																					
8		2					3	**171**	2																				3
9				2				2	**321**		2	6	3				4										14		60
10									4								5										3		2
11			2						22	**12**	9		2										2						21
12									49	9	**40**												3				5		41
13									13								3										3		17
14				4					5		2		**346**				28	3											95
15									9		2						14						1				4		14
16				1					2					**53**									1				12		32
17				1					4						**20**		2						1						24
18				2					7				4				5		3										34
19									1							**152**	6										29		57
20									2							10													
21			1	8					4			3	36	3		3	**86**	7		1						2	17		190
22									4		1						26	**23**	12	2							29		62
23				2					16								4	2	**24**								58		51
24																													2
25								2																					
26									2								2			**1**							16		19
27																				1	**7**						5		15
28																													2
29																											2	1	
30									2								8						**100**	1			1		45
31				2													1						5	**6**					23
32				2									2										2		**41**		2		19
33				1					4								7						2			**4**	8		33
34				11					6			2					35	17	12				2				**282**	4	125
35				3									1					4				1					11	**15**	54
36	1	1	6	1				2	46				54	1	5	7	56	13	4	1			11	2	2		72	5	3,481

The National Information and Communications Technology, Australia Speech to Clinical Text Demonstration System

To demonstrate the SR and IE system design and workflow, we implemented a Web-app, written in the HyperText Markup Language, version 5 to allow any Web-browser to use it (Figure 7 show this) [64]. In particular, this means that the app is iPad compatible.

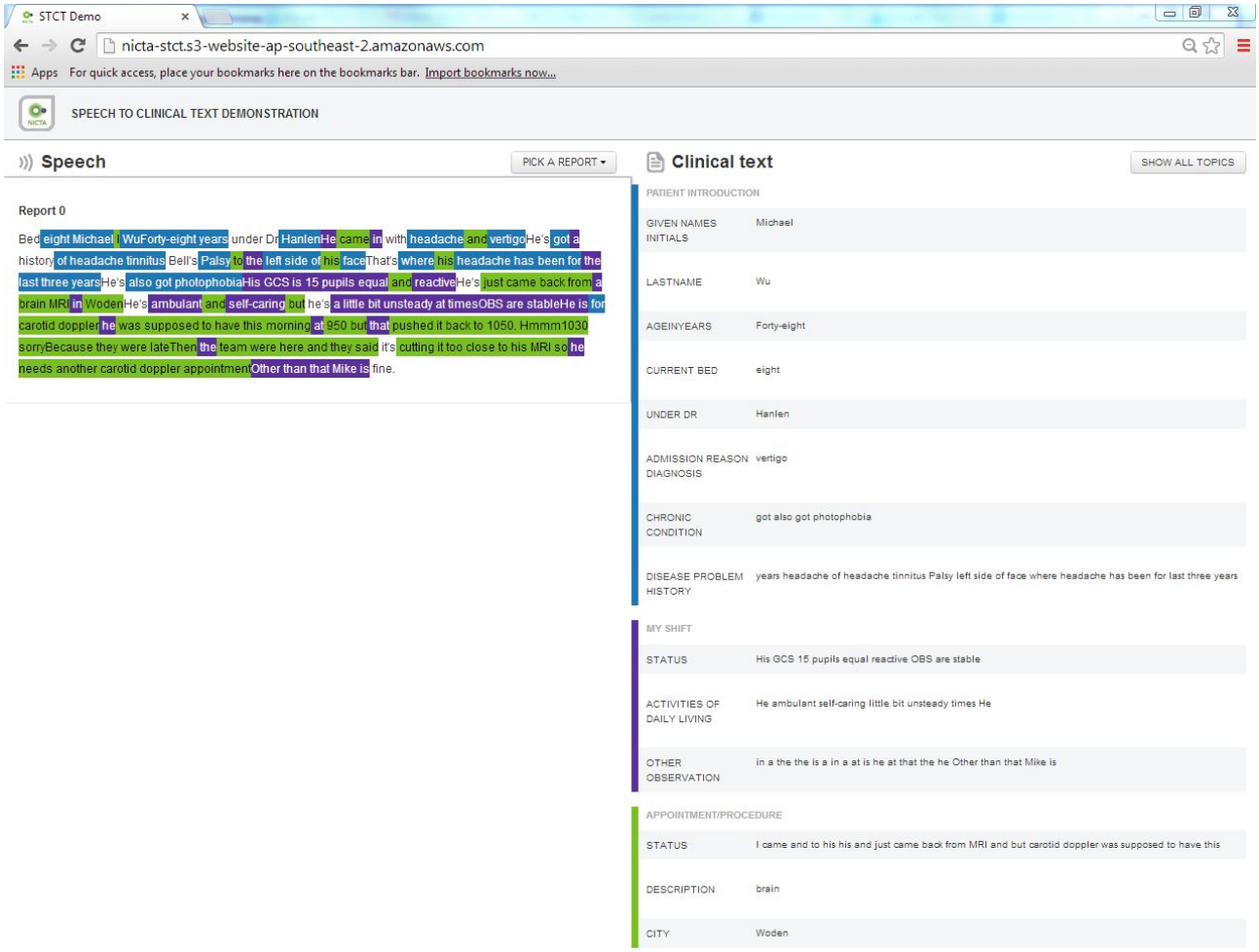
As an input, the app receives a form structure and an XML document, which includes all information needed to fill out this form. That is, the input has typed or speech-recognized text documents and their word-by-word classification with respect to the form categories.

The user (eg, a nurse) can choose a report to be structured from the "Pick a report" menu, see this written, free-form text on the left-hand side, and the filled-out form is given on the right-hand side. The report text is highlighted with respect to the headings of the form. In this way, the full text context never gets lost. The user can choose to see either the entire form (ie, "Show all topics") or only the subheadings and subsubheadings with extracted content (ie, "Only show available topics").

Extending the app to other IE tasks is straightforward by simply updating the input. However, we need to emphasize that this app performs visualization and not processing. That is, the spoken documents need to be converted to writing (by typing or SR) and classified with respect to the form structure (by manual highlighting or automated IE) in advance.

SR has not been included in the app. This is mainly because of the licensing constraints related to using a domain-specialized SR method (for a Microsoft Windows computer) that also needs to be trained to each speaker individually. However, also the aspect of being able to demonstrate the app in a noisy conference, technology festival, and other showcase environments led us to not include SR in the app.

Figure 7. National Information and Communications Technology, Australia (NICTA) speech to clinical text demonstration system that visualizes the example record.



Discussion

Principal Results

Cascaded SR and IE to fill out a handover form for clinical proofing and sign-off provide a way to make clinical documentation more effective and efficient. This way also improves accessibility and availability of existing documents in clinical judgment, situational-awareness, and decision making. Thereby, it contributes to the health care quality and people's health.

This cascading also evokes fruitful research challenges. First, conducting SR at clinical wards with noisy background and accented speakers is much more difficult than in a peaceful office. Second, its errors multiply when cascaded with IE. Third, every system error may have severe implications in clinical decision making. However, neither shared evaluation sets, nor baseline methods exist for this task.

In this paper, we have opened realistic, but synthetic data, methods, and evaluations related to clinical handover, SR, and IE to the research community in order to stimulate research and track continuous performance improvements in time. We have also introduced a Web app to demonstrate the system design and workflow.

Limitations

Setting for the Study

A real hospital setting cannot be idealized or modeled in a laboratory. Although we have attempted to capture the main components of a nursing handover scenario, there are several limitations in the data.

These limitations represent opportunities for future data gathering exercises. First, we used a single narrative voice rather than a team environment. In order to further develop any real system, collection of multiple voices communicating in a group setting is needed. Second, we did not include patient responses. In the recorded data from real nursing scenarios, patients rarely contributed to the conversation. Third, the data comprises 100 full verbatim documents. This provides a low power to any statistical analysis, and hence more data are always beneficial.

Performance Evaluation and Error Analysis

A detailed performance evaluation and error analysis of the system as a whole (ie, extrinsic evaluation) and each of its components (ie, intrinsic evaluation) is a crucial step in the development of cascaded pipeline apps [65,66]. At their best, SR can be only a percentage from perfect, and according to our findings, only a quarter of substitution errors could be considered as correction candidates. Similarly with our IE

component, the category-specific performance is at its best perfect, and altogether three-fourths (6349) of all 8487 words are correctly classified by our best system. The system performance is also convincing in filtering out irrelevant text (ie, 0.86 F1).

These rates of sound-alike SR-errors and slightly incorrect highlighting boundaries are not likely to harm a document's human readability. This is because the context around the highlighted text snippets is likely to assist in reading the text correctly. However, the extrinsic performance of this cascaded system remains to be formally evaluated.

Every corrected error is one less potential error in clinical decision making and in SR, a substantial amount of errors occur with words that are phonetically similar to each other. Based on our error analysis, the correction method should consider the following five characteristics: (1) PS between words or word sequences; (2) detection and correction of errors in proper names, by using, for example, other parts of a given patient's record; (3) difference between single-word and multi-word errors; (4) proofing for spelling and grammar; and (5) clear marking of automatically corrected words and possibility to choose a correction candidate interactively from a ranked list.

Comparison With Prior Work

Clinical SR has resulted in 1.3-5.7 times faster turnover-time in scientific studies [62]. The impact of SR on documentation time has been studied at two US emergency departments with a report turnover-time of less than 4 minutes, and proofing-time of 3 minutes, 39 seconds [67]. For transcription by hand, the respective times are nearly 40 minutes, and 3 minutes, 46 seconds. Similar conclusions on freeing up time have been published from three US military medical teaching facilities (ie, 19 hours vs 89 hours) [68], over forty US radiology practices (ie, 16 hours vs 48 hours) [69], a Finnish radiology department (ie, 12 hours vs 25 hours) [70], and 5011 US surgical pathology reports (ie, 72 hours vs 96 hours) [71]. When comparing the clinical workflows of SR to transcription by hand followed by proofing and sign off, the capability to use SR produces nearly two-thirds of the signed-off reports in less than an hour at the aforementioned Finnish radiology department, while this proportion is a third for transcription by hand [70]. This efficiency gain is evident also in the aforementioned longitudinal study on 5011 US surgical pathology reports [71], SR with proofing by hand increases the proportion of the reports signed off in less than a day from a fifth for time before SR, through a quarter during the first 35 months of SR use, to over a third after this initialization period. The respective proportions for the reports signed off in less than two days are over half, nearly two-thirds, and over two-thirds.

Clinical SR achieves an impressive word correctness percentage of 90-99, with only 30 to 60 minutes of training to a given clinician's speech. In other words, correcting SR errors by hand as a part of proofing is not likely to be time consuming. This recognition rate is supported by studies using the speech of twelve US-English male physicians on two medical progress notes, one assessment summary, and one discharge summary [72]; two US-English physicians' speech on 47 emergency-department charts [67]; and the speech of seven Canadian-English pathologists, and one foreign-accented researcher on 206 surgical pathology reports [73]. In our previous study [36] that uses the speech of two Australian-English female nurses and one Australian-English male physician on six nursing handover documents, the correctness is up to 0.79, while now it was 0.73. Differences in the correctness across different systems are negligible (ie, 0.91-0.93 for IBM ViaVoice 98, General Medicine; 0.85-0.87 for L&H Voice Xpress for Medicine 1.2, General Medicine; and 0.85-0.86 for Dragon Medical 3.0) [72]. In comparison, the report-wise error rate in word correctness is 0.4 for transcribing clinical text by hand and 6.7 for SR [73].

Similarly to the good correctness of clinical SR, clinical IE has gradually improved to exceed F1 of 0.90 in 1995-2008 [10]. It is most commonly used for content extraction, structuring, and enrichment to support diagnosis coding, decision making, and surveillance in health care. Other typical applications are deidentifying records for research purposes and managing clinical terminologies. This processing focuses on processing chest and other types of radiography reports, discharge summaries, echocardiogram reports, and pathology reports. However, the 170 reviewed studies do not address handover. Our performance is comparable to this; when considering the 50 mutually exclusive categories in IE, our performance is 0.86 for irrelevant text and up to perfect (ie, 1.00) for the remaining 35 nonempty form categories. Our performance is superior to our previous study [36] on 150 Australian handover documents and five main headings, F1 is slightly (ie, +0.01) better now, while the macro-averaged F1 for the form categories is the same.

The benefits of the combined use of SR and IE for handover documentation are twofold [36]. First, this approach stores all information along the workflow of having the verbal handover, using SR in real time to transcribe the recording, storing the content also as an audio recording, using IE in real time to fill out the handover form from the transcription for proofing, tracking the proofing changes, and signing off the document. In this way, clinicians can always keep the context of information, track changes, and perform searches on both the transcriptions and forms. The editing history can also be used to improve SR and IE correctness. Second, the approach makes the record drafts available and accessible almost instantly to everyone with an authorized access to a particular patient's documents. The speech-recognized transcription for a minute of verbal handover (approximately 160 words) is available only 20 seconds after finishing the handover with real time SR. Automated structuring through IE is almost instant and avoids problems related to subjectivity when structuring by hand. In comparison, clinicians would need to wait for almost four minutes for the hand-written transcription if they had a ward clerk to write the notes as they speak. This approach of using a clerk, either in real time or later on by the end of the shift, is also more prone to errors than clinicians, supported by a SR and IE system, writing the notes themselves in real time; if interpolating from the rate of information loss percentage from 60 to 100 after 3-5 shifts if notes are taken by hand, or not taken at all [4,74], more than an eighth of the information gets lost during one shift.

Acknowledgments

The Australian Government, through the Department of Communications, and the Australian Research Council, through the Information and Communications Technology Centre of Excellence Program, fund NICTA. 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. We used the Protégé resource, which is supported by grant GM10331601 from the National Institute of General Medical Sciences of the United States National Institutes of Health. We express our gratitude to Maricel Angel, RN at NICTA, for helping us to create the datasets for SR and IE. We acknowledge the contribution of Andrea Lau and Jack Zhao at Small Multiples for implementing our demonstration system. The first and third authors conceptualized the study, justified its significance, defined the form categories, and developed our demonstration system. The first author designed and supervised the process of creating realistic, but synthetic datasets as well as performed all SR experiments. Together with the last author, she analyzed SR errors and feasibility of phonetic similarity for their correction. The first two authors conducted all work related to IE. Together with the third author, the first author drafted the manuscript, and after this, all authors critically commented and revised it.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Additional illustrations and guidelines for creating showcase data.

[PDF File (Adobe PDF File), 2MB - medinform_v3i2e19_app1.pdf]

References

1. Glaser SR, Zamanou S, Hacker K. Measuring and interpreting organizational culture. Management Communication Quarterly 1987 Nov 01;1(2):173-198. [doi: 10.1177/0893318987001002003]
2. Patterson ES, Roth EM, Woods DD, Chow R, Gomes JO. Handoff strategies in settings with high consequences for failure: Lessons for health care operations. Int J Qual Health Care 2004 Apr;16(2):125-132 [FREE Full text] [doi: 10.1093/intqhc/mzh026] [Medline: 15051706]
3. Tran DT, Johnson M. Classifying nursing errors in clinical management within an Australian hospital. Int Nurs Rev 2010 Dec;57(4):454-462. [doi: 10.1111/j.1466-7657.2010.00846.x] [Medline: 21050197]
4. Matic J, Davidson PM, Salamonson Y. Review: Bringing patient safety to the forefront through structured computerisation during clinical handover. J Clin Nurs 2011 Jan;20(1-2):184-189. [doi: 10.1111/j.1365-2702.2010.03242.x] [Medline: 20815861]
5. Finlayson SG, LePendu P, Shah NH. Building the graph of medicine from millions of clinical narratives. Sci. Data 2014 Sep 16;1:140032. [doi: 10.1038/sdata.2014.32]
6. Australian NSW Department of Health. 2009. Implementation toolkit: Standard key principles for clinical handover URL: http://www.archi.net.au/documents/resources/qs/clinical/clinical-handover/implementation-toolkit.pdf [accessed 2015-02-04] [WebCite Cache ID 6W4wf1AzA]
7. Western Australian Department of Health. 2013. WA health clinical handover policy URL: http://www.safetyandquality.health.wa.gov.au/docs/initiative/CLINICAL_HANDOVER_Policy.pdf [accessed 2015-02-04] [WebCite Cache ID 6W4wNBaKD]
8. National Health Medical Research Council, Australian Research Council, Australian Vice-Chancellors' Committee. National statement on ethical conduct in human research. Canberra, ACT, Australia: Australian National Health and Medical Research Council; 2007. URL: https://www.nhmrc.gov.au/guidelines-publications/e72 [accessed 2015-04-01] [WebCite Cache ID 6XTHQ5iOr]
9. Friedlin FJ, McDonald CJ. A software tool for removing patient identifying information from clinical documents. J Am Med Inform Assoc 2008;15(5):601-610 [FREE Full text] [doi: 10.1197/jamia.M2702] [Medline: 18579831]
10. Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: A review of recent research. Yearb Med Inform 2008:128-144. [Medline: 18660887]
11. Neamatullah I, Douglass MM, Lehman LWH, Reisner A, Villarroel M, Long WJ, et al. Automated de-identification of free-text medical records. BMC Med Inform Decis Mak 2008;8:32 [FREE Full text] [doi: 10.1186/1472-6947-8-32] [Medline: 18652655]
12. El EK, Arbuckle L, Koru G, Eze B, Gaudette L, Neri E, et al. De-identification methods for open health data: The case of the Heritage Health Prize claims dataset. J Med Internet Res 2012;14(1):e33 [FREE Full text] [doi: 10.2196/jmir.2001] [Medline: 22370452]

13. Pathak MA. Privacy-preserving machine learning for speech processing. Berlin Heidelberg, Germany: Springer Theses; 2013:978-971.
14. Hrynaszkiewicz I, Norton ML, Vickers AJ, Altman DG. Preparing raw clinical data for publication: Guidance for journal editors, authors, and peer reviewers. BMJ 2010;340:c181 [FREE Full text] [Medline: 20110312]
15. BioGrid Australia: about us. URL: https://www.biogrid.org.au/page/3/about-us [accessed 2015-02-04] [WebCite Cache ID 6W5BweRCc]
16. I2b2: Informatics for integrating biology & the bedside, data sets. URL: https://www.i2b2.org/NLP/DataSets/Main.php [accessed 2015-03-25] [WebCite Cache ID 6XIGjCDeA]
17. Linguistic Data Consortium. Linguistic data consortium, data access. URL: https://www.ldc.upenn.edu/language-resources/data/obtaining [accessed 2015-02-04] [WebCite Cache ID 6W4yL9sG8]
18. Dunn AG, Day RO, Mandl KD, Coiera E. Learning from hackers: Open-source clinical trials. Sci Transl Med 2012 May 2;4(132):132cm5 [FREE Full text] [doi: 10.1126/scitranslmed.3003682] [Medline: 22553248]
19. Morita M, Kono Y, Ohkuma T. Overview of the NTCIR-10 MedNLP task. In: Proceedings of the 10th NTCIR Conference. Tokyo, Japan: NII Testbeds and Community for Information access Research (NTCIR); 2013 Presented at: 10th NTCIR; 2013 June 18-21; Tokyo, Japan p. 696-701 URL: http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings10/pdf/NTCIR/MedNLP/04-NTCIR10-MEDNLP-ImachiH.pdf
20. Poissant L, Pereira J, Tamblyn R, Kawasumi Y. The impact of electronic health records on time efficiency of physicians and nurses: A systematic review. J Am Med Inform Assoc 2005;12(5):505-516 [FREE Full text] [doi: 10.1197/jamia.M1700] [Medline: 15905487]
21. Hakes B, Whittington J. Assessing the impact of an electronic medical record on nurse documentation time. Comput Inform Nurs 2008;26(4):234-241. [doi: 10.1097/01.NCN.0000304801.00628.ab] [Medline: 18600132]
22. Banner L, Olney CM. Automated clinical documentation: Does it allow nurses more time for patient care? Comput Inform Nurs 2009;27(2):75-81. [doi: 10.1097/NCN.0b013e318197287d] [Medline: 21685832]
23. Zweigenbaum P, Demner-Fushman D, Yu H, Cohen KB. Frontiers of biomedical text mining: Current progress. Brief Bioinform 2007 Sep;8(5):358-375 [FREE Full text] [doi: 10.1093/bib/bbm045] [Medline: 17977867]
24. Demner-Fushman D, Chapman WW, McDonald CJ. What can natural language processing do for clinical decision support? J Biomed Inform 2009 Oct;42(5):760-772 [FREE Full text] [doi: 10.1016/j.jbi.2009.08.007] [Medline: 19683066]
25. Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: An introduction. J Am Med Inform Assoc 2011;18(5):544-551 [FREE Full text] [doi: 10.1136/amiajnl-2011-000464] [Medline: 21846786]
26. Friedman C, Rindflesch TC, Corn M. Natural language processing: State of the art and prospects for significant progress, a workshop sponsored by the National Library of Medicine. J Biomed Inform 2013 Oct;46(5):765-773 [FREE Full text] [doi: 10.1016/j.jbi.2013.06.004] [Medline: 23810857]
27. Friedman C, Elhadad N. Natural language processing in health care and biomedicine. In: Shortliffe EH, Cimino JJ, editors. Biomedical Informatics: Computer Applications in Health Care and Biomedicine. London, UK: Springer-Verlag; 2014:255-284.
28. Suominen H. Text mining and information analysis of health documents. Artif Intell Med 2014 Jul;61(3):127-130. [doi: 10.1016/j.artmed.2014.06.001] [Medline: 24998391]
29. Chapman WW, Nadkarni PM, Hirschman L, D'Avolio LW, Savova GK, Uzuner O. Overcoming barriers to NLP for clinical text: The role of shared tasks and the need for additional creative solutions. J Am Med Inform Assoc 2011;18(5):540-543 [FREE Full text] [doi: 10.1136/amiajnl-2011-000465] [Medline: 21846785]
30. Voorhees EM, Hersh W. NIST Special Publication 500-298: The Twenty-First Text REtrieval Conference Proceedings (TREC 2012). Gaithersburg, MD, USA: Department of Commerce, National Institute of Standards and Technology; 2012. Overview of the TREC 2012 medical records track URL: http://trec.nist.gov/pubs/trec21/papers/MED12OVERVIEW.pdf [accessed 2015-03-28] [WebCite Cache ID 6XMdUZS3E]
31. Suominen H, Salanterä S, Velupillai S, Chapman MM, Savova G, Elhadad N. In: Forner P, Müller H, Paredes R, Rosso P, Stein B, editors. Information access evaluation. Multilinguality, multimodality, and visualization, lecture notes in computer science 8138. Berlin, Germany: Springer-Verlag; 2013. Overview of the ShARe/CLEF eHealth Evaluation Lab 2013 URL: http://www.springer.com/us/book/9783642408014 [accessed 2015-04-01] [WebCite Cache ID 6XTIz8aom]
32. The patient record. Health design challenge URL: http://healthdesignchallenge.com/ [accessed 2015-01-23] [WebCite Cache ID 6VmhTcVol]
33. Kuniavsky M. Observing the user experience: A practitioner's guide to user research. San Francisco, CA, USA: Morgan Kaufmann Publishers; 2003.
34. Australian Government, Department of Health. Chronic disease: Chronic diseases are leading causes of death and disability in Australia. 2015. URL: http://www.health.gov.au/internet/main/publishing.nsf/Content/chronic [accessed 2015-01-23] [WebCite Cache ID 6Vmix5u6i]
35. Chaboyer W. Clinical handover. Slides.: NHMRC Centre of Resear Excellence in Nursing Care for Hospitalised Patients; 2011. URL: http://www.health.qld.gov.au/psq/handover/docs/ch_presentation2.pdf [accessed 2015-02-04] [WebCite Cache ID 6W5IQSZY2]

36. Suominen H, Johnson M, Zhou L, Sanchez P, Sirel R, Basilakis J, et al. Capturing patient information at nursing shift changes: Methodological evaluation of speech recognition and information extraction. J Am Med Inform Assoc 2014 Oct 21. [doi: 10.1136/amiajnl-2014-002868] [Medline: 25336589]
37. American Nurses Association. 2015. The nursing process URL: http://www.nursingworld.org/EspeciallyForYou/What-is-Nursing/Tools-You-Need/ [accessed 2015-01-23] [WebCite Cache ID 6VmkclWzr]
38. Ogren PV. Knowtator: A Protégé plug-in for annotated corpus construction. Stroudsburg, PA, USA: Association for Computational Linguistics; 2006. Proceedings of the 2006 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology; New York, NY, USA; 2006 Jun URL: http://www.aclweb.org/anthology/N06-4006 [accessed 2015-03-27] [WebCite Cache ID 6XNQAerTu]
39. Audacity® is free, open source, cross-platform software for recording and editing sounds. URL: http://audacity.sourceforge.net/ [accessed 2015-02-04] [WebCite Cache ID 6W5If78xf]
40. US National Institute of Standards and Technology (NIST), Information Technology Laboratory, Information Access Division (IAD). Tools. URL: http://www.itl.nist.gov/iad/mig/tools/ [accessed 2015-02-04] [WebCite Cache ID 6W5ImdeUH]
41. Mermelstein P. Pattern Recognition and Artificial Intelligence. 1976. Distance measures for speech recognition – psychological and instrumental URL: http://web.haskins.yale.edu/sr/SR047/SR047_07.pdf [accessed 2015-03-28] [WebCite Cache ID 6XNPoVqG1]
42. Kaki S, Sumita E, Iidar H. Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics (ACL 1998) and 17th International Conference on Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics; 1998. A method for correcting errors in speech recognition using the statistical features of character co-occurrence URL: http://www.aclweb.org/anthology/P98-1107 [accessed 2015-03-28] [WebCite Cache ID 6XNPgyBjd]
43. Jeong M, Kim B, Lee GG. Proceedings of the HLTNAACL special workshop on Higher-Level Linguistic Information for Speech Processing. Stroudsburg, PA, USA: Association for Computational Linguistics; 2004. Using higher-level linguistic knowledge for speech recognition error correction in a spoken Q/A dialog URL: http://www.aclweb.org/anthology/W04-3009.pdf [accessed 2015-03-28] [WebCite Cache ID 6XNPQAgm3]
44. Pucher M, Tu¨rk A, Ajmera J. Proceedings of the 3rd Congress of the Alps Adria Acoustics Association. Graz, Austria: Alps Adria Acoustics Association; 2007. Phonetic distance measures for speech recognition vocabulary and grammar optimization URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.396.4647&rep=rep1&type=pdf [WebCite Cache ID 6XTJp0TPl]
45. Philips L. Computer Language. 1990. Hanging on the metaphone URL: ftp://ftp.math.utah.edu/pub/tex/bib/toc/complang.html#7(12):December:1990 [accessed 2015-04-13] [WebCite Cache ID 6Xkxwa4PQ]
46. Philips L. C/C++ Users Journal. 2000. The double metaphone search algorithm URL: http://dl.acm.org/citation.cfm?id=349132 [accessed 2015-04-01] [WebCite Cache ID 6XTK6JzYc]
47. Class metaphone. URL: http://commons.apache.org/proper/commons-codec/apidocs/org/apache/commons/codec/language/Metaphone.html [accessed 2015-02-04] [WebCite Cache ID 6W6Kt5T8G]
48. SourceForge. SimMetrics. URL: http://sourceforge.net/projects/simmetrics/ [accessed 2015-02-04] [WebCite Cache ID 6W6KwJ8XY]
49. Lafferty JD, McCallum A, Pereira FCN. Proceedings of the 18th International Conference on Machine Learning, ICML 2001; Williamstown, MA, USA. Burlington, MA, USA: Morgan Kaufmann; 2001. Conditional random fields: Probabilistic models for segmenting and labelling sequence data URL: http://repository.upenn.edu/cgi/viewcontent.cgi?article=1162&context=cis_papers [WebCite Cache ID 6XNOQY9i4]
50. CRF++, Yet another CRF Toolkit. URL: http://taku910.github.io/crfpp/ [accessed 2015-04-01] [WebCite Cache ID 6XTKSIu3X]
51. Manning CD, Surdeanu M, Bauer J, Finkel J, Bethard SJ, McClosky D. Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations; Baltimore, MA, USA. Stroudsburg, PA, USA: Association for Computational Linguistics; 2014 Jun. The Stanford CoreNLP natural language processing toolkit URL: http://nlp.stanford.edu/pubs/StanfordCoreNlp2014.pdf [accessed 2015-03-27] [WebCite Cache ID 6XNOAYHPB]
52. MetaMap – A tool for recognizing UMLS concepts in text. URL: http://metamap.nlm.nih.gov/ [accessed 2015-02-04] [WebCite Cache ID 6W6MDPZVh]
53. Ontoserver. URL: http://ontoserver.csiro.au:8080/ [accessed 2015-03-25] [WebCite Cache ID 6XIIHDUMk]
54. Suominen H, Pyysalo S, Hiissa M. Performance evaluation measures for text mining. In: Song M, Wu YFB, editors. Handbook of Research on TextWeb Mining Technologies. Hershey, PA, USA: IGI Global; 2008:724-747.
55. Cardie C, Daelemans W, Nédellec C, Tjong Kim San E, editors. Introduction to the CoNLL-2000 shared task: Chunking. In: Proceedings of CoNLL-2000LLL-2000, Lisbon, Portugal. Stroudsburg, PA, USA: Association for Computational Linguistics; 2000:127-132.
56. The NICTA synthetic nursing handover dataset. Open NICTA: Datasets for download. URL: http://www.opennicta.com/datasets [accessed 2015-01-23] [WebCite Cache ID 6Vmoi8MNy]
57. Attribution-noncommercial-noderivatives 4.0 international (CC BY-NC-ND 4.0). URL: http://creativecommons.org/licenses/by-nc-nd/4.0/ [accessed 2015-01-23] [WebCite Cache ID 6VmpgnGnB]

58. Attribution 4.0 international (CC BY 4.0). URL: http://creativecommons.org/licenses/by/4.0/ [accessed 2015-01-23] [WebCite Cache ID 6VmpZB1a0]
59. CLEFeHealth 2015: Lab overview. 2015. URL: https://sites.google.com/site/clefehealth2015/ [accessed 2015-01-23] [WebCite Cache ID 6Vmpr4eWy]
60. Johnson M, Sanchez P, Suominen H, Basilakis J, Dawson L, Kelly B, et al. Comparing nursing handover and documentation: Forming one set of patient information. Int Nurs Rev 2014 Mar;61(1):73-81. [doi: 10.1111/inr.12072] [Medline: 24308444]
61. Dawson L, Johnson M, Suominen H, Basilakis J, Sanchez P, Estival D, et al. A usability framework for speech recognition technologies in clinical handover: A pre-implementation study. J Med Syst 2014 Jun;38(6):56. [doi: 10.1007/s10916-014-0056-7] [Medline: 24827759]
62. Johnson M, Lapkin S, Long V, Sanchez P, Suominen H, Basilakis J, et al. A systematic review of speech recognition technology in health care. BMC Med Inform Decis Mak 2014;14:94 [FREE Full text] [doi: 10.1186/1472-6947-14-94] [Medline: 25351845]
63. Suominen H, Ferraro G. In: Karimi S, Verspoor K, editors. Proceedings of the Australasian Language Technology Association Workshop 2013 (ALTA 2013); Brisbane, QLD, Australia. Stroudsburg, PA, USA: Association for Computational Linguistics; 2013 Dec. Noise in speech-to-text voice: Analysis of errors and feasibility of phonetic similarity for their correction URL: http://aclweb.org/anthology/U/U13/U13-1006.pdf [accessed 2015-03-28] [WebCite Cache ID 6XNNIzEvh]
64. NICTA speech to clinical text demonstration. URL: http://nicta-stct.s3-website-ap-southeast-2.amazonaws.com/ [accessed 2015-01-23] [WebCite Cache ID 6VmuZsah1]
65. Spärck JK, Galliers JR. Evaluating natural language processing systems: An analysis and review. In: Lecture Notes in Computer Science 1083. New York, NY, USA: Springer-Verlag; 1996.
66. Hirschman L, Thompson HS. Overview of evaluation in speech and natural language processing. In: Cole R, editor. Survey of the State of the Art in Human Language Technology. New York, NY, USA: Cambridge University Press; 1997:409-414.
67. Zick RG, Olsen J. Voice recognition software versus a traditional transcription service for physician charting in the ED. Am J Emerg Med 2001 Jul;19(4):295-298. [doi: 10.1053/ajem.2001.24487] [Medline: 11447517]
68. Callaway EC, Sweet CF, Siegel E, Reiser JM, Beall DP. Speech recognition interface to a hospital information system using a self-designed visual basic program: Initial experience. J Digit Imaging 2002 Mar;15(1):43-53 [FREE Full text] [doi: 10.1007/s10278-001-0007-y] [Medline: 12134214]
69. Langer SG. Impact of speech recognition on radiologist productivity. J Digit Imaging 2002 Dec;15(4):203-209 [FREE Full text] [doi: 10.1007/s10278-002-0014-7] [Medline: 12415463]
70. Kauppinen T, Koivikko MP, Ahovuo J. Improvement of report workflow and productivity using speech recognition--a follow-up study. J Digit Imaging 2008 Dec;21(4):378-382 [FREE Full text] [doi: 10.1007/s10278-008-9121-4] [Medline: 18437491]
71. Singh M, Pal TR. Voice recognition technology implementation in surgical pathology: Advantages and limitations. Arch Pathol Lab Med 2011 Nov;135(11):1476-1481. [doi: 10.5858/arpa.2010-0714-OA] [Medline: 22032576]
72. 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(5):462-468 [FREE Full text] [Medline: 10984465]
73. Al-Aynati MM, Chorneyko KA. Comparison of voice-automated transcription and human transcription in generating pathology reports. Arch Pathol Lab Med 2003 Jun;127(6):721-725. [doi: 10.1043/1543-2165(2003)127<721:COVTAH>2.0.CO;2] [Medline: 12741898]
74. Pothier D, Monteiro P, Mooktiar M, Shaw A. Pilot study to show the loss of important data in nursing handover. Br J Nurs 2005;14(20):1090-1093. [doi: 10.12968/bjon.2005.14.20.20053] [Medline: 16301940]
75. Matic J, Davidson PM, Salamonson Y. Review: Bringing patient safety to the forefront through structured computerisation during clinical handover. J Clin Nurs 2011 Jan;20(1-2):184-189. [doi: 10.1111/j.1365-2702.2010.03242.x] [Medline: 20815861]

Abbreviations

ANA: American Nurses Association
CLEF: Conference and Labs of the Evaluation Forum
CRF: conditional random field
CV: cross validation
IE: information extraction
LOO: leave one out
NA: not applicable
NER: named entity recognition
NICTA: National Information and Communications Technology, Australia
NLP: natural language processing
POS: part of speech
PS: phonetic similarity

RN: registered nurse
RS: reference standard
SR: speech recognition
WAV: WAVeform (audio format)
WMA: Windows Media Audio

Edited by G Eysenbach; submitted 06.02.15; peer-reviewed by B Seidel; accepted 07.03.15; published 27.04.15.

Please cite as:
Suominen H, Zhou L, Hanlen L, Ferraro G
Benchmarking Clinical Speech Recognition and Information Extraction: New Data, Methods, and Evaluations
JMIR Med Inform 2015;3(2):e19
URL: http://medinform.jmir.org/2015/2/e19/
doi:10.2196/medinform.4321
PMID:25917752

Original Paper

Web-Based Textual Analysis of Free-Text Patient Experience Comments From a Survey in Primary Care

Inocencio Daniel Maramba[1], BS, MD, MSc; Antoinette Davey[1], BA(Hons), MSc, MPhil; Marc N Elliott[2], BA, MA, PhD; Martin Roberts[3], BA (Hons), Cert Ed, MSc; Martin Roland[4], MA, BM, BCh, DM, FRCGP, FMedSci; Finlay Brown[5], BMBS; Jenni Burt[4], BA (Hons), MSc, PhD; Olga Boiko[1], MSc, MPhil, PhD; John Campbell[1], MBChB, MD, FRCGP

[1]Primary Care, University of Exeter Medical School, University of Exeter, Exeter, United Kingdom

[2]RAND Corporation, Santa Monica, CA, United States

[3]Collaboration for the Advancement of Medical Education Research and Assessment (CAMERA), Peninsula Schools of Medicine and Dentistry, Plymouth University, Plymouth, United Kingdom

[4]Institute of Public Health, University of Cambridge, Cambridge, United Kingdom

[5]Peninsula College of Medicine and Dentistry, Universities of Plymouth and Exeter, Plymouth, United Kingdom

Corresponding Author:
Inocencio Daniel Maramba, BS, MD, MSc
Primary Care
University of Exeter Medical School
University of Exeter
JS01 Smeall Building
St Luke's Campus University of Exeter, Magdalen Road
Exeter, EX1 2LU
United Kingdom
Phone: 44 1392 722895
Fax: 44 1392 722894
Email: I.D.C.Maramba@exeter.ac.uk

Abstract

Background: Open-ended questions eliciting free-text comments have been widely adopted in surveys of patient experience. Analysis of free text comments can provide deeper or new insight, identify areas for action, and initiate further investigation. Also, they may be a promising way to progress from documentation of patient experience to achieving quality improvement. The usual methods of analyzing free-text comments are known to be time and resource intensive. To efficiently deal with a large amount of free-text, new methods of rapidly summarizing and characterizing the text are being explored.

Objective: The aim of this study was to investigate the feasibility of using freely available Web-based text processing tools (text clouds, distinctive word extraction, key words in context) for extracting useful information from large amounts of free-text commentary about patient experience, as an alternative to more resource intensive analytic methods.

Methods: We collected free-text responses to a broad, open-ended question on patients' experience of primary care in a cross-sectional postal survey of patients recently consulting doctors in 25 English general practices. We encoded the responses to text files which were then uploaded to three Web-based textual processing tools. The tools we used were two text cloud creators: TagCrowd for unigrams, and Many Eyes for bigrams; and Voyant Tools, a Web-based reading tool that can extract distinctive words and perform Keyword in Context (KWIC) analysis. The association of patients' experience scores with the occurrence of certain words was tested with logistic regression analysis. KWIC analysis was also performed to gain insight into the use of a significant word.

Results: In total, 3426 free-text responses were received from 7721 patients (comment rate: 44.4%). The five most frequent words in the patients' comments were "doctor", "appointment", "surgery", "practice", and "time". The three most frequent two-word combinations were "reception staff", "excellent service", and "two weeks". The regression analysis showed that the occurrence of the word "excellent" in the comments was significantly associated with a better patient experience (OR=1.96, 95%CI=1.63-2.34), while "rude" was significantly associated with a worse experience (OR=0.53, 95%CI=0.46-0.60). The KWIC

results revealed that 49 of the 78 (63%) occurrences of the word “rude” in the comments were related to receptionists and 17(22%) were related to doctors.

Conclusions: Web-based text processing tools can extract useful information from free-text comments and the output may serve as a springboard for further investigation. Text clouds, distinctive words extraction and KWIC analysis show promise in quick evaluation of unstructured patient feedback. The results are easily understandable, but may require further probing such as KWIC analysis to establish the context. Future research should explore whether more sophisticated methods of textual analysis (eg, sentiment analysis, natural language processing) could add additional levels of understanding.

(JMIR Med Inform 2015;3(2):e20) doi:10.2196/medinform.3783

KEYWORDS

patient experience; patient feedback; free-text comments; quantitative content analysis; textual analysis

Introduction

Patient experience is an important component of quality of health care, and questionnaires capturing patient experience have been widely used to provide insight into the quality of primary health care provision [1-3]. Feedback from survey results has been proposed as a cost-effective method to support and facilitate quality improvement [4,5].

In addition to capturing responses via closed questionnaire items, open-ended questions eliciting free-text comments have also been widely adopted [6,7] as the exclusive use of quantitative data limits the potential of surveys to improve practice [8]. Analysis of free-text comments can provide deeper or new insight, identify areas for action, and initiate further investigation [9]. Also, they may be a promising way to progress from documentation of patient experience to achieving quality improvement [9,10]. Free-text comments have been evaluated using methods such as content analysis [11,12], thematic analysis [9,13,14], and the Holsti Method [15]. However, these approaches can be resource intensive [6,15,16]. To efficiently deal with a large amount of free-text, new methods of rapidly summarizing and characterizing the text are being explored [17].

Text clouds are visual representation of a body of text, where the more frequently occurring words appear larger in the "cloud"[18,19]. The first widespread use of text clouds was “tag clouds”, which originated as a representation of the "tags" or keywords that users would assign to a Web resource [20,21]. Tag clouds have been used in health related websites to counter biased information processing [22].

The same technology that creates tag clouds may also be used to create word clouds from texts and textual data in general [23]. Text clouds differ from tag clouds in that their purpose is predominantly comprehension of the text rather than navigation of webpages [23]. Text clouds can be used to rapidly summarize textual data, revealing textual messages in a pictorial form [24]. Text clouds may have utility in supporting searching and browsing of webpages, as well as impression formation and recognition/matching of textual data [25].

Web applications such as TagCrowd, Many Eyes, Wordle, and Tagxedo are commonly used to generate text clouds. The majority of these are free for nonprofit use.

Text clouds have been used in a wide range of health related areas, such as examining the differences between various versions of a General Medical Council document [24] as well as a UK Government White Paper [26], survey responses on ehealth [27], survey of pharmacists’ perceptions [28], patients’ use of online message forums [29], and to analyze the responses of multiple sclerosis sufferers to open-ended questions [30].

Other uses of computerized textual analysis in health include: automatic analysis of online discussions related to diabetes [31], content analysis of the free text comments in multi-source feedback about specialist registrars [32], automatic drug side effect discovery by analysis of online patient submitted reviews [33], keyword analysis of an online survey investigating nurses' perceptions of spirituality and spiritual care [34], and uncovering signs and symptoms of opiate exposure from comments posted on YouTube [35].

We aimed to investigate the feasibility of using Web-based textual analysis for extracting useful information from large amounts of free-text patient comments, and to identify key issues or topics that would be revealed by computerized text processing, using tools that are currently available at no cost on the Web.

Methods

Participants and Procedure

The data was collected as part of the “Improve” study, a research program funded by the National Institute for Health Research (NIHR) [36], exploring various aspects of patient experience in primary care. One of the projects involved a post-consultation postal survey using a modified version of the English GP Patient Survey (GPPS) questionnaire. The GPPS is the largest survey program of patients registered with an English general practice. A random sample of patients from each English practice-~2.6 million patients each year in total-is invited to take part in the survey [37,38]. A particular change made to the GPPS questionnaire (at the request of participating practices) was the inclusion of a free-text comments question worded as follows: "*Your [general] practice has asked that we collect any further comments you would like to make about the service they provide*."

Detailed survey methods have been previously reported in the paper by Roberts et al [39], which are briefly summarized here. Following a recent face-to-face consultation between November

2011 and June 2013 with one of 105 doctors from 25 practices in six areas of England (Cornwall, Devon, Bristol, Bedfordshire, Cambridgeshire, Peterborough, and North London), patients were sent a questionnaire regarding their experiences of care. One reminder was sent to nonrespondents. Free-text comments were anonymized during data entry, extracted from the database and exported to a text file. Approval for the study was obtained from the South West 2 Research Ethics Committee on January 28th 2011 (ref: 09/H0202/65). Return of a completed questionnaire was taken to indicate patient consent to participate in the study.

Textual Analysis Methods

Free-text comments were analyzed using three Web-based textual analysis tools: TagCrowd v.10/02/2011 [40], Many Eyes v.1.0 [41], and Voyant Tools v.1.0 [42], which were chosen for their ease of use and range of functionalities.

TagCrowd is a Web application for visualizing word frequencies in any text by creating what is popularly known as a word cloud, text cloud or tag cloud. We created text clouds based on an aggregated corpus of free-text patient comments.

We used the following parameters in TagCrowd: (1) frequently occurring English words and connectives (eg, “a”, “in”, “is”, “it”, and “you”) were ignored; (2) the tag cloud was created from the 50 most frequently occurring single words; (3) a stemming algorithm combined related words (eg, learn, learned, learning -> learn). The 50 word limit was chosen as it has been used in previous work using text clouds to examine health information [24,26]. We also tried generating a 60 word text cloud but found the result to be difficult to read.

Many Eyes is a Web-based data visualization application created by IBM [41]. Fundamentally the software incorporates the capacity to create and view various forms of text visualization and representation. We chose to use Many Eyes because of its capability of creating text clouds from the most frequent two-word combinations. We hypothesized that two-word combinations might give a more nuanced insight into the meanings behind the most frequently used words as some of their associations would be preserved.

Voyant Tools is a Web-based reading and analysis environment for digital texts. It was created as part of a collaborative project to develop and theorize text analysis tools and text analysis rhetoric [42]. In addition to calculating word frequencies and creating text clouds, Voyant Tools performs other textual analysis functions, such as identifying distinctive words in the documents that make up a text corpus. To investigate the validity of the distinctive words component, we divided the comments into separate text files depending on whether the patients reported if they were either “satisfied” or “not satisfied” with their experience of care. The question was “In general, how satisfied are you with the care you get at this GP surgery or health center?”. Patients were given five options to rate their satisfaction with the practice: “very satisfied”, “fairly satisfied”, “neither satisfied nor dissatisfied”, “fairly dissatisfied”, and “very dissatisfied”. In this analysis, the “very satisfied” and “fairly satisfied” responses were recoded as “satisfied” and the last three options as “not satisfied”. We used the “distinctive words” function to identify the words that occurred more frequently in comments originating from patients who were “satisfied” and words which occurred more frequently in comments from patients who were “not satisfied”.

Statistical Analysis

We used logistic regression to investigate the occurrence/nonoccurrence of words within individual patient comments. The words were selected from the results of the distinctive word analysis. We obtained and compared the frequency of use of the five most distinctive words from the comments classified as originating from either the “satisfied” or “not satisfied” patients (ten words in total).

Logistic regression was used to predict the presence or absence of each of these words in a comment from the standardized scores (z-scores) of the patients’ responses to the survey question on satisfaction. We used the following formula for the standardized scores (z-scores):

$$z=(x-\mu)/(\sigma)$$

Two additional models were run for each word, predicting its presence or absence from the z-scores of the patients’ ratings of their confidence and trust in the doctor and their ratings of the doctors’ communication skills. We derived these scores from the patients’ responses to two other structured questions that were asked in the questionnaire. These variables were chosen as we hypothesized that confidence and trust in the doctor, as well as the communication skills of the doctor could influence the words used by the patients in their comments. Statistical analyses were performed in STATA version SE13.1 for Windows. We then plotted the odds ratios for the selected words against their standardized frequencies. The standardized frequency is calculated in the same way as a z-score, where x is the frequency of a particular word, μ is the mean frequency of all words in the patient comments, and σ is their standard deviation.

Keyword in Context Analysis

Voyant Tools provides a Keyword in Context (KWIC) function. KWIC involves searching for a particular keyword in the text and analyzing its local meaning in relation to a fixed number of words immediately preceding and following it [43]. KWIC can help identify underlying connections that are being implied by the text [44]. KWIC analysis had been used in content analysis of blogs about female incontinence [45], as well as in content analysis of audiology service improvement documentation [46]. The KWIC function in Voyant tools can quickly display the KWIC for a selected keyword and the results can be exported to a format suitable for further analysis. For this analysis we selected 15 words that preceded and followed the word “rude”. The resulting text was then manually examined to determine the context of the use of “rude”, and the results were tabulated.

Results

Textual Analysis Methods

From 7721 respondents, we collected 3426 individual comments (comment rate: 44.4%). The comments came to a total of

150,699 words of which 6867 are unique words. The average length of response is 43.98 words. There are 273 instances of 90 unique, non-English terms (mostly misspellings). Figure 1 shows the text cloud resulting from all the free-text comments as generated by TagCrowd. The five most frequent words were: “doctor”, “appointment”, “surgery”, “practice”, and “time”. Included in the 50 most frequent words were those that have a positive connotation such as: “helpful” and “excellent”. Words with a negative connotation, such as “difficult” and “problem” were also present, but were less frequent.

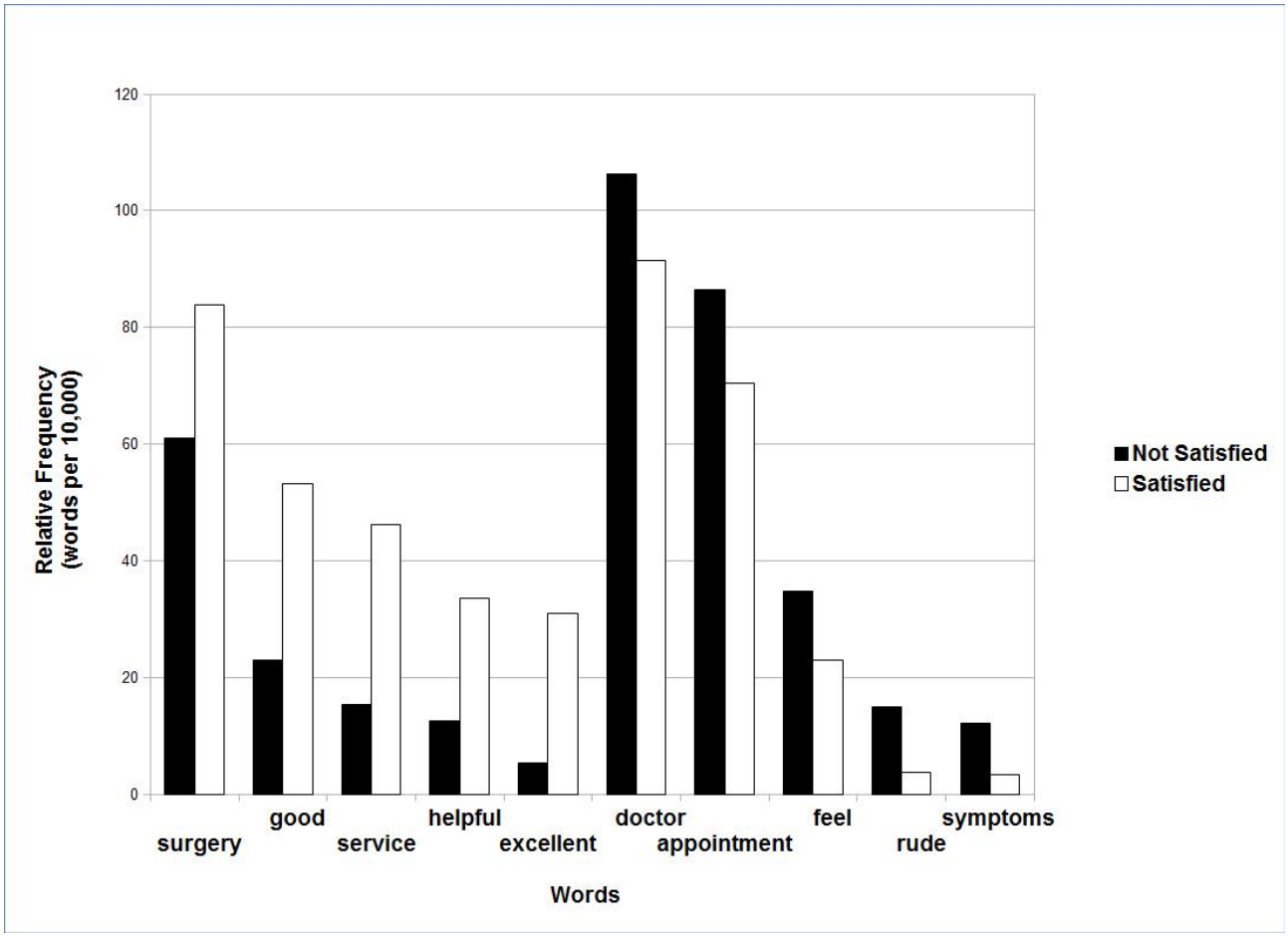
The two-word text cloud generated by Many Eyes is shown in Figure 2, displaying the 200 most frequent two-word phrases (bigrams). The five most frequent bigrams were: “reception staff”, “excellent service”, “two weeks”, “medical centre” and “good service”.

Figure 3 shows the results of the Voyant Tools distinctive words component when applied to comments categorized as originating from satisfied or dissatisfied patients. The words “surgery”, ”excellent”, “service”, “good”, and “helpful” were the five most distinctive words from satisfied patients, while the words “doctor”, “feel “, “appointment”, “rude”, and “symptoms” were the five most distinctive words in the comments from dissatisfied patients.

Figure 1. Single-word text cloud created in TagCrowd from all free text comments.

Figure 2. Two-word text cloud created in Many Eyes.

Figure 3. Word frequency by patient satisfaction.



Statistical Analysis

From the logistic regression models, odds ratios were calculated for the distinctive words. In this analysis, the odds ratio indicates the amount by which the odds of a particular word occurring at least once in a comment are multiplied for every point increase in the z-score. Table 1 reports the results of the logistic regression for the 10 distinctive words and the scores for satisfaction, doctor-patient communication, and confidence and trust of the patient in the doctor.

Table 1. Odds ratio (95% CI, R^2 value) for the occurrence of distinctive words corresponding to a one standard deviation increase in measures of patient experience.

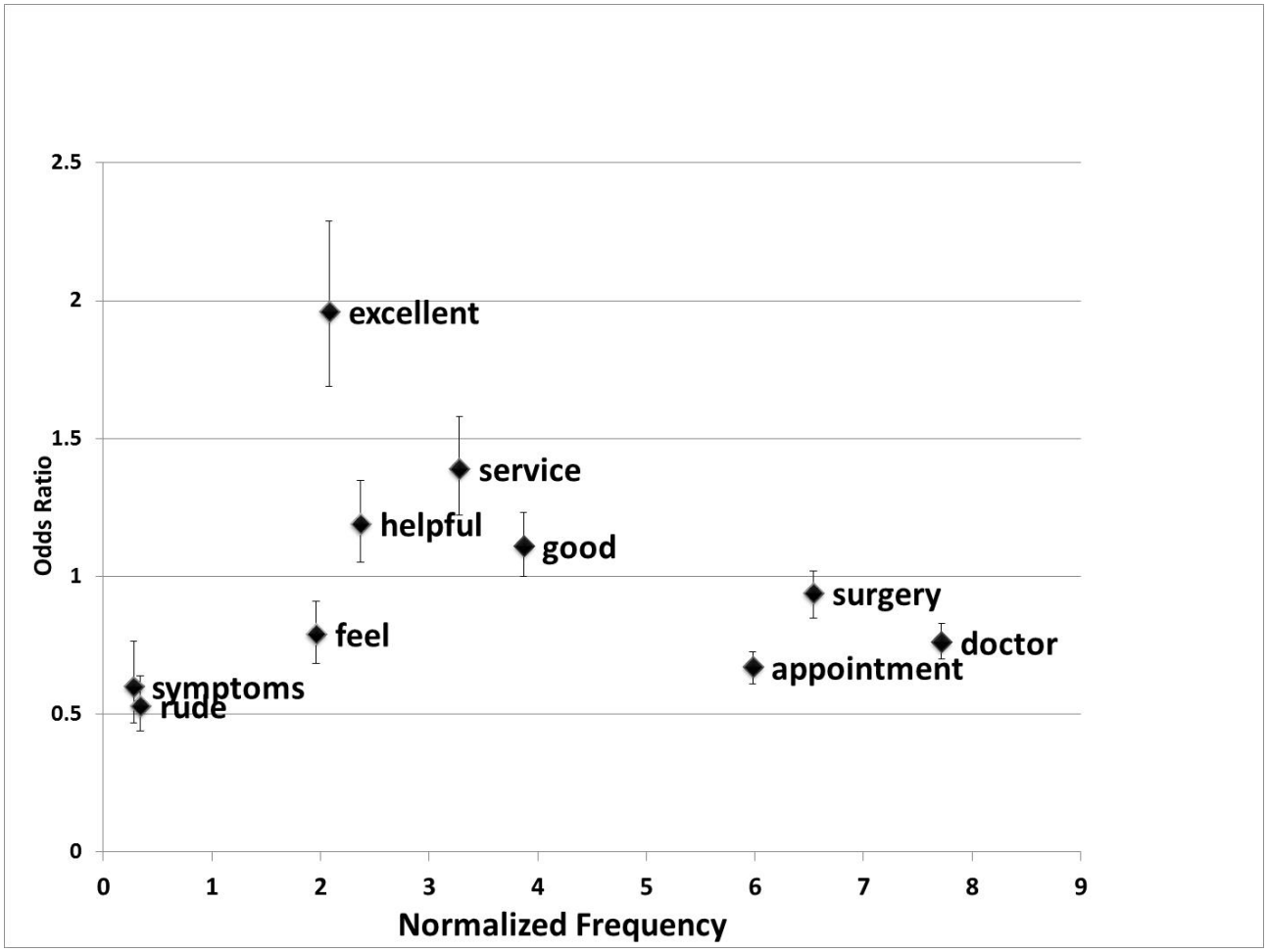
	Overall satisfaction N=3134	Doctor's communication skills N=3062	Confidence and trust in the doctor N=3066
Word	OR (CI, R^2)	OR (CI, R^2)	OR (CI, R^2)
Service	1.39 (1.25-1.54, .02)	1.25 (1.13-1.39, .007)	1.29 (1.16-1.43, .009)
Good	1.11 (1.02-1.21, .002)	1.09 (0.99-1.2, .001)	1.06 (0.97-1.16, .0005)
Excellent	1.96 (1.63-2.34, .04)	2.09 (1.69-2.58, .09)	1.76 (1.45-2.15, .003)
Surgery	0.94 (0.88-1.01, .0008)	0.98 (0.90-1.05, .0001)	1.00 (0.93-1.08, .00)
Helpful	1.19 (1.07-1.32, .005)	1.24 (1.10-1.40, .006)	1.10 (0.99-1.22, .001)
Appointment	0.67 (0.63-0.72, .04)	0.77 (0.77-0.82, .01)	0.80 (0.74-0.86, .01)
Doctor	0.76 (0.71-0.81, .02)	0.81 (0.75-0.87, .01)	0. 81 (0.76-0.88, .008)
Feel	0.79 (0. 72-0.87, .01)	0.78 (0.71-0.86, .01)	0.77 (0.70-0.85, .01)
Rude	0.53 (0.46-0.60, .10)	0.63 (0.55-0.74, .04)	0.60 (0.51-0.70, .05)
Symptoms	0.60 (0.51-0.70, .06)	0.61 (0.52-0.72, .05)	0.64 (0.54-0.77, .03)

As shown in the table, the regression for the word "excellent" results in an OR of 1.96, for the bivariate model for patient satisfaction; that is, an increase of one standard deviation in the patient satisfaction is associated with almost twice the odds of the word "excellent" occurring in the comments. There is also a significant association of the occurrence of "excellent" in the comments with the z-scores for doctor communication skills and confidence in the doctor which have odds ratios of 2.09 and 1.76 respectively.

In contrast, the word "rude" has an OR of 0.53 in the bivariate model for patient satisfaction, indicating that an increase in one standard deviation in the satisfaction score almost halves the odds that the word "rude" will appear in the comments. The OR is also significantly lower for the occurrence of "rude" when scores for doctor communication skills or confidence in the doctor are higher. To summarize, the words "service" and "excellent" had a significant positive association for overall satisfaction, doctor's communication skills scores, and confidence and trust in the doctor scores. The word "helpful" had a significant positive association for overall satisfaction and doctor's communication skill scores, but was not significant for confidence and trust. The words "rude" and "symptoms" had a significant negative association with all three scores.

Figure 4 shows a plot of the odds ratios for the occurrence of a word due to a one standard deviation increase in patient satisfaction score as calculated by the bivariate model. The odds ratios for ten distinctive words (five most distinctive words each from satisfied and dissatisfied patients) are plotted against their standardized frequencies (x-axis).

Figure 4. Frequency of selected words and the odds ratios (with 95% confidence intervals) associated with the z-scores for satisfaction.



Keywords in Context (KWIC) Analysis

We chose to look at the context of the usage of the word "rude" using Keyword in Context (KWIC) analysis. We examined 15 words to the left and right of the keyword in question because we felt that was the minimum amount where we could satisfactorily establish the context of use of the word in question. Earlier attempts using 10 words on each side gave results wherein the context was still ambiguous in some of the comments. We manually reviewed the output from the KWIC tool, and established the context of the various instances of the adjective "rude". We then constructed a table listing the sources of the rude behavior and their frequency of mention (Table 2).

Overall, "reception staff" was mentioned in 63% of the occurrences of the word "rude", while doctors accounted for 22% of occurrences. Among the patients who were dissatisfied, the proportion of doctors being associated with occurrences of the word "rude" increased to 30% compared to 15% in satisfied patients. Reception staff had a larger proportion of association with occurrences of the word "rude" among satisfied patients, 72%, than in patients who were dissatisfied: 54%.

Table 2. Keyword in Context (KWIC) analysis for the word "rude". Frequency of occurrence (% within patient type) by patient satisfaction and subject.

	Subject of adjective "rude"						
	Frequency of occurrence (% within patient type)						
Patients	Doctor	Nurse	Practice manager	Reception staff	Staff	Patient	Total
Not satisfied	11 (30)	1 (3)	2 (5)	20 (54)	3 (8)	0 (0)	37
Satisfied	6 (15)	3 (7)	1 (2)	29 (72)	1 (2)	1 (2)	41
All	17 (22)	4 (5)	3 (4)	49 (63)	4 (5)	1(1)	78

Discussion

Principal Results

We found the three textual analysis tools easy to use and the results were generated very quickly, considering the volume of text that was processed (approximately 150,699 words). The tools used the standard ASCII text file format, which most data analysis software can easily export to. The text clouds did give a concise summary of what the majority of comments were about, but it was difficult to establish the exact context of the use of the most frequent words. The distinctive word analysis

gave more insight by showing the different usage of words by differing sources of comments (satisfied vs dissatisfied patients). Some of the words in the output of the distinctive word analysis showed significant associations with satisfaction scores, notably, “excellent” and “rude”. The word “excellent” was associated with high patient satisfaction scores. The high frequency of the two-word phrase “excellent service” showed that patients used it in relation to their rating of the quality of service they received.

The word “rude” occurred much more frequently in comments from dissatisfied patients, suggesting that rude behavior encountered by patients may trigger dissatisfaction. The KWIC analysis showed that the word “rude”, was most commonly associated with the reception staff. Receptionists have been recognized as crucial members of the primary health care team [47], and recent work has suggested that the historical perception of the receptionist as a “dragon behind the desk” has been getting in the way of understanding the role of receptionists and thus improving patient care [48]. Also worth noting is that the proportion of rude actions being attributed to doctors was higher amongst the patients who were dissatisfied with their practice. Winsted has identified rudeness from doctors and other forms of negative behavior as being a “dissatisfier” in medical encounters [49].

The increased relative frequency of the word “feel” in the comments from dissatisfied patients might indicate an emotional reaction being a component of patient dissatisfaction. However, the word may also be used in other contexts, for instance “I feel that the doctor should”, as opposed to “I feel disappointed.” The recurrence of the word “symptoms” in the comments from dissatisfied patients could indicate a relationship between dissatisfaction and the perception of poor health, as has been reported previously by Xiao et al [50]. It may also provide a comment on the perceived thoroughness of the clinical encounter. One point of interest is that the words positively associated with patient satisfaction focus on the system (eg, “excellent service”), while those associated with dissatisfaction highlight some of the interpersonal aspects of care (eg, “rude”, “feel”).

Limitations

While the textual analysis applications are easy to use and give results quickly, one limitation is that an internet connection is required for all the software tools to work. However, a high-speed connection is not necessary, and the software runs on any modern operating system with an updated Web browser.

When we attempted to identify the messages contained in the text cloud, we found it difficult to ascertain the significance of the high frequency of the words “doctor”, “practice”, “surgery”, “appointment”, and “time”. This is due to the text cloud showing the words dissociated from their original context, making it difficult to discern the meaning behind the high frequencies of these words. This loss of context due to the dissociation of the words from one another is a major limitation in the interpretation of the results of the text cloud. When words are separated from one another, and only their frequencies rather than their relationships are scrutinized, there is a danger of overlooking subtle and important nuances and meanings formed by the synergy of the words [24]. The software tools are also limited in that they are unable to group together words that are synonymous, (eg, “doctor”, “dr”, and “gp”), unless the software is specifically instructed to group these synonyms. In addition to the individual words, meaning is also conveyed by the patterns that words form.

Another consequence of dissociation is that our method does not automatically deal with negation of terms. However, for the words “good”, “excellent”, “helpful” “rude”, and “symptoms”, we examined the results of the keyword in context extraction to see if they contained instances of negation. We were satisfied that all mentions of those words did not contain instances of negation. A more sophisticated approach using natural language processing and machine learning is required to automatically deal with negation. Sentiment analysis, which is a more sophisticated textual analysis technique, is one method that takes the patterns of words, and not just their frequencies, into account. Research reports are emerging in which sentiment analysis has been used to examine free text comments from patients [51-57].

A further limitation of this study is related to the nature of the question being presented to elicit the comments. The very broad nature of the request for comments means that the patients’ responses were, almost inevitably, quite varied. The wide spectrum of issues raised in the comments make them quite difficult to neatly categorize and characterize. The quality of information gleaned from patient responses could be improved by focusing the wording of the request for comments to address central issues of interest [9]. This focus of interest could also be coupled to particular quantitative questions, to give insight into why the patient answered the question in that particular way.

Further Research

Web-based textual analysis shows promise as a means of rapidly summarizing the messages contained in free-text comments from primary care patients. Text clouds are a feasible means of presenting the most frequent words used in free-text comments from patients. However, text clouds are limited by an inability to provide a contextually meaningful summary of the original corpus of comments. This is commonly encountered when relying primarily on a simple, mechanistic algorithm, in this case, word frequency. Words convey meaning by working together, and there is a synergy created through the combination of various words [24]. A more accurate way of capturing the messages contained in the free-text comments by a computer mediated approach is through KWIC analysis. The use of more sophisticated technologies, such as machine learning, natural language processing, neural networks, and sentiment analysis may address some of these shortcomings. Future research needs to be done around generating sentence level summarization using the techniques from the NLP community [58], such as latent Dirichlet allocation [59,60] . For a wider uptake, a user-friendly (preferably open source) application needs to be developed to fill this gap. This would enable practices to make better use of the large amounts of free-text feedback that they have collected. In addition, careful attention needs to be paid to formulating focused and precise requests for comments which

might be expected to yield feedback that could provide a substantial basis for computer mediated textual analysis. Finally, mixed methods approaches as well as sociocybernetics methods have also been proposed as a way of completing the picture of patient experience [61,62].

Conclusions

Our study has shown that by using Web-based text processing tools to extract information from patient comments, we can discover words that the patients have used in their comments that have significant associations with quantitative measurements of patient experience. The logistic regression revealed strong positive and negative associations between the satisfaction scores and the occurrence of certain words. KWIC analysis was then used to examine the context of the uses of words, which yielded useful information; for example, the sources of rude behavior that is associated with patient dissatisfaction. This approach could help practices in formulate policies to increase patient satisfaction. Sequential use of these methods may prove useful in documenting how patients' experience of care changes over time, similar to the method used by Gill et al in revealing the longitudinal changes in the document "Good Medical Practice" produced by the General Medical Council [24]. An approach that examines the key words in the context is useful in deriving insights from the free-text comments. Further research is necessary in refining these methods, so that the results would be comparable to traditional techniques of content analysis.

Acknowledgments

This paper presents independent research funded by the National Institute for Health Research (NIHR) under its Programme Grants for Applied Research Programme (Grant Reference Number RP-PG-0608-10050). The views expressed are those of the authors and not necessarily those of the National Health Service England (NHS England), the NIHR or the Department of Health.

We would like to thank the patients, practice managers, GPs, and other staff of the general practices who kindly agreed to participate in this study and without whom the study would not have been possible. Thanks also to Gary Abel, Natasha Elmore, Emily Taylor, Jenny Newbould, Emma Whitton, Amy Gratton, Charlotte Paddison, Mary Carter, and Dawn Swancutt for invaluable help with study set-up, practice recruitment, data collection, and data entry. We confirm that all personal identifiers have been removed or disguised so that study participants are not identifiable and cannot be identified through the provided details.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Screenshots of Web based textual analysis tools.

[PDF File (Adobe PDF File), 1MB - medinform_v3i2e20_app1.pdf]

Multimedia Appendix 2

Survey questionnaire.

[PDF File (Adobe PDF File), 290KB - medinform_v3i2e20_app2.pdf]

References

1. Roland M, Elliott M, Lyratzopoulos G, Barbiere J, Parker R, Smith P, et al. Reliability of patient responses in pay for performance schemes: analysis of national General Practitioner Patient Survey data in England. BMJ 2009 Sep 29;339(sep29 3):b3851-b3851 [FREE Full text] [doi: 10.1136/bmj.b3851]
2. Jenkinson C, Coulter A, Bruster S. The Picker Patient Experience Questionnaire: development and validation using data from in-patient surveys in five countries. Int J Qual Heal Care Internet 2002;14(5):353-358 [FREE Full text]
3. Giordano LA, Elliott MN, Goldstein E, Lehrman WG, Spencer PA. Development, implementation, and public reporting of the HCAHPS survey. Med Care Res Rev 2010 Feb;67(1):27-37. [doi: 10.1177/1077558709341065] [Medline: 19638641]
4. Cleary PD. The increasing importance of patient surveys. Qual Health Care 1999 Dec;8(4):212 [FREE Full text] [Medline: 10847881]
5. Cleary P. Expanding the use of patient reports about patient-centered care. Isr J Health Policy Res Internet Israel Journal of Health Policy Research 2013 Jan [FREE Full text]
6. Wright C, Richards SH, Hill JJ, Roberts MJ, Norman GR, Greco M, Taylor Matthew R S, et al. Multisource feedback in evaluating the performance of doctors: the example of the UK General Medical Council patient and colleague questionnaires. Acad Med 2012 Dec;87(12):1668-1678. [doi: 10.1097/ACM.0b013e3182724cc0] [Medline: 23095930]
7. Graham C, Woods P. Patient experience surveys. In: Ziebland S, Coulter A, Calabrese JD, Locock L, editors. Understanding and Using Health Experiences: Improving patient care. Oxford: Oxford University Press; 2013.

8. Asprey A, Campbell JL, Newbould J, Cohn S, Carter M, Davey A, et al. Challenges to the credibility of patient feedback in primary healthcare settings: a qualitative study. Br J Gen Pract 2013 Mar;63(608):e200-e208 [FREE Full text] [doi: 10.3399/bjgp13X664252] [Medline: 23561787]
9. Riiskjær E, Ammentorp J, Kofoed P. The value of open-ended questions in surveys on patient experience: number of comments and perceived usefulness from a hospital perspective. Int J Qual Health Care 2012 Oct;24(5):509-516 [FREE Full text] [doi: 10.1093/intqhc/mzs039] [Medline: 22833616]
10. Liu GC, Harris MA, Keyton SA, Frankel RM. Use of unstructured parent narratives to evaluate medical student competencies in communication and professionalism. Ambul Pediatr 2007;7(3):207-213. [doi: 10.1016/j.ambp.2007.03.001] [Medline: 17512880]
11. Shuyler K, Knight K. What are patients seeking when they turn to the Internet? Qualitative content analysis of questions asked by visitors to an orthopaedics Web site. J Med Internet Res 2003 Oct 10;5(4):e24 [FREE Full text] [doi: 10.2196/jmir.5.4.e24] [Medline: 14713652]
12. Tang PC, Black W, Young CY. Proposed criteria for reimbursing eVisits: content analysis of secure patient messages in a personal health record system. AMIA Annu Symp Proc 2006:764-768 [FREE Full text] [Medline: 17238444]
13. Burner ER, Menchine MD, Kubicek K, Robles M, Arora S. Perceptions of successful cues to action and opportunities to augment behavioral triggers in diabetes self-management: qualitative analysis of a mobile intervention for low-income Latinos with diabetes. J Med Internet Res 2014;16(1):e25 [FREE Full text] [doi: 10.2196/jmir.2881] [Medline: 24476784]
14. Ashley L, Jones H, Thomas J, Newsham A, Downing A, Morris E, et al. Integrating patient reported outcomes with clinical cancer registry data: a feasibility study of the electronic Patient-Reported Outcomes From Cancer Survivors (ePOCS) system. J Med Internet Res 2013;15(10):e230 [FREE Full text] [doi: 10.2196/jmir.2764] [Medline: 24161667]
15. Richards SH, Campbell JL, Walshaw E, Dickens A, Greco M. A multi-method analysis of free-text comments from the UK General Medical Council Colleague Questionnaires. Med Educ 2009 Aug;43(8):757-766. [doi: 10.1111/j.1365-2923.2009.03416.x] [Medline: 19659489]
16. Frohna A, Stern D. The nature of qualitative comments in evaluating professionalism. Med Educ 2005 Aug;39(8):763-768. [doi: 10.1111/j.1365-2929.2005.02234.x] [Medline: 16048618]
17. Verhoef LM, Van de Belt TH, Engelen L, Schoonhoven L, Kool RB. Social media and rating sites as tools to understanding quality of care: a scoping review. J Med Internet Res 2014;16(2):e56 [FREE Full text] [doi: 10.2196/jmir.3024] [Medline: 24566844]
18. Hearst M, Rosner D. Tag Clouds: Data Analysis Tool or Social Signaller? : Data Analysis Tool or Social Signaller? Proc 41st Annu Hawaii Int Conf Syst Sci (HICSS 2008) Internet Ieee; 2008 Presented at: Hawaii International Conference on System Sciences, Proceedings of the 41st Annual; 7-10 Jan. 2008; Waikoloa, HI p. 160 URL: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4438863
19. Sinclair J, Cardew-Hall M. The folksonomy tag cloud: when is it useful? Journal of Information Science 2007 May 31;34(1):15-29 [FREE Full text] [doi: 10.1177/0165551506078083]
20. Thomas M, Caudle D, Schmitz C. To tag or not to tag? Library Hi Tech 2009 Sep 04;27(3):411-434 [FREE Full text] [doi: 10.1108/07378830910988540]
21. Schweier R, Romppel M, Richter C, Hoberg E, Hahmann H, Scherwinski I, et al. A web-based peer-modeling intervention aimed at lifestyle changes in patients with coronary heart disease and chronic back pain: sequential controlled trial. J Med Internet Res 2014;16(7):e177 [FREE Full text] [doi: 10.2196/jmir.3434] [Medline: 25057119]
22. Schweiger S, Oeberst A, Cress U. Confirmation bias in web-based search: a randomized online study on the effects of expert information and social tags on information search and evaluation. J Med Internet Res 2014;16(3):e94 [FREE Full text] [doi: 10.2196/jmir.3044] [Medline: 24670677]
23. Lamantia J. Internet 2007 cited. 2014 Apr 16. Text clouds: A new form of tag cloud URL: http://www.joelamantia.com/blog/archives/tag_clouds/text_clouds_a_new_form_of_tag_cloud.html [accessed 2014-08-13] [WebCite Cache ID 6RnT6KYBL]
24. Gill D, Griffin A. Good Medical Practice: what are we trying to say? Textual analysis using tag clouds. Med Educ 2010 Mar;44(3):316-322. [doi: 10.1111/j.1365-2923.2009.03588.x] [Medline: 20444063]
25. Rivadeneira AW, Gruen DM, Muller MJ, Millen DR. Getting our head in the clouds: toward evaluation studies of tagclouds. USA: ACM; 2007 Presented at: CHI '07 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems; 28 April - 3 May, 2007; San Jose, California p. 995-998.
26. Park S, Griffin A, Gill D. Working with words: exploring textual analysis in medical education research. Med Educ 2012 Apr;46(4):372-380. [doi: 10.1111/j.1365-2923.2011.04184.x] [Medline: 22429173]
27. Moen A, Hackl W, Hofdijk J, Van Gemert-Pijnen L, Ammenwerth E, Nykänen P, et al. eHealth in Europe - Status and Challenges. Yearb Med Inform 2013;8(1):59-63. [Medline: 23974549]
28. Al Hamarneh Yazid N, Rosenthal M, McElnay JC, Tsuyuki RT. Pharmacists' perceptions of their practice: a comparison between Alberta and Northern Ireland. Int J Pharm Pract 2012 Feb;20(1):57-64. [doi: 10.1111/j.2042-7174.2011.00163.x] [Medline: 22236181]

29. O'Grady L, Wathen CN, Charnaw-Burger J, Betel L, Shachak A, Luke R, et al. The use of tags and tag clouds to discern credible content in online health message forums. Int J Med Inform 2012 Jan;81(1):36-44. [doi: 10.1016/j.ijmedinf.2011.10.001] [Medline: 22030035]
30. Osborne LA, Noble JG, Lockhart-Jones HM, Middleton R, Thompson S, Maramba IDC. Sources of discovery, reasons for registration, and expectations of an internet-based register for Multiple Sclerosis: Visualisations and explorations of word uses and contexts. International Journal of Healthcare Information Systems and Informatics (IJHISI) 2012;7(3):27-43.
31. Hamon T, Gagnayre R. Improving knowledge of patient skills thanks to automatic analysis of online discussions. Patient Educ Couns 2013 Aug;92(2):197-204. [doi: 10.1016/j.pec.2013.05.012] [Medline: 23769423]
32. Archer J, McGraw M, Davies H. Assuring validity of multisource feedback in a national programme. Arch Dis Child 2010 May;95(5):330-335. [doi: 10.1136/adc.2008.146209] [Medline: 20457700]
33. Liu J, Li A, Seneff S. Automatic Drug Side Effect Discovery from Online Patient-Submitted Reviews: Focus on Statin Drugs. 2011 Presented at: IMMM 2011, The First International Conference on Advances in Information Mining and Management; October 23, 2011 to October 29, 2011; Barcelona, Spain p. 91-96 URL: http://www.thinkmind.org/index.php?view=article&articleid=immm_2011_5_10_20050
34. McSherry W, Jamieson S. The qualitative findings from an online survey investigating nurses' perceptions of spirituality and spiritual care. J Clin Nurs 2013 Nov;22(21-22):3170-3182. [doi: 10.1111/jocn.12411] [Medline: 24118520]
35. Chary M, Park EH, McKenzie A, Sun J, Manini AF, Genes N. Signs & Symptoms of Dextromethorphan Exposure from YouTube. PLoS One Internet 2014 Jan [FREE Full text] [doi: 10.1371/journal.pone.0082452]
36. Improve Study Team. 2014. Improve Study Website Internet URL: https://improve.exeter.ac.uk/ [accessed 2014-08-15] [WebCite Cache ID 6RqYPViDo]
37. Campbell J, Smith P, Nissen S, Bower P, Elliott M, Roland M. The GP Patient Survey for use in primary care in the National Health Service in the UK--development and psychometric characteristics. BMC Fam Pract Internet 2009 Jan [FREE Full text] [doi: 10.1186/1471-2296-10-57]
38. gp-patient. uk. co URL: https://gp-patient.co.uk/about [accessed 2014-08-13] [WebCite Cache ID 6RnSCcfIL]
39. Roberts MJ, Campbell JL, Abel GA, Davey AF, Elmore NL, Maramba I, et al. Understanding high and low patient experience scores in primary care: analysis of patients' survey data for general practices and individual doctors. BMJ 2014 Nov 11;349(nov11 3):g6034-g6034 [FREE Full text] [doi: 10.1136/bmj.g6034]
40. Steinbock D. 2014. What is TagCrowd? Internet URL: http://www.tagcrowd.com/blog/about/ [accessed 2014-08-13] [WebCite Cache ID 6RnSQLTIu]
41. IBM. 2014. Many Eyes Internet URL: http://www-958.ibm.com/software/analytics/manyeyes/page/Tour.html [accessed 2014-08-13] [WebCite Cache ID 6RnSWkipS]
42. Sinclair S. ca – The Rhetoric of Text Analysis Internet. Voyeur Tools: See Through Your Texts | Hermeneuti URL: http://hermeneuti.ca/voyeur [accessed 2014-08-13] [WebCite Cache ID 6RnSnhJw9]
43. Baskarada S, Koronios A. Data, Information, Knowledge, Wisdom (DIKW): A Semiotic Theoretical and Empirical Exploration of the Hierarchy and its Quality Dimension. Australasian Journal of Information Systems 2013 Jun 6;18(1) [FREE Full text]
44. Leech NL, Onwuegbuzie AJ. An array of qualitative data analysis tools: A call for data analysis triangulation. School Psychology Quarterly 2007;22(4):557-584. [doi: 10.1037/1045-3830.22.4.557]
45. Saiki LS, Cloyes KG. Blog Text About Female Incontinence: Presentation of Self, Disclosure, and Social Risk Assessment. Nurs Res Internet(2) 2014;63 [FREE Full text] [doi: 10.1097/NNR.0000000000000016]
46. Barker F, de LS, Baguley D, Gagne JP. An evaluation of audiology service improvement documentation in England using the chronic care model and content analysis. Int J Audiol 2014 Jun;53(6):377-382. [doi: 10.3109/14992027.2013.860242] [Medline: 24313709]
47. Arber S, Sawyer L. The role of the receptionist in general practice: A ‘dragon behind the desk’? Social Science & Medicine 1985 Jan;20(9):911-921 [FREE Full text] [doi: 10.1016/0277-9536(85)90347-8]
48. Hammond J, Gravenhorst K, Funnell E, Beatty S, Hibbert D, Lamb J. Slaying the dragon myth: an ethnographic study of receptionists in UK general practice. British Journal of General Practice 2013 [FREE Full text] [doi: 10.3399/bjgp13X664225]
49. Frazer Winsted K. Patient satisfaction with medical encounters – a cross‐cultural perspective. International Journal of Service Industry Management 2000 Dec;11(5):399-421. [doi: 10.1108/09564230010360137]
50. Xiao H, Barber JP. The effect of perceived health status on patient satisfaction. Value Health 2008 Jul;11(4):719-725. [doi: 10.1111/j.1524-4733.2007.00294.x] [Medline: 18179667]
51. Alemi F, Torii M, Clementz L, Aron DC. Feasibility of real-time satisfaction surveys through automated analysis of patients' unstructured comments and sentiments. Qual Manag Health Care 2012;21(1):9-19. [doi: 10.1097/QMH.0b013e3182417fc4] [Medline: 22207014]
52. Greaves F, Ramirez-Cano D, Millett C, Darzi A, Donaldson L. Use of sentiment analysis for capturing patient experience from free-text comments posted online. J Med Internet Res 2013;15(11):e239 [FREE Full text] [doi: 10.2196/jmir.2721] [Medline: 24184993]

53. Cambria E, Hussain A, Durrani T, Havasi C, Eckl C, Munro J. Sentic Computing for patient centered applications. : IEEE; 2010 Presented at: IEEE 10th Int Conf SIGNAL Process Proc Internet; 2010; Beijing p. 1279-1282 URL: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5657072 [doi: 10.1109/ICOSP.2010.5657072]
54. Sokolova M, Bobicev V. Sentiments and Opinions in Health-related Web Messages. In: Proceedings of Recent Advances in Natural Language Processing. 2011 Presented at: RANLP 2011; 12-14 September 2011; Hissar, Bulgaria p. 132-139.
55. Sokolova M, Bobicev V. What Sentiments Can Be Found in Medical Forums? In: Proceedings of Recent Advances in Natural Language Processing 2013. 2013 Presented at: RANLP; 7-13 September 2013; Hissar, Bulgaria p. 639.
56. Bobicev V, Sokolova M, Jafer Y, Schramm D. Learning Sentiments from Tweets with Personal Health Information. : Springer Berlin Heidelberg; 2012 Presented at: 25th Canadian conference on Advances in Artificial Intelligence; May 28-30 2012; Toronto, Canada p. 37-48. [doi: 10.1007/978-3-642-30353-1_4]
57. Bobicev V, Sokolova M, Oakes M. Recognition of Sentiment Sequences in Online Discussions. In: Proceedings of the Second Workshop on Natural Language Processing for Social Media (SocialNLP). 2014 Presented at: SocialNLP; August 24 2014; Dublin, Ireland p. 44-49 URL: http://www.google.ca/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=0CCMQFjAA&url=http%3A%2F%2Fwww.aclweb.org%2Fanthology%2FW14-5907&ei=jbfGVKjALcuxggTw8oGgBw&usg=AFQjCNGx-VGr1kuO6RT6xdqGH_nP_VJOoA&sig2=TCX_gu8OwCseEFNFBrMoFw
58. Das D, Martins AFT. A survey on automatic text summarization. Lit Surv Lang Stat II course C 2007;4:192-195 [FREE Full text]
59. Blei DM, Ng AY, Jordan MI. Latent dirichllocation. Journal of Machine Learning Research 3 2003:993-102 [FREE Full text]
60. Wallace BC, Paul MJ, Sarkar U, Trikalinos TA, Dredze M. A large-scale quantitative analysis of latent factors and sentiment in online doctor reviews. J Am Med Inform Assoc 2014;21(6):1098-1103. [doi: 10.1136/amiajnl-2014-002711] [Medline: 24918109]
61. Lyles CR, Sarkar U. Additional considerations for 'Harnessing the cloud of patient experience'. BMJ Qual Saf 2013 Aug;22(8):698 [FREE Full text] [doi: 10.1136/bmjqs-2013-001893] [Medline: 23476069]
62. Dijkum C, Lam N, Verheul W. The challenge of modeling complexity in the social sciences illustrated with an example: the communication between GP and Patients. Methodol Rev Appl Res Internet 2013 [FREE Full text]

Abbreviations

ASCII: American Standard Code for Information Interchange
CI: confidence interval
GPPS: General Practice Patient Survey
KWIC: keyword in context
NHS England: National Health Service England
NIHR: National Institute for Health Research
NLP: natural language processing
OR: odds ratio

Edited by G Eysenbach; submitted 22.08.14; peer-reviewed by M Sokolova, WY Wang; comments to author 21.10.14; revised version received 02.12.14; accepted 20.12.14; published 06.05.15.

Please cite as:
Maramba ID, Davey A, Elliott MN, Roberts M, Roland M, Brown F, Burt J, Boiko O, Campbell J
Web-Based Textual Analysis of Free-Text Patient Experience Comments From a Survey in Primary Care
JMIR Med Inform 2015;3(2):e20
URL: http://medinform.jmir.org/2015/2/e20/
doi:10.2196/medinform.3783
PMID:25947632

Original Paper

Prioritization of Free-Text Clinical Documents: A Novel Use of a Bayesian Classifier

Mark Singh[1*], BE(Elec), M.D.; Akansh Murthy[2*], BS; Shridhar Singh[3*], BS (Current)

[1]Carnegie Mellon University, University of Massachusetts Medical School, Braintree, MA, United States

[2]Massachusetts Institute of Technology, Cambridge, MA, United States

[3]Carnegie Mellon University, Pittsburgh, PA, United States

[*]all authors contributed equally

Corresponding Author:
Akansh Murthy, BS
Massachusetts Institute of Technology
77 Mass Ave
Cambridge, MA, 02139
United States
Phone: 1 6172531000
Email: ambshun@mit.edu

Related Article:

This is a corrected version. See correction statement: https://medinform.jmir.org/2020/6/e21379

Abstract

Background: The amount of incoming data into physicians' offices is increasing, thereby making it difficult to process information efficiently and accurately to maximize positive patient outcomes. Current manual processes of screening for individual terms within long free-text documents are tedious and error-prone. This paper explores the use of statistical methods and computer systems to assist clinical data management.

Objective: The objective of this study was to verify and validate the use of a naive Bayesian classifier as a means of properly prioritizing important clinical data, specifically that of free-text radiology reports.

Methods: There were one hundred reports that were first used to train the algorithm based on physicians' categorization of clinical reports as high-priority or low-priority. Then, the algorithm was used to evaluate 354 reports. Additional beautification procedures such as section extraction, text preprocessing, and negation detection were performed.

Results: The algorithm evaluated the 354 reports with discrimination between high-priority and low-priority reports, resulting in a bimodal probability distribution. In all scenarios tested, the false negative rates were below 1.1% and the recall rates ranged from 95.65% to 98.91%. In the case of 50% prior probability and 80% threshold probability, the accuracy of this Bayesian classifier was 93.50%, with a positive predictive value (precision) of 80.54%. It also showed a sensitivity (recall) of 98.91% and a F-measure of 88.78%.

Conclusions: The results showed that the algorithm could be trained to detect abnormal radiology results by accurately screening clinical reports. Such a technique can potentially be used to enable automatic flagging of critical results. In addition to accuracy, the algorithm was able to minimize false negatives, which is important for clinical applications. We conclude that a Bayesian statistical classifier, by flagging reports with abnormal findings, can assist a physician in reviewing radiology reports more efficiently. This higher level of prioritization allows physicians to address important radiologic findings in a timelier manner and may also aid in minimizing errors of omission.

(JMIR Med Inform 2015;3(2):e17) doi:10.2196/medinform.3793

KEYWORDS

clinical reports; prioritization; Bayesian classifier; radiology; natural language processing

Introduction

Data Concerns

In today's environment with electronic medical records (EMR) gaining prevalence in hospitals, urgent care clinics, and specialist facilities, primary care physicians are receiving more clinical reports on a daily basis. These electronic systems typically generate more pages per report than in the past. A Brigham and Women's Hospital study reports that full-time primary care physicians on average review 930 pieces of chemistry/hematology data and 60 pathology/radiology reports in a given week [1]. Also, with the increasing utilization of new imaging modalities, such as computerized axial tomography (CAT) scans, magnetic resonance images (MRI), and pet scans in addition to traditional plain film studies, physicians have to process more types of reports, manage incidental findings, as well as significant findings that may require follow-up over an interval of a few weeks to even years. To compound matters, there are also more numbers of insured patients coming into the medical care community [2]. Given the existing data load and a potential increase in data [3], it will be challenging for a physician to keep up with the workload efficiently. Consequently, it is not uncommon even now for a clinician to overlook or fail to address an abnormal result. In the outpatient settings, between 8% and 26% of abnormal test results, including those suspicious for malignancy, are not followed up in a timely manner [4]. Failing to do so can result in patient morbidity and mortality, as well as possible costly malpractice litigation.

In fact, failure to review and follow up on an outpatient test result compromises patient safety and raises malpractice concerns in the order of billions of dollars annually [5]. A regional Veterans Administration health care network study indicated that almost 65% of diagnostic errors are due to abnormal test results that were missed and not addressed appropriately [6,7]. Despite the greater availability of EMR with test result transmission and notification availability, the problem of missed test results has not been eliminated. This missing of abnormal results was true even when one or more providers read the results. Alert fatigue, an inevitable presence with multiple electronic systems, is also a huge concern, especially since it results in physicians ignoring vital alerts about patients [8]. Overlooking these key recommendations or findings contained in a report, such as detection of an early cancer or a new medical condition results in adverse patient outcomes, annually, more than 100,000 patient deaths [5]. Thus, there exists a critical need for a more reliable method of clinical report management.

Literature Review

Currently, patient clinical data are both structured and unstructured. Structured patient data are typically a laboratory test containing discrete numerical values. An example can be a patient's potassium result, which could have a value of "4.2". Discrete results can easily be identified and traced by automatic systems. Urgent or critical laboratory values can be detected by performing a simple numerical comparison. A problem exists, however, with free-text reports such as radiology results. These reports have to be read by the physician and important findings need to be noted and logged for proper tracking [9]. Automatic interpretation of these free-text reports and determination of whether they contain a critical finding has been challenging for computers.

There are many existing applications that use natural language processing (NLP) to extract patient medical information from free-text reports for purposes of medical billing or populating a patient's health record. Several studies have also demonstrated the ability of NLP to extract clinical information, such as pneumonia cases, from radiology reports [10,11]. Another study validated the use of a Bayesian classifier to identify the diagnosis of appendicitis from radiology reports based on training data [12]. Further experiments have demonstrated the feasibility of using statistical text classification to detect severe extreme-risk events in clinical incident reports [13,14]. However, current literature does not contain within it an application that classifies a real-time stream of incoming free-text radiology reports, automatically flags critical reports as high-priority, and learns from the physician's actions.

By performing an initial screen of incoming data and flagging reports as potentially low-priority or high-priority, a classifier can aid physicians in better prioritization of his or her stack of clinical data that is to be reviewed. The intent is not to replace the manual review and signing off of each report, but rather to assist the physician by providing a level of prioritization to the stack of unordered documents awaiting review, an additional safety net. The benefits of such a system would be, at the very least, quicker notification of results to a patient and fewer missed or overlooked findings, resulting in better patient outcomes and possibly even less malpractice exposure.

Naive Bayes Approach

A statistical approach using the Bayes theorem was developed to classify free clinical reports as low- or high-priority. A naive Bayesian classifier is a probabilistic classifier based on Bayes' theorem that makes a strong independence assumption. In context, the classifier assumes that all features of a document are independent of one another. The presence or absence of one feature is assumed to have no effect on the presence or absence of any other feature. When classifying text, each feature is an individual word in the text.

A supervised learning approach is used to enable a naive Bayes classifier to differentiate a document into different categories. In the case of classifying clinical reports, the classifier categorizes reports as low- or high-priority. The classifier is trained using a corpus of documents that is already categorized. The corpus of documents is tokenized, and each word from the documents is assigned a probability of appearing in a high-priority report. Each word, or feature, is represented by "f_i". The probability of a report being high-priority is given by the Bayes theorem shown here in generic form,

$$P(H \mid f_i) = [P(f_1, f_2, f_3 ... f_i \mid H)\ P(H)]\ /$$

$$[P(H)\ P((f_1, f_2, f_3 ... f_i \mid H) + (1-P(H))\ (1-P(f_1, f_2, f_3 ... f_i \mid H))].$$

The term P(H) represents the prior probability or the probability of any given document being high-priority. $P(f_i \mid H)$ is the probability of the feature, f_i, appearing in a document given that

the document in question is high-priority. The denominator is the probability of f_i appearing in any given document, or $P(f_i)$. Thus, the equation can be simplified into,

$P(H \mid f) = [P(f, f,f...f \mid H)\ P(H)] /$

$[P(f_1, f_2,f_3...f_i)].$

The experiment is based on the premise that there are patterns within clinical reports that influence a physician's determination of the reports severity, and these patterns can be detected by a computer based on the relative presence of certain words in documents. If true, then a computer could use principles of statistics and machine learning to prioritize free-text clinical reports.

Methods

Study Site

Blue Hills Medical Associates is an internal medicine practice, consisting of 2 physicians and 1 nurse practitioner situated with the encatchment area of three community hospitals, each affiliated with a separate major Massachusetts health system. The practice sees over 60 patients daily and receives over 5000 pages of clinical reports each month in the form of faxes, paper mail, and electronic results via a health level-7 (HL7) interface. These reports include consult reports, laboratory results, hospital admission and discharge reports, as well as radiology reports. The focus of this study was on the management of radiology results. There were 2 primary care physicians who reviewed the reports.

Datasets

There were two sets of data that were used in this study. Both sets of data were extracted from clinical reports stored in the EMR at the practice site used for this study. The first set, the training data, was used to train the Bayesian classifier to detect physicians' definitions of what constitutes a high-priority report. The second set of data, the test data, was a set of documents independent from the training set used to test and validate the classifier against the physicians' own categorization.

Radiology reports usually have an Impression section, summarizing the report's key findings. The Impression section is the interpreting radiologist's summarization and prioritization of the report's key findings. Focusing on this section allows for easier data processing, since the reports have been preprioritized by level of importance by the radiologist. The Impression section was extracted from each report for inclusion in the corpora based on the assumption that it contained the key information that distinguishes a low-priority report from a high-priority report. In the few cases where a report does not contain an Impression section or its equivalent, the entire report body was processed. The described extraction limits the amount of extraneous data.

Training Data

There were one hundred reports, 50 from each category generated between the years of 2011 and 2013, that were selected from the EMR that were representative of the types of low- and high-priority reports seen in study site. These were then categorized into low- and high-priority by the physicians. Figure 1 shows examples of deidentified high-priority and low-priority reports in common text format.

Figure 1. Deidentified high-priority (top) and low-priority (bottom) patient reports in text format.

abnormal

```
Name:      xxxxxxxxxx                                                    Ref MD: xxxxxxxxx

MRN: xxxxxxx                Pt Type: REG CLI              Loo: DISTE      Account:    xxxxxxxxxx

Age: 54                     DOB:                                          Service Date:

Clinical History: RT UPPER QUAD PAIN & EPIGASTRIC PAIN

Modality: US
Exam: Abdominal Complete Ultrasound                                       Order Date and #:
CPT #: 76700


History:      RT UPPER QUAD PAIN EPIGASTRIC PAIN

Technique;     Grayscale and color ultrasound of the abdomen

Comparison:      None

Findings: The liver demonstrates  normal echotexture. No sonographic evidence for liver mass. There is no
biliary dilatation. The common bile duct measures 0.4 cm.

There is a 1.5 cm polypoid echogenic structure associated with the gallbladder wall which does not
shadow most compatible with a polyp. There are a few additional subcentimeter similar appearing
structures also most compatible small polyps. There are no gallstones or gallbladder wall thickening.

The kidneys are normal in size and echotexture The right kidney measures 1 1.8 x 4.4 cm x 4.9 cm. The
left kidney measures 11.5 cm x 5.4 cm x 4.4 cm. There is no hydronephrosis. The visualized pancreas is
unremarkable. The spleen is normal in size measuring 11.1 cm. There is no ascites. The abdominal aorta is
normal in caliber. The IVC is unremarkable.

Impression:

1. Multiple gallbladder polyps with the largest measuring 1.5 cm. surgical consultation is suggested.
```

normal

```
Hospital Diagnostic Imaging

Name: xxxxxx                        Ref MD: xxxxx
MRN: xxxx                           Account: xxxxx
Age: xxxx                           Pt Type: REG CLI
DOB: xxxx                           Service Date: xx/xx/xx                Loc: xxx

Clinical History: COUGH

Modality: DX
Exam: Chest Pa and Lat CXR                                                Order Date and #: xxxx
CPT #: 71020

History:
Cough.

Technique:
Chest, PA and lateral dated xxxx, xxxx at 1114 hours.

Findings:
The lungs are clear bilaterally. The cardiac silhouette is normal. There are degenerative changes in the
thoracic spine.

Impression:
1. No acute pathology.
```

Test Data

There were three hundred and fifty four radiology and diagnostic reports, ordered by the practice and generated between the years 2011 and 2013, that were selected randomly out of 4800 reports to test the classifier trained by the training dataset. These reports include CAT scans of the head abdomen, MRIs of the head and neck cervical spine lumbar spine abdomen, and plain X-ray films of the chest abdomen and various extremities (Table 1). They were not limited by a particular specialty, since a primary care practice patient panel is broad based and not limited by specialty.

Table 1. Distribution of the types of reports used in the test dataset.

Type of report	Percentage, n (%)
Mammograms	35/354 (9.9)
CAT scans	36/354 (10.2)
Plain radiology films	71/354 (20.1)
Ultrasounds	70/354 (19.8)
MRIs	142/354 (40.1)

High-Priority Reports

Reports flagged as high-priority are those reports that require further follow-up by the primary care physician. An example is a finding of a renal cyst, which may require a 6-month follow-up ultrasound. Another example is a lung nodule, which may require a 4-month follow-up CAT scan.

Processing Steps

Figure 2 shows the main components of the system. First, the document was retrieved from the EMR, and the Impression section was extracted. The resulting data were then processed to remove any protected health information (PHI). By extracting just the Impression section of the report, much of the PHI was automatically excluded. However, in some cases, there was remaining PHI, such as patient identifying information or the name of the health care facility. Using lexical look-up tables, regular expressions, and simple heuristics described in [9], any remaining PHI was removed.

The next step was text processing and feature extraction, which began with cleanup routines such as conversion of all characters to lowercase type and the removal of stop words. Stop words are words that do not have any value in determining the priority level of a document. Examples of stop words include “the”, “it”, “of”, and “a”. Removing stop words in the preprocessing step is a common practice in artificial intelligence. Doing so minimizes the overall processing load and memory requirements, and results in a narrow set of clinically relevant terms [15]. Then, terms that were negated by negation terms were removed. Negation terms have a large effect on the meaning of sentences. For example, a high-priority report may contain the phrase “acute lung disease”, and a low-priority report may contain the phrase “no acute lung disease”. A naive Bayes classifier cannot differentiate between such distinctions, making the difference between these low-priority and high-priority reports ambiguous.

In addition, clinical reports often contain common phrases such as “otherwise normal chest” that can distinguish a high-priority report from a low-priority report. A naive Bayes classifier only extracts individual words from documents and assumes that the probability of each word being in different document categories is independent of the probabilities of other words. However, if a document contains a phrase such as “otherwise normal chest”, the individual probabilities assigned to each word in the phrase are clinically dependent on each other. Thus, common phrases were identified, and white spaces contained within these phrases in clinical reports were removed to create a single term that the classifier could recognize. A list of common phrases used in this study is provided in Table 2 below. An example of white space removal is shown below in Table 3.

Figure 2. System architecture.

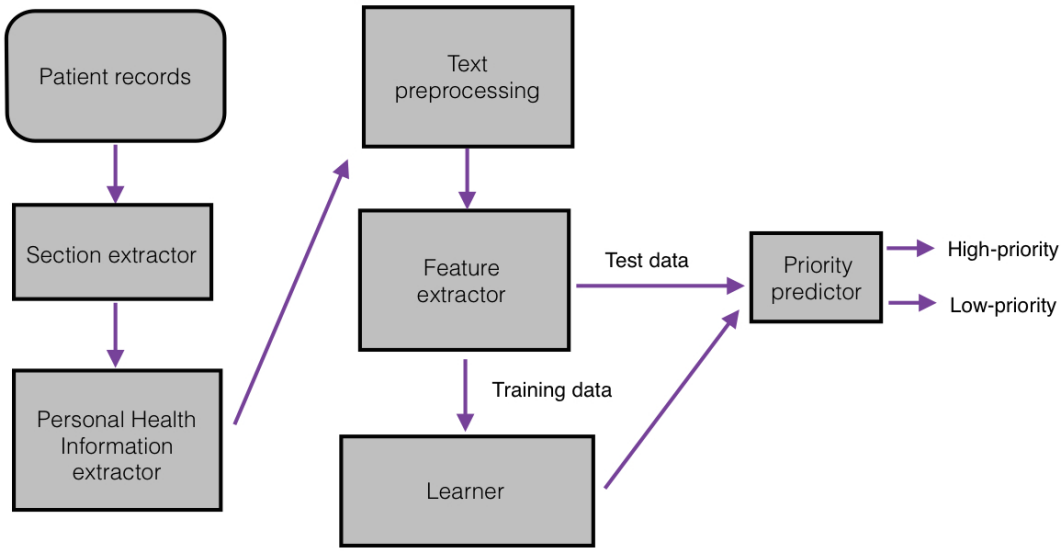


Table 2. Examples of common phrases used in the data cleaning process.

Common phrases
“No significant abnormality is identified”
“No mammographic change or evidence of malignancy”
“No acute cardiopulmonary process”
“No acute pulmonary process”
“Within normal limits”
“Normal abdominal ultrasound”
“No acute intracranial process”
“Appropriate for age”
“Routine annual screening mammogram”
“No acute pathology”
“Correlation recommended”
“Biopsy should be performed”
“Surgical consultation is suggested”
“Appear significantly changed”

Table 3. Common phrase and white space removal depiction.

Common phrase before white space removal	Common phrase after white space removal
“within normal limits”	“withinnormallimits”
“Normal abdominal ultrasound”	“Normalabdominalultrasound”

“Bag of Words”

The remaining words were stored as a “bag of words”, which is a representation of text as an unordered collection of terms that disregards word order or grammar. The naive Bayesian classifier treats each term in this “bag of words” independently from the others. The average number of total unique words in the “bag of words” per report was 684.

P_p and P_{th}

For the implementation of the naive Bayesian classifier, an open source, C# implementation of a spam filter algorithm [16] was repurposed. A spam filter was used as the initial code base because it is essentially a Bayesian classifier that is trained to detect text messages that a user considers to be spam based on training data. After the Bayesian filter was trained on clinical report training dataset, it was tested on the clinical report test dataset. The P_p and P_{th} values from the Bayesian equation were used as parameters in this study to facilitate the calculation of the precision, recall, F-measure, and accuracy values. P_p is the prior probability distribution as defined in the Bayesian equation [17,18]. It represents the probability that a received report is important based on past experience. Similarly, it can be thought of as a percentage that represents the level of suspicion that a document is important. This value, which can be set by the user, affects the misclassification rate of a report, because increasing its value will increase the likelihood that a report will be classified as important. To minimize the false negative rate, the prior probability should be set at a higher value, thus biasing the classifier toward classifying a given report as a positive one.

P_{th} is the threshold probability distribution as defined in the Bayesian equation. It represents the probability cutoff where a document is classified as high-priority. Since a high P_{th} would result in higher false negatives, the manipulation of that parameter in this study was important. The cost of a misclassified important report, or a false negative, is much greater than a misclassification of a routine report. P_{th} also indicates the minimal probability at which a report is classified as important. The user can set this threshold value, typically to levels greater than 50%. For the purposes of this study, the value was set to a level that minimized false negatives, while keeping the false positives at a tolerable level, thereby not missing important reports, but also not contributing to alarm fatigue. P_p and P_{th} were used because their individual effects, when properly adjusted, could be used to compare sensitivity, and consequently, performance, of the classifier.

Precision, Recall, F-Measure, and Accuracy

The performance of this Bayesian classifier implementation was evaluated using precision, recall, F-measure, and accuracy, standard performance measures for classification and machine learning tasks [19]. Precision is the ratio of true positives to the total number of documents classified as positives,

$$\text{Precision} = TP/(TP+FP),$$

where TP is true positive and FP is false positive.

Recall is the proportion of actual positives that are correctly identified as such,

$$\text{Recall} = TP/(TP+FN),$$

where FN is false negative.

F-Measure is the harmonic mean of the precision and recall measures,

F-Measure = (2)(precision)(recall/(precision+recall)).

Accuracy is the percentage of true positives and true negatives to the total number of reports processed.

Accuracy = TP+TN/(TP+TN+FP+FN).

Results

Randomly Selected Reports

In this study, 354 radiology reports, randomly selected from the date range of 2011 to 2013, were tested to evaluate the performance of the Bayesian classifier in detecting high-priority reports. The classifier was trained on data, preclassified by physicians, whose interrater reliability was a Cohen's kappa value of 0.86. This training set consisted of 50 low-priority and 50 high-priority radiology reports randomly selected from the same time range. The performance of the algorithm was tested under 2 independent conditions, the prior probability of a report being high-probability or $P_{p,}$ and the probability threshold P_{th}, at which a report is classified as high-probability. Tests were run for 2 possible values for each of these variables, giving a total of 4 sets of results for analysis.

P_p and P_{th}

The probability of each report being high-priority was determined using Bayes formula as described in the Introduction. The distribution of the probabilities of each report being high-priority is shown below for each of the prior probabilities (P_p) (Figure 3 shows this).

The frequency distribution of the radiology reports for each calculated probability range shows a clustering of reports at both extremes, with the majority of reports having a probability of 0 or 1, and the fewest number of reports being in the mid probability ranges from 0.2000 to 0.6999.

Precision, Recall, F-Measure, and Accuracy Values for P_p=10%

Table 4 lists the classifier metrics for prior probability of 10%, of being a document classified as high-priority, and for each cutoff threshold probability of 50% and 80%.

Figure 3. Distribution of reports in each probability range. The x-axis represents probability and the y-axis represents number of reports from the test set.

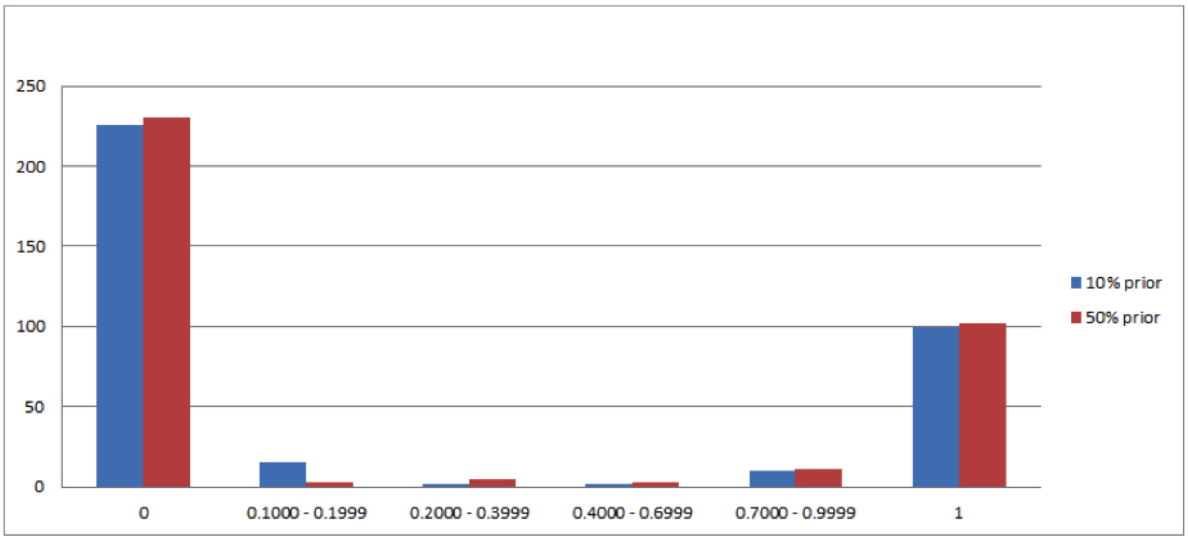


Table 4. Precision, recall, F-measure, and accuracy values for the classifier with P_p= 10%.

	50% P_{th}	80% P_{th}
10% P_p of high-priority report	TP 89, FP 22, TN 240, FN 3	TP 88, FP 21, TN 241, FN4
Precision, %	80.18	80.73
Recall, %	96.74	95.65
F-measure, %	87.66	87.56
Accuracy, %	92.94	92.94

Precision, Recall, F-Measure, and Accuracy Values for P_p=50%

As can be seen, the precision, recall, F-measure, and accuracy values were similar for either probability thresholds given that the P_p was 10%. The accuracy rates for both thresholds are above 90%, but in both situations, there are a few false negatives. Table 5 lists these measures for prior probability of 50%, of being a document being classified as high-priority, and for each cutoff threshold of 50% and 80%.

In the situation of P_p being 50%, the precision, recall, F-measure, and accuracy values are noticeably different. The number of false negatives was the lowest at P_p of 50% for both values of $P_{th.}$ In fact, there was only 1 false negative in each threshold probability scenario resulting in a 0.28% false negative rate. This is also reflected in the recall rate, which decreases as the number of false negatives increases, as seen when comparing Tables 4 and 5.

Table 5. Precision, recall, F-measure, and accuracy values for the classifier with P_p= 50%.

	50% P_{th}	80% P_{th}
50% P_p of report being high-priority	TP 91, FP 25,TN 237, FN 1	TP 91, FP 22, TN 240, FN 1
Precision, %	78.45	80.53
Recall, %	98.91	98.91
F-measure, %	87.50	88.78
Accuracy, %	92.66	93.50

Discussion

Principal Results

The results indicate that a naive Bayesian classifier works remarkably well in classifying radiology reports as low-priority or high-priority. The recall rate varied from 95.65% to 98.91%. This signifies that the classifier succeeded in accurately detecting high-priority reports, also known as true positives, while minimizing false negatives. The rate of false negatives was very low in this study, with the number of false negatives varying from 1 to 4. As indicated earlier, a lower false negative rate is desirable in clinical contexts and the current application. The precision rate, however, was lower, varying from 78.45% to 80.73%. In other words, the classifier had a higher rate of false positives. That is, it classified a greater number of documents as being high-priority, even though they were actually low-priority.

The actual magnitude of change in performance was not too dramatic for the 2 values of P_{th} (50% and 80%). The reason for this is seen by observing the nearly bimodal distribution of reports (Figure 2) falling under the extremes at probabilities of 0 and 1, with few in between.

Similar observations can be made about varying the P_p. This is the prior probability used in the Bayes formula to calculate the probability that a report is high-priority. Increasing this value increases the likelihood that a report is high-priority. Choosing a higher P_p will have the effect of potentially increasing the false positive rate in the same way as it did when the threshold, P_{th}, was increased. The data also demonstrated this, but again, modestly. The bimodal distribution described earlier again shows why this was the case.

This study showed a clear distinction between low-priority and high-priority reports. Why was there such a clear distinction? Low-priority reports are the normal reports. They typically have text in the Impression section such as “no evidence of fracture” or “no acute disease of the chest”. Our negation algorithm removed all negated terms. So these normal reports were presented as empty text to the Bayesian classifier. The text from a high-priority report would typically have language such as “nodule identified”, “possible developing mass”, or “small infiltrate suggesting early pneumonia”. Since these terms are not seen in a low-priority report, the classifier assigns a very high probability to reports containing these terms. Furthermore, by removing negated terms, we greatly improved the scores of the training dataset. Removing stop words had minimal impact on the document scores, but it rendered a cleaner “bag of words” to study and debug.

A closer review of the reports indicates the reason for a higher number of false positives. A false positive report for a mammogram read,

> *...there is no mammographic evidence of malignancy; routine follow-up mammogram in 1 year is recommended; bi-rads category 1. negative according to the nci model the patients lifetime risk of developing breast cancer is 3.6%...* [Patient laboratory report]

Although this report should have been classified as low-priority, there was language used by the radiologist to provide general guidance to the ordering physician, *...according to the nci model the patients lifetime risk of developing breast cancer is 3.6%*. The classifier identified the term, “cancer”, and assigned a high probability to the report. In another case, a report read, “chest is without evidence of pneumonia”. Our classifier did not properly detect the negation term “without”, and thus the term “pneumonia” resulted in a false positive.

Limitations

More robust negation detection should be developed as a part of any future enhancements. Additionally, use of NLP and/or common phrase detection may enhance the ability of the classifier to better distinguish if terms mentioned are part of a

patient's report findings, as in the case described above. Additional statistical methods, in addition to Bayesian statistics, could also provide a stronger classification system.

Although this study had a relatively low number of documents to test and a small number of reviewers, the superior results obtained provide assurance for the performing of future studies and make intuitive sense for the nature of the primary care setting. Patients in this setting are generally healthy and tend to have normal radiology reports. The distinction between a normal and an abnormal report is usually obvious due to the presence of key words, making it easy for a classifier to detect an abnormal report and denote it as high-priority.

Differences From Prior Work

The results of this study highlight the promise of using statistical classifiers, such as this Bayesian implementation, in prioritizing a primary care physician's workload across electronic systems in real-time with an ability to be trained, a marked difference from the retrospective and static analyses done by many of the prior studies in the literature. Due to the EMR-agnostic design of this classifier, it is generalizable to any EMR system or patient data interface for that matter. The real-time incoming data feed to an EMR can consist of various entry points, such as HL7, faxes, scanned documents, and Web services. This classifier might also offset the chance of radiologists not electronically coding radiology reports as normal or abnormal, as these specialists are typically required by the American College of Radiology to electronically code radiology reports as normal or abnormal when communicating with primary care physicians [20]. Even in the case of proper coding, this classifier can act as an additional layer of safety and clinical intelligence with minimal infrastructure and integration costs that are typical of many of the reviewed software systems of the past. Ultimately, use of this tool for prioritizing the physician's workload and aiding in the detection of abnormal radiologic, as well as other findings, can greatly enhance patient safety.

While the scope of documents was limited in this study to radiology, we believe the classifier can be adapted to other verticals within health care. Implementing it on a greater number of radiology reports and testing it on other report types, such as pathology and microbiology reports, will further test the effectiveness of the Bayesian classifier in this study.

Conclusions

In conclusion, a Bayesian classifier can be used, in conjunction with other available methods, to detect high-priority radiology reports and improve primary care provider efficiency in addressing these reports. This novel study showed, for the first time to our knowledge, that the Bayesian system, used on this representative sample of free text, unstructured radiology reports received in a primary care setting, displayed a high rate of success in detecting true positives. Use of this type of technology has the potential to improve patient safety, as well as minimize physician malpractice exposure.

Future work may include studying the effectiveness of this classifier in a different practice setting, such as a specialist's office. For example, in an oncology or cardiology practice, given the nature of each specialty, a greater number of patient reports are expected to be abnormal, and yet may be classified by the specialist as low-priority. It would be interesting to see how this classifier would perform. It is possible that more advanced techniques such as NLP, in combination with a statistical classifier, would be required in order to have a satisfactory rate of high-priority detection. Furthermore, search engine capabilities could be a future extension as specific terms within reports can be identified, leading to a more connected experience for the patient [21]. Such an application might be able to assist in recording and analysis of a long-term view of high-priority events or even disease maps based on the terms that have been flagged, resulting in better visualizations for value-based care or pharmaceutical drug targeting. More immediately, this study makes clear that the intersection of computer science, statistics, and health care can have huge implications that can improve efficiency, patient safety, and quality of care.

Acknowledgments

The authors acknowledge the Blue Hills Medical Associates facility for serving as the testing site for this study and Drs Maggio and Chakrabarti for reviewing patient reports. Internal funding supported this work.

Conflicts of Interest

Authors MS and AM worked for Hermes Clinical, Inc during the writing of this paper. Hermes Clinical, Inc is a health care information technology company that utilizes some of the technologies described in this paper.

References

1. Poon EG, Wang SJ, Gandhi TK, Bates DW, Kuperman GJ. Design and implementation of a comprehensive outpatient results manager. J Biomed Inform 2003;36(1-2):80-91. [Medline: 14552849]
2. Sommers BD, Musco T, Finegold K, Gunja MZ, Burke A, McDowell AM. Health reform and changes in health insurance coverage in 2014. N Engl J Med 2014 Aug 28;371(9):867-874. [doi: 10.1056/NEJMsr1406753] [Medline: 25054609]
3. Wang W, Krishnan E. Big data and clinicians: A review on the state of science. JMIR Med Inform 2014;2(1):e1. [doi: 10.2196/medinform.2913]
4. Davis GT, Singh H. Should patients get direct access to their laboratory test results? An answer with many questions. JAMA 2011 Dec 14;306(22):2502-2503. [doi: 10.1001/jama.2011.1797] [Medline: 22122864]

5. Saber Tehrani Ali S, Lee H, Mathews SC, Shore A, Makary MA, Pronovost PJ, et al. 25-year summary of US malpractice claims for diagnostic errors 1986-2010: An analysis from the National Practitioner Data Bank. BMJ Qual Saf 2013 Aug;22(8):672-680. [doi: 10.1136/bmjqs-2012-001550] [Medline: 23610443]
6. Wahls T, Haugen T, Cram P. The continuing problem of missed test results in an integrated health system with an advanced electronic medical record. Jt Comm J Qual Patient Saf 2007 Aug;33(8):485-492. [Medline: 17724945]
7. Wahls T. Diagnostic errors and abnormal diagnostic tests lost to follow-up: A source of needless waste and delay to treatment. J Ambul Care Manage 2007;30(4):338-343. [doi: 10.1097/01.JAC.0000290402.89284.a9] [Medline: 17873665]
8. Smith M, Murphy D, Laxmisan A, Sittig D, Reis B, Esquivel A, et al. Developing software to "track and catch" missed follow-up of abnormal test results in a complex sociotechnical environment. Appl Clin Inform 2013;4(3):359-375 [FREE Full text] [doi: 10.4338/ACI-2013-04-RA-0019] [Medline: 24155789]
9. Neamatullah I, Douglass MM, Lehman LWH, Reisner A, Villarroel M, Long WJ, et al. Automated de-identification of free-text medical records. BMC Med Inform Decis Mak 2008;8:32 [FREE Full text] [doi: 10.1186/1472-6947-8-32] [Medline: 18652655]
10. Bejan CA, Xia F, Vanderwende L, Wurfel MM, Yetisgen-Yildiz M. Pneumonia identification using statistical feature selection. J Am Med Inform Assoc 2012;19(5):817-823 [FREE Full text] [doi: 10.1136/amiajnl-2011-000752] [Medline: 22539080]
11. Hripcsak G, Austin John H M, Alderson PO, Friedman C. Use of natural language processing to translate clinical information from a database of 889,921 chest radiographic reports. Radiology 2002 Jul;224(1):157-163. [doi: 10.1148/radiol.2241011118] [Medline: 12091676]
12. Elkin PL, Froehling D, Wahner-Roedler D, Trusko B, Welsh G, Ma H, et al. NLP-based identification of pneumonia cases from free-text radiological reports. AMIA Annu Symp Proc 2008:172-176 [FREE Full text] [Medline: 18998791]
13. Pyrros A, Nikolaidis P, Yaghmai V, Zivin S, Tracy JI, Flanders A. A Bayesian approach for the categorization of radiology reports. Acad Radiol 2007 Apr;14(4):426-430. [doi: 10.1016/j.acra.2007.01.028] [Medline: 17368211]
14. Ong MS, Magrabi F, Coiera E. Automated identification of extreme-risk events in clinical incident reports. J Am Med Inform Assoc 2012 Jun;19(e1):e110-e118 [FREE Full text] [doi: 10.1136/amiajnl-2011-000562] [Medline: 22237865]
15. Zhang R, Pakhomov S, McInnes BT, Melton GB. Evaluating measures of redundancy in clinical texts. AMIA Annu Symp Proc 2011;2011:1612-1620 [FREE Full text] [Medline: 22195227]
16. Kester J. Code Project. 2008. A naive Bayesian spam filter for C# URL: http://www.codeproject.com/Articles/23472/A-Naive-Bayesian-Spam-Filter-for-C [accessed 2014-12-22] [WebCite Cache ID 6V1FHh4qe]
17. SAS Institute. Prior probabilities. 2014. SAS Institute URL: http://support.sas.com/documentation/cdl/en/emxndg/64759/HTML/default/viewer.htm [accessed 2015-03-10] [WebCite Cache ID 6WvppvYmT]
18. StatSoft Inc. Tulsa, OK: StatSoft; 2013. Electronic statistics textbook URL: http://www.statsoft.com/textbook/naive-bayes-classifier [accessed 2015-04-02] [WebCite Cache ID 6XUCgudEd]
19. Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. Information Processing & Management 2009 Jul;45(4):427-437. [doi: 10.1016/j.ipm.2009.03.002]
20. Singh H, Thomas EJ, Mani S, Sittig D, Arora H, Espadas D, et al. Timely follow-up of abnormal diagnostic imaging test results in an outpatient setting: Are electronic medical records achieving their potential? Arch Intern Med 2009 Sep 28;169(17):1578-1586 [FREE Full text] [doi: 10.1001/archinternmed.2009.263] [Medline: 19786677]
21. Celi LA, Zimolzak AJ, Stone DJ. Dynamic clinical data mining: Search engine-based decision support. JMIR Med Inform 2014;2(1):e13 [FREE Full text] [doi: 10.2196/medinform.3110] [Medline: 25600664]

Abbreviations

CAT: computerized axial tomography
EMR: electronic medical records
HL7: health level-7
MRI: magnetic resonance images
NLP: natural language processing
PHI: protected health information

Edited by G Eysenbach; submitted 19.08.14; peer-reviewed by MS Ong, A Al-Mutairi; comments to author 05.11.14; revised version received 22.12.14; accepted 21.01.15; published 10.04.15.

Please cite as:
Singh M, Murthy A, Singh S
Prioritization of Free-Text Clinical Documents: A Novel Use of a Bayesian Classifier
JMIR Med Inform 2015;3(2):e17
URL: http://medinform.jmir.org/2015/2/e17/
doi:10.2196/medinform.3793
PMID:25863643

Original Paper

A Web-Based Tool for Patient Triage in Emergency Department Settings: Validation Using the Emergency Severity Index

Pierre Elias[1], BA; Ash Damle[2], BS; Michael Casale[3], PhD; Kim Branson[2], PhD; Nick Peterson[2], PhD; Chaitanya Churi[4], MBBS, MPH; Ravi Komatireddy[5], MD, MHS; Jamison Feramisco[2], MD, PhD

[1]Duke Clinical Research Institute, Duke University School of Medicine, Durham, NC, United States

[2]Lumiata, Inc, San Mateo, CA, United States

[3]Clinical Research Division, West Health Institute, La Jolla, CA, United States

[4]University of Texas Health Science Center, Houston, TX, United States

[5]Clinical Research Division, Scripps Translation Science Institute, La Jolla, CA, United States

Corresponding Author:
Pierre Elias, BA
Duke Clinical Research Institute
Duke University School of Medicine
DCRI 7th Floor
2400 Pratt St
Durham, NC, 27705
United States
Phone: 1 407 782 2266
Fax: 1 407 782 2266
Email: pierre.elias@duke.edu

Related Article:

This is a corrected version. See correction statement: http://medinform.jmir.org/2015/3/e24/

Abstract

Background: We evaluated the concordance between triage scores generated by a novel Internet clinical decision support tool, Clinical GPS (cGPS) (Lumiata Inc, San Mateo, CA), and the Emergency Severity Index (ESI), a well-established and clinically validated patient severity scale in use today. Although the ESI and cGPS use different underlying algorithms to calculate patient severity, both utilize a five-point integer scale with level 1 representing the highest severity.

Objective: The objective of this study was to compare cGPS results with an established gold standard in emergency triage.

Methods: We conducted a blinded trial comparing triage scores from the ESI: A Triage Tool for Emergency Department Care, Version 4, Implementation Handbook to those generated by cGPS from the text of 73 sample case vignettes. A weighted, quadratic kappa statistic was used to assess agreement between cGPS derived severity scores and those published in the ESI handbook for all 73 cases. Weighted kappa concordance was defined a priori as almost perfect (kappa > 0.8), substantial (0.6 < kappa < 0.8), moderate (0.4 < kappa < 0.6), fair (0.2 < kappa< 0.4), or slight (kappa < 0.2).

Results: Of the 73 case vignettes, the cGPS severity score matched the ESI handbook score in 95% of cases (69/73 cases), in addition, the weighted, quadratic kappa statistic showed almost perfect agreement (kappa = 0.93, 95% CI 0.854-0.996). In the subanalysis of 41 case vignettes assigned ESI scores of level 1 or 2, the cGPS and ESI severity scores matched in 95% of cases (39/41 cases).

Conclusions: These results indicate that the cGPS is a reliable indicator of triage severity, based on its comparison to a standardized index, the ESI. Future studies are needed to determine whether the cGPS can accurately assess the triage of patients in real clinical environments.

(JMIR Med Inform 2015;3(2):e23) doi:10.2196/medinform.3508

KEYWORDS

triage; emergency severity index; differential diagnosis; clinical decision support

Introduction

Emergency Department Medical Triage for Patients

Accurate medical triage is critical to patient management in environments such as urgent care centers and emergency departments. Previous research has shown that matching the supply and demand of medical resources within the emergency department (ED) is a complex task with many competing variables [1,2]. Errors in the initial clinical evaluation of patients can potentially lead to severe consequences such as a misdiagnosis, delayed treatment, disproportionate health care resource utilization, and increased costs [3,4]. Over the past three decades, two particular developments have significantly contributed to improved triage.

The Use of Standardized Triage Protocols

The first such development toward improved ED triage was the introduction of standardized protocols such as the Ipswich triage scale, the Australasian triage scale, the Canadian Association of Emergency Physicians triage scale, and the Emergency Severity Index (ESI), which have provided triage templates aimed at consistency and reproducibility across a wide array of patient presentations [5-8]. The ESI, a 5-level triage scoring system, provides a standardized and experimentally validated method of assigning risk severity to patients based upon assessment of their complaints, relevant history, and vital signs when appropriate. Additionally, the anticipated resource utilization is considered [9]. The ESI 5-level triage methodology has been validated when considering resource utilization in diagnostic testing, consultation, and admission to an inpatient setting, as well as 6-month post clinical-evaluation morbidity and mortality [10]. For this evaluation, the ESI was treated as the reference standard for assigning appropriate triage.

Electronic Tools and Patient Data

The second development was the creation and adoption of electronic tools in the form of electronic health records, utilization review software, and clinical decision support (CDS) systems. These tools have changed the way medical professionals manage patient data, communicate with patients and other health care providers, and consider diagnostic and therapeutic options. Preliminary research on CDS technology suggests that this type of medical guidance may play a significant role in patient management and triage, particularly in clinical areas such as EDs, which are characterized by high patient volume, time pressure, and varied pathologies and severities [11,12].

There is significant potential to improve the triaging process through automation. Triaging represents a costly bottleneck in hospital throughput; a 2009 study of Pennsylvania ED directors found that 83% agreed that ED overcrowding was a problem in their hospitals [13]. Such challenges have worsened in recent years; from 1995 to 2005, annual ED visits in the United States increased by 20% (from 96.5 to 115.3 million), and the ED utilization rate increased by 7% (from 36.9 to 39.6 ED visits per 100 persons) [14]. Despite increasing ED visits, the number of hospital EDs decreased by 381, and total hospital beds decreased by 134,000 during the same decade [14].

In part by automating time-intensive tasks, computerized CDS tools for triage aim to improve expediency, patient outcomes, and hospital throughput. Numerous systems have attempted to develop computerized CDS for triage with some success [11,15]. The Taiwan Triage and Acuity Scale was able to significantly decrease overtriaging and medical resource consumption in one study of its implementation [16]. Yet, despite significant advancements, current computerized CDS triage tools suffer significant limitations. First, they are only able to incorporate structured data. This represents a significant workflow restriction; health care provider notes, which provide some of the most valuable information for triaging, are not utilized. Second, agreement of these tools compared to chart review is often poor to moderate. A systematic review found kappa ranges from 0.2 to 0.87, with most below 0.5 [15].

In this study, we evaluated concordance between an established triage severity scoring system, the ESI, and a 5-level triage score created by a novel CDS tool, Clinical GPS v2.0 from Lumiata Inc (cGPS). Although cGPS is designed to be used with electronic health records utilizing an application programming interface (API), it is not yet available to the public, since it is still in the development stage and it has not yet been evaluated by the US Food and Drug Administration. Physician providers are currently testing it in multiple health care settings; the publication of the results of those validation studies is planned.

Using a proprietary database of physician-curated medical information, the cGPS produces a triage severity score based upon a patient's demographics, clinical objective signs, subjective symptoms, vital signs, objective laboratory data, past medical history, and medications (note that laboratory data, past medical history, and medications are not required inputs in the triage setting). The cGPS tool aggregates these data, performs its analysis, and then constructs a list of probable diagnoses from a graph-structure database, each of which has an associated triage score range. All clinical inputs and differential diagnosis lists are utilized to arrive at a single whole-number triage score. In this study, we sought to pursue an independent evaluation against a reference standard in triage (ESI) to validate the potential for future clinical use of cGPS in actual health care settings. This blinded study compared cGPS-generated triage scores for 73 sample case vignettes from the Emergency Severity Index, Version 4: Implementation Handbook (Agency for Healthcare Research and Quality, 2005) [17] against the "gold-standard" ESI-created scores found in the Handbook.

Methods

Methodology of the Clinical GPS and Emergency Severity Index Algorithms

The cGPS database was created through physician curation, which attempts to connect the data inputs physicians receive

directly from patients with all the knowledge, data, and experience health professionals have acquired over the years. Multiple physicians began with a list of signs and symptoms associated with a given diagnosis. Each physician would then remove and add symptoms based on their experience, available data, and published guidelines. This process was followed by adding associated ranges of severity and frequency (eg, severe cough is a common symptom of the diagnosis asthma).

The cGPS system then uses multi-dimensional probability distribution to build graph representations of how illnesses and patients are connected. The cGPS algorithm is based on a probabilistic graphical model, or graph analysis. Graph analysis is a technique for making sense of large datasets, primarily by determining how similar data points are among a range of parameters. The graph's nodes include diagnoses, objective signs, symptoms, laboratory test results, vital signs, and other common inputs to medical decision making. The edges between nodes are probabilistic and based on demographic factors, including age, gender, race, and duration of symptoms. For example, the baseline probability of "abdominal pain" being a presentation of the diagnosis "appendicitis" increases or decreases depending on factors such as the patient's age, gender, and the duration of the symptom.

Each sign and symptom in the cGPS system was assigned a range of potential severity scores from 1-5 by a physician using the working definition highlighted in Table 1. Physicians also manually curated the relationships between diagnoses, signs, and symptoms, and assigned a frequency category, as seen in Table 2. The physician-curated frequency and severity categories are then combined to generate probability distribution scores per diagnosis. For example, "tearing chest pain" may be considered critical and assigned a score of 2, but because it is a rare symptom for appendicitis, it would not heavily contribute to the overall cGPS score of the diagnosis. Another example is the diagnosis of "pneumonia" presenting with the symptom "confusion". This combination is significantly more common in younger patients and older patients, and less prevalent for ages in between. While the severity of confusion as a symptom of pneumonia would be input as an equally severe score of 3 across all ages, it would be input as more common for the young and elderly.

The ESI algorithm categorizes ED patients by first assessing acuity level, followed by expected resource needs. Acuity is determined by the stability of vital functions and the potential threat to life, limb, or organ. Expected resource needs are defined as the number of resources a patient is expected to consume in order to arrive at a disposition decision (discharge, admission, or transfer). Triage personnel work through an algorithm of four decision points to arrive at the ESI severity score and triage level [17].

Both the ESI and cGPS utilize a 5-point scale, with 1 representing the highest severity level and 5 representing the lowest. The definitions of each triage score are similar between the ESI and cGPS, and thus were assumed to be roughly equivalent. The 5-point ESI and cGPS scales are detailed in Table 1. The descriptions for the cGPS severity scores were chosen during the initial development of the program. A key difference between the scales is that cGPS allows fractional scores (eg, 4.3) in the preliminary stage. All such fractions were converted to integer values before comparing the scores (see the Study Methodology subsection below; Figure 1 shows the algorithm below and Figure 2 shows the cGPS interface.).

Table 1. Working definitions used to describe each level of severity.

Severity score	ESI (text descriptors are extrapolated from Figure 2-1A) [17]	cGPS
1	Immediate lifesaving intervention required	Revive/unstable
2	High-risk situation or confused/lethargic/disoriented or severe pain/distress	Critical
3	Urgent, complex (2 or more resources)	Urgent
4	Nonurgent, less complex (1 resource)	Nonurgent
5	Nonurgent (no resources)	Referred

Figure 1. Overview of the algorithm used to derive the triage score.

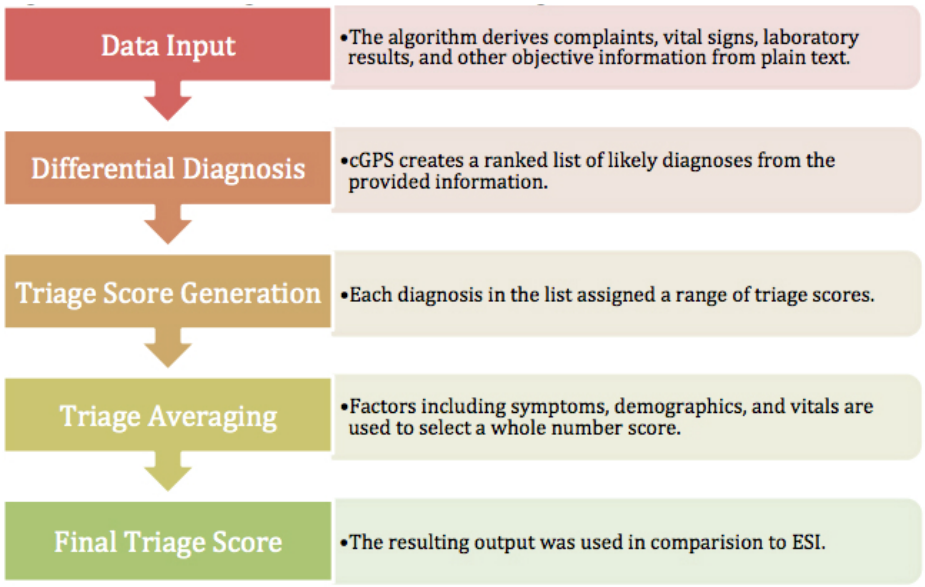
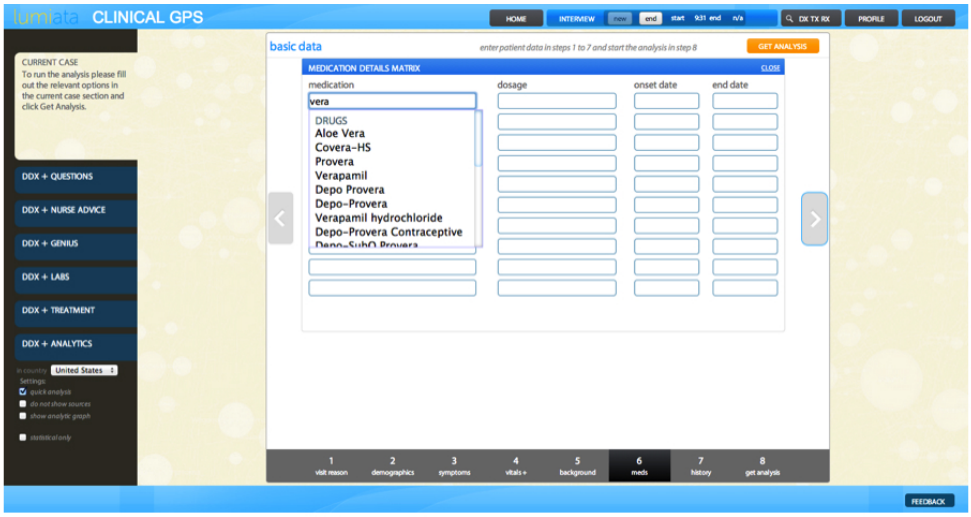


Figure 2. The clinical GPS v2.0 (cGPS) Web-based tool takes clinicians through an 8-step process that supports natural language entry (A) and uses autosuggestions and "quick picks" to maximize efficiency (B).



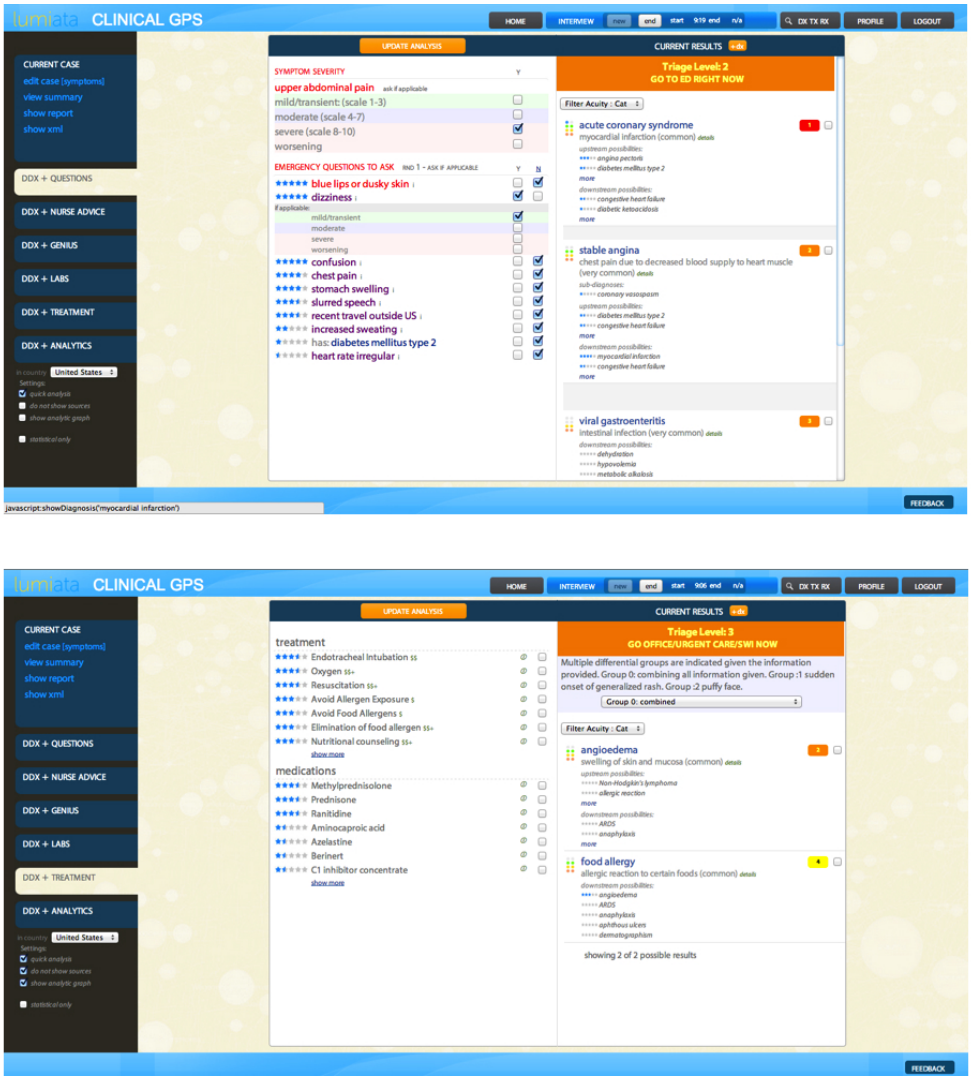
Clinical GPS and Emergency Severity Index Algorithm Differences

As detailed above, the methods used by the ESI and cGPS to arrive at these scores are fundamentally different. The ESI score utilizes acuity information in addition to projected resource utilization to arrive at a triage score. In contrast, for each set of signs and symptoms presented in the input, the cGPS produces a list of differential diagnoses (Figure 3 shows this) using the algorithm detailed below (shown in Figure 1). Because the approaches are fundamentally untethered, no ESI data were used for training the cGPS algorithm prior to the study.

Table 2. cGPS's physician-curated signs and symptoms frequency categories.

Frequency category	Description
Key	Required for diagnosis
Very common	Occurs in >50% of presentations for diagnosis
Common	Occurs in 10-50% of presentations for diagnosis
Uncommon	Occurs in 1-10% of presentations for diagnosis
Rare	Occurs in <1% of presentations for diagnosis

Figure 3. The clinical GPS v2.0 (cGPS) generates differential diagnoses with severity scores and upstream and downstream possibilities, and follow-up questions and tests, including associated costs (C & D).



Study Methodology

For this study, 73 sample patient case vignettes from the ESI handbook were entered verbatim into the cGPS. No case vignettes were excluded from entry. Next, the corresponding ESI score for each vignette was taken directly from the ESI handbook, and the blinded scores were compared. As no human subject or actual patient medical information was used in this analysis, institutional review board approval and patient consent were not applicable.

For each case vignette, the ESI handbook provides a severity score that was obtained using the ESI triage methodology. Vignette topics cover a wide array of potential pathologies and severities. The ESI handbook does not list specific differential diagnoses for each vignette, nor do the vignettes present objective laboratory data; however, all vignettes include the patient's age, gender, and at least one sign or symptom. Whenever present, patient's complaints, past medical history, vital signs, and physical exam values were entered into the cGPS program exactly as written in the handbook. The narrative nature of the handbook vignettes was compatible with the cGPS user interface that allows users to input core information in a format similar to the history and physical note that physicians commonly use to document care (Figures 2 and 4 show the interface). Thus, for each ESI vignette, the data available in the handbook were entered directly into the cGPS tool, and the cGPS algorithm then generated a triage score.

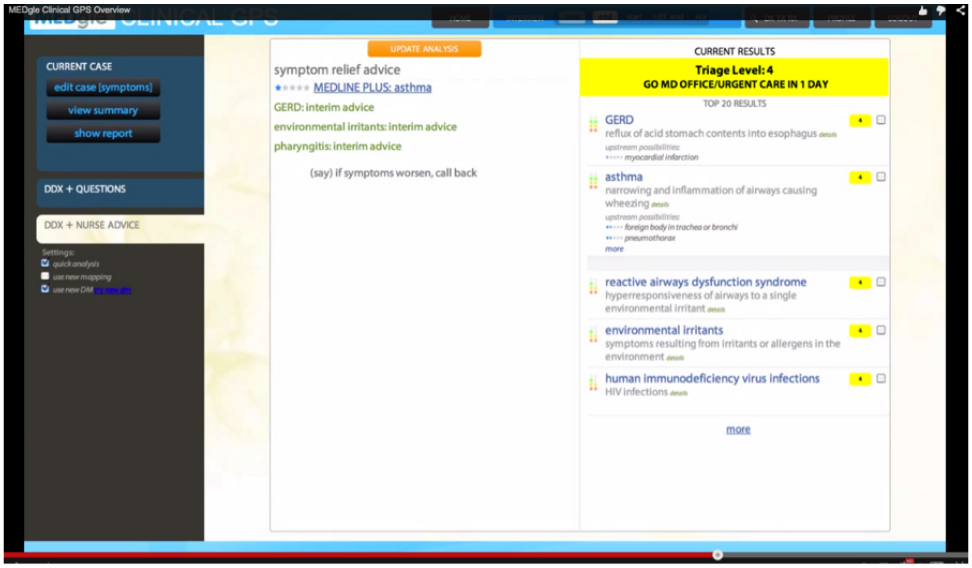
This preliminary severity score generated by the cGPS algorithm was a value ranging from 1 through 5 that included a fractional component and was converted to a whole number value. Average scores with a fraction that fell within the middle range of 0.40-0.60, for example 3.5, required an extra step for rounding. For these middle range scores in the cGPS, the list of recommended diagnostic tests and the duration of symptoms were examined.

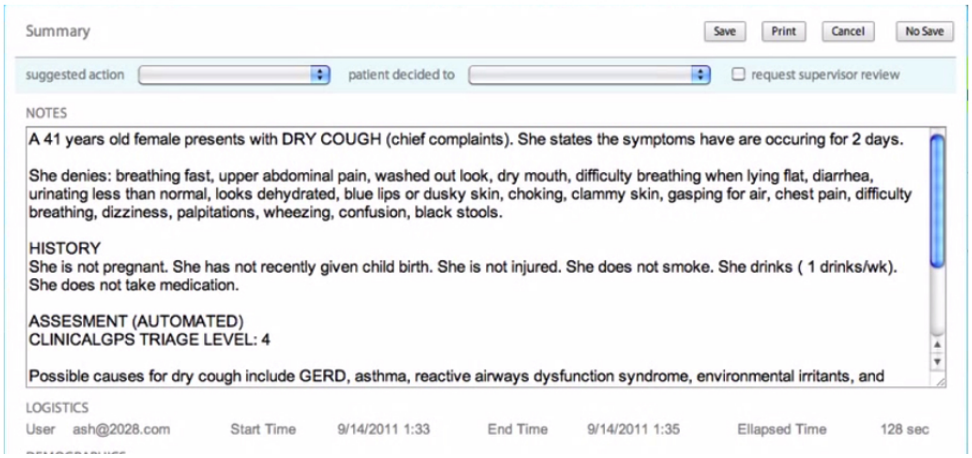
The presence of one or more testing modalities considered to be critical with regard to a specific urgent/emergent diagnosis (eg, electrocardiogram for chest pain), as well as a short duration of symptoms, caused the average severity score to be rounded to the next highest triage level. Conversely, the lack of any suggestions for critical testing modalities or a long duration of symptoms resulted in the score being rounded down to the next lower triage level. Figure 1 shows an overview of the cGPS algorithm used to derive a triage score. The cGPS utilized the individual whole number severity scores for the top 100 diagnoses in the differential diagnosis list to create an average severity score. It was estimated that using 100 diagnoses would account for most physicians' lists of differential diagnoses with a $P<.001$. The individual integer severity scores for each diagnosis were weighted by the baseline prevalence of each diagnosis combined with the likelihood it would present with the given signs, symptoms, and vitals.

Determinants of the severity score for each diagnosis as well as a list of suggested testing modalities were determined a priori using a medical database constructed using expert level knowledge and clinical experience. Data were collected in a spreadsheet and subsequently analyzed. The fraction of exact matches between both severity scores was calculated for all 73 cases. Additionally, a weighted, quadratic kappa statistic was calculated to assess agreement between the cGPS and ESI handbook for all 73 cases, as well as a subset of 41 cases determined to be severity level 1 and 2 in the handbook.

A quadratic kappa statistic was used because the end value of agreement is adjusted based upon the degree of disparity between the two scores. Triage scores that were several categories apart, for example, were considered to exhibit nonlinear disagreement. For example, a situation with an assessed severity of 3, but actual severity of 1, is a potentially life-ending mistake. As such, we chose to use quadratic kappa to account for the significant impact of errors several categories apart. Levels of agreement for the weighted, quadratic kappa analysis were defined a priori based upon previous research to facilitate comparisons; values < 0 indicated no agreement, 0-0.20 slight, 0.21-0.40 fair, 0.41-0.60 moderate, 0.61-0.80 substantial, and 0.81-1.0 almost perfect agreement [18-24]. Although the highest level of agreement possible was the overall goal, we considered a level of agreement above "moderate" (kappa > 0.6) as evidence of sufficient potential to pursue further improvement and testing of this triage algorithm.

Figure 4. The clinical GPS v2.0 (cGPS) interfaces directly with the electronic health record (E & F).





Results

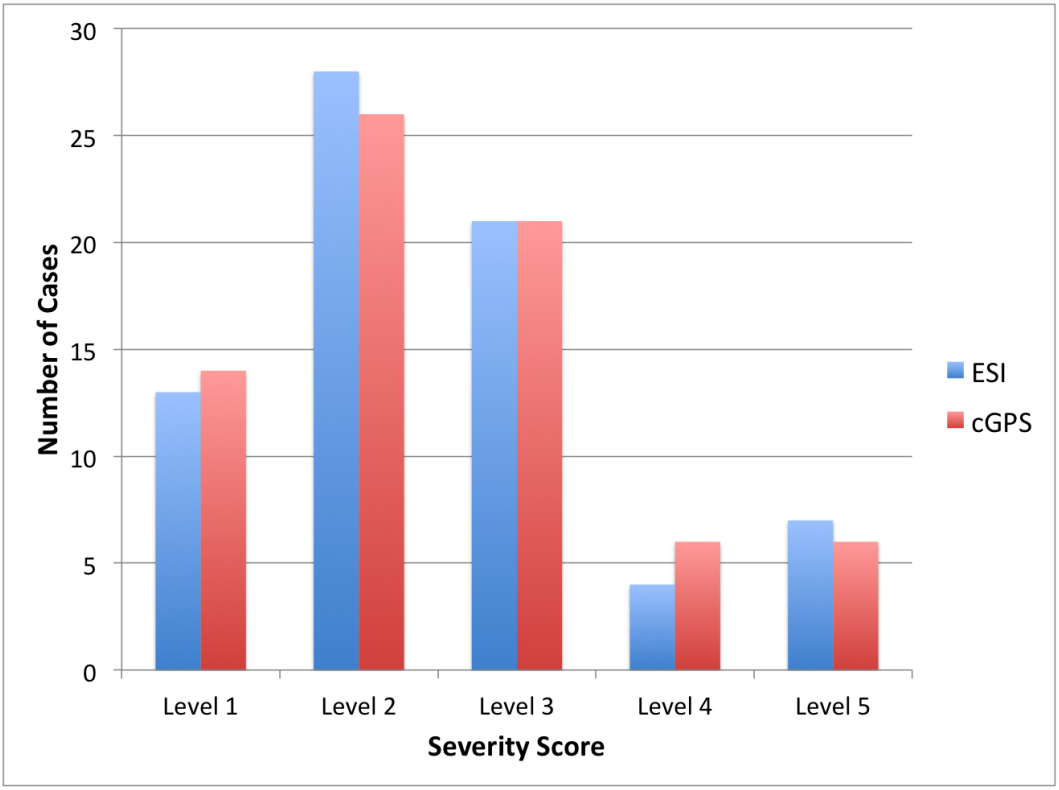
Clinical GPS and Emergency Severity Index Algorithm Scores

Of the 73 ESI handbook clinical case vignettes, the cGPS severity score perfectly matched the ESI handbook severity score in 95% of the cases (69/73 cases). The weighted, quadratic kappa statistic was kappa = 0.93 (95% CI 0.854-0.996), as Figure 5 illustrates, and as summarized in Table 3.

Table 3. Results of the analysis using all case vignettes (n=73).

Results	
Matched: cGPS = ESI	69
Unmatched: cGPS ≠ ESI	4
Total number of cases	73
Percentage of cases with identical severity score	95%
Weighted, quadratic kappa	0.933 (95% CI 0.854-0.996)

Figure 5. Distribution of Emergency Severity Index (ESI) and clinical GPS v2.0 (cGPS) severity scores for the case vignettes (n=73).



A Subanalysis of Case Vignettes

A subanalysis of 41 case vignettes that were assigned the two highest severity scores by the ESI, levels 1 and 2, showed that the cGPS and ESI scores matched exactly in 95% of cases (39/41 cases). A weighted, quadratic kappa statistic for this subgroup was 0.85 (95% CI 0.750-1.037), as shown in Table 4. However, although the quadratic kappa statistic for all 73 cases and the 41 cases in the high level subgroup were greater than 0.8, the lower limit of the CI for the subgroup falls below 0.8, rendering it not significantly better than "substantial" agreement. There were four clinical case vignettes that differed in severity scores between the two systems. In two cases, the cGPS assigned a higher severity score than that of the ESI handbook; it assigned a lower severity score than the ESI handbook in the remaining two cases (Table 5).

Table 4. Subgroup analysis of case vignettes determined to be severity level 1 or 2 by the ESI system (n=41).

Results	
Matched: cGPS = ESI	39
Unmatched: cGPS ≠ ESI	2
Total number of cases	41
Percentage of cases where scores were identical	0.947
Weighted, quadratic kappa	0.851 (95% CI 0.750-1.037)

Table 5. A matrix representation of severity score distribution between the ESI and calculated triage score.

		ESI				
		1	2	3	4	5
cGPS	1	13	1	0	0	0
	2	0	26	0	0	0
	3	0	0	20	0	1
	4	0	1	1	4	0
	5	0	0	0	0	6

Discussion

Principal Findings

Lumiata's cGPS v2.0 application aims to provide patient-specific diagnostic and treatment information based upon both subjective and objective patient data. A novel feature of this software package is its ability to infer a 5-level, ESI-equivalent triage score for patients with a broad range of clinical pathologies and severities. The cGPS exhibited a high level of agreement with the ESI when applied to the 73 sample case vignettes, despite relying upon a very different core algorithm to derive triage scores. Specifically, the cGPS generated an average triage score by analyzing an explicitly created differential diagnosis based on data input into the Internet system. Each diagnosis was assigned an individual triage score. As described in the Methods section, these scores were averaged to produce a final triage score. The cGPS produced reasonable, reliable triage scores when applied to sample cases.

In this investigation, the excellent agreement between the severity indices suggests that, in a clinical environment, the cGPS would assign a triage score similar to that assigned by the ESI methodology the majority of the time using a similar 5-point severity scale. Furthermore, there was also good agreement in triage scores for the subgroup of level 1 and 2 cases, which suggests the cGPS shows promise in identifying cases with life-threatening pathologies that require urgent or emergent medical attention. To that end, the cGPS shows significant clinical potential for evaluating triage. Further investigation is required to determine the accuracy and efficacy in a clinical setting where it could augment clinical judgment or existing triage tools such as the ESI. Importantly, the clinical consequences of assigning an incorrect triage score, even in a minority of cases, could result in significant diagnostic and treatment delays. This is especially concerning for clinical cases incorrectly assigned a lower triage severity. Similarly, an incorrect assignment of a case with low severity to a high severity category could result in inappropriate resource utilization and overall delays in the diagnosis and treatment for other patients in a queue. Although this analysis was not designed to test the accuracy of the differential diagnoses generated by cGPS, the high level of agreement between the cGPS and ESI suggests the individual triage scores assigned for each item in the differential diagnosis were accurate. It is unclear whether this suggests relevancy of the differential diagnosis itself.

We examined the four case vignettes in which the cGPS assigned a different triage score than the ESI. For the two cases where the cGPS assigned a greater severity level than the ESI handbook, one case was due to a difference in sensitivity with regard to the interpretation of vital signs, where cGPS assumed greater severity due to a minor aberrance in heart rate. The other case was due to a lower threshold of safety assumed by the cGPS in a pediatric patient. Both of these cases were considered borderline between two discrete levels of severity, which resulted in the cGPS erring on a higher severity score for presumed safety. Research on the validity and reliability of the ESI in pediatric cases has found that pediatric patients are more often incorrectly triaged than adult patients and, overall, patients under age 18 show the largest variation in triage decision [25]. These results highlight the importance of high-quality, reliable CDS to back up ED personnel, who may feel less comfortable making triage acuity decisions about children, especially infants. In the two cases where the GPS assigned a lower severity than the ESI handbook, the underlying reason appeared to be related to a lack of information in the cGPS database that correlates specific signs or symptoms to more severe diagnoses. For example, one of the cases involved the eye of a construction worker being exposed to concrete. The cGPS database did not categorize concrete exposure as a chemical (alkali) splash to eye, a time-sensitive threat to life or organ, which would constitute a very-high-priority level-2 patient in both the cGPS and ESI systems. Because the cGPS database did not include concrete exposure as an alkali splash, it was unable to correctly identify it as a case requiring a higher severity score than it was assigned. Continual improvements to the underlying database of medical pathologies, as well as the incorporation of stricter rules regarding the interpretation of vital signs, would help reduce the frequency of similar inappropriate triage in the future.

ED staff deal with issues that range from overcrowding to emerging infectious diseases and natural disasters. It is important that they have access to a reliable, accurate, easy-to-use triage system. The cGPS tool meets those needs, while fitting well within the existing workflow of triage management for multiple reasons. Nurses, both in the ED, as well as over the phone, currently do most triage. The tool allows nurses to quickly and effectively document a patient's current state within minutes, as they normally would, while receiving much more robust CDS than is currently available within most available systems. In testing, use of the cGPS tool averages 6 minutes, which is similar to other CDS systems [26] and is less than the amount of time that entering documentation into an electronic health record (EHR) typically takes [27]. Furthermore, the fact that cGPS follows standard documentation pathways, specifically the history and physical notes physicians write, makes the use and integration of such a system significantly easier. Perhaps the most important aspect for usability is the recognition that most institutions are not interested in stand-alone dashboards. If a CDS is not embedded within the EHR, it is unlikely to find interest or support. The cGPS tool is meant to be used within EHRs, utilizing an API, with future plans to communicate using Fast Healthcare Interoperability Resources to allow standardized communication with all mainstream EHRs.

Limitations

There are several important limitations to this analysis. The case vignettes presented in the ESI handbook were assumed to represent reasonable examples of real-life encounters with patients in a triage environment, although this assumption was not specifically validated. A potential confounder includes data entry into the cGPS. To maintain consistency, only one researcher was used for data entry; however, this also creates an opportunity to introduce systematic error into the data entry process. Also, there are several different methods we could have used to round the average severity scores from values with fractions to a whole number value from 1 through 5. We used

what we considered to be a reasonable algorithm that would theoretically err on the side of safety and resolve ambiguity by rounding to the next higher severity. However, other methods and any associated differences in results were not explored. In addition, the generalizability of these results to an actual patient population is limited given the inherent medical complexities present in a real clinical environment, as demonstrated in previous research with electronic triage tools and simulated cases [8,9]. Finally, this study assumes that the 5-point scales used by the cGPS and the ESI are roughly equivalent with respect to identifying levels of illness and patients' needs, although the clinical features that form the boundaries of each triage level may be subject to end-user interpretation. For the subgroup analysis of level 1 and 2 cases, we feel that increasing the sample size would aid in better characterizing the significance of the kappa statistic in this patient population.

Conclusions

Despite a high level of agreement among the 73 cases and the subgroup with the most severe cases, the clinical consequences of incorrectly triaging even a minority of cases do not justify the use of cGPS in a clinical environment without experienced clinician guidance and comparative measures. We would expect that prior to any use in an actual clinical setting, all triage subgroups would meet at least an almost perfect level of triage score agreement with an established standard such as the ESI. In addition, the use of cGPS in any clinical setting would also require establishing its safety and reliability with a controlled clinical pilot evaluating patient cases in parallel with an established triage tool and in a prospective manner with different users and patient populations as well as clinician oversight.

This initial investigation suggests that an automated tool providing computerized CDS has potential to serve as an independent triage tool for use by providers in urgent care and emergency room settings. However, additional prospective pilot clinical studies and database improvements will be required to determine the triage accuracy with regard to real patient cases, various score algorithms, data entry procedures, performance and usability within clinical environments, and interobserver agreement between different users in clinical environments.

Acknowledgments

We'd like to acknowledge Hanumanth Reddy, Francisco Grajales, and Kara McArthur for their important feedback regarding this manuscript.

Conflicts of Interest

JF is Chief Scientific Officer of Lumiata Inc, KB is Chief Data Scientist of Lumiata Inc, AD is Chief Executive Officer of Lumiata Inc, PE is a consultant for Lumiata Inc, NP is an employee of Lumiata Inc, and CC is an employee of Lumiata Inc.

References

1. Wuerz R, Fernandes CM, Alarcon J. Inconsistency of emergency department triage. Emergency Department Operations Research Working Group. Ann Emerg Med 1998 Oct;32(4):431-435. [Medline: 9774926]
2. Göransson KE, Ehrenberg A, Marklund B, Ehnfors M. Emergency department triage: Is there a link between nurses' personal characteristics and accuracy in triage decisions? Accid Emerg Nurs 2006 Apr;14(2):83-88. [doi: 10.1016/j.aaen.2005.12.001] [Medline: 16540319]
3. Tanabe P, Travers D, Gilboy N, Rosenau A, Sierzega G, Rupp V, et al. Refining Emergency Severity Index triage criteria. Acad Emerg Med 2005 Jun;12(6):497-501. [doi: 10.1197/j.aem.2004.12.015] [Medline: 15930399]
4. Baker DW, Stevens CD, Brook RH. Regular source of ambulatory care and medical care utilization by patients presenting to a public hospital emergency department. JAMA 1994;271(24):1909-1912. [Medline: 8201734]
5. FitzGerald G, Jelinek GA, Scott D, Gerdtz MF. Emergency department triage revisited. Emerg Med J 2010 Feb;27(2):86-92. [doi: 10.1136/emj.2009.077081] [Medline: 20156855]
6. Beveridge R. CAEP issues. The Canadian Triage and Acuity Scale: A new and critical element in health care reform. Canadian Association of Emergency Physicians. J Emerg Med 1998;16(3):507-511. [Medline: 9610988]
7. Grouse AI, Bishop RO, Bannon AM. The Manchester Triage System provides good reliability in an Australian emergency department. Emerg Med J 2009 Jul;26(7):484-486. [doi: 10.1136/emj.2008.065508] [Medline: 19546267]
8. Yousif K, Bebbington J, Foley B. Impact on patients triage distribution utilizing the Australasian Triage Scale compared with its predecessor the National Triage Scale. Emerg Med Australas 2005;17(5-6):429-433. [doi: 10.1111/j.1742-6723.2005.00773.x] [Medline: 16302934]
9. Wuerz RC, Milne LW, Eitel DR, Travers D, Gilboy N. Reliability and validity of a new five-level triage instrument. Acad Emerg Med 2000 Mar;7(3):236-242. [Medline: 10730830]
10. Wuerz R. Emergency severity index triage category is associated with six-month survival. ESI Triage Study Group. Acad Emerg Med 2001 Jan;8(1):61-64. [Medline: 11136151]
11. Lyman JA, Cohn WF, Bloomrosen M, Detmer DE. Clinical decision support: Progress and opportunities. J Am Med Inform Assoc 2010;17(5):487-492 [FREE Full text] [doi: 10.1136/jamia.2010.005561] [Medline: 20819850]
12. Dong SL, Bullard MJ, Meurer DP, Blitz S, Ohinmaa A, Holroyd BR, et al. Reliability of computerized emergency triage. Acad Emerg Med 2006 Mar;13(3):269-275. [doi: 10.1197/j.aem.2005.10.014] [Medline: 16495428]

13. Fee C, Weber EJ, Maak CA, Bacchetti P. Effect of emergency department crowding on time to antibiotics in patients admitted with community-acquired pneumonia. Ann Emerg Med 2007 Nov;50(5):501-509. [doi: 10.1016/j.annemergmed.2007.08.003] [Medline: 17913300]
14. Nawar EW, Niska RW, Xu J. National Hospital Ambulatory Medical Care Survey: 2005 emergency department summary. Adv Data 2007 Jun 29(386):1-32. [Medline: 17703794]
15. Farrohknia N, Castrén M, Ehrenberg A, Lind L, Oredsson S, Jonsson H, et al. Emergency department triage scales and their components: A systematic review of the scientific evidence. Scand J Trauma Resusc Emerg Med 2011;19:42 [FREE Full text] [doi: 10.1186/1757-7241-19-42] [Medline: 21718476]
16. Ng CJ, Yen ZS, Tsai JCH, Chen LC, Lin SJ, Sang YY, TTAS national working group. Validation of the Taiwan triage and acuity scale: A new computerised five-level triage system. Emerg Med J 2011 Dec;28(12):1026-1031. [doi: 10.1136/emj.2010.094185] [Medline: 21076055]
17. Gilboy N, Tanabe P, Travers DA, Rosenau AM, Eitel DR. AHRQ Publication No. 05-0046-2. Rockville, MD: Agency for Healthcare Research and Quality. May. 2012. Emergency Severity Index, version 4: Implementation handbook URL: http://www.ahrq.gov/professionals/systems/hospital/esi/esihandbk.pdf [accessed 2015-06-05] [WebCite Cache ID 6Z42a1VDc]
18. Makam AN, Auerbach AD, Steinman MA. Blood culture use in the emergency department in patients hospitalized with respiratory symptoms due to a nonpneumonia illness. J Hosp Med 2014 Aug;9(8):521-524. [doi: 10.1002/jhm.2205] [Medline: 24753399]
19. Carpenter CR. Kappa statistic. CMAJ 2005 Jul 5;173(1):15-16 [FREE Full text] [doi: 10.1503/cmaj.1041742] [Medline: 15997024]
20. van der Wulp Ineke, van Stel Henk F. Calculating kappas from adjusted data improved the comparability of the reliability of triage systems: A comparative study. J Clin Epidemiol 2010 Nov;63(11):1256-1263. [doi: 10.1016/j.jclinepi.2010.01.012] [Medline: 20430580]
21. Dallaire C, Poitras J, Aubin K, Lavoie A, Moore L, Audet G. Interrater agreement of Canadian Emergency Department Triage and Acuity Scale scores assigned by base hospital and emergency department nurses. CJEM 2010 Jan;12(1):45-49 [FREE Full text] [Medline: 20078918]
22. van der Wulp Ineke, van Stel Henk F. Adjusting weighted kappa for severity of mistriage decreases reported reliability of emergency department triage systems: A comparative study. J Clin Epidemiol 2009 Nov;62(11):1196-1201. [doi: 10.1016/j.jclinepi.2009.01.007] [Medline: 19398298]
23. Gravel J, Gouin S, Manzano S, Arsenault M, Amre D. Interrater agreement between nurses for the Pediatric Canadian Triage and Acuity Scale in a tertiary care center. Acad Emerg Med 2008 Dec;15(12):1262-1267. [doi: 10.1111/j.1553-2712.2008.00268.x] [Medline: 18945238]
24. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977 Mar;33(1):159-174. [Medline: 843571]
25. Travers DA, Waller AE, Katznelson J, Agans R. Reliability and validity of the emergency severity index for pediatric triage. Acad Emerg Med 2009 Sep;16(9):843-849. [doi: 10.1111/j.1553-2712.2009.00494.x] [Medline: 19845551]
26. Wagholikar KB, Hankey RA, Decker LK, Cha SS, Greenes RA, Liu H, et al. Evaluation of the effect of decision support on the efficiency of primary care providers in the outpatient practice. J Prim Care Community Health 2015 Jan;6(1):54-60 [FREE Full text] [doi: 10.1177/2150131914546325] [Medline: 25155103]
27. Poissant L, Pereira J, Tamblyn R, Kawasumi Y. The impact of electronic health records on time efficiency of physicians and nurses: A systematic review. J Am Med Inform Assoc 2005;12(5):505-516 [FREE Full text] [doi: 10.1197/jamia.M1700] [Medline: 15905487]

Abbreviations

API: application programming interface
CDS: clinical decision support
cGPS: clinical GPS v2.0
ED: emergency department
EHR: electronic health record
ESI: Emergency Severity Index

Edited by G Eysenbach; submitted 02.05.14; peer-reviewed by Q Li, D Travers; comments to author 04.11.14; revised version received 29.01.15; accepted 27.04.15; published 10.06.15.

Please cite as:
Elias P, Damle A, Casale M, Branson K, Peterson N, Churi C, Komatireddy R, Feramisco J
A Web-Based Tool for Patient Triage in Emergency Department Settings: Validation Using the Emergency Severity Index
JMIR Med Inform 2015;3(2):e23
URL: http://medinform.jmir.org/2015/2/e23/
doi:10.2196/medinform.3508
PMID:

Original Paper

A Telesurveillance System With Automatic Electrocardiogram Interpretation Based on Support Vector Machine and Rule-Based Processing

Te-Wei Ho[1*], PhD; Chen-Wei Huang[2*], PhD; Ching-Miao Lin[3], MSc; Feipei Lai[1], PhD; Jian-Jiun Ding[2], PhD; Yi-Lwun Ho[4], MD, PhD; Chi-Sheng Hung[4], MD

[1]National Taiwan University, Graduate Institute of Biomedical Electronics and Bioinformatics, Taipei, Taiwan

[2]National Taiwan University, Graduate Institute of Communication Engineering, Taipei, Taiwan

[3]National Taiwan University, Department of Electrical Engineering, Taipei, Taiwan

[4]National Taiwan University Hospital, Telehealth Center, Taipei, Taiwan

[*]these authors contributed equally

Corresponding Author:
Jian-Jiun Ding, PhD
National Taiwan University
Graduate Institute of Communication Engineering
No. 1, Section 4, Roosevelt Road
Taipei, 10617
Taiwan
Phone: 886 233669652
Fax: 886 233663662
Email: jjding@ntu.edu.tw

Abstract

Background: Telehealth care is a global trend affecting clinical practice around the world. To mitigate the workload of health professionals and provide ubiquitous health care, a comprehensive surveillance system with value-added services based on information technologies must be established.

Objective: We conducted this study to describe our proposed telesurveillance system designed for monitoring and classifying electrocardiogram (ECG) signals and to evaluate the performance of ECG classification.

Methods: We established a telesurveillance system with an automatic ECG interpretation mechanism. The system included: (1) automatic ECG signal transmission via telecommunication, (2) ECG signal processing, including noise elimination, peak estimation, and feature extraction, (3) automatic ECG interpretation based on the support vector machine (SVM) classifier and rule-based processing, and (4) display of ECG signals and their analyzed results. We analyzed 213,420 ECG signals that were diagnosed by cardiologists as the gold standard to verify the classification performance.

Results: In the clinical ECG database from the Telehealth Center of the National Taiwan University Hospital (NTUH), the experimental results showed that the ECG classifier yielded a specificity value of 96.66% for normal rhythm detection, a sensitivity value of 98.50% for disease recognition, and an accuracy value of 81.17% for noise detection. For the detection performance of specific diseases, the recognition model mainly generated sensitivity values of 92.70% for atrial fibrillation, 89.10% for pacemaker rhythm, 88.60% for atrial premature contraction, 72.98% for T-wave inversion, 62.21% for atrial flutter, and 62.57% for first-degree atrioventricular block.

Conclusions: Through connected telehealth care devices, the telesurveillance system, and the automatic ECG interpretation system, this mechanism was intentionally designed for continuous decision-making support and is reliable enough to reduce the need for face-to-face diagnosis. With this value-added service, the system could widely assist physicians and other health professionals with decision making in clinical practice. The system will be very helpful for the patient who suffers from cardiac disease, but for whom it is inconvenient to go to the hospital very often.

(JMIR Med Inform 2015;3(2):e21) doi:10.2196/medinform.4397

KEYWORDS

telehealth care; telesurveillance system; electrocardiogram; ECG classification; support vector machine

Introduction

Telehealth care is a global trend affecting clinical practice around the world. It allows for the remote care of patients at a distance using information and communication technology (ICT). Telehealth care is a continuous, automatic, real-time, and home-based remote monitoring system of patients that provides person-centered facilities to support individual health care. A previous study has reported that telehealth care may help patients and families to optimize adherence to therapy and may promote early intervention of abnormal signs by long-term telehealth monitoring [1]. In addition, several surveys in telehealth programs revealed beneficial results in clinical outcomes. A study of telemonitoring programs indicated that the all-cause mortality, the length of hospital stay, and the hospitalization rate were significantly reduced in telehealth users [2]. For these reasons, recent developments in Web-based telehealth care systems were designed to continually monitor the health status of chronic disease patients and elderly people [3-5]. People with heart disease problems, especially, should be warned to take particular care in daily life. However, it is difficult to follow up the situation of patients in real time and to provide early intervention in emergency cases. Fortunately, with the progress and development of telecommunication technologies, particularly in networks and electrical signal devices, telecom facilities have afforded telehealth care as an appropriate approach for disease management [6-9]. A real-time, computer-based support system is suitable for patients and health care providers in clinical practice [10]. Generally, information and communication technology has been recognized as an important tool in helping reduce health care costs while maintaining a high level of quality. A general-purpose telehealth care system must fully integrate remote management programs, including wireless telecommunication, a sensor network, a user interactive platform, and the information technology to deliver the synchronous service. Therefore, to provide ubiquitous monitoring and to offer value-added services, a comprehensive, reliable, and efficient data-reporting and analyzing system and its extendable modules must be established.

The electrocardiogram (ECG) is commonly used to detect abnormal heart rhythms and investigate the cause of heart abnormalities. An ECG, which can be acquired by a noninvasive procedure, is a transthoracic interpretation of the heart electrical activity over a period of time by using electrodes attached to the surface of the skin. In clinical practice, an ECG is a critical tool in diagnosing and identifying heart abnormalities by several features. Some features observed by ECGs are the RR interval (ie, the time measurement between the R waves of two heartbeats), the QRS complexes (ie, the duration of ventricular depolarization from Q wave to S wave), the ST segment (ie, the interval between ventricular depolarization and repolarization, between S wave and T wave), the T wave (ie, repolarization of the ventricles), and the amplitude of R-wave peaks (ie, electrical stimulus passing through the ventricular walls). An ECG also gives important information about human heart status related to critical healthy or unhealthy parameters. Most heart diseases can be detected by analyzing the ECG signal. The ECG is characterized by a cyclic occurrence of patterns with different frequency contents. A good ECG analysis method can accurately detect the morphological characteristics of the QRS complexes as well as the peaks. In the ECG analysis process, one of the most important procedures is to detect R-wave peaks. When the position of the R-wave peak is found, the locations of other feature points of ECG signals, such as Q peaks and S peaks, can be found by the relative position to the R-wave peak. Therefore, the accuracy of R-wave peak detection in ECG signals becomes very important. There have been several R-wave peak detection algorithms proposed in the past decades. Generally, these algorithms can be categorized into time-based detection algorithms [11-14], which are easy to implement but sometimes sensitive to noise, and frequency-based detection algorithms [15-18], which require more computation time but have better detection performance because of good robustness-to-interference, or noise, ratio.

In recent years, there were some research studies about atrial premature contraction (APC) heartbeat detection from ECG signals. Most algorithms of APC detection are time based and use the QRS morphology information for APC heartbeat classification [19-23]. On the other hand, some APC detection algorithms [24-26] are frequency based and adopt the Fourier transform or the wavelet transform. In these R-wave peak detection and APC detection algorithms, the support vector machine (SVM), the rule-based decision tree, the artificial neural network, or fuzzy logic are used as classifiers.

Over the past decades, many studies have put effort into ECG peak identification and heartbeat classification. However, few of them specifically focused on multidisease interpretation from ECG signals. Additionally, despite the numerous classification approaches in the literature, no study has convincingly demonstrated the hybrid model using a large, real-world ECG database.

In general, interpretation of ECG signals is a complicated and time-consuming task for cardiologists, especially when the data size is very large. Hence, to mitigate the increasing workload of cardiologists, and to provide continuous telehealth care and offer value-added service, the aim of this study was to construct a clinical decision support system (CDSS) with a knowledge-based ECG recognition program based on the support vector machine and rule-based processing approaches. The proposed software was designed to aid medical practitioners in decision making and clinical practice. The entire system included the automatic mechanism of data transmission, data storage, signal processing, and classification analysis. With the information from electronic medical records and analysis results, medical staff could use this telesystem to provide ubiquitous health care for patients.

Methods

Electrocardiogram Signal Analysis Using the Proposed Telesurveillance System

The data flow of ECG signal analysis is illustrated in Figure 1. In this study, we divided the flowchart into two parts. The first part represents the data flow on the patient side. The flowchart shows how we derived the ECG signal from patients. Patients can use the ECG recorder, which is similar in size to handheld mobile phones, to derive single-lead ECG signals as independently as possible. The recorder can securely and quickly transmit the measured data to the hospital server over Ethernet connection or the wireless local area network (WLAN). The other part of the flowchart shows the data processing on the hospital side. Data preprocessing is an important process for data analysis. We adopted the finite impulse response (FIR) filter to remove noise and the drift caused from the baseline. After noise reduction, we extracted the key features of the ECG waveforms and used SVM or rule-based processing to construct a classification model, which could suggest diagnoses. Finally, the medical practitioners were able to make decisions with the help of the suggested diagnoses from the system.

The interpretation mechanism is the critical part of an automatic classification system. The process of the automatic ECG recognition algorithm is shown in Figure 2. We divided the process into four sections: noise reduction, peak estimation, feature extraction, and diagnosis interpretation. Noise reduction could enhance the signal part of the ECG from a contaminated record. Peak estimation was used to detect the locations of the P, Q, R, S, and T peaks for further analysis. Feature extraction was used to extract the key information of signals as the interpretation criteria of classifiers. Finally, we used the classifiers for the purpose of heartbeat status monitoring in this study.

The clinical decision support system was implemented using the C# language in the ASP.NET Model-View-Controller (MVC) architecture. The model is an application object and the controller is a function between the user interface and input. The concept of MVC (see Figure 3) is to connect the human's mental model with the digital model, which exists in the computer. At the very least, the concept was adopted as a design pattern which is able to separate different sections. First, the user interface, including representation and the input control, is designed. Second, users can view and manage the data. Finally, the data bank will be updated. Microsoft Structured Query Language (SQL) Server 2008 was used for data computation and analysis. For the purpose of timely transmission and efficient delivery of the needed data to the user, the system was developed using the asynchronous JavaScript and XML (AJAX) technology and service-oriented architecture (SOA). AJAX, a group of client-side technologies, is based on existing standards that allows asynchronous communication by exchanging small amounts of data with the server in the background. The main purpose of AJAX is to enhance the speed, performance, and usability of Web applications. SOA is basically a collection of services that may be under the control of different ownership domains, and is able to interact, share, and exchange information without knowing the inner mechanism of the different systems. In this study, to provide individualized health management, we used the Web service to derive electronic medical records (EMRs) from the National Taiwan University Hospital (NTUH), which included such information as prescriptions, allergy records, laboratory data, and comorbidities.

Before analyzing ECG signals, the process of noise reduction was applied, as in Figure 2. Noise reduction was used to remove the interference and the baseline from signals. Its purpose is to address ECG enhancement and to accurately interpret a contaminated ECG signal. In this study, the denoising approach was based on a finite impulse response filter, which has become one of the most effective and popular denoising methods in many biomedical signal fields in recent years [27,28]. A band-pass FIR filter can reduce the noise and remove the baseline. The ECG signal has always suffered from the baseline drifting problem, which may lead to misdiagnosis if the drifting is severe. Therefore, baseline removal was very important to the ECG signal analysis. After removing the baseline, the locations and amplitudes of the P, Q, R, S, and T peaks can be determined accurately. Instead of using the median filter, which was adopted by many existing algorithms, we applied an innovative method to remove the baseline based on a gradient weighting function and a baseline ratio index [29]. These functions could improve the detection accuracy of the ECG R-wave peak for feature extraction, as discussed in the next section.

Figure 1. Flowchart of ECG signal analysis in the telesurveillance system. Patients use the handheld recorder to obtain the single-lead ECG signal, which will be automatically transmitted to the Telehealth Center at the NTUH for monitoring.

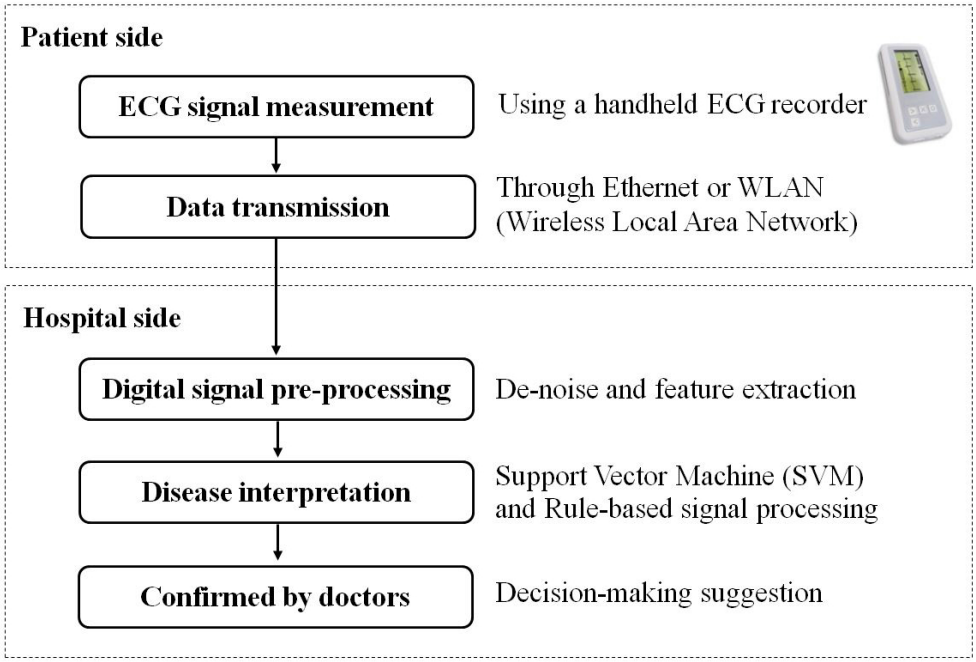


Figure 2. Flowchart of the automatic ECG recognition algorithm. Several preprocessing steps (ie, denoising, baseline removal, and feature extraction) and the classifiers of SVM and rule-based processing are applied to analyze the ECG signal.

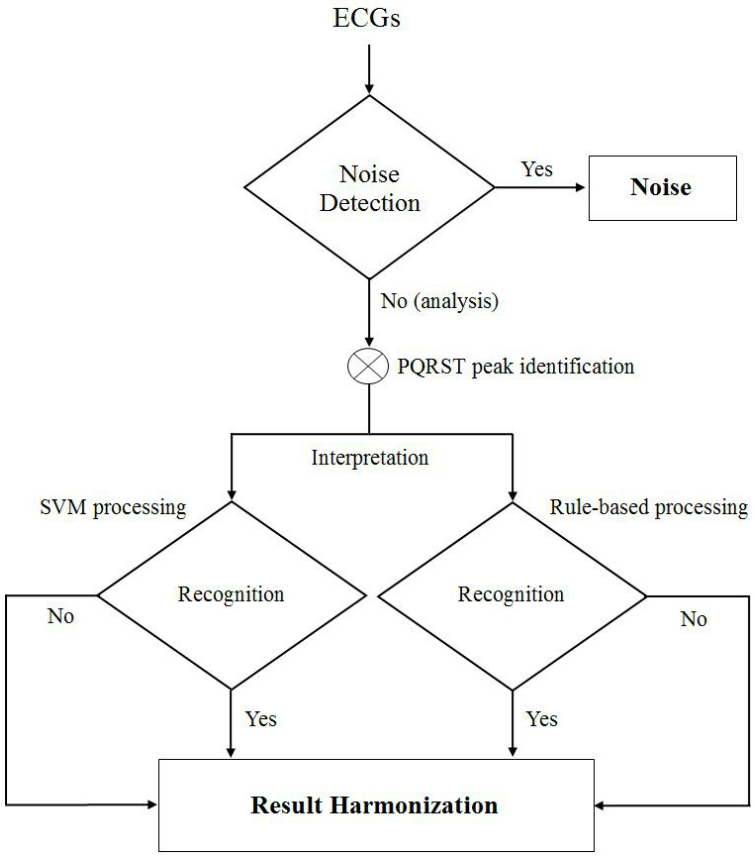
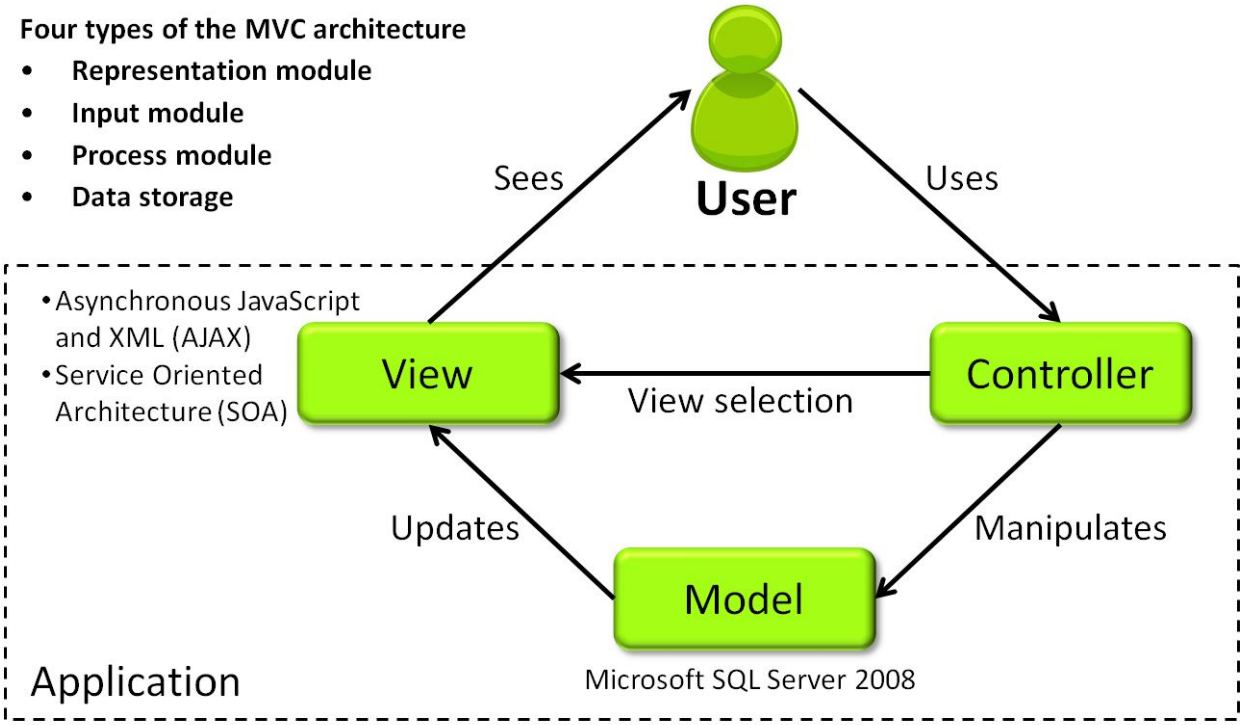


Figure 3. The high-level description of the user-environment system architecture, Model-View-Controller (MVC). Based on the MVC architecture, the modules of the platform can be clean, flexible, reusable, and extendable for programmers.



Peak Estimation and Feature Extraction

The R-wave peak of an ECG complex signal is a dominant and essential characteristic which usually has the greatest height in a QRS complex. In this paper, R-wave peak candidates were identified by using the local maximum or minimum in a sliding window [30]. To enhance the detection accuracy, we took advantage of an adaptive peak-height thresholding method and a search-back method for sifting through R-wave peaks precisely. Moreover, the Q, S, P, and T peaks are also representative characteristic features. Their locations highly influence the accuracy of feature extraction. In the proposed system, several techniques were applied to estimate the P, Q, S, and T peaks accurately and efficiently. For efficiency, the sliding detection window technique and the second-order difference method were applied. For accuracy, the Mexican hat function was applied as a template-matching filter to approximate the PQRST complex. Based on the proposed algorithm, the Q, S, P, and T peaks can be detected in an accurate way even if the ECG signal suffers highly from noise.

The detection performance of R-wave peaks is based on the Massachusetts Institute of Technology-Beth Israel Hospital (MIT-BIH) arrhythmia database, which contains 48 half-hour, two-channel ambulatory ECG records. The characteristics of the MIT-BIH arrhythmia database include 11-bit resolution and a sampling frequency of 360 Hz. There are total 650,000 sampling points per ECG signal record. All 48 records, including 2546 atrial premature contraction heartbeats, 7130 ventricular premature contraction (VPC) heartbeats, and a total number of 109,494 heartbeats, were evaluated in the proposed method.

The features adopted in the classifiers of the entire system are summarized in Table 1. They were grouped into three parts. First, we employed a general extraction method based on the wavelet transform. It can extract both the detailed and the large-scale information. In this study, we applied three types of wavelet transforms: the spline 5/3 wavelet, the Cohen-Daubechies-Feauveau (CDF) 9/7 wavelet, and the Daubechies wavelet. Second, as another part of feature extraction, we calculated peak segments. For example, we first detected R-wave peaks, then the R-wave peak was utilized to calculate the vector between nearby peaks. Hence, we could derive all peak points of the ECG signal. Third, we acquired the features by computing the correlation and the segment lengths among these peak points. Most importantly, we established several features for specific diseases. For example, atrial fibrillation (AF) is the most common abnormal heart rhythm disease. The irregularity of RR intervals and the absence of P waves are used as the features to identify AF. Hence, we used the variant RR interval lengths to detect the irregular RR intervals and used fake P waves to detect the absence of P waves.

Furthermore, in a rule-based processing classifier, to detect the morphological characteristics of the ECGs we also applied the wave pattern and the time-based features among peaks, as seen in previous studies [31,32].

Table 1. Features of classifiers.

Model and extraction methods	Descriptions
Support vector machine	
Wavelet transform 5/3, 9/7, and Daubechies	The maximum, minimum, mean, and variance using each wavelet transform. The number of features extracted by the three wavelet transforms is 12.
Peak-segment features	Local maximums of RR interval widths/R-wave peak amplitudes in different scales
	Local minimums of RR interval widths/R-wave peak amplitudes in different scales
	Local means of RR interval widths/R-wave peak amplitudes in different scales
	Local variances of RR interval widths in different scales
	The number of local maximums between two R-wave peaks
	The rate of the case where the P peak does not exist
	Local mean of PR-segment lengths
	Local mean of QT-segment lengths
	Local mean of ST-segment lengths
	Local mean of P-wave widths
	Local mean of T-wave widths
	Local mean of QRS-complex width
	Local mean of P-wave amplitudes
	Local mean of T-wave amplitudes
Rule based	
Amplitude and time analysis	The amplitude of the R-wave peak
	The amplitude of the S peak
	The amplitude ratio of the R-wave peak and the maximal amplitude of the S peak and the Q peak
	The distance between the Q and S peaks in a QRS complex
	The distance between the R and S peaks in a QRS complex
	The distance between the Q- and R-wave peaks in a QRS complex
	The ratio of the current RR interval to the local average RR interval

Data Collection

The ECG data were collected from the telehealth program of the Telehealth Center at the National Taiwan University Hospital from February 14, 2012 to December 31, 2014. The dataset contained 213,420 ECGs from 530 patients. For classification, we divided the data into the training dataset and the validation dataset. We selected the data from 2012 as the training dataset, and the remaining data as the validation dataset. Out of 213,420 ECGs, the training dataset and validation dataset contained 26,181 (12.27%) and 187,239 (87.73%) ECGs, respectively. The training data were used to construct the SVM classification model, whereas the validation data were used to validate the accuracy of the models. The parameters of the ECG signals in this study were as follows: the time to acquire a continuous ECG signal was 15 seconds, the sampling frequency was 256 samples/second, the input dynamic range was ±2 mV, and the bandwidth was 0.004 Hz to 40 Hz.

Diagnosis and Electrocardiogram Heartbeat Classification

For the purpose of ECG signal classifications, we applied the algorithm that is a combination of support vector machine and rule-based processing. The SVM [33] is a nonlinear classification method. It is a supervised learning model with automatic learning algorithms that analyze data patterns for classification and discrimination analysis. The concept of the SVM method is to transfer the input features into a multiple-dimensional space. In this space, a set of hyperplanes is constructed by the attributes transformed from the features. The ultimate goal of the SVM method is to generate the optimal hyperplanes that are used as the classification principles to separate all subjects [34]. The SVM method has become more and more popular in signal and image processing [35-37]. In our system, the radial basis function (RBF) kernel was applied for constructing SVM models, and the model parameter for the slack variable was set to 100.

The classification method of rule-based processing is to interpret ECGs using expert knowledge in designing. The method is generally suitable for analyzing the morphological characteristics of ECGs. For this reason, we could discriminate abnormal heartbeats, such as APC and VPC, using the QRS-wave pattern and the RR interval. Generally, these kinds of heartbeats do not have a normal morphology and impose an arrhythmic change in normal ECG patterns. Thus, we applied a rule-based, weighted Bayesian classifier to detect abnormal heartbeats. According to the medical definition of heartbeats, we applied the following rules for classification: (1) the current RR interval is smaller than the average RR interval, (2) the current QRS-complex width is larger than the average QRS-complex width, and (3) the amplitudes of the current R, S, and Q peaks are higher than those of other heartbeats.

Due to the diversity of the ECG waveforms and the purpose of optimization classification, we constructed an integrated surveillance decision algorithm that integrates the SVM model and rule-based processing according to their discrimination performance. In addition, the overfitting problem should be avoided as it may overrepresent the performance of models. Therefore, we did not discriminate the data using the principle "OR" among classifier results.

To test the performance of the proposed algorithm, the statistical indicators of sensitivity (SE), the positive prediction rate (+P), the detection error rate (DER), specificity (SP), and accuracy (ACC) were adopted for evaluating the results. An accurate algorithm will have higher SE, +P, SP, and ACC values and a smaller DER value. The formulas of SE, +P, and DER are listed in equations (1) and (2) where true positive (TP) is the number of the true cases that are successfully recognized as true cases, false negative (FN) is the number of true cases that are regarded as false cases, false positive (FP) is the number of false cases that are treated as true cases, and true negative (TN) is the number of false cases that are validly identified as false cases.

$$SE(\%) = TP/(TP+FN), +P(\%) = TP/(TP+FP), DER(\%) = (FP+FN)/(TP+FN) \quad (1)$$

$$SP(\%) = TN/(TN+FP), ACC(\%) = (TP+TN)/(TP+TN+FP+FN) \quad (2)$$

To validate the capability of the proposed classification models, we conducted a retrospective study using the confirmation ECG data that were used for diagnosis by cardiologists as the gold standard to verify the models.

Results

Telesurveillance System

To provide ubiquitous telehealth care, a telesurveillance system at the Telehealth Center of the NTUH was deployed under an SOA framework with the Health Level Seven (HL7) standard. By providing long-term informative interaction and long-term health monitoring, the presented telehealth care system is more than a health monitoring system—it is also helpful for clinical decision making. The service provided must be able to take care of routines and subroutines and act as a health information center to share the data among heterogeneous platforms, such as hospital information systems and health information systems. Hence, our system successfully provides a continuous, real-time, secure, Web-based telehealth care service for both patients and medical staff. A screenshot of our system is shown in Figure 4. The menu on the left side of the screen includes a patient list with the personal serial number and name of each patient. The right-hand section contains the following menu items: (1) VitalSign, which contains uploaded biometric data, including single-lead ECGs, blood pressure, heart rate, oximetry, and glucometry—in diabetic patients with impaired fasting glucose and impaired glucose tolerance, (2) Plan, which contains telephone interview records and health care planning, (3) Profile, which contains the patient's individual profile, (4) Report, which is the monthly statistical report, (5) EMR, which is the electronic medical record, including the history of prescriptions, medications, allergies, and laboratory data at the NTUH, (6) Image, which contains uploaded wound photos, (7) Feedback, which contains the user's satisfaction survey and nutritional assessment, and (8) Renew, which refreshes the page and data. For example, the section VitalSign illustrates the main uploaded information, including the recording date/time, uploading date/time, and estimated heartbeat. It also provides the list of patients' ECGs. Users are able to access the required information by clicking the tabs. Moreover, in order to reduce the workload of medical practitioners, the user can switch between sinus and disease ECGs. These tags are labeled by the automatic classification mechanism. Sinus data are normal rhythm ECGs and disease data are ECGs associated with any disease. After selection, the ECG data will be displayed on the platform.

To provide value-added service, the system is equipped with an automatic interpretation function to help medical personnel in clinical practice. We designed a Web-based user interface for medical staff, which can review the ECG data on the platform and make a decision with corresponding classification suggestions. The diagnostic interface of an ECG record is illustrated in Figure 5. The left-hand section shows a continuous 15-second ECG signal with the common standard unit. In a standard ECG, the width of a single, small square represents 0.2 seconds and the height of the square is 0.5 mV. Moreover, users can click on the bottom, left-hand buttons to indicate the R, P, and T peaks, and the baseline of the ECG. By the same token, they can not only switch the height of the ECG figure to 15 mm/mV or to 10 mm/mV, but they can also use the filter to eliminate the frequency noise. Items on the right-hand side include the date/time of uploading and measuring, estimated heartbeats, diagnosis selection, and the marked area. If "Show in patient's report" is selected, the ECG and judgment will appear in the patient's monthly statistical report. At the Telehealth Center of the NTUH, there are 20 types of ECG diagnoses built into the database, such as sinus rhythm, atrial fibrillation, and first-degree atrioventricular block (AVB1). We employ an icon to easily represent the diagnosis suggestion by classification models. The information is relayed to the cardiologist who makes the final clinical decision and health care suggestions. For example, in Figure 5, the "AF" cell is labeled with a blue dot by the model in this ECG case. Hence, the physician could pay more attention to this icon, whether the suggestion is consistent with their diagnosis or not. In particular, the supported information is very important with some complicated data—it could provide assistance to physicians in enhancing the accuracy of decision making. After diagnosis, for high quality of care, the serious abnormal data would immediately alert case managers, who could then make phone calls to patients or their caregivers.

Figure 4. A screenshot of the telesurveillance system. Users are able to access the required information on the platform, such as patients' biometric data, electronic medical records, and monthly statistical reports.

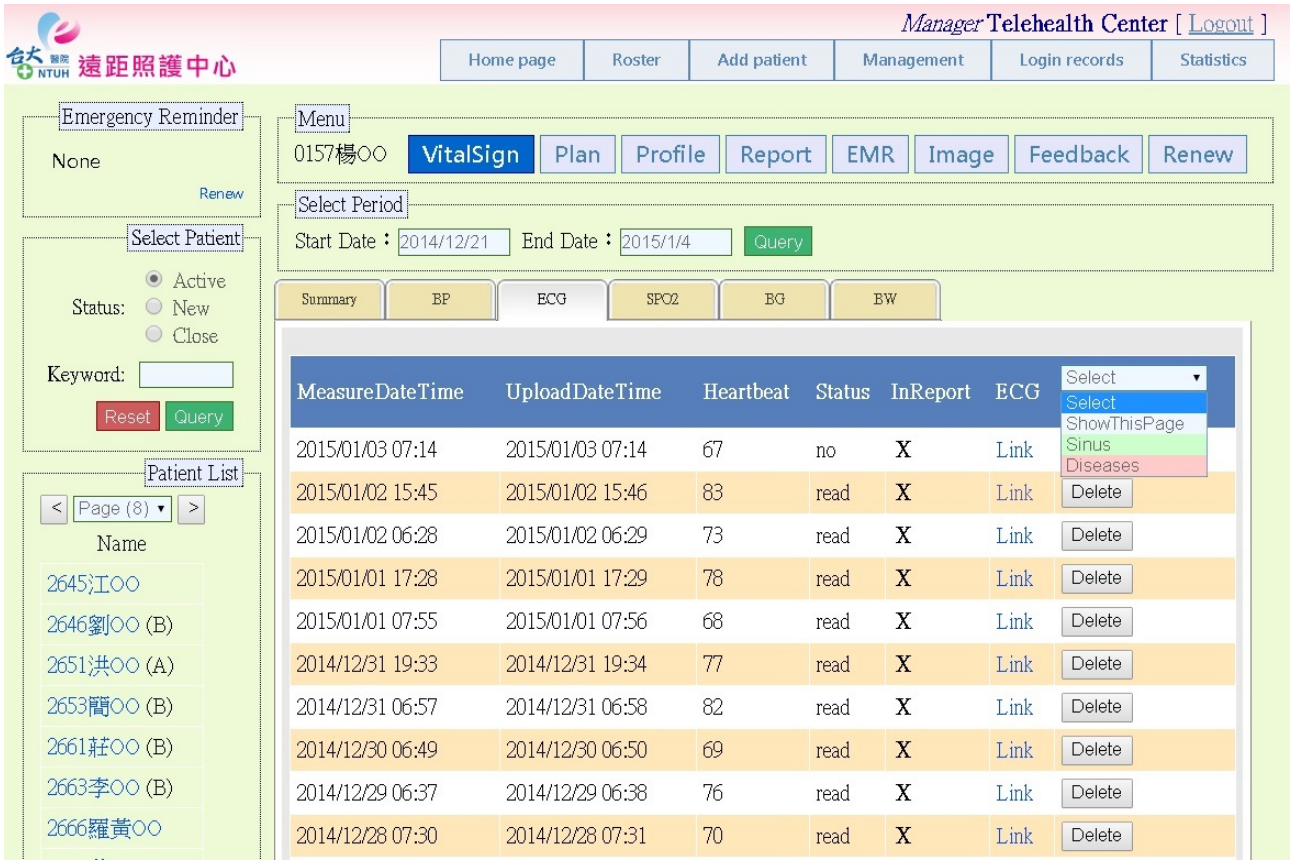
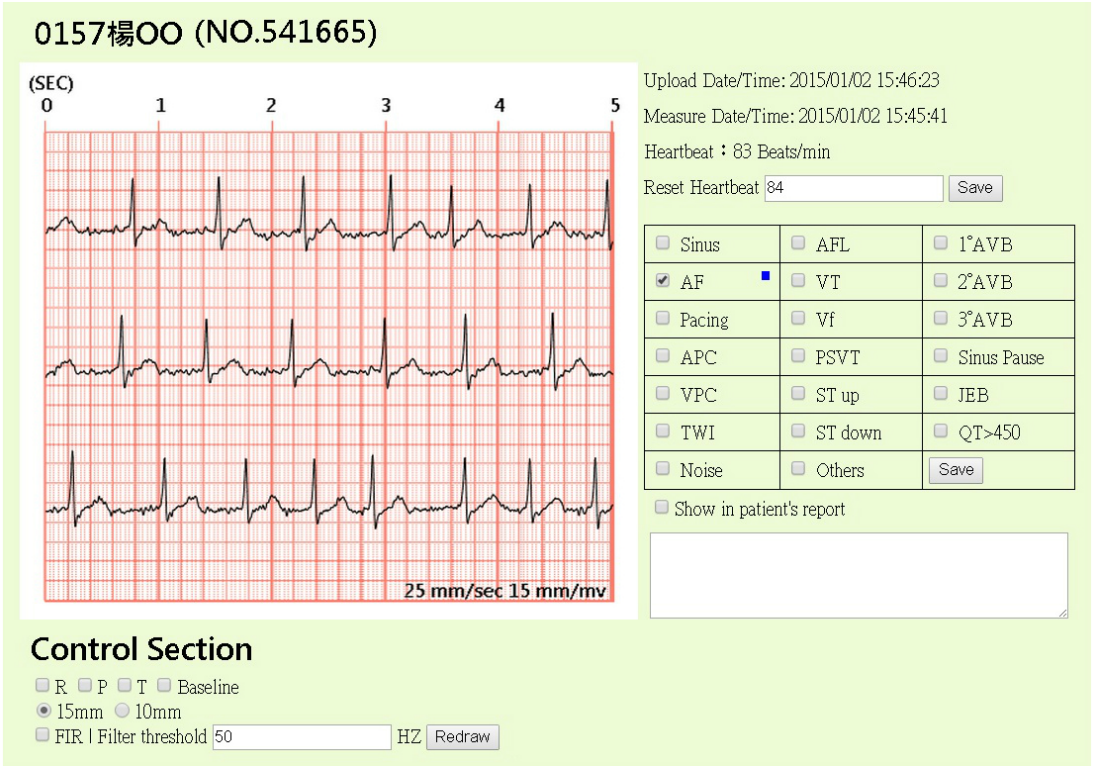


Figure 5. A screenshot of ECG diagnosis using the telesurveillance system. The ECG waveform and the corresponding classification suggestions are revealed on the screen. The suggested heartbeat classification is marked with a blue dot. Health professionals can make decisions using this information in clinical practice.



Peak Evaluation Results

The performance of the proposed R-wave peak detection algorithm, based on a total of 48 records from the MIT-BIH arrhythmia database, was analyzed. Out of a total number of 109,494 heartbeats in the MIT-BIH arrhythmia database, the algorithm detected 73 false negatives and 134 false positives, giving it a detection error rate of 0.19%. The sensitivity of the algorithm was 99.93% (109,371 true positives/[109,371 true positives plus 73 false negatives]). The algorithm had a positive prediction rate of 99.88% (109,371 true positives/[109,371 true positives plus 134 false positives]) (see Table 2). In particular, for healthy and semihealthy cases, the average detection error rate was 0.1% (SD 0.002). One thing to notice is that the R-wave peak detection algorithm was very simple to implement without using any transform-domain methods, such as the Fourier transform and the wavelet transform. This algorithm also used the adaptive threshold method to increase the ECG R-wave peak detection accuracy rate and reduce the numbers of false positives and false negatives by considering the cases of irregularity and noise-like peaks on ECG signals. When implemented by MATLAB, the average detection time for each 30-minute ECG dataset in the MIT-BIH database was less than 0.65 seconds.

Table 2. Performance of R-wave peak detection algorithm.

Characteristics of dataset and algorithm		n or %
Type of beats, n		
	Total beats	109,494
	True positives	109,371
	False negatives	73
	False positives	134
Algorithm performance measure, %		
	Detection error rate	0.19
	Positive prediction rate	99.88
	Sensitivity	99.93

Descriptive Statistics

The automatic ECG classification mechanism proposed by this study was evaluated using the diagnostic data at the Telehealth Center of the NTUH. The distribution of the ECG data is shown in Table 3. There were 213,420 heartbeats from 530 patients measured between February 14, 2012 and December 31, 2014. Overall, the number of sinus, disease, and noise cases from the entire dataset of 213,420 heartbeats were 151,040 (70.77%), 54,218 (25.40%), and 10,514 (4.93%), respectively. Additionally, the heartbeat problem that occurred most often was atrial fibrillation (21,580/213,420, 10.11%). Other common problems—sorted by the number of cases—out of 213,420 heartbeats were atrial premature contraction (11,181, 5.24%), atrial flutter (7858, 3.68%), first-degree atrioventricular block (6304, 2.95%), and pacemaker rhythm (6040, 2.83%). To compare the differences between the training dataset and the validation dataset, the proportion of the sinus, disease, and noise cases in the training dataset was made to be similar to that of the validation dataset.

Table 3. Electrocardiogram dataset from 530 patients from the Telehealth Center of the National Taiwan University Hospital.

Diagnosis	Number of heartbeats, n (%)		
	Total (n=213,420)	Training dataset (n=26,181)	Validation dataset (n=187,239)
All	213,420 (100)	26,181 (100)	187,239 (100)
Sinus	151,040 (70.77)	18,429 (70.39)	132,611 (70.82)
Uncertain	8162 (3.82)	1593 (6.08)	6569 (3.51)
Disease [a]			
All[b]	54,218 (25.40)	6159 (23.52)	48,059 (25.67)
AF	21,580 (10.11)	2232 (8.53)	19,348 (10.33)
AFL	7858 (3.68)	800 (3.06)	7058 (3.77)
Pacemaker rhythm	6040 (2.83)	1433 (5.47)	4607 (2.46)
APC	11,181 (5.24)	1234 (4.71)	9947 (5.31)
VPC	732 (0.34)	107 (0.41)	625 (0.33)
TWI	3064 (1.44)	307 (1.17)	2757 (1.47)
ST-segment down	1156 (0.54)	92 (0.35)	1064 (0.57)
AVB1	6304 (2.95)	341 (1.30)	5963 (3.18)
JEB	263 (0.12)	39 (0.15)	224 (0.12)
QT>450	5 (0)	0 (0)	5 (0)
Noise	10,514 (4.93)	1904 (7.27)	8610 (4.60)

[a]Atrial fibrillation (AF), atrial flutter (AFL), atrial premature contraction (APC), ventricular premature contraction (VPC), T-wave inversion (TWI), first-degree atrioventricular block (AVB1), junctional escape beat (JEB), QT-segment length is more than 450 milliseconds (QT>450).

[b]Since two or more problems may occur at a heartbeat at the same time, the sum of individual disease heartbeats is more than the number of all disease heartbeats combined.

Automatic Electrocardiogram Classification Results

In this study, we used the validation data to obtain an objective performance evaluation with several indicators. The capability of the proposed ECG classification mechanism is shown in Table 4. The experimental results show that the accuracy, sensitivity, and specificity in sinus (ie, normal rhythms) cases were 53.32% ([47,036 true positive plus 52,804 true negative]/187,239 total), 35.47% (47,036 true positive/[47,036 true positive plus 85,575 false negative]), and 96.67% (52,804 true negative/[1824 false positive plus 52,804 true negative]), respectively.

Since we hope that, when the disease case occurs, the clinician can be informed, it is important to prevent the classification model from missing any possible disease data. Therefore, the model with higher specificity for sinus cases and higher sensitivity for disease cases is preferred. Table 4 shows that, in the disease case, our model yielded a sensitivity of 98.50% (47,339 true positive/[47,339 true positive plus 720 false negative]). In the sinus case, the model yielded a specificity of 96.67%.

For the detection performances of specific diseases, the recognition models generated sensitivity values of 92.70% (17,935 true positive/[17,935 true positive plus 1413 false negative]) in atrial fibrillation, 89.10% (4105 true positive/[4105 true positive plus 502 false negative]) in pacemaker rhythm, 88.60% (8813 true positive/[8813 true positive plus 1134 false negative]) in atrial premature contraction, 72.98% (2012 true positive/[2012 true positive plus 745 false negative]) in T-wave inversion, 62.21% (4391 true positive/[4391 true positive plus 2667 false negative]) in atrial flutter, and 62.57% (3731 true positive/[3731 true positive plus 2232 false negative]) in first-degree atrioventricular block. Moreover, the accuracy, sensitivity, and specificity to detect the noise cases were 81.17% ([6984 true positive plus 144,995 true negative]/187,239 total), 81.11% (6984 true positive/[6984 true positive plus 1626 false negative]), and 81.17% (144,995 true negative/[33,634 false positive plus 144,995 true negative]), respectively. Since the noisy ECG signals could be identified by the algorithm accurately, it could be adjusted by denoising approaches to yield good-quality ECG signals.

Table 4. Electrocardiogram classification performance for the dataset from the Telehealth Center of the National Taiwan University Hospital.

Diagnosis	Characteristics of dataset and algorithm						
	Type of beats, n				Algorithm performance measure, %		
	True positive	False negative	False positive	True negative	Accuracy	Sensitivity	Specificity
Sinus	47,036	85,575	1824	52,804	53.32	35.47	96.66
Disease [a]							
All	47,339	720	94,842	44,338	48.96	98.50	31.86
AF	17,935	1413	20,357	147,534	88.37	92.70	87.88
AFL	4391	2667	12,530	167,651	91.88	62.21	93.05
Pacemaker rhythm	4105	502	117,876	64,756	36.78	89.10	35.46
APC	8813	1134	48,838	128,454	73.31	88.60	72.45
VPC	317	308	4595	182,019	97.38	50.72	97.54
TWI	2012	745	22,623	161,859	87.52	72.98	87.74
ST-segment down	471	593	10,007	176,168	94.34	44.27	94.63
AVB1	3731	2232	15,771	165,505	90.39	62.57	91.30
JEB	30	194	4698	182,317	97.39	13.39	97.49
QT>450	1	4	10,630	176,604	94.32	20.00	94.32
Noise	6984	1626	33,634	144,995	81.17	81.12	81.17

[a]Atrial fibrillation (AF), atrial flutter (AFL), atrial premature contraction (APC), ventricular premature contraction (VPC), T-wave inversion (TWI), first-degree atrioventricular block (AVB1), junctional escape beat (JEB), QT-segment length is more than 450 milliseconds (QT>450).

Discussion

Principal Findings

In this study, a telesurveillance system with automatic recognition of the ECG in real time was implemented. Our system was intentionally designed for monitoring and classifying the ECG signals of telehealth users who are being cared for at home. Ultimately, ECGs could not only be transmitted to the hospital over the telecommunication system, but could also be recognized using automatic ECG classifiers for offering a suggestion for diagnosis. Therefore, the system provides the 24-hour service every day. It can automatically identify abnormal ECGs and send alarms to health care providers. In ECG preprocessing, we used a denoising approach based on an FIR filter and performed baseline drift removal with a gradient weighting function. Both techniques can enhance the signal portion of a contaminated ECG record and improve the accuracy of feature extraction. Next, a fixed sliding window, an adaptive peak-height thresholding scheme, and a search-back method were applied for ECG peak detection. According to the preliminary results of R-wave peak evaluation from the MIT-BIH arrhythmia database, our algorithm achieved a detection error rate of 0.19%, a sensitivity of 99.93%, and a positive prediction rate of 99.88%. Moreover, wavelet transforms, relative locations, matched filters, and the regularity test were also employed for feature extraction. For abnormal heartbeat classification, we adopted the interpretation approaches of the support vector machine and rule-based processing. The experimental results of the proposed ECG classification mechanism showed the classifiers yielded a specificity of 96.66% for normal heartbeats, a sensitivity of 98.50% for disease cases, and an accuracy of 81.17% for noise cases. For diagnosing specific heartbeat problems, the interpretation model generated sensitivities of 92.70% for atrial fibrillation, 89.10% for pacemaker rhythm, 88.60% for atrial premature contraction, 72.98% for T-wave inversion, 62.21% for atrial flutter, and 62.57% for first-degree atrioventricular block. For medical staff, they would be able to upload the ECG signals of patients through this clinical decision support system. Then, the immediately automatic interpretation of the ECG could provide physicians with a suggested diagnosis to help them make a decision accurately. This system is very helpful especially when the data size is very large. Moreover, we also integrated electronic medical records into the system, which include such information as prescriptions, food allergies, and drug allergies. With this information, the medical staff could provide more adequate advice to patients.

Limitations

There were some limitations to this study. First, the SVM model is not suitable to use with the imbalanced data, since it tends to classify the instances into the majority class. To overcome this problem, we first applied the rule-based approach to recognize the minority class. The rule-based classifier could immediately detect the disease cases using some specific features. Second, we adopted a genetic algorithm to generate the most relevant features for constructing SVM models, whereas the total features were selected as input features for training in order to create optimal classifiers. As well, additional rule-based features were required to augment the current automated classification models to consider all of the features for classifying. For the rule-based

classifier, all selected features were determined after discussions with ECG-domain knowledge experts (ie, hospital doctors), and also from in-depth consultation of several ECG textbooks. Third, we classified abnormal heartbeats with specific features. Hence, for these classifiers it is hard to identify heartbeat problems without significant features. For example, ventricular tachycardia (VT) and ventricular fibrillation (VF) usually do not have normal waves, complexes, and segments due to improper electrical activity and the uncoordinated contraction of the cardiac muscle. Moreover, the number of cases of these diseases is fairly small and it is not suitable to construct SVM models. These kinds of ECGs may usually be classified into the noise class. Fortunately, with the progress and development of the implantable cardioverter defibrillator (ICD), this therapy could save patients with sudden cardiac disease. Finally, the accuracy of ECG diagnosis depends on the coding of cardiologists [38,39]. This is an innate disadvantage of big database analysis. Nevertheless, these kinds of studies reveal real-world information that can be used for medical science research studies, and they offer a meaningful contribution in the form of generating evidence to solve current medical issues. Besides, this study resulted in 96.18% (205,258/213,420) readable ECGs—the ECGs that were not classified as uncertain cases in Table 3. We believe that the reliability of this data is sufficient for conducting research studies and for making diagnosis suggestions for physicians.

Comparison With Prior Work

With the advances in modern telecommunication technologies, telehealth care is one of the trends in medical treatment. Previous studies have confirmed that telehealth care is an efficient approach in disease management [2,40-42]. The telehealth care system in this study is not only a health monitoring system, but also a tool that assists in decision making. Fortunately, our previous studies have shown that the Telehealth Center of the National Taiwan University Hospital has provided effective telehealth care for chronic cardiovascular disease patients and has reduced medical costs and the burden on caregivers [43-45]. A previous study has also indicated that the data analytics in the telehealth care system could assist clinicians at the point of care [46]. In this study, we established an automatic mechanism for ECG signal collection, transmission, and processing, and then used this massive amount of data to implement a clinical decision support system, which was codesigned by the clinicians at the NTUH.

ECG R-wave peak detection is one of the most important parts of a fully automated ECG analysis algorithm. Many R-wave peak detection algorithms have been proposed. The methods in some previous studies [13,15,47] are time-domain based, and those in two other studies [48,49] are transform-domain based. Cui [13] proposed an algorithm based on zero-crossing counting. It achieved a sensitivity of 99.8% and a detection error rate of 0.6%. The algorithm by Chen et al [15] mainly applied morphology and background noise removal, and achieved a sensitivity of 99.7% and a detection error rate of 0.7%. Wang et al [47] proposed another QRS-detection algorithm—it generated a sensitivity of 99.8% and a detection error rate of 0.5%. Arzeno et al [48] proposed an algorithm that is based on the discrete wavelet transform (DWT) with a sigma-delta modulator—it achieved a sensitivity of 98.0% and a detection error rate of 2.8%. The algorithm by Hamilton and Tompkins [49] used the biorthogonal spline wavelet and applied the Mallat algorithm to detect feature points—it had a sensitivity of 99.7% and a detection error rate of 0.5%. By contrast, the real-time R-wave peak detection algorithm adopted in our system used slopes to find the local maxima or minima within a fixed time slot. The QRS R-wave peak usually happened at the local maxima or minima with the largest change of slope. In addition, an adaptive thresholding scheme, the regularity of heartbeats, the matched filter, and the sharpness of the peak were also adopted for R-wave peak detection. Evaluation results showed that the proposed method achieved a positive prediction rate of 99.88%, a sensitivity of 99.93%, and a detection error rate of 0.19% when applied to data from the MIT-BIH arrhythmia database, which indicates better performance than that of the other methods. Moreover, since R-wave peak candidate sifting is applied, our algorithm can be implemented in an efficient way.

In recent years, there were several related research studies about detecting cardiac anomalies from ECG signals. The first study proposed an arrhythmia disease diagnosis method based on the artificial neural network (ANN) classifier using the University of California at Irvine (UCI) 12-lead arrhythmia data. Their model classified ECGs into normal or abnormal (ie, arrhythmia) cases. They obtained a sensitivity and a specificity of 93.8% and 93.1%, respectively [50]. Another method that applied the feed-forward artificial neural network to identify normal, VPC, and other heartbeats was proposed by Ince et al [51]. For the MIT-BIH arrhythmia database, Ince et al's method achieved 99.4% sensitivity and 98.9% specificity for identifying normal heartbeats, 93.4% sensitivity and 93.3% specificity for determining VPC heartbeats, and 87.5% sensitivity and 97.8% specificity for other heartbeats. Sankari and Adeli [52] proposed a mobile cardiac monitoring system for identifying three cardiac pathologies: atrial fibrillation, atrioventricular block, and myocardial infarction. The system yielded a sensitivity of 95.0%—detecting 95.0% of the pathologies—and a specificity of 100%. However, the system was tested using 60 simulated pathologic ECG datasets rather than a big database. A recent study that investigated the autoregressive model for atrial fibrillation screening was proposed by Parvaresh and Ayatollahi [53]. The experimental results using the MIT-BIH AF database showed that the model's sensitivity and specificity were 96.1% and 93.2%, respectively [53]. Similarly, in another research study, Lian et al [54] developed an AF detector based on the change of RR intervals. It yielded 94.3% sensitivity and 95.1% specificity when applied to the MIT-BIH atrial fibrillation database, and 98.1% sensitivity and 77.0% specificity when applied to the MIT-BIH arrhythmia database. However, it only generated a specificity of 84.1% for non-AF detection when applied to the MIT-BIH normal sinus rhythm database. In fact, we also tested the performance of our algorithm using the MIT-BIH database. Most of the VPC heartbeats were detected successfully with an average sensitivity value of 98.08% and an average specificity value of 99.31%. For APC feature extraction and classification, the proposed algorithm yielded an average sensitivity value of 97.45% and an average specificity value of 99.52%. Compared with the previous studies,

our methods have an even better performance when applied to the MIT-BIH database.

Although these studies have algorithms that achieve good performance for ECG classification, they are generally not suitable for multiple heartbeat problem diagnoses. In addition, the performance of these models was evaluated using the MIT-BIH database rather than real-world data, which can be significantly affected by various environmental factors and can be much more complicated to analyze.

To make the proposed telesurveillance system really helpful to practical clinics, we developed an automatic ECG interpretation algorithm using real-world, multiple-diagnosed ECG data from the telehealth care program. The proposed system yielded a much higher specificity for normal cases and a much higher sensitivity for disease cases than those of other algorithms. As a result, our mechanism is reliable enough to obviate the need for the physician's diagnosis and confirmation.

Conclusions

Via the telesurveillance system, the telehealth care and communication devices, and the automatic ECG interpretation mechanism, telehealth users can be monitored and cared for at home anytime, whereby real-time ECG signals are collected, transmitted, and displayed, and the corresponding classification suggestions are revealed on the system. Furthermore, this paper presents several methods for ECG signal preprocessing and classification. Traditional techniques aim at identifying heartbeats and adjusting the waveforms of ECG signals. In contrast, our proposed interpretation mechanism combines SVM and rule-based processing, and is intentionally designed to automatically analyze the ECG signals of patients in the telehealth care service system. With this value-added service, this intelligent system could widely assist physicians and other health professionals with decision-making tasks in clinical practice, which is important for making users accept remote medical assistance technologies in general.

Acknowledgments

The authors would like to thank all medical staff for their work and assistance at the Telehealth Center of the National Taiwan University Hospital. The authors would also like to thank the National Taiwan University for funding this study (Grant No. NTU-CESRP-104R7608-3).

Conflicts of Interest

None declared.

Multimedia Appendix 1

CONSORT-EHEALTH checklist V1.6.2 [55].

[PDF File (Adobe PDF File), 84KB - medinform_v3i2e21_app1.pdf]

References

1. Paré G, Jaana M, Sicotte C. Systematic review of home telemonitoring for chronic diseases: the evidence base. J Am Med Inform Assoc 2007 Jun;14(3):269-277 [FREE Full text] [doi: 10.1197/jamia.M2270] [Medline: 17329725]
2. Inglis SC, Clark RA, McAlister FA, Ball J, Lewinter C, Cullington D, et al. Structured telephone support or telemonitoring programmes for patients with chronic heart failure. Cochrane Database Syst Rev 2010(8):CD007228. [doi: 10.1002/14651858.CD007228.pub2] [Medline: 20687083]
3. Chiu TM, Ku BP. Moderating effects of voluntariness on the actual use of electronic health records for allied health professionals. JMIR Med Inform 2015 Feb;3(1):e7 [FREE Full text] [doi: 10.2196/medinform.2548] [Medline: 25720417]
4. Zhang HW, Lin YJ, Su YH, Chen SJ, Chen HS. Pilot study on a community-based ubiquitous healthcare system for current and retired university employees. In: Proceedings of the IEEE International Conference on Communications Workshops. 2009 Jun Presented at: IEEE International Conference on Communications Workshops; June 14-18, 2009; Dresden, Germany p. 1-5. [doi: 10.1109/ICCW.2009.5208085]
5. Dhillon JS, Ramos C, Wunsche BC, Lutteroth C. Designing a web-based telehealth system for elderly people: An interview study in New Zealand. In: Proceedings of the 24th International Symposium on Computer-Based Medical Systems (CBMS). 2011 Jun Presented at: 24th International Symposium on Computer-Based Medical Systems (CBMS); June 27-30, 2011; Bristol, UK p. 1-6. [doi: 10.1109/CBMS.2011.5999157]
6. Davies N. Reducing inequalities in healthcare provision for older adults. Nurs Stand 2011;25(41):49-55; quiz 58. [doi: 10.7748/ns2011.06.25.41.49.c8573] [Medline: 21815517]
7. Shany T, Hession M, Pryce D, Galang R, Roberts M, Lovell N, et al. Home telecare study for patients with chronic lung disease in the Sydney West Area Health Service. Stud Health Technol Inform 2010;161:139-148. [Medline: 21191167]
8. El-Menyar A, AlMahmeed W. Heart failure in 2010. Expert Rev Cardiovasc Ther 2010 Sep;8(9):1231-1234. [doi: 10.1586/erc.10.118] [Medline: 20828344]

9. Klug S, Krupka K, Dickhaus H, Katus HA, Hilbel T. Displaying computerized ECG recordings and vital signs on Windows Phone 7 smartphones. In: Proceedings of Computing in Cardiology. 2010 Presented at: Computing in Cardiology; September 26-29, 2010; Belfast, Northern Ireland, UK p. 1067-1070 URL: http://www.cinc.org/2010/program/106.pdf
10. Anker SD, Koehler F, Abraham WT. Telemedicine and remote management of patients with heart failure. Lancet 2011 Aug 20;378(9792):731-739. [doi: 10.1016/S0140-6736(11)61229-4] [Medline: 21856487]
11. Pan J, Tompkins WJ. A real-time QRS detection algorithm. IEEE Trans Biomed Eng 1985 Mar;32(3):230-236. [doi: 10.1109/TBME.1985.325532] [Medline: 3997178]
12. Friesen GM, Jannett TC, Jadallah MA, Yates SL, Quint SR, Nagle HT. A comparison of the noise sensitivity of nine QRS detection algorithms. IEEE Trans Biomed Eng 1990 Jan;37(1):85-98. [doi: 10.1109/10.43620] [Medline: 2303275]
13. Cui X. A new real-time ECG R-wave detection algorithm. In: Proceedings of the 6th International Forum on Strategic Technology (IFOST). 2011 Aug Presented at: 6th International Forum on Strategic Technology (IFOST); Aug 22-24, 2011; Harbin, China p. 1252-1255. [doi: 10.1109/IFOST.2011.6021247]
14. Zheng H, Wu J. Real-time QRS detection method. In: Proceedings of the 10th International Conference on e-health Networking, Applications and Services (HealthCom). 2008 Jul Presented at: 10th International Conference on e-health Networking, Applications and Services (HealthCom); July 7-9, 2008; Singapore p. 169-170. [doi: 10.1109/HEALTH.2008.4600130]
15. Chen SW, Chen HC, Chan HL. A real-time QRS detection method based on moving-averaging incorporating with wavelet denoising. Comput Methods Programs Biomed 2006 Jun;82(3):187-195. [doi: 10.1016/j.cmpb.2005.11.012] [Medline: 16716445]
16. Schuck A, Wisbeck JO. QRS detector pre-processing using the complex wavelet transform. In: Proceedings of the 25th Annual International Conference of the IEEE on Engineering in Medicine and Biology Society. 2003 Sep 17 Presented at: 25th Annual International Conference of the IEEE on Engineering in Medicine and Biology Society; September 17-19, 2003; Cancun, Mexico p. 2590-2593. [doi: 10.1109/IEMBS.2003.1280445]
17. Benitez DS, Gaydecki PA, Zaidi A, Fitzpatrick AP. A new QRS detection algorithm based on the Hilbert transform. In: Proceedings of Computers in Cardiology. 2000 Sep 24 Presented at: Computers in Cardiology; September 24-27, 2000; Cambridge, MA p. 379-382. [doi: 10.1109/CIC.2000.898536]
18. Pan T, Zhang Z, Zhou S. Detection of ECG characteristic points using biorthogonal spline wavelet. In: Proceedings of the 3rd International Conference on Biomedical Engineering and Informatics (BMEI). 2010 Oct 16 Presented at: 3rd International Conference on Biomedical Engineering and Informatics (BMEI); October 16-18, 2010; Yantai, China p. 858-863. [doi: 10.1109/BMEI.2010.5639905]
19. Zong W, Mukkamala E, Mark RG. A methodology for predicting paroxysmal atrial fibrillation based on ECG arrhythmia feature analysis. In: Proceedings of Computers in Cardiology. 2001 Sep 23 Presented at: Computers in Cardiology; September 23-26, 2001; Rotterdam, Netherlands p. 125-128. [doi: 10.1109/CIC.2001.977607]
20. Ge D, Srinivasan N, Krishnan SM. Cardiac arrhythmia classification using autoregressive modeling. Biomed Eng Online 2002 Nov;1(5):1-12. [doi: 10.1186/1475-925X-1-5]
21. Hickey B, Heneghan C. Screening for paroxysmal atrial fibrillation using atrial premature contractions and spectral measures. In: Proceedings of Computers in Cardiology. 2002 Sep 22 Presented at: Computers in Cardiology; September 22-25, 2002; Memphis, TN p. 217-220. [doi: 10.1109/CIC.2002.1166746]
22. Yeh Y, Wang W, Chiou CW. Cardiac arrhythmia diagnosis method using linear discriminant analysis on ECG signals. Measurement 2009 Jun;42(5):778-789. [doi: 10.1016/j.measurement.2009.01.004]
23. Yeh YC, Lin HJ. Cardiac arrhythmia diagnosis method using fuzzy C-means algorithm on ECG signals. In: Proceedings of the International Symposium on Computer Communication Control and Automation (3CA). 2010 May 05 Presented at: International Symposium on Computer Communication Control and Automation (3CA); May 5-7, 2010; Tainan, Taiwan p. 272-275. [doi: 10.1109/3CA.2010.5533831]
24. Ebrahimzadeh A, Khazaee A. An efficient technique for classification of electrocardiogram signals. AECE 2009;9(3):89-93. [doi: 10.4316/AECE.2009.03016]
25. Martis RJ, Acharya UR, Ray AK, Chakraborty C. Application of higher order cumulants to ECG signals for the cardiac health diagnosis. In: Proceedings of the Engineering in Medicine and Biology Society, Annual International Conference of the IEEE. 2011 Aug 30 Presented at: Engineering in Medicine and Biology Society, Annual International Conference of the IEEE; August 30, 2011-September 3, 2011; Boston, MA p. 1697-1700. [doi: 10.1109/IEMBS.2011.6090487]
26. Chudacek V, Petrik M, Georgoulas G, Cepek M, Lhotska L, Stylios C. Comparison of seven approaches for holter ECG clustering and classification. In: Proceedings of the Engineering in Medicine and Biology Society, Annual International Conference of the IEEE. 2007 Aug 22 Presented at: Engineering in Medicine and Biology Society, Annual International Conference of the IEEE; August 22-26, 2007; Lyon, France p. 3844-3847. [doi: 10.1109/IEMBS.2007.4353171]
27. Li Y, Yan H, Hong F, Song J. A new approach of QRS complex detection based on matched filtering and triangle character analysis. Australas Phys Eng Sci Med 2012 Sep 01;35(3):341-356. [doi: 10.1007/s13246-012-0149-x]
28. Lin CC. Analysis of unpredictable components within QRS complex using a finite-impulse-response prediction model for the diagnosis of patients with ventricular tachycardia. Comput Biol Med 2010 Jul;40(7):643-649. [doi: 10.1016/j.compbiomed.2010.05.002] [Medline: 20605138]

29. Chen YJ, Ding JJ, Huang CW, Ho YL, Hung CS. ECG baseline extraction by gradient varying weighting functions. In: Proceedings of the Signal and Information Processing Association Annual Summit and Conference (APSIPA). 2013 Oct 29 Presented at: Signal and Information Processing Association Annual Summit and Conference (APSIPA); October 29, 2013-November 1, 2013; Kaohsiung, Taiwan p. 1-4. [doi: 10.1109/APSIPA.2013.6694129]
30. Ding JJ, Huang CW, Ho YL, Hung CS, Lin YH, Chen YH. An efficient selection, scoring, and variability ratio test algorithm for ECG R-wave peak detection. Exp Clin Cardiol 2014;20(8):4256-4263 [FREE Full text]
31. PT AS, Joseph PK, Jacob J. Automated diagnosis of diabetes using heart rate variability signals. J Med Syst 2012 Jun;36(3):1935-1941. [doi: 10.1007/s10916-011-9653-x] [Medline: 21271353]
32. Yilmaz B, Asyali MH, Arikan E, Yetkin S, Ozgen F. Sleep stage and obstructive apneaic epoch classification using single-lead ECG. Biomed Eng Online 2010;9:39 [FREE Full text] [doi: 10.1186/1475-925X-9-39] [Medline: 20723232]
33. Chang C, Lin C. LIBSVM: A library for support vector machines. ACM Trans Intell Syst Technol 2011 Apr 01;2(3):1-27. [doi: 10.1145/1961189.1961199]
34. Chen WH, Hsieh SL, Hsu KP, Chen HP, Su XY, Tseng YJ, et al. Web-based newborn screening system for metabolic diseases: machine learning versus clinicians. J Med Internet Res 2013;15(5):e98 [FREE Full text] [doi: 10.2196/jmir.2495] [Medline: 23702487]
35. Gokgoz E, Subasi A. Effect of multiscale PCA de-noising on EMG signal classification for diagnosis of neuromuscular disorders. J Med Syst 2014 Apr;38(4):31. [doi: 10.1007/s10916-014-0031-3] [Medline: 24696395]
36. Tachibana RO, Oosugi N, Okanoya K. Semi-automatic classification of birdsong elements using a linear support vector machine. PLoS One 2014;9(3):e92584 [FREE Full text] [doi: 10.1371/journal.pone.0092584] [Medline: 24658578]
37. Ho TW, Lai HY, Wang YJ, Chen WH, Lai F, Ho YL, et al. A clinical decision and support system with automatically ECG classification in telehealthcare. In: Proceedings of the IEEE 16th International Conference on e-Health Networking, Applications and Services (Healthcom). 2014 Oct 15 Presented at: IEEE 16th International Conference on e-Health Networking, Applications and Services (Healthcom); October 15-17, 2014; Natal-RN, Brazil p. 293-297. [doi: 10.1109/HealthCom.2014.7001857]
38. Agarwal SK, Soliman EZ. ECG abnormalities and stroke incidence. Expert Rev Cardiovasc Ther 2013 Jul;11(7):853-861. [doi: 10.1586/14779072.2013.811980] [Medline: 23895029]
39. Kannel WB, Wolf PA, Benjamin EJ, Levy D. Prevalence, incidence, prognosis, and predisposing conditions for atrial fibrillation: population-based estimates. Am J Cardiol 1998 Oct 16;82(8A):2N-9N. [Medline: 9809895]
40. Clark RA, Inglis SC, McAlister FA, Cleland JG, Stewart S. Telemonitoring or structured telephone support programmes for patients with chronic heart failure: systematic review and meta-analysis. BMJ 2007 May 5;334(7600):942 [FREE Full text] [doi: 10.1136/bmj.39156.536968.55] [Medline: 17426062]
41. Winkler S, Koehler F. A meta-analysis of remote monitoring of heart failure patients. J Am Coll Cardiol 2010 Apr 6;55(14):1505-1506; author reply 1506 [FREE Full text] [doi: 10.1016/j.jacc.2009.12.028] [Medline: 20359604]
42. Clarke M, Shah A, Sharma U. Systematic review of studies on telemonitoring of patients with congestive heart failure: a meta-analysis. J Telemed Telecare 2011;17(1):7-14. [doi: 10.1258/jtt.2010.100113] [Medline: 21097564]
43. Ho YL, Yu JY, Lin YH, Chen YH, Huang CC, Hsu TP, et al. Assessment of the cost-effectiveness and clinical outcomes of a fourth-generation synchronous telehealth program for the management of chronic cardiovascular disease. J Med Internet Res 2014;16(6):e145 [FREE Full text] [doi: 10.2196/jmir.3346] [Medline: 24915187]
44. Chen YH, Lin YH, Hung CS, Huang CC, Yeih DF, Chuang PY, et al. Clinical outcome and cost-effectiveness of a synchronous telehealth service for seniors and nonseniors with cardiovascular diseases: quasi-experimental study. J Med Internet Res 2013;15(4):e87 [FREE Full text] [doi: 10.2196/jmir.2091] [Medline: 23615318]
45. Chiang LC, Chen WC, Dai YT, Ho YL. The effectiveness of telehealth care on caregiver burden, mastery of stress, and family function among family caregivers of heart failure patients: a quasi-experimental study. Int J Nurs Stud 2012 Oct;49(10):1230-1242. [doi: 10.1016/j.ijnurstu.2012.04.013] [Medline: 22633448]
46. Simpao AF, Ahumada LM, Gálvez JA, Rehman MA. A review of analytics and clinical informatics in health care. J Med Syst 2014 Apr;38(4):45. [doi: 10.1007/s10916-014-0045-x] [Medline: 24696396]
47. Wang Y, Deepu CJ, Lian Y. A computationally efficient QRS detection algorithm for wearable ECG sensors. In: Proceedings of the 33rd Annual International Conference of the IEEE on Engineering in Medicine and Biology Society (EMBC). 2011 Presented at: 33rd Annual International Conference of the IEEE on Engineering in Medicine and Biology Society (EMBC); August 30-September 3, 2011; Boston, MA p. 5641-5644. [doi: 10.1109/IEMBS.2011.6091365]
48. Arzeno NM, Deng ZD, Poon CS. Analysis of first-derivative based QRS detection algorithms. IEEE Trans Biomed Eng 2008 Feb;55(2 Pt 1):478-484 [FREE Full text] [doi: 10.1109/TBME.2007.912658] [Medline: 18269982]
49. Hamilton PS, Tompkins WJ. Quantitative investigation of QRS detection rules using the MIT/BIH arrhythmia database. IEEE Trans Biomed Eng 1986 Dec;33(12):1157-1165. [Medline: 3817849]
50. Jadhav S, Nalbalwar S, Ghatol A. Artificial neural network models based cardiac arrhythmia disease diagnosis from ECG signal data. Int J Comput Appl 2012 Apr 30;44(15):8-13. [doi: 10.5120/6338-8532]
51. Ince T, Kiranyaz S, Gabbouj M. Automated patient-specific classification of premature ventricular contractions. In: Proceedings of the Engineering in Medicine and Biology Society, Annual International Conference of the IEEE. 2008 Aug

20 Presented at: Engineering in Medicine and Biology Society, Annual International Conference of the IEEE; August 20-25, 2008; Vancouver, BC p. 5474-5477. [doi: 10.1109/IEMBS.2008.4650453]
52. Sankari Z, Adeli H. HeartSaver: a mobile cardiac monitoring system for auto-detection of atrial fibrillation, myocardial infarction, and atrio-ventricular block. Comput Biol Med 2011 Apr;41(4):211-220. [doi: 10.1016/j.compbiomed.2011.02.002] [Medline: 21377149]
53. Parvaresh S, Ayatollahi A. Automatic atrial fibrillation detection using autoregressive modeling. In: Proceedings of the International Conference on Biomedical Engineering and Technology. 2011 Jun 4 Presented at: International Conference on Biomedical Engineering and Technology; June 4-5, 2011; Kuala Lumpur, Malaysia p. 105-108 URL: http://www.webcitation.org/6XJsrvi2y
54. Lian J, Wang L, Muessig D. A simple method to detect atrial fibrillation using RR intervals. Am J Cardiol 2011 May 15;107(10):1494-1497. [doi: 10.1016/j.amjcard.2011.01.028] [Medline: 21420064]
55. Eysenbach G, Consort- E. CONSORT-EHEALTH: improving and standardizing evaluation reports of Web-based and mobile health interventions. J Med Internet Res 2011;13(4):e126 [FREE Full text] [doi: 10.2196/jmir.1923] [Medline: 22209829]

Abbreviations

+P: positive prediction rate
ACC: accuracy
AF: atrial fibrillation
AFL: atrial flutter
AJAX: asynchronous JavaScript and XML
ANN: artificial neural network
APC: atrial premature contraction
AVB1: first-degree atrioventricular block
CDF: Cohen-Daubechies-Feauveau
CDSS: clinical decision support system
DER: detection error rate
DWT: discrete wavelet transform
ECG: electrocardiogram
EMR: electronic medical record
FIR: finite impulse response
FN: false negative
FP: false positive
HL7: Health Level Seven
ICD: implantable cardioverter defibrillator
ICT: information and communication technology
JEB: junctional escape beat
MIT-BIH: Massachusetts Institute of Technology-Beth Israel Hospital
MVC: Model-View-Controller
NTUH: National Taiwan University Hospital
QT>450: QT-segment length is more than 450 milliseconds
RBF: radial basis function
SE: sensitivity
SP: specificity
SOA: service-oriented architecture
SQL: Structured Query Language
SVM: support vector machine
TN: true negative
TP: true positive
TWI: T-wave inversion
UCI: University of California at Irvine
VF: ventricular fibrillation
VPC: ventricular premature contraction
VT: ventricular tachycardia
WLAN: wireless local area network

Edited by G Eysenbach; submitted 04.03.15; peer-reviewed by HP Ma; comments to author 19.03.15; accepted 03.04.15; published 07.05.15.

Please cite as:
Ho TW, Huang CW, Lin CM, Lai F, Ding JJ, Ho YL, Hung CS
A Telesurveillance System With Automatic Electrocardiogram Interpretation Based on Support Vector Machine and Rule-Based Processing
JMIR Med Inform 2015;3(2):e21
URL: http://medinform.jmir.org/2015/2/e21/
doi:10.2196/medinform.4397
PMID:25953306

Publisher:
JMIR Publications
130 Queens Quay East.
Toronto, ON, M5A 3Y5
Phone: (+1) 416-583-2040
Email: support@jmir.org