JMIR Medical Informatics

Contents

Viewpoint

Original Papers

Corrigenda and Addendas

Viewpoint

Optimizing Patient Preparation and Surgical Experience Using eHealth Technology

Amy Waller[1], BA Psych (Hons), PhD; Kristy Forshaw[1], BMedSci, MPH; Mariko Carey[1], BSc, PhD (Clin Psychol); Sancha Robinson[2,3], MBBS, FANZCA; Ross Kerridge[2,3], MBBS, FRCA, FANZCA; Anthony Proietto[4], BMBS, FRANZCOG; Rob Sanson-Fisher[1], BPsych, MPsych, PhD

[1]University of Newcastle & Hunter Medical Research Institute, Health Behaviour Research Group, Callaghan, Australia

[2]John Hunter Hospital, Perioperative Service, Newcastle, Australia

[3]University of Newcastle, School of Medicine and Public Health, Callaghan, Australia

[4]Hunter New England Local Health District, Cancer Services and Cancer Network, Newcastle, Australia

Corresponding Author:
Amy Waller, BA Psych (Hons), PhD
University of Newcastle & Hunter Medical Research Institute
Health Behaviour Research Group
W4 HMRI Building
University of Newcastle
Callaghan, 2305
Australia
Phone: 61 2 4042 0708
Fax: 61 2 4042 0044
Email: amy.waller@newcastle.edu.au

Abstract

With population growth and aging, it is expected that the demand for surgical services will increase. However, increased complexity of procedures, time pressures on staff, and the demand for a patient-centered approach continue to challenge a system characterized by finite health care resources. Suboptimal care is reported in each phase of surgical care, from the time of consent to discharge and long-term follow-up. Novel strategies are thus needed to address these challenges to produce effective and sustainable improvements in surgical care across the care pathway. The eHealth programs represent a potential strategy for improving the quality of care delivered across various phases of care, thereby improving patient outcomes. This discussion paper describes (1) the key functions of eHealth programs including information gathering, transfer, and exchange; (2) examples of eHealth programs in overcoming challenges to optimal surgical care across the care pathway; and (3) the potential challenges and future directions for implementing eHealth programs in this setting. The eHealth programs are a promising alternative for collecting patient-reported outcome data, providing access to credible health information and strategies to enable patients to take an active role in their own health care, and promote efficient communication between patients and health care providers. However, additional rigorous intervention studies examining the needs of potential role of eHealth programs in augmenting patients' preparation and recovery from surgery, and subsequent impact on patient outcomes and processes of care are needed to advance the field. Furthermore, evidence for the benefits of eHealth programs in supporting carers and strategies to maximize engagement from end users are needed.

Global Burden of Surgical Conditions

Approximately 234 million surgical operations take place each year globally. Depending on the procedure, there may be substantial direct costs for consumers, including specialist consultations and hospitalization, postoperative care, and medications, as well as indirect costs, including travel and lost productivity [1]. Personal costs include pain, suffering, and premature mortality. Hospital costs can vary according to the length of stay, surgical procedure performed, and the care needs of the patient [2]. Patients undergoing surgery are increasingly older, often have complex comorbidities, and require more

XSL·FO
RenderX

efficient surgical care [3]. It is expected that with population growth and aging, the demand for surgical services will escalate [4-6]. The health system faces considerable pressure to increase the level and quality of surgical care within finite health care resources.

Demands for High Quality, Patient-Centered Care Across the Surgical Pathway Are Not Being Met

The surgical care pathway is characterized by multiple phases of care, from the decision to have surgery to discharge from hospital and follow-up care. Providing optimal care across the different phases of the surgical pathway has become increasingly challenging, due to the complexity of procedures, increasing time pressures on staff, and the demand for a patient-centered approach [7]. Breakdowns in one phase can affect other phases, which in turn can cause delays, cancellations, and complications. For example, minimum standards for informed consent and decision making are not always achieved [8]. This eventually results in unnecessary or unwanted procedures or preventable harm [9]. Patients report inadequate preparation, resulting in surgical cancellations and delays, undiagnosed medical problems, and anxiety, as well as increased length of hospital stay, analgesic requirements, and cost of surgical care [3,10]. Discharge planning may be compromised by a lack of guidelines and systems in hospitals, poor information recall, or limited involvement of patients in the discharge process, as well as a shortage of caregiver and community resources to support recovery. Patients do not always receive detailed instructions at the time of discharge, and this increases the risk of an unnecessarily prolonged recovery, thereby reducing quality of life and increasing costs [11]. Novel strategies are thus needed to address these challenges to produce effective and sustainable improvements in surgical care across the care pathway.

Using eHealth to Address Current Challenges Across the Surgical Pathway

Overview

The World Health Organization defines eHealth as "the transfer of health-related resources and health care by electronic means, including information, support resources, assessments, interventions, and health care records" [12]. Endorsed as part of a strategic plan to improve quality of health care, one of the key recommendations made by the Institute of Medicine was the use of eHealth programs [13]. The eHealth programs have the potential to support care delivery models, engage providers and patients, and deliver self-assessment and self-management tools [14]. The key functions of eHealth programs can be categorized as information gathering, transfer, and exchange. The aim of this discussion paper is to describe these key functions, and outline how such features can be applied to presurgical and postsurgical care. Advantages and challenges posed by the use of eHealth as well as key gaps in the evidence base are discussed.

Information Gathering

Variation in the type and quality of information obtained by clinicians during clinical interviews occurs as a consequence of time and resource constraints, as well as individual clinicians' bias [15]. Utilizing self-report assessments of eHealth programs via tablets can improve data integrity by standardizing information collected by clinicians. To reduce complexity and data-collection time, algorithms can be built-in to the software so that items can be auto-populated or skipped based on responses. Programs can be developed so that patients can access and complete assessments outside the clinic environment before surgical consultations.

Information Transfer

The eHealth programs can connect patients with credible, standard information and support regardless of geographic location, the clinician providing care, or the resources of the institution. A credible single source of information is critical given the quantity and variable quality of information available on the Internet [16]. When evidence-based practice recommendations change, information can be updated easily and quickly. Patients can control the number of times they access eHealth programs and the level of information they search and obtain. Providing information tailored to an individual's knowledge and preferences reduces anxiety, improves information comprehension, and recall [17].

Information Exchange

Health information exchanges (ie, electronic health records) are available as a platform for key information to be made available to authorized health care providers across care settings to promote continuity. This is especially relevant for older patients and those with multiple comorbidities, given the range of health care providers they may encounter. For example, health information exchanges have the potential to support the electronic sharing of clinical data across organizations, offering timely and complete medical records at the point of care. Immediate access to medical records or investigation results can increase satisfaction and treatment compliance [18] and reduce medical errors and complaints against services [19].

Potential of eHealth to Improve Care and Outcomes Across the Surgical Pathway

The phases of surgical care are conceptualized as follows: the "preoperative phase," which refers to care delivered prior to surgery; the "intraoperative phase" when surgery is performed; and the "postoperative phase," which is the period from surgery completion/patient recovery to discharge from hospital. Within each phase are critical steps that patients encounter as they progress through the pathway. We have used these steps as a framework to illustrate examples where eHealth programs could improve outcomes in the preoperative and postoperative phases of care.

Preoperative Care

Step 1: Enhancing Decision-Making Process and Streamlining Informed Consent

Ideally, patients should have a complete understanding of the risks, benefits, and potential outcomes of the procedure before consent. eHealth programs can augment standard face-to-face informed consent processes by conveying supplementary information, meeting patients' preferences, and exploring understanding of information once it has been delivered [20]. Evidence-based features, such as decision aids, can be incorporated and accessed by the patient before the consultation to help focus discussions [21]. Nonbiased presentation of the risks and benefits of relevant options, a table of pros and cons for easy comparison, value-clarification exercises, and targeted assessments can help clarify patient understanding, identify gaps in knowledge, and reduce decisional regret [22-24]. Programs can also act as a point of reference for patients to access after the consent consultation to consolidate and re-explore information.

Step 2: Collecting Medical History Data, Delivering Information, and Optimizing Preoperative Preparation

Traditionally, there has been only a short timeframe for providing perioperative care [25]. More recently, models of care have been employed in which patient assessment, preparation, and discharge planning begin at the time of booking itself [25]. The eHealth programs enable patients to complete their medical history online at home, or in the waiting room before their surgical consultation using a tablet. This information can then be transferred to the provider in real time so that it is readily accessible and clarified by staff at the preoperative consultation. The eHealth programs can also alleviate some of the burden on providers by delivering written and audiovisual information about the potential risks of anesthesia, the procedure, and preparation requirements [26,27]. Preoperative education programs have reduced length of stay, postoperative medication usage, complications, and anxiety [28]. Providing both procedural and sensory information offers additional benefits [29-31].

Step 3: Streamlining Admission Procedures

Information should be provided to the patient regarding where they need to go in the hospital, dietary and other preparation requirements, and the processes involved from the point of arrival at the hospital to recovering back in the ward after the procedure. Short message services or email can be used to prompt patients about what to bring with them, including consent forms, test and imaging results, medication lists, and Medicare and health fund details. Electronic reminders can also be used to prompt providers to collect specific information from the patient and/or perform a specific clinical action during admission. Electronic reminders can reduce cancellation rates and increase compliance with instructions.

Postoperative Care

Step 4: Delivering Individually Tailored Postoperative Care Plans

Nowadays, postoperative hospital stays are becoming increasingly shorter as a consequence of novel interventions, such as minimally invasive techniques and fast-track programs. Although this can increase patient satisfaction and reduce health care utilization and costs, a major disadvantage is that there is less opportunity for patient education [32]. Using tablets, patients can complete symptom assessments electronically, and during recovery the results can be transmitted through electronic alerts to their care team [14]. Additional information on pain and expected length of stay, as well as evidence-based strategies to self-manage identified symptoms, side effects, and aspects of recovery can be provided to patients using multiple formats. For example, education about the benefits of early mobilization and less reliance on strong analgesics may be particularly important in facilitating early recovery [33].

Step 5: Promote Effective Discharge Planning

Discharge planning that includes appropriate and useful information for patients and their caregivers reduces length of hospital stay and unplanned hospital readmissions, improves quality of inpatient and home care, and increases patient satisfaction [34]. The eHealth programs enable discharge plans to be readily accessible to patients at their own convenience. Information and links to available services and support resources can be tailored to the patient's condition, location, and procedure. Information about whom to contact and when to contact particular health care providers in the event of complications can also be incorporated.

Step 6: Optimizing Rehabilitation and Long-Term Follow-Up

The need to undergo additional surgery to manage complications can be minimized through continuity and timeliness of follow-up care. Patients self-reporting symptoms from home through eHealth programs can result in earlier symptom detection, improve communication, and provide an efficient means to capture data evaluating the effects of procedures on health-related quality of life. Interactive health communication apps combine health information with social support, decision support, or behavior change support and can improve knowledge, social support, and behavioral and clinical outcomes [35]. Programs can be designed to enable goal setting, monitoring of progress, and tailoring of recommendations regarding activities and resources that may be helpful to achieve goals.

These programs can also reduce the burden associated with travel and accommodation for follow-up care. For example, the current practice of routine, face-to-face follow-up of patients who received asymptomatic total joint replacement may be excessively costly and unnecessary. In this situation, tele-rehabilitation via Web-based communication following the surgery may be an alternative option [36], especially for patients who are located remotely. It enables a surgeon to conduct a

follow-up consultation without being physically present using a mobile remote videoconferencing equipment.

Challenges of eHealth and Future Directions

Overview

While promising, a number of potential disadvantages to eHealth programs have been raised in relation to inequity in access to the Internet, poor health literacy, and concerns over privacy and costs. The notion of a "digital divide" in relation to access has been highlighted for particular subgroups, such as those residing in rural areas [37]. Similarly, older people report lower rates of Internet use [38]. As the demand for orthopedic, cardiovascular, and cancer surgery increases as a consequence of an aging population, these access issues must be considered when proposing eHealth programs [39-41]. Others express concern that some groups might have less capacity for eHealth programs. Poor health literacy and cognitive deficits in end users may be particularly challenging. However, integrating features, such as presenting information in a range of accessible formats such as video and audio clips, may help overcome these issues. Familiarity with e-technology is increasing, with growing mobile phone and tablet ownership, which suggests its acceptability in day-to-day life. Research also shows that these are acceptable to people from a variety of health care settings, including surgical patients.

Internet-Based Interventions Are Promising but More Evidence Is Needed

The Internet has been touted as promising for diverse applications in surgical patients' care, such as real-time monitoring lifestyle behaviors among candidates for bariatric surgery [42], and educating breast augmentation patients regarding treatments, medications, and surgical options [43]. However, there is limited evidence of the impact of such approaches on patient outcomes. This may to some extent reflect reluctance to test online interventions in those cases where the evidence for the intervention delivered by more conventional means (eg, face to face) is mixed or ambiguous. For example, there is mixed evidence that face-to-face and telephone-delivered preoperative interventions for surgical patients can improve a number of outcomes such as knowledge, pain, recovery time, and anxiety [26,30,44-46]. The mixed nature of research findings likely suggests that the specific nature of the intervention (content and dose) and the specific patient population need to be considered when making judgments about intervention effectiveness.

There have also been limited studies that evaluated the impact of online preparatory interventions on patient outcomes or processes of care. One randomized controlled trial showed that orthopedic patients who received Internet-based education on anesthesia options before surgery had greater knowledge of anesthesia and were more likely to choose neuraxial rather than general anesthesia compared with the control group [47]. Similarly, although there is emerging evidence that interactive eHealth interventions have positive effects on knowledge, social support, and potentially on behavioral and clinical outcomes

for people with chronic diseases [48], few studies have examined the impact of Internet-delivered interventions for improving self-management and recovery in the perioperative period.

The current generation of mobile phones provides access to Internet [49] with wireless capabilities enabling users to have continuous access from any location [50]. Such continuous connectivity holds immense potential for use in health care [49] and the use of mobile technology in patient care is particularly appealing [51] because of its portability, continuous uninterrupted data stream, and capability to support multimedia software apps [49]. The mobile app industry is also rapidly evolving [51] with a huge potential for interventions to benefit health and health service delivery processes. For example, a previous study reported that for low-risk postoperative ambulatory patients, use of a mobile app for follow-up care was suitable [52]. Although a range of surgical mobile phone apps exist that could benefit both surgeons and patients [53], systematic reviews on the impact of such technologies on health outcomes remain scarce [50]. Interdisciplinary collaboration is thus essential for future advances in this field [51].

Gap Between the Interest in eHealth Educational Tools and Real-World Usage

The eHealth programs have the potential to enable a dramatic transformation in the delivery of surgical care, making it safer, more effective, and more efficient. However, in order for eHealth interventions to achieve these goals, they must be accessible to and used as intended by consumers. Therefore, it is imperative that strategies to maximize consumer engagement and uptake of eHealth programs be considered in any intervention trials. When designing such eHealth programs for surgical patients, key learnings from other areas in which eHealth has been successfully applied may be useful to consider. For example, a meta-analysis showed that online health behavior interventions that are brief, goal oriented, and include tools or strategies to show users the consequences of their actions, assist them in meeting goals, and apply normative social pressures are more likely to be adhered to than those without these features [54]. Another review found that eHealth interventions that include greater interaction with a health care provider, greater dialogue support (eg, praise, peer examples), and more frequent updates were likely to be adhered to by participants [55]. While the impact of eHealth programs is usually measured based on a specific population (eg, people undergoing knee replacement surgery), it is important that the influence of other factors, such as geographical location, are also considered, as these may confound findings. Although it is unclear whether such factors will influence surgical patients' adherence and engagement to eHealth interventions in the absence of surgery-specific studies, these provide a useful starting point.

Augmenting Surgical Care Across the Entire Surgical Care Pathway

Most research on eHealth has focused on improving care during one specific phase of the surgical care pathway, such as preoperative preparation or discharge planning. Segmenting surgical care in this manner does not mirror the patient's experience. Poor patient outcomes may be a consequence of the type of care received during a particular phase on the

continuum (eg, suboptimal consent process) or the transition between different phases (eg, transfer between hospital and home/community services). Targeting improvement strategies to a single phase does not acknowledge the interdependence between each phase. Thus, eHealth programs that promote a holistic model of care across the entire surgical care pathway should be considered.

Promoting a Dyadic Approach to Surgical Care

Despite the increased reliance on family and friends to provide informal care for surgical patients, carers often feel unprepared for the patient's transition from hospital to home. Inadequate preparation results in poorer physical health and high levels of perceived strain and disruptions to family and social life. The eHealth programs can deliver information about strategies that the carer can implement to assist the patient, including how to assist with daily living activities, monitor emotional well-being, and when to contact services for help. Programs can also provide information and activities that the carer can utilize to help manage their own well-being.

Conclusions

The eHealth platforms have the potential to address gaps in the gathering and transfer of information across the 6 phases of the surgical journey. Rather than approaching each of these phases as separate entities, interventions should strive to address each of the phases to promote continuity and holistic care. Rigorous intervention studies are needed to determine the impact of these programs on patient outcomes and processes of care. Studies examining the role of eHealth programs in supporting carers, and strategies to maximize engagement from end users are also needed.

Conflicts of Interest

None declared.

References

1. Australian Day Surgery Council. Day Surgery in Australia: Report and Recommendations on Day Surgery. East Melbourne, VIC: Royal Australasian College of Surgeons; 2004. URL: http://www.anzca.edu.au/resources/college-publications/pdfs/books-and-publications/Day%20Surgery%20in%20Australia.pdf [accessed 2015-08-14] [WebCite Cache ID 6alZVY5KL]

2. Australian Institute of Health and Welfare. Australia's Health 2012. Canberra, Australia: Australian Institute of Health and Welfare; Jun 21, 2012.

3. Bouamrane M, Mair FS. A study of clinical and information management processes in the surgical pre-assessment clinic. BMC Med Inform Decis Mak 2014;14:22 [FREE Full text] [doi: 10.1186/1472-6947-14-22] [Medline: 24666471]

4. Etzioni D, Liu J, Maggard M, Ko C. The aging population and its impact on the surgery workforce. Ann Surg 2003 Aug;238(2):170-177. [doi: 10.1097/01.SLA.0000081085.98792.3d] [Medline: 12894008]

5. Liu J, Etzioni D, O'Connell J, Maggard M, Ko C. The increasing workload of general surgery. Arch Surg 2004 Apr;139(4):423-428. [doi: 10.1001/archsurg.139.4.423] [Medline: 15078711]

6. Royal Australasian College of Surgeons. Surgical Workforce Projection to 2025 (Volume 1). Melbourne, VIC: Royal Australasian College of Surgeons; 2011. URL: https://www.surgeons.org/media/437871/rpt_racs_workforce_projection_to_2025.pdf [accessed 2015-08-13] [WebCite Cache ID 6alaBQfA4]

7. Nehme J, El-Khani U, Chow A, Hakky S, Ahmed AR, Purkayastha S. The use of multimedia consent programs for surgical procedures: A systematic review. Surg Innov 2013 Feb;20(1):13-23. [doi: 10.1177/1553350612446352] [Medline: 22589017]

8. Ankuda C, Block S, Cooper Z, Correll D, Hepner D, Lasic M, et al. Measuring critical deficits in shared decision making before elective surgery. Patient Educ Couns 2014 Mar;94(3):328-333. [doi: 10.1016/j.pec.2013.11.013] [Medline: 24382399]

9. Gogos A, Clark RB, Bismark MM, Gruen RL, Studdert DM. When informed consent goes poorly: A descriptive study of medical negligence claims and patient complaints. Med J Aust 2011 Sep 19;195(6):340-344. [Medline: 21929499]

10. Caumo W, Schmidt A, Schneider C, Bergmann J, Iwamoto C, Adamatti L, et al. Risk factors for postoperative anxiety in adults. Anaesthesia 2001 Aug;56(8):720-728. [Medline: 11493233]

11. Bouwsma EV, Anema JR, Vonk Noordegraaf A, Knol DL, Bosmans JE, Schraffordt Koops SE, et al. The cost effectiveness of a tailored, web-based care program to enhance postoperative recovery in gynecologic patients in comparison with usual care: Protocol of a stepped wedge cluster randomized controlled trial. JMIR Res Protoc 2014;3(2):e30 [FREE Full text] [doi: 10.2196/resprot.3236] [Medline: 24943277]

12. Oh H, Rizo C, Enkin M, Jadad A. What is eHealth (3): A systematic review of published definitions. J Med Internet Res 2005;7(1):e1 [FREE Full text] [doi: 10.2196/jmir.7.1.e1] [Medline: 15829471]

13. Ortiz E, Clancy C, AHRQ. Use of information technology to improve the quality of health care in the United States. Health Serv Res 2003 Apr;38(2):xi-xxii [FREE Full text] [Medline: 12785557]

14. Cook D, Manning D, Holland D, Prinsen S, Rudzik S, Roger V, et al. Patient engagement and reported outcomes in surgical recovery: Effectiveness of an e-health platform. J Am Coll Surg 2013 Oct;217(4):648-655. [doi: 10.1016/j.jamcollsurg.2013.05.003] [Medline: 23891066]

15. Ahmadian L, Cornet R, Van Klei WA, DE Keizer NF. Data collection variation in preoperative assessment: A literature review. Comput Inform Nurs 2011 Nov;29(11):662-670. [doi: 10.1097/NCN.0b013e31820660bb] [Medline: 21270635]

16. McLean T, Delbridge L. Comparison of consumer information on the Internet to the current evidence base for minimally invasive parathyroidectomy. World J Surg 2010 Jun;34(6):1304-1311. [doi: 10.1007/s00268-009-0306-x] [Medline: 20012291]

17. Gaston CM, Mitchell G. Information giving and decision-making in patients with advanced cancer: A systematic review. Soc Sci Med 2005 Nov;61(10):2252-2264. [doi: 10.1016/j.socscimed.2005.04.015] [Medline: 15922501]

18. Ha J, Longnecker N. Doctor-patient communication: A review. Ochsner J 2010;10(1):38-43 [FREE Full text] [Medline: 21603354]

19. Levinson W, Roter D, Mullooly J, Dull V, Frankel R. Physician-patient communication. The relationship with malpractice claims among primary care physicians and surgeons. JAMA 1997 Feb 19;277(7):553-559. [Medline: 9032162]

20. Braddock C3, Hudak PL, Feldman JJ, Bereknyei S, Frankel RM, Levinson W. "Surgery is certainly one good option": Quality and time-efficiency of informed decision-making in surgery. J Bone Joint Surg Am 2008 Sep;90(9):1830-1838 [FREE Full text] [doi: 10.2106/JBJS.G.00840] [Medline: 18762641]

21. National Breast Cancer Centre, National Cancer Centre Initiative. Clinical Practice Guidelines for the Psychosocial Care of Adults With Cancer. Camperdown, NSW: National Breast Cancer Centre; 2003 Apr. URL: https://www.nhmrc.gov.au/_files_nhmrc/publications/attachments/cp90_psychosocial_care_adults_cancer_131223.pdf [accessed 2015-08-13] [WebCite Cache ID 6alb7kVPB]

22. Abhyankar P, Volk RJ, Blumenthal-Barby J, Bravo P, Buchholz A, Ozanne E, et al. Balancing the presentation of information and options in patient decision aids: An updated review. BMC Med Inform Decis Mak 2013;13 Suppl 2:S6 [FREE Full text] [doi: 10.1186/1472-6947-13-S2-S6] [Medline: 24625214]

23. Fagerlin A, Pignone M, Abhyankar P, Col N, Feldman-Stewart D, Gavaruzzi T, et al. Clarifying values: An updated review. BMC Med Inform Decis Mak 2013;13 Suppl 2:S8 [FREE Full text] [doi: 10.1186/1472-6947-13-S2-S8] [Medline: 24625261]

24. Leighl NB, Shepherd FA, Zawisza D, Burkes RL, Feld R, Waldron J, et al. Enhancing treatment decision-making: Pilot study of a treatment decision aid in stage IV non-small cell lung cancer. Br J Cancer 2008 Jun 3;98(11):1769-1773 [FREE Full text] [doi: 10.1038/sj.bjc.6604395] [Medline: 18506180]

25. Lee A, Kerridge RK, Chui PT, Chiu CH, Gin T. Perioperative systems as a quality model of perioperative medicine and surgical care. Health Policy 2011 Oct;102(2-3):214-222. [doi: 10.1016/j.healthpol.2011.05.009] [Medline: 21680044]

26. Lee A, Chui PT, Gin T. Educating patients about anesthesia: A systematic review of randomized controlled trials of media-based interventions. Anesth Analg 2003 May;96(5):1424-1431. [Medline: 12707146]

27. Schenker Y, Fernandez A, Sudore R, Schillinger D. Interventions to improve patient comprehension in informed consent for medical and surgical procedures: A systematic review. Med Decis Making 2011;31(1):151-173. [doi: 10.1177/0272989X10364247] [Medline: 20357225]

28. Snowdon D, Haines T, Skinner E. Preoperative intervention reduces postoperative pulmonary complications but not length of stay in cardiac surgical patients: A systematic review. J Physiother 2014 Jun;60(2):66-77 [FREE Full text] [doi: 10.1016/j.jphys.2014.04.002] [Medline: 24952833]

29. Johnston M, Vögele C. Benefits of psychological preparation for surgery: A meta-analysis. Ann Behav Med 1993;154:245-256.

30. Suls J, Wan C. Effects of sensory and procedural information on coping with stressful medical procedures and pain: A meta-analysis. J Consult Clin Psychol 1989 Jun;57(3):372-379. [Medline: 2738210]

31. Hathaway D. Effect of preoperative instruction on postoperative outcomes: A meta-analysis. Nurs Res 1986;35(5):269-275. [Medline: 3638606]

32. McMurray A, Johnson P, Wallis M, Patterson E, Griffiths S. General surgical patients' perspectives of the adequacy and appropriateness of discharge planning to facilitate health decision-making at home. J Clin Nurs 2007 Sep;16(9):1602-1609. [doi: 10.1111/j.1365-2702.2006.01725.x] [Medline: 17727581]

33. Henriksen MG, Jensen MB, Hansen HV, Jespersen TW, Hessov I. Enforced mobilization, early oral feeding, and balanced analgesia improve convalescence after colorectal surgery. Nutrition 2002 Feb;18(2):147-152. [Medline: 11844646]

34. Bauer M, Fitzgerald L, Haesler E, Manfrin M. Hospital discharge planning for frail older people and their family. Are we delivering best practice? A review of the evidence. J Clin Nurs 2009 Sep;18(18):2539-2546. [doi: 10.1111/j.1365-2702.2008.02685.x] [Medline: 19374695]

35. Lithner M, Johansson J, Andersson E, Jakobsson U, Palmquist I, Klefsgard R. Perceived information after surgery for colorectal cancer: An explorative study. Colorectal Dis 2012 Nov;14(11):1340-1350. [doi: 10.1111/j.1463-1318.2012.02982.x] [Medline: 22329948]

36. Russell T, Buttrum P, Wootton R, Jull G. Internet-based outpatient telerehabilitation for patients following total knee arthroplasty: A randomized controlled trial. J Bone Joint Surg Am 2011 Jan 19;93(2):113-120. [doi: 10.2106/JBJS.I.01375] [Medline: 21248209]

37. Curtin J. A Digital Divide in Rural and Regional Australia. Canberra, Australia: Department of Parliamentary Library; 2001 Aug 7. URL: http://www.aph.gov.au/About_Parliament/Parliamentary_Departments/Parliamentary_Library/Publications_Archive/CIB/cib0102/02CIB01 [accessed 2015-08-14] [WebCite Cache ID 6ale1Eb7E]

38. Campbell R, Nolfi D. Teaching elderly adults to use the Internet to access health care information: Before-after study. J Med Internet Res 2005;7(2):e19 [FREE Full text] [doi: 10.2196/jmir.7.2.e19] [Medline: 15998610]

39. Korc-Grodzicki B, Downey RJ, Shahrokni A, Kingham TP, Patel SG, Audisio RA. Surgical considerations in older adults with cancer. J Clin Oncol 2014 Jul 28;32(24):2647-2653. [doi: 10.1200/JCO.2014.55.0962]

40. Zhang Y, Jordan JM. Epidemiology of osteoarthritis. Clin Geriatr Med 2010 Aug;26(3):355-369 [FREE Full text] [doi: 10.1016/j.cger.2010.03.001] [Medline: 20699159]

41. Bettelli G. Preoperative evaluation in geriatric surgery: Comorbidity, functional status and pharmacological history. Minerva Anestesiol 2011 Jun;77(6):637-646 [FREE Full text] [Medline: 21617627]

42. Thomas J, Bond D, Sarwer D, Wing R. Technology for behavioral assessment and intervention in bariatric surgery. Surg Obes Relat Dis 2011;7(4):548-557 [FREE Full text] [doi: 10.1016/j.soard.2011.02.009] [Medline: 21514246]

43. Gladfelter J. The Internet as an educational tool for breast augmentation. Plast Surg Nurs 2003;23(3):121-123. [Medline: 14666807]

44. Danielsen A, Burcharth J, Rosenberg J. Patient education has a positive effect in patients with a stoma: A systematic review. Colorectal Dis 2013 Jun;15(6):e276-e283. [doi: 10.1111/codi.12197] [Medline: 23470040]

45. Johansson K, Salanterä S, Heikkinen K, Kuusisto A, Virtanen H, Leino-Kilpi H. Surgical patient education: Assessing the interventions and exploring the outcomes from experimental and quasiexperimental studies from 1990 to 2003. Clin Eff Nurs 2004 Jun;8(2):81-92. [doi: 10.1016/j.cein.2004.09.002]

46. McDonnell A. A systematic review to determine the effectiveness of preparatory information in improving the outcomes of adult patients undergoing invasive procedures. Clin Eff Nurs 1999 Mar;3(1):4-13. [doi: 10.1016/S1361-9004(99)80003-X]

47. Groves N, Humphreys H, Williams A, Jones A. Effect of informational internet web pages on patients' decision-making: Randomised controlled trial regarding choice of spinal or general anaesthesia for orthopaedic surgery. Anaesthesia 2010 Mar;65(3):277-282. [Medline: 20336817]

48. Murray E, Burns J, See T, Lai R, Nazareth I. Interactive health communication applications for people with chronic disease. Cochrane Database Syst Rev 2005 Oct 19(4):CD004274. [doi: 10.1002/14651858.CD004274.pub4] [Medline: 16235356]

49. Boulos MN, Brewer AC, Karimkhani C, Buller DB, Dellavalle RP. Mobile medical and health apps: State of the art, concerns, regulatory control and certification. Online J Public Health Inform 2014;5(3):229 [FREE Full text] [doi: 10.5210/ojphi.v5i3.4814] [Medline: 24683442]

50. Fiordelli M, Diviani N, Schulz PJ. Mapping mHealth research: A decade of evolution. J Med Internet Res 2013;15(5):e95 [FREE Full text] [doi: 10.2196/jmir.2430] [Medline: 23697600]

51. Becker S, Miron-Shatz T, Schumacher N, Krocza J, Diamantidis C, Albrecht U. mHealth 2.0: Experiences, possibilities, and perspectives. JMIR Mhealth Uhealth 2014;2(2):e24 [FREE Full text] [doi: 10.2196/mhealth.3328] [Medline: 25099752]

52. Armstrong KA, Semple JL, Coyte PC. Replacing ambulatory surgical follow-up visits with mobile app home monitoring: Modeling cost-effective scenarios. J Med Internet Res 2014;16(9):e213 [FREE Full text] [doi: 10.2196/jmir.3528] [Medline: 25245774]

53. Kulendran M, Lim M, Laws G, Chow A, Nehme J, Darzi A, et al. Surgical smartphone applications across different platforms: Their evolution, uses, and users. Surg Innov 2014 Aug;21(4):427-440. [doi: 10.1177/1553350614525670] [Medline: 24711263]

54. Cugelman B, Thelwall M, Dawes P. Online interventions for social marketing health behavior change campaigns: A meta-analysis of psychological architectures and adherence factors. J Med Internet Res 2011;13(1):e17 [FREE Full text] [doi: 10.2196/jmir.1367] [Medline: 21320854]

55. Kelders SM, Kok RN, Ossebaard HC, Van Gemert-Pijnen JE. Persuasive system design does matter: A systematic review of adherence to web-based interventions. J Med Internet Res 2012;14(6):e152 [FREE Full text] [doi: 10.2196/jmir.2104] [Medline: 23151820]

XSL•FO
RenderX

Original Paper

Context-Sensitive Spelling Correction of Consumer-Generated Content on Health Care

Xiaofang Zhou[1,2*]; An Zheng[2*]; Jiaheng Yin[2,3*]; Rudan Chen[2]; Xianyang Zhao[2]; Wei Xu[2]; Wenqing Cheng[2]; Tian Xia[2,4], PhD; Simon Lin[5], MD

[1]The Department of Ophthalmology, Wuhan Central Hospital, Wuhan, China

[2]Internet Technology and Engineering Research and Development Center, School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan, China

[3]Department of Biostatistics and Computational Biology, School of Life Science, Fudan University, Shanghai, China

[4]Northwestern University Biomedical Informatics Center (NUBIC), Feinberg School of Medicine, Northwestern University, Chicago, IL, United States

[5]Research Institute, Nationwide Children's Hospital, Columbus, OH, United States

[*]these authors contributed equally

Corresponding Author:
Tian Xia, PhD
Internet Technology and Engineering Research and Development Center
School of Electronic Information and Communications
Huazhong University of Science and Technology
Nanyi, Huazhong University of Science and Technology
1037 Luoyu Road, Hongshan
Wuhan, 430074
China
Phone: 86 02787544704
Fax: 86 02787544704
Email: tianxia@hust.edu.cn

Abstract

Background: Consumer-generated content, such as postings on social media websites, can serve as an ideal source of information for studying health care from a consumer's perspective. However, consumer-generated content on health care topics often contains spelling errors, which, if not corrected, will be obstacles for downstream computer-based text analysis.

Objective: In this study, we proposed a framework with a spelling correction system designed for consumer-generated content and a novel ontology-based evaluation system which was used to efficiently assess the correction quality. Additionally, we emphasized the importance of context sensitivity in the correction process, and demonstrated why correction methods designed for electronic medical records (EMRs) failed to perform well with consumer-generated content.

Methods: First, we developed our spelling correction system based on Google Spell Checker. The system processed postings acquired from MedHelp, a biomedical bulletin board system (BBS), and saved misspelled words (eg, sertaline) and corresponding corrected words (eg, sertraline) into two separate sets. Second, to reduce the number of words needing manual examination in the evaluation process, we respectively matched the words in the two sets with terms in two biomedical ontologies: RxNorm and Systematized Nomenclature of Medicine -- Clinical Terms (SNOMED CT). The ratio of words which could be matched and appropriately corrected was used to evaluate the correction system's overall performance. Third, we categorized the misspelled words according to the types of spelling errors. Finally, we calculated the ratio of abbreviations in the postings, which remarkably differed between EMRs and consumer-generated content and could largely influence the overall performance of spelling checkers.

Results: An uncorrected word and the corresponding corrected word was called a spelling pair, and the two words in the spelling pair were its members. In our study, there were 271 spelling pairs detected, among which 58 (21.4%) pairs had one or two members matched in the selected ontologies. The ratio of appropriate correction in the 271 overall spelling errors was 85.2% (231/271). The ratio of that in the 58 spelling pairs was 86% (50/58), close to the overall ratio. We also found that linguistic errors took up 31.4% (85/271) of all errors detected, and only 0.98% (210/21,358) of words in the postings were abbreviations, which was much lower than the ratio in the EMRs (33.6%).

XSL•FO
RenderX

Conclusions: We conclude that our system can accurately correct spelling errors in consumer-generated content. Context sensitivity is indispensable in the correction process. Additionally, it can be confirmed that consumer-generated content differs from EMRs in that consumers seldom use abbreviations. Also, the evaluation method, taking advantage of biomedical ontology, can effectively estimate the accuracy of the correction system and reduce manual examination time.

Introduction

Background

In the last two decades, spelling correction methods for clinical texts have been studied extensively. Nevertheless, the majority of related studies mainly focused on the electronic medical record (EMR) [1], but largely ignored consumer-generated content which has accumulated rapidly because of the development of online media and social networks. The consumers mentioned here include those who describe their symptoms and seek online medical assistance, and those who have been successfully cured and willing to share their treatment process experience on public websites or forums. Although there is no doubt that the EMR content is worthy of in-depth study, information in consumer-generated content is equally useful and informative, which has been discussed in a US National Research Council Committee Framework [2] and in Zeng et al [3]. Mining information in consumer-generated content based on large-scale text analysis becomes increasingly important in the context where social networks have become pervasive in recent years. For example, the useful relationship information between biomedical terms can be inferred based on texts extracted from postings in various online health communities written by patients. Obviously, the accuracy of these inferences relies on correctly spelled text. Therefore, the development of spelling correction methods for consumer-generated content is critical for ensuring the accuracy and efficiency of downstream text analysis.

Related Work

Spelling Correction

Numerous approaches for correcting spelling errors, such as Levenshtein edit distance [4,5] and semantic correction [1,6], have been proposed. The Levenshtein edit distance model demonstrates a method to measure the edit distance of converting one string to another, which is calculated by counting the number of four-letter operations—deletions, insertions, transpositions, and substitutions—during the conversion. For example, when correcting "plls" to "pills," we need to insert the letter "i" which increases the edit distance by one. The candidate with the lowest edit distance will be recognized as the best replacement for the misspelled word. The semantic correction model utilizes context-sensitive detection and has been widely applied to studies using natural language processing (NLP). For example, Wong and Glance [1] developed a robust system using semantic correction to correct misspelled words, especially abbreviation disambiguation, in progress notes. In addition, according to a study proposed by Ruch et al [7], these

two models can be combined: first, the Levenshtein edit distance is computed and the resulting candidate words are ranked according to the edit distance. Each word is then examined according to the context using semantic correction. Finally, the best suitable candidate is picked according to both edit distance and semantic meaning. There are other extensively used methods such as the Soundex system proposed by Odell and Russell [8,9] and the n-gram model [7,10,11]. Some studies applied an integrated spelling correction application programming interface (API), such as GNU Aspell, Yahoo API, etc. Wong and Glance [1] adopted and mixed GNU Aspell and Yahoo API corrective interfaces in their systems for real-time abbreviation disambiguation, which has achieved good results. These interfaces have become highly sound and mature after a long period of development.

Evaluation Methods

The mainstream evaluation methods for spelling correction systems can be ascribed into two types: horizontal comparison and longitudinal comparison. Horizontal comparison means that researchers test several different correction models with the same input, and then compare their performance and accuracy to prove the strength of the newly designed model. For example, in the study by Ruch et al , they compared the correction results of four different correction models derived from NLP. Longitudinal comparison is generally applied in evaluating methods which are used to improve and perfect existing spelling correction systems. This comparison mainly focuses on the difference between the spelling error correction rate before and after the improvement, as in Crowell et al [12].

Currently, most of the prevalent evaluation methods for spelling correction are based on manual inspection [1]. Although it is accurate, the manual evaluation is time consuming, and not feasible to be applied in large-scale experiments. Therefore, we explored the use of formal ontologies to evaluate the effectiveness of spelling correction.

Spelling Error Classifications

Spelling errors are usually divided into different categories. Ruch et al classified misspelled words in EMRs into two categories. The first category, called typographical error, refers to spelling mistakes which lead to misspelled words becoming nonexistent in the dictionary. For instance, when a consumer spells "plls" instead of "pills," there is no chance of finding "plls" in a lexicon. The second category, called *linguistic error*, refers to typing errors which cause a word's original meaning to change, but the misspelled word still exists in the dictionary (eg, spelling "three pills" as "tree pills"). Syntactic and semantic spelling errors are included in this category. Similarly, many

other studies, such as those of Jurafsky and James [13] and Wilbur et al [14], classified spelling errors according to whether misspelled words needed *isolated-word error correction* or *context-dependent error correction*. Our system followed Ruch's classification method, categorizing spelling errors into *typographical* and *linguistic errors*.

Limitations With Existing Approach

There are several limitations that exist within the current approaches and hinder the correction process from achieving highly efficient performance. To begin with, some existing approaches will become less efficient and require an abundance of training data when processing large amounts of text. For example, according to Ruch et al, correction systems using Levenshtein edit distance require extremely large amounts of training data, which can be scarcely satisfied in real-world situations. Also, the semantic correction process is highly complex when the correction system needs to detect both typographical and linguistic errors [7].

In addition, the context-related errors make up a large ratio of spelling errors in consumer-generated content (shown in the following sections). If we only focus on typographical errors [15] in order to achieve high efficiency, then the accuracy of the correction system will be largely sacrificed, and overall system performance will appear much less desirable than approaches considering both linguistic and typographical errors.

Moreover, unique features of consumer-generated content should also be taken into consideration in the correction process. Consumer-generated content differs from EMR content, in that there are many abbreviations written by clinical professionals in EMRs, which are rarely shown in consumer-generated content. EMRs contain abbreviated terms such as "VSS" (vital signs stable), "PVCs" (premature ventricular contractions), NTG" (nitroglycerin), and gtt" (guttae) to describe patients' physical and mental conditions in a quantitative and professional fashion, while consumers prefer to describe their conditions using common language such as "depressed," "pain," and "feel

better." This distinct feature leads to differences in spelling correction strategies between EMR and consumer-generated content [1].

Our Approach

We proposed a spelling correction system based on Google Spell Checker, which is not only able to automatically correct both typographical and linguistic errors, but is also highly efficient thanks to Google Spell Checker's core algorithms [16]. Our system focuses on correcting spelling errors in daily medical vocabularies, rather than professional, but not commonly used, terminology like the methods proposed by Wang et al [1], Doan et al [17], and Patrick et al [18]. It is a real-time and high-performance method that can be easily applied to studies requiring automatic correction of misspelled words.

In order to shorten the evaluation period and preserve the reliability of the evaluation, we narrowed down the range of words being examined by matching these words with biomedical ontology items, and then manually examining the matched words. Ontologies consist of words and phrases describing and annotating concepts in many fields, such as biomedical informatics *and artificial intelligence. To evaluate our system, we selected two* biomedical ontologies: Systematized Nomenclature of Medicine -- Clinical Terms (SNOMED CT), which is focused on diseases and symptoms, and RxNorm, which is focused on drugs.

Methods

Dataset

In this study, we randomly selected 150 postings (21,358 words in total; Multimedia Appendix 1) from MedHelp's bulletin board system (BBS) [19]. This set of postings is related to a drug named *Zoloft* and contains consumers' descriptions of their symptoms and suggestions from others, such as doctors, pharmacists, and patients, who have already used Zoloft. Figure 1 shows one example from the 150 postings.

Figure 1. Screenshot of a sample post from MedHelp's bulletin board system.

Tools Used in Our Study

Google Spell Checker

We based our system on Google Spell Checker, a state-of-art spelling correction tool which is embedded in Google Search and utilizes the Web pages as corpus. Our system can upload text segments, which need spelling checked, onto Google Search and spelling suggestions will be automatically generated by Google Spell Checker. Google Spell Checker's high accuracy and efficiency have been proven by Jacquemont et al and Islam and Inkpen who applied Google's search engine and Google Web 1T n-gram—a language model extracting nearly 1 trillion words from Web pages—into the spelling correction process.

National Center for Biomedical Ontology Annotator

To reduce the amount of manual work in the evaluation process, we used the National Center for Biomedical Ontology (NCBO) Annotator [20] to match texts with formal ontologies. The NCBO is a website which contains all biomedical ontologies and relevant knowledge; ontology is a set of terms related to a certain subject, such as biochemistry and movement (eg, "*Amino Acid Ontology*" and "*Cell Ontology*"). The NCBO's Annotator is used to search annotations of biomedicine-related texts in the given ontologies. After selecting ontologies and submitting original texts, users will obtain matched terms from the Annotator; terms exist in the designated ontologies. In addition, there is no need for the users to manually submit text one by one in NCBO's website. A Web service is provided for all users to accomplish the text-mining jobs programmatically [21].

Framework

Construction of Our Spelling Correction System

We developed our spelling correction system based on Google Spell Checker. The system works in three steps: text segmentation, text spelling correction, and text reconstruction.

In the first step—text segmentation—content (eg, a post from MedHelp) is automatically grouped into sets of less than 32 words, since Google's search engine can only process 32 words at a time in the correction program. It is worth mentioning that, although our system divides the postings automatically, it does not destroy the complete structure of one sentence. According to the online data [22], the average sentence length is 15 to 20 words, which is less than the 32-word requirement in the Google search engine. Additionally, the Google Spell Checker is able to consider the context of the candidates' suggestions, and evolves in accordance with the update of millions of Web pages [16]. All the segments processed are saved in our database. In this way, when context-sensitive texts are separated, this will prevent changes to their original meaning.

In the second step—text spelling correction—our system uploads the segments saved in the database onto Google Search and downloads the feedback generated by Google Spell Checker. Google Spell Checker not only corrects typographical errors but also proposes suggestions for linguistic errors according to relations of context, including syntactic and semantic relations. The syntactic relation helps in correcting grammatical errors. For example, in some posts, "*had*" was misspelt as "*has,*" but it turns out that "*had*" was more suitable in the contexts. In these circumstances, our system can find this type of problem and deliver the correct output. The semantic relation is used in correcting consumer mistakes that may produce ambiguity (eg, mistakenly writing "*three*" as "*tree*"). These problems can be resolved using the Google Spell Checker because it can intelligently conclude the most probable text candidate according to the sentence meaning. After correcting the whole text, the system will output and save the corrected text. Table 1 uses the sentence "I tooj tree pills last night before bad time" as an example, and explains how our system works on sentences. Each row shows how our system corrects a single word each time. The number in the second row, such as "-1" and "+3," shows the position of each word in this sentence. For example,

if we are presently focusing on the word "tooj," then "-1" corresponds to "I" and "+3" corresponds to "last." The column "Correction" shows the corrected results. The last column, "Error type," is manually classified, which will be discussed in the Error Classification section.

In the third step—text reconstruction—our system reconstructs full-text segments in accordance to their original order. Throughout the above three-step operation, our system will successfully correct the input postings and save both the uncorrected and corrected texts into our database.

Thus, we entered the consumer-generated postings collected from MedHelp and followed the steps above. After the correction process, we obtained both misspelled and corresponding corrected words, *respectively* saved into the uncorrected (U) *set and the corrected (C) set*. For example, after processing the sentence "I tooj tree pills last night before bad time," "tooj," "tree," and "bad" will be saved in set U and "took," "three," and "bed" in set C.

Table 1. The spelling correction process.

Step number	Word position[a]			Misspelled word	Word position[a]			Correction	Error type
	-3	-2	-1		+1	+2	+3		
1			I	tooj	tree	pills	last	took	typographical
2		I	tooj	tree	pills	last	night	three	linguistic
3	last	night	before	bad	time			bed	linguistic

[a]The number represents the position of each word in the sentence relative to the word presently being focused on.

Evaluation Process

During the evaluation of our system's correction quality, first we used the NCBO Annotator Web service to decrease the number of words examined manually; we input corrected words from set C into the NCBO Annotator, selected the RxNorm and SNOMED CT ontologies, and then ran the Annotator search. The reason we selected these two ontologies was because the former, RxNorm, contains all of the terminologies of drugs available on the US market [23], and the latter, SNOMED CT, contains a collection of clinical terms and is recognized as the most comprehensive health care terminology resource in the world [24]. After the data had been completely scanned and processed, the NCBO Annotator presented the words which could be matched in RxNorm and SNOMED CT in a downloadable Web page (see Figure 2). We then downloaded and saved the matched words. Similarly, we also input uncorrected words from set U, acquired the words which could be matched in RxNorm and SNOMED CT, and then saved them into our database (see Figure 3). After this preprocessing, instead of examining all the words in set C and set U, we could only manually examine the matched words, count the number of words which were appropriately corrected, and then calculate the ratio of these corrections.

Figure 2. Screenshot of the NCBO Annotator presenting words before spelling correction.

Figure 3. Screenshot of the NCBO Annotator presenting words after spelling correction.

Error Classification

We manually classified the results into two sets of errors—typographical errors (set T) and linguistic errors (set L)—and invited two clinical doctors and a medical researcher to confirm the correctness of our classification.

Abbreviation Counts

In accordance with the definition in Wong and Glance [1], abbreviations in this study refer to shortened forms of words, including acronyms, initialisms, and so on. Following this definition, we manually counted the number of abbreviations in the postings.

Results

Our spelling correction system detected 271 spelling errors in the selected postings (see Multimedia Appendix 2). For ease of explanation, we called an uncorrected word and its corresponding corrected word a spelling pair, and the two words

XSL•FO
RenderX

in the spelling pair are its members. For example, "tooj" and "took" compose, and are the members of, the spelling pair. A total of 271 spelling pairs were detected, among which we found that 58 (21.4%) spelling pairs contained one or two matched members in the selected ontologies—a member able to be matched in the ontologies is called a matched word, and its pair is called a matched pair (see Multimedia Appendix 2). We ascribed the 58 matched pairs into two groups—positive and negative impact—to evaluate the accuracy of our system.

Positive refers to misspelled words corrected appropriately and negative refers to those corrected inappropriately. The two impacts contain several different situations and their definitions are shown in Table 2.

In the correction process, we respectively recorded the number of situations defined above through manual inspection (see Table 3) and we asked three senior medical professionals—two clinical doctors and a medical researcher—to verify the correctness our classification.

Table 2. Definition for positive and negative impacts.

Impact	Situation	Definition
Positive		
	New match identified	Words cannot be found in the ontology before correction, but can be found after correction, and the corrected word is suitable in context.
	Wrong match identified	Words can be found in the ontology before correction, and cannot be found after correction, but the uncorrected word is unsuitable in context.
	Better match identified	Both words before and after correction can be found in the ontology and the corrected word is more suitable in context.
Negative	Right match missed	Either or both words before and after correction can be found in the ontology, but the corrected word is inappropriate in context.

Table 3. Results of spelling correction experiment (n=58).

Impact	Example	Effect	Ontology, n (%)	Representative letter
New match identified/ positive	"converts to seretonin" → "converts to serotonin"	A match of "serotonin" is found	37 (64)	A
Wrong match identified/ positive	"I took tree pills" → "I took three pills"	The improper match of "tree" is avoided	8 (14)	B
Better match identified/ positive	"last night before bad time" → "last night before bedtime"	A better match of "bedtime" replaces "bad time"	5 (9)	C
Right match missed/ negative	"I'm no chemist" → "I'm no chemistry"	A wanted match of "chemist" disappears	8 (14)	D
Total			58 (100)	F

The first column gives the situations that we defined in Table 2, including detailed types and their impact. The second column shows one example for each situation about how our system corrects the spelling errors. The third column explains the effect of the correction process on the sentences. The fourth column presents the number of matched pairs that conform to the corresponding situation. The letters in the fifth column represent the corresponding number in the fourth column (ie, A=37, B=8, C=5, D=8, and F=58).

From the results, it shows that 64% (A divided by F, 37/58) of the words could not be found in the ontology before correction, could be found in the ontology after correction, and the corrected words were suitable in the context by the situation definition.

Similar to Wong and Glance [1], we calculated the following expression to explain the performance of this system:

Accuracy = (A+B+C)/F

The accuracy—the ratio of misspelled words appropriately corrected in the 58 spelling pairs—was 86% (50/58). Also, we calculated the ratio of appropriate correction in the 271 overall spelling errors to be 85.2% (231/271). We also did a series of random sampling experiments; we randomly sampled 58 spelling pairs each time from the 271 spelling pairs. The trends of mean value and standard deviation are shown in Figure 4. The figure shows that, as the number of experiments increased, both the trends of mean value and standard deviation gradually became stable, respectively approaching 85.3% and 0.047.

After the classification according to the types of spelling errors, from a total of 271 errors our system detected 186 (68.6%) typographical errors (saved in set T) and 85 (31.4%) linguistic errors (saved in set L). In addition, there were a total of 210

abbreviations, making up 0.98% of all words in the postings (n=21,358).

Figure 4. Trends of mean value and standard deviation with change in sample size.
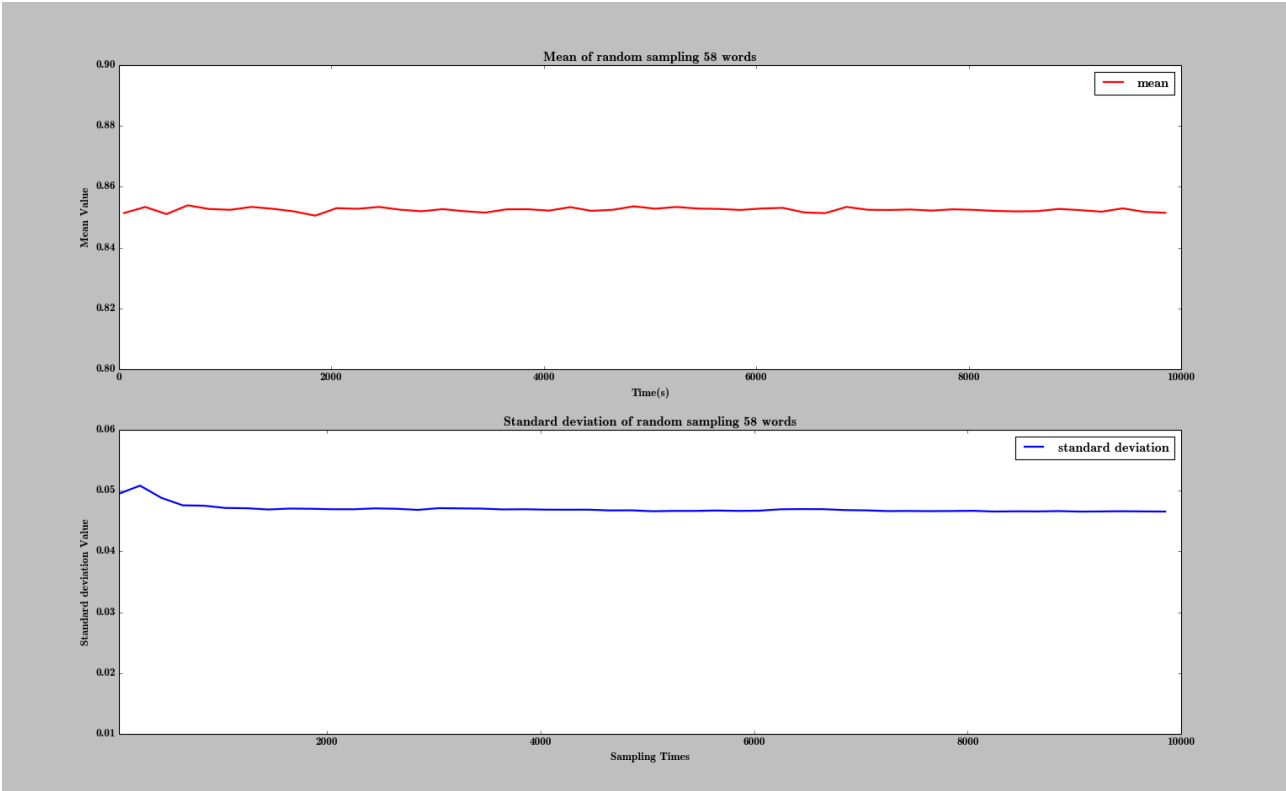
Discussion

System Performance

From the correction results, we found that 64% (37/58) of the matched words were newly found, which proved that our correction process exerted a positive effect on increasing the accuracy of downstream biomedical research, such as NLP research. Using the same corpus, our system's accuracy (50/58, 86%) was higher than that of most of the commonly used spelling checkers, including medical dictionary-based Aspell [25], Microsoft Office Word 2013, and Jazzy Spell Checker [26]. The result is shown in Table 4 and the detailed data are included in Multimedia Appendix 2. This illustrates that our spelling correction system is a suitable and high-performance tool for consumer-generated content.

Table 4. Comparison of spell checking tools for finding correct words for misspelled words.

Spell checking tool	Correct words found, n/n (%)
Our method	50/58 (86)
Aspell [25] with general dictionary	304/763 (39.8)
Aspell [25] with medical dictionary	353/564 (62.6)
Microsoft Office Word 2013	313/431 (72.6)
Jazzy Spell Checker [26]	240/574 (41.8)

It is noteworthy that, compared with other spelling checkers which usually provide several spelling suggestions to choose from for a spelling error, our method is more convenient and can directly provide the optimal candidate according to its context. In addition, in contrast to traditional spelling checkers such as Aspell, the corpus in our method does not need manual updates due to its Web page-based corpus. These characteristics are highly meaningful, especially for the automatic spell checking of big data.

Moreover, unlike the method applied by Ruch et al [7] in which spelling errors were artificially added into spelling error-free texts, our system obtained the original text directly from a health forum, which more objectively reflected the real situation of consumer-generated content.

Classification

From the classification results of spelling error types, it can be observed that errors in set L took up 31.4% of all spelling errors, which shows that correcting linguistic errors is indispensable during processing consumer-generated content. The systems that only focused on the correction of typographical errors, such as that of Peterson [15], ignore a large number of the spelling errors.

The Number of Abbreviations

In EMRs, the ratio of abbreviations is 33.6% [1], much higher than the ratio in consumer-generated content (0.98%). Therefore, detecting and correcting abbreviations in consumer-generated content appears to be much less important than in EMRs. Instead, from the results of classifying 271 spelling errors according to the meaning of corrected words (see Table 5), the correction systems for consumer-generated content should focus more on common vocabularies.

In Table 5, common vocabulary refers to those words people frequently use in daily life (eg, "good," "hadn't," and "loose"). Medical vocabulary refers to words that cannot be defined as a symptom, drug, or disease but are still used in the medical field such as "hygiene." Extra space refers to situations in which consumers enter extra spaces between words (eg, "weight__is" where there are two spaces between "weight" and "is").

Table 5. Classification of misspelled words (n=271).

Type of word, or issue	Number of words, n (%)
Common vocabulary	151 (55.7)
Symptom	8 (3.0)
Drug	12 (4.4)
Medical vocabulary	14 (5.2)
Disease	1 (0.4)
Extra space	85 (31.4)

Evaluation

During the evaluation process, only 58 words matched in the ontology, which was only about one-fifth (21.4%) of the number originally needed to process (n=271) and largely reduced the manual inspection time. This is the reason why we put forward the idea of using the NCBO Annotator to pick out the words related to the biomedical fields.

The ratio of misspelled words, which were appropriately corrected in the 58 matched pairs, was close to the overall spelling errors (the difference was 0.97%, less than 1%), and the accuracy (50/58, 86%) fell within the reliable range, within one standard deviation from the mean value of 85.2% (range 80.6% to 89.9%). For these reasons, the NCBO Annotator can well represent the overall performance of our system.

Future Work

In future work, different types of ontologies are needed to test and verify whether our evaluation method can be applied in other fields. Moreover, we will add and mix more correction tools in addition to Google Spell Checker to promote the overall performance of our spelling correction system.

Conclusions

From this study, the following can be confirmed:

1. Our system is suitable for spelling correction in consumer-generated content. The unique features in consumer-generated content have been identified and taken into consideration. Google Spell Checker displays high performance in spelling error detection and correction in consumer-generated content.

2. Context sensitivity is indispensable in the correction process.

3. Our evaluation method, taking advantage of biomedical ontology, can effectively evaluate the correction system and reduce manual inspection time on a large scale.

4. In consumer-generated content, consumers rarely use abbreviations, unlike in EMRs.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Original data: 150 postings.

[XLSX File (Microsoft Excel File), 103KB - medinform_v3i3e27_app1.xlsx]

Multimedia Appendix 2

Result file.

XSL•FO
RenderX

[PDF File (Adobe PDF File), 338KB - medinform_v3i3e27_app2.pdf]

References

1. Wong W, Glance D. Statistical semantic and clinician confidence analysis for correcting abbreviations and spelling errors in clinical progress notes. Artif Intell Med 2011 Nov;53(3):171-180. [doi: 10.1016/j.artmed.2011.08.003] [Medline: 21924593]
2. National Research Council (US) Committee on A Framework for Developing a New Taxonomy of Disease. Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease. Washington, DC: National Academies Press (US); 2011.
3. Zeng Q, Kogan S, Ash N, Greenes RA. Patient and clinician vocabulary: how different are they? Stud Health Technol Inform 2001;84(Pt 1):399-403. [Medline: 11604772]
4. Ristad E, Yianilos P. Learning string-edit distance. IEEE Trans Pattern Anal Mach Intell 1998 May;20(5):522-532. [doi: 10.1109/34.682181]
5. Levenshtein VI. Binary codes capable of correcting deletions, insertions and reversals. Soviet Physics Doklady 1966 Feb;10:707.
6. Courtin J, Dujardin D, Kowarski I, Genthial D, De Lima VL. Towards a complete detection/correction system. In: Proceedings of the International Conference on Current Issues in Computational Linguistics. 1991 Presented at: International Conference on Current Issues in Computational Linguistics; 1991; Penang, Malaysia p. 158-173.
7. Ruch P, Baud R, Geissbühler A. Using lexical disambiguation and named-entity recognition to improve spelling correction in the electronic patient record. Artif Intell Med 2003 Oct;29(1-2):169-184. [Medline: 12957786]
8. Russell RC. United States Patent and Trademark Office. Washington, DC: United States Patent Office; 1922 Nov 14. Patent number: US001435663 URL: https://www.google.com.hk/patents/US1435663?dq=1,435,663&hl=en&sa=X&ved=0CB0Q6AEwAGoVChMI78js_4GAxwIVZOCmCh3IYADK [accessed 2015-07-19] [WebCite Cache ID 6a99fFxoF]
9. Russell RC. United States Patent and Trademark Office. Washington, DC: United States Patent Office; 1918 Apr 02. Patent number: US001261167 URL: https://www.google.com.hk/patents/US1261167?dq=1,261,167&hl=en&sa=X&ved=0CB0Q6AEwAGoVChMIh-mFtIKAxwIVZCemCh3n2wUx [accessed 2015-07-17] [WebCite Cache ID 6a5p0MoqO]
10. Golding A, Schabes Y. Combining trigram-based and feature-based methods for context-sensitive spelling correction. In: Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics; 1996 Presented at: 34th Annual Meeting of the Association for Computational Linguistics; June 24-27, 1996; Santa Cruz, CA p. 71-78 URL: https://aclweb.org/anthology/P/P96/P96-1000.pdf
11. Stevenson M, Guo Y, Gaizauskas R, Martinez D. Disambiguation of biomedical text using diverse sources of information. BMC Bioinformatics 2008;9 Suppl 11:S7 [FREE Full text] [doi: 10.1186/1471-2105-9-S11-S7] [Medline: 19025693]
12. Crowell J, Zeng Q, Ngo L, Lacroix E. A frequency-based technique to improve the spelling suggestion rank in medical queries. J Am Med Inform Assoc 2004 Jun;11(3):179-185 [FREE Full text] [doi: 10.1197/jamia.M1474] [Medline: 14764616]
13. Jurafsky D, Martin JH. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition. Upper Saddle River, NJ: Pearson Education; Jan 26, 2000.
14. Wilbur WJ, Kim W, Xie N. Spelling correction in the PubMed search engine. Inf Retr Boston 2006 Nov;9(5):543-564 [FREE Full text] [doi: 10.1007/s10791-006-9002-8] [Medline: 18080004]
15. Peterson JL. Computer programs for detecting and correcting spelling errors. Commun ACM 1980 Dec;23(12):676-687.
16. Jacquemont S, Jacquenet F, Sebban M. Correct your text with Google. In: Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence. Washington, DC: IEEE; 2007 Presented at: IEEE/WIC/ACM International Conference on Web Intelligence; November 2-5, 2007; Fremont, CA p. 170-176.
17. Doan S, Bastarache L, Klimkowski S, Denny JC, Xu H. Integrating existing natural language processing tools for medication extraction from discharge summaries. J Am Med Inform Assoc 2010 Oct;17(5):528-531 [FREE Full text] [doi: 10.1136/jamia.2010.003855] [Medline: 20819857]
18. Patrick J, Li M. High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge. J Am Med Inform Assoc 2010 Oct;17(5):524-527 [FREE Full text] [doi: 10.1136/jamia.2010.003939] [Medline: 20819856]
19. MedHelp. Zoloft URL: http://www.medhelp.org/tags/show/7661/Zoloft?section=subjects [accessed 2015-07-16] [WebCite Cache ID 6a3pOlVQn]
20. BioPortal. Annotator URL: http://bioportal.bioontology.org/annotator [accessed 2015-07-16] [WebCite Cache ID 6a3pQXekb]
21. Xu W, Wang H, Cheng W, Fu D, Xia T, Kibbe WA, et al. A framework for annotating human genome in disease context. PLoS One 2012;7(12):e49686 [FREE Full text] [doi: 10.1371/journal.pone.0049686] [Medline: 23251346]
22. Nirmaldasan. Readability Monitor. 2008 Jul 28. The average sentence length URL: https://strainindex.wordpress.com/2008/07/28/the-average-sentence-length/ [accessed 2015-07-16] [WebCite Cache ID 6a3pM1oPP]
23. US National Library of Medicine. RxNorm URL: http://www.nlm.nih.gov/research/umls/rxnorm/ [accessed 2015-07-16] [WebCite Cache ID 6a3pRpbXj]

XSL•FO
RenderX

24.   International Health Terminology Standards Development Organisation. SNOMED CT URL: http://www.ihtsdo.org/snomed-ct/ [accessed 2015-05-06] [WebCite Cache ID 6YJiIBBek]
25.   Aspell.net. GNU Aspell URL: http://aspell.net/ [accessed 2015-05-06] [WebCite Cache ID 6YJhUb44E]
26.   sourceforge.net. Jazzy - Java spell check API URL: http://jazzy.sourceforge.net/ [accessed 2015-05-06] [WebCite Cache ID 6YJiDnQAK]

Abbreviations

API: application programming interface
BBS: bulletin board system
C: corrected
EMR: electronic medical record
gtt: guttae
L: linguistic error
NCBO: National Center for Biomedical Ontology
NLP: natural language processing
NTG: nitroglycerin
PVCs: premature ventricular contractions
SNOMED CT: Systematized Nomenclature of Medicine -- Clinical Terms
T: typographical error
U: uncorrected
VSS: vital signs stable

XSL•FO
RenderX

Original Paper

Using MEDLINE Elemental Similarity to Assist in the Article Screening Process for Systematic Reviews

Xiaonan Ji[1], MS; Po-Yin Yen[1], RN, PhD

The Ohio State University, Department of Biomedical Informatics, Columbus, OH, United States

Corresponding Author:
Po-Yin Yen, RN, PhD
The Ohio State University
Department of Biomedical Informatics
1800 Canon Drive
Columbus, OH,
United States
Phone: 1 503 889 6181
Email: po-yin.yen@osumc.edu

Abstract

Background: Systematic reviews and their implementation in practice provide high quality evidence for clinical practice but are both time and labor intensive due to the large number of articles. Automatic text classification has proven to be instrumental in identifying relevant articles for systematic reviews. Existing approaches use machine learning model training to generate classification algorithms for the article screening process but have limitations.

Objective: We applied a network approach to assist in the article screening process for systematic reviews using predetermined article relationships (similarity). The article similarity metric is calculated using the MEDLINE elements title (TI), abstract (AB), medical subject heading (MH), author (AU), and publication type (PT). We used an article network to illustrate the concept of article relationships. Using the concept, each article can be modeled as a node in the network and the relationship between 2 articles is modeled as an edge connecting them. The purpose of our study was to use the article relationship to facilitate an interactive article recommendation process.

Methods: We used 15 completed systematic reviews produced by the Drug Effectiveness Review Project and demonstrated the use of article networks to assist article recommendation. We evaluated the predictive performance of MEDLINE elements and compared our approach with existing machine learning model training approaches. The performance was measured by work saved over sampling at 95% recall (WSS95) and the F-measure (F_1). We also used repeated analysis over variance and Hommel's multiple comparison adjustment to demonstrate statistical evidence.

Results: We found that although there is no significant difference across elements (except AU), TI and AB have better predictive capability in general. Collaborative elements bring performance improvement in both F_1 and WSS95. With our approach, a simple combination of TI+AB+PT could achieve a WSS95 performance of 37%, which is competitive to traditional machine learning model training approaches (23%-41% WSS95).

Conclusions: We demonstrated a new approach to assist in labor intensive systematic reviews. Predictive ability of different elements (both single and composited) was explored. Without using model training approaches, we established a generalizable method that can achieve a competitive performance.

Introduction

Systematic reviews provide summaries of evidence from high quality studies to answer specific research questions. They are regularly used in health care [1-4] and for health policy making [5]. Evidence-based medicine (EBM) relies heavily on the use of synthesized, up-to-date research evidence to make decisions. Systematic reviews are considered the highest quality source of evidence for EBM [6]. However, systematic reviews require a series of very resource and time intensive steps [4] that

typically require several months to complete [7]. Such workload and resource challenges can limit the tractability of an individual review, the ability to fund a review, and also the ability to respond to new evidence that may require an update to an existing review.

MEDLINE is a biomedical literature database that stores and indexes a variety of relevant publications and is a primary resource for identifying studies for systematic reviews targeting the health sciences. However, the size of MEDLINE increases at a rate of over 12,000 articles per week, including reports related to over 300 randomized trials [8]. The Cochrane Collaboration is an international organization dedicated to producing up-to-date systematic reviews, with more than 15,000 people participating in the work [9]. According to The Cochrane Collaboration, more than 10,000 systematic reviews are needed for existing effectiveness research [9]. In addition, a recent study reported that 23% of reviews require updates within 2 years [10]. With the need to conduct a large amount of original and updated systematic reviews, it is essential to improve the efficiency of producing systematic reviews and their incumbent synthesized knowledge.

A systematic review is commonly conducted by domain experts who are able to draft systematic review scopes, retrieve relevant citations, assess study quality, and synthesize evidence. The process can be broken down into 15 steps [11]. Expert reviewers first identify the systematic review scope and research questions, and then generate search strategies to explore related databases (eg, MEDLINE). The search result is a list of citations usually organized in reference management software (eg, EndNote, RefWorks). Before synthesizing relevant evidence, expert reviewers need to classify articles based on the title and abstract. Then through the article screening (or article selection) process, relevant articles will proceed to the full-text level. In most systematic reviews, expert reviewers include a small portion ranging from 2% to 30% of citations at the title and abstract level; 1.6% to 27% get included at the full-text level [7]. In other words, expert reviewers spend most of their effort excluding non-relevant or low quality studies. To accelerate this process, several machine learning approaches (ie, naive Bayes and support vector machine) [7,12-15] were proposed to focus on facilitating and enhancing the title and abstract level triage, abstracts screening [11], or article screening, which is crucial and time-consuming as it requires expert reviewers to screen a large amount of literature. The intelligent article selection process can be also called citation classification or citation screening.

In this paper, we proposed to use established and predetermined article relationships and incorporate the concept of active machine learning to iteratively recommend articles and receive feedback from human reviewers. Although the idea of integrating human judgment sounds similar to the active learning approach implemented in Wallace's work [13,14], our approach uses a different strategy. We do not formulate a classification model. Instead, we generate an article network representing the relationships between articles. We use the articles classified by human reviewers as a reference set to recommend the next similar article. There is no model trained during the recommendation procedure. The approach is similar to relevance feedback, a feature in some information retrieval systems. In general, users classify documents as relevant or irrelevant and provide the feedback to the information retrieval system. The information retrieval system then uses this information to retrieve documents similar to the relevant documents. Relevance feedback is commonly used as an automatic technique for queries modification. The process of relevance feedback is executed as a cycle of activity that refines queries in each iteration of feedback collection [16,17].

The predetermined relationships between articles can be conceptualized as an article network, which is different from the traditional citation network. A traditional citation network uses the *citing* and *cited by* of an article to build the network [18]. We build article networks based on the similarity of any paired articles. Our similarity metric is calculated using data elements [19] from an article, such as title, abstract, medical subject heading, author, and publication type. Under this concept, each article is modeled as a node in the network and the relationship (similarity) between two articles is modeled as an edge connecting them. Although the network method is not novel in the general document clustering area, we are the first to use the approach to facilitate systematic reviews and demonstrate its strength. Figure 1 shows an illustrated network of a real systematic review (Urinary Incontinence) displayed in an aesthetically pleasing force-directed graphic layout. Theoretically, the network should be a complete graph in which every pair of articles has an edge representing the similarity between them. For visualization purposes, we eliminated the edges with lower similarity scores to provide a more human readable network.

During our preliminary experiments, we found that a similarity score composed of all MEDLINE elements does not work well for every systematic review. We suspected that some elements (eg, title, abstract, publication type, MeSH, author) are better predictors for recommendations than others. Therefore, the purpose of our study was to answer two research questions. When an article is classified as included, what element(s) are better to use to calculate the similarity score to predict the next relevant article? Since every element plays a different role and should be weighted accordingly, what combinations and weights of elements are better to predict the next relevant article?

Figure 1. Illustrated article network.
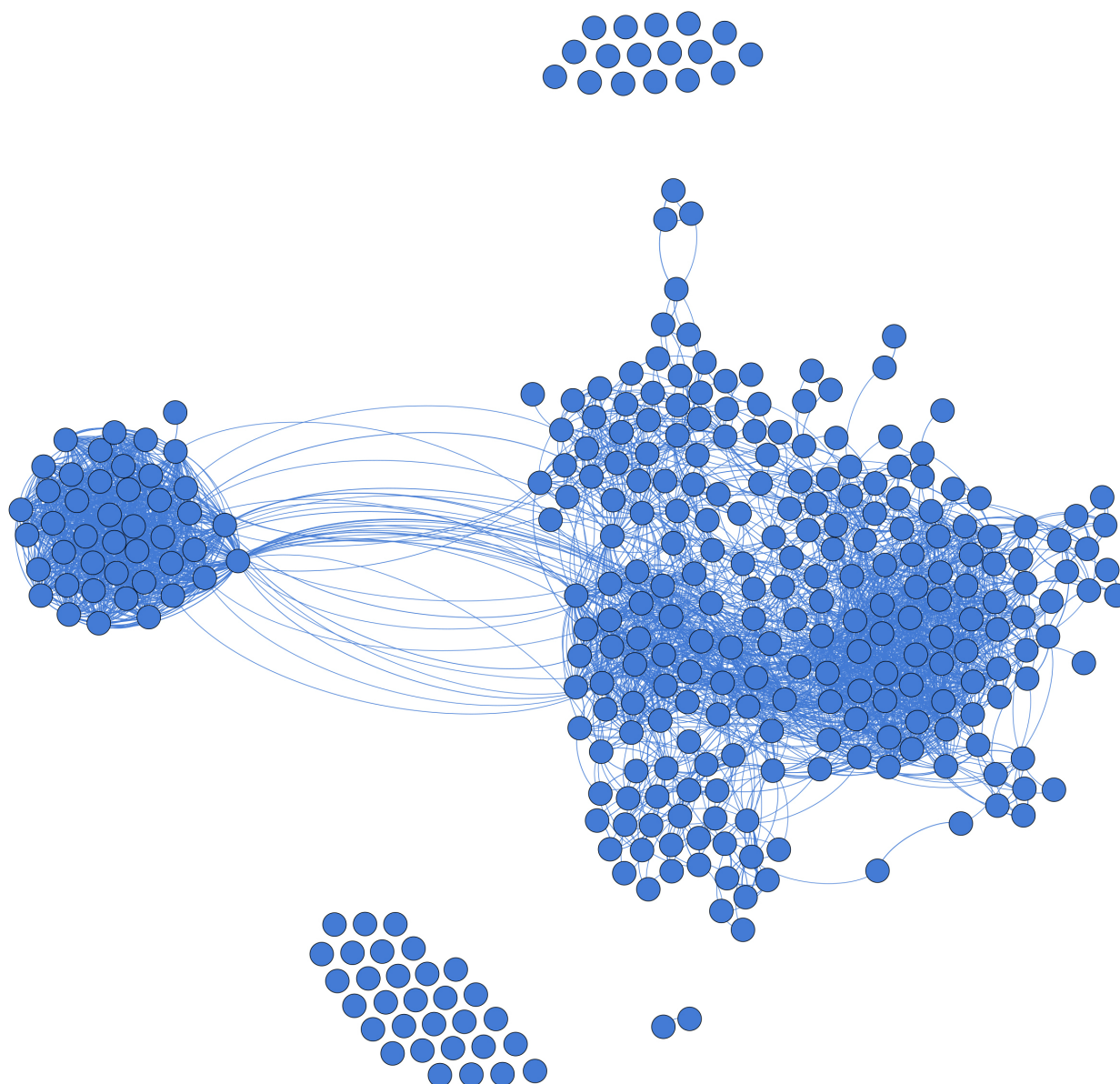
Methods

Data Source

To evaluate our approach, we used 15 publicly available completed systematic review samples produced by the Drug Effectiveness Review Project (DERP) (coordinated by the Center for Evidence-Based Policy at Oregon Health and Science University) [20]. These 15 systematic reviews were completed by experienced and knowledgeable human expert reviewers, with inclusion and exclusion decisions made by at least two expert reviewers. Table 1 shows the number and percentage of articles included at abstract level decision and full-text level decision.

For instance, the review for ACE Inhibitors has a total of 2544 citations. Based on the abstracts, 183 (7.19%) were included; after full-text reading, 41 (1.61%) were included in the ACE Inhibitor systematic review. The final inclusion rates range from 0.78% to 27.04%. The 15 systematic reviews are also the same test collection previously used and made publicly available by Cohen et al [7]. Using the PubMed Identifier (PMID), we downloaded the full record in MEDLINE format [19] and extracted the data elements title, abstract, publication types, author and medical subject heading (MeSH) as the input.

Table 1. Total article numbers and rates of inclusion.

SR report topic	Total	Abstract n (%)	Full text n (%)
ACE inhibitors	2544	183 (7.19)	41(1.61)
ADHD	851	84 (9.87)	20 (2.35)
Antihistamines	310	92 (29.68)	16 (5.16)
Atypical antipsychotics	1120	363 (32.41)	146 (13.04)
Beta blockers	2072	302 (14.58)	42 (2.03)
Calcium channel blockers	1218	279 (22.91)	100 (8.21)
Estrogens	368	80 (21.74)	80 (21.74)
NSAIDS	393	88 (22.39)	41 (10.43)
Opioids	1915	48 (2.51)	15 (0.78)
Oral hypoglycemics	503	139 (27.63)	136 (27.04)
Proton pump inhibitors	1333	238 (17.85)	51 (3.83)
Skeletal muscle relaxants	1643	34 (2.07)	9 (0.55)
Statins	3465	173 (4.99)	85 (2.45)
Triptans	671	218 (32.49)	24 (3.58)
Urinary incontinence	327	78 (23.85)	40 (12.23)

MEDLINE Elements

MEDLINE elements are the fields in the MEDLINE format that document the major pieces of information of a publication (article) [19]. The MEDLINE display format is used in PubMed MEDLINE records. As the most informative elements, title (TI), abstract (AB) and MeSH (MH) elements are widely used in related work to build feature spaces for machine learning algorithms. Publication type (PT) is also selected by some studies [15,21] as it could be a key factor for inclusion or exclusion decisions. In our preliminary work, we found that author information also has some predictive value in the article screening process. Therefore, in this study, in addition to TI, AB, MH, and PT, we also included the author (AU) element in our experiments.

Similarity Score

We calculate the similarity score using cosine similarity [22]. Cosine similarity is widely applied to text mining and measures the cosine of the angle between a pair of vectors. It reflects the degree of similarity based on the presence and frequency of words or terms in each text. For every pair of AUs, PTs, and MHs, we simply compared them by exact string matching, because a minor difference may completely alter the outcome. For example, even if two author names are very similar, they may be two different people. However, TI and AB are free text. To calculate the similarity between two TIs and between two ABs, we preprocessed TIs and ABs by removing some common words from the PubMed stop word list [23] (eg, the, is, are) that appear frequently in text, stemming each word by the classic Porter Stemmer algorithm [24]. This approach, named alphabetic features, also has been verified to be an effective method to represent an article [25]. The resulting similarity score ranges from 0 to 1 for each element, where 0 indicates independence

and 1 means exactly the same. In summary, the similarity score is the sum of the MEDLINE element similarities.

Simulated Interactive Recommendation Process

Overview

In this study, there is no human reviewer in this experimental process. The interactive recommendation process is simulated using the 15 completed DERP systematic reviews.

After identifying a list of articles to be screened for a systematic review, the recommendation process starts with calculating the similarity scores of any pairs of articles. This process constructs the relationship of the articles and builds a conceptualized article network. The first recommended article is selected based on key questions and search strategies of the systematic review. Once a recommended article is classified as included (IN) or excluded (EX), an IN list and an EX list are created (in this study, we used completed systematic reviews, which have predetermined decisions to simulate this step). We then iteratively recommend relevant articles based on the similarity to the IN. Assuming V is the set of all articles and U is the set of articles that have never been recommended, U is defined as U=V–IN–EX. Therefore, the sum of similarity scores represents the similarity between an article v with article(s) x in IN (see Figure 2).

Figure 2. Calculation of the similarity between articles.

$$ \text{Similarity} = \arg max_{v \in U} \left(\sum_{x \in IN} \overrightarrow{vx} \cdot \vec{w} \right) $$

U is the set of articles that have never been recommended.
v is one article from U.
IN is the set of articles that have been included.
x is one article from IN.
$\overrightarrow{vx}$ is a similarity vector representing the similarity of MEDLINE element(s) between x and v.
$\vec{w}$ is a weight parameter that controls the contribution of each element similarity vector in the overall similarity score.

XSL•FO
RenderX

In the formula, vx is a similarity vector representing the similarity of MEDLINE element(s) between x and v. The weight parameter w controls the contribution of each element similarity vector in the overall similarity score. We recommend articles with the highest overall similarity scores.

Figure 3 illustrates the simulated interactive recommendation process: (a) Process articles and extract data elements; (b) Calculate similarity scores (this will establish a conceptualized article network. Weight parameters are optional); (c) Recommend article(s) with the highest similarity to included articles list (in this simulation, one article is recommended per each round); (d) Human reviewers classify the recommended article as included or excluded (again, in this study, we used completed systematic review reports, which have predetermined decisions to simulate this step); (e) Create and update the included and excluded article list; Steps (c), (d), and (e) repeat until the article screening process is completed.

To evaluate our performance, we used two performance measures: work saved over sampling at 95% recall (WSS95) and F-measure. These measures are commonly used for evaluating similar work [7,12,15]. We also used repeated analysis of variance (ANOVA) and post hoc analysis with Hommel multiple comparison adjustment to further explore statistical evidences. Hommel's method demonstrated type I error protection with good power and is considered a better approach than Bonferroni or Homl [26,27].

Figure 3. Simulated interactive recommendation process.
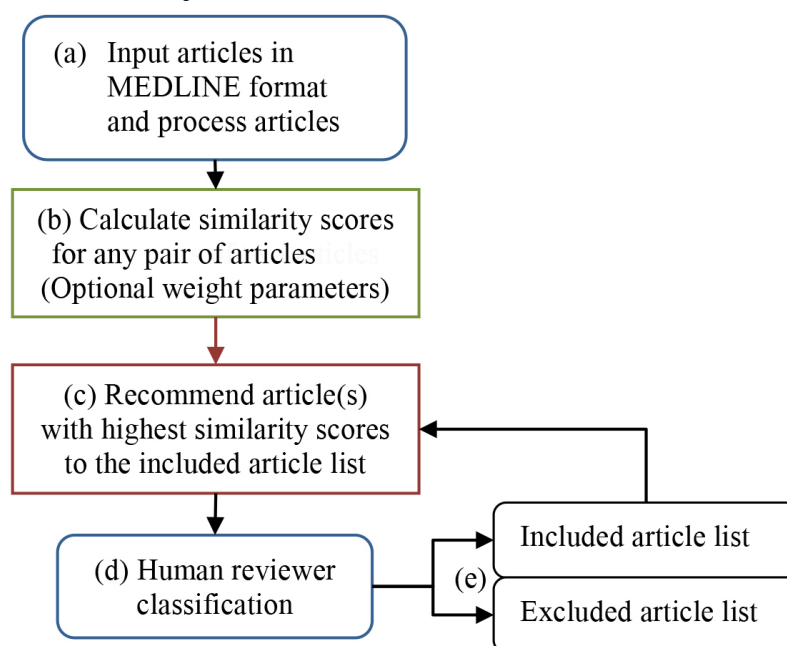
Work Saved Over Sampling at 95% Recall

WSS95 is a performance measure first proposed by Cohen [7] to calculate the overall labor saving while maintaining the recall at 95%. This assumes that a recall higher than 0.95 is necessary for a document classification system. Precision should be as high as possible, as long as recall is at least 0.95. WSS95 is calculated with the equations in Figure 4.

TP is the number of true positive (relevant) articles, TN is the number of true negatives (irrelevant) articles, FN is the number of false negative (relevant) articles, and N is the total number of articles in each report.

Figure 4. Formulas of precision, recall, and F_1.

$$precision = \frac{|\{relevant\ documents\} \cap \{retrieved\ documents\}|}{|\{retrieved\ documents\}|}$$

$$recall = \frac{|\{relevant\ documents\} \cap \{retrieved\ documents\}|}{|\{relevant\ documents\}|}$$

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

F-Measure

F-measure is a measure of information retrieval accuracy. It considers both precision and recall and commonly combines them into a weighted harmonic mean. When they are weighted equally, the balanced F-measure is also called F_1, where it reaches its best value at 1 and the worst value at 0. As a general measure of accuracy, F_1 has been widely used in previous works for the evaluation of classification performance, such as Cohen 2006 [7], Bekhuis 2010 [28], Kastrin 2010 [29], and Frunza 2010 [30]. For our performance evaluation purposes, when we recommend one article each time, the immediate recall, precision, and F_1 are dynamically changed each time (see Figure 5).

Since F_1 is dynamically changed over time, we can detect the highest F_1 from the steepness of the performance curve. That means if the higher F_1 scores occur during the early stage of the recommendation process (ie, before 50% of articles are screened), we are more likely to save more workload (high accuracy). We use F_1 to help us evaluate how accurate and how quickly we can make recommendations on the relevant articles.

XSL•FO

RenderX

Figure 5. Formulas of WSS and WSS95.

$$WSS = \frac{(TN + FN)}{N} - 1 + \frac{TP}{TP + FN}$$

$$WSS95 = \frac{(TN + FN)}{N} - 0.05$$

TP is the number of true positive (relevant) articles.
TN is the number of true negatives (irrelevant) articles.
FN is the number of false negative (relevant) articles.
N is the total number of articles in each report.

Results

Single Element Performance

The single MEDLINE element performance results are shown in [Table 2](). TI gets the best average WSS95 performance

(34.01%), followed by PT (33.41%), and AB (33.30%). MH has a much lower WSS95 than other elements (25.31%). AU receives 0% workload saved due to the dispersion among articles' authorship. If there is no authorship similarity between articles, we are not able to recommend relevant articles based solely on AU element. Using PT also brings good performance; we speculate it is a key consideration when conducting system reviews. However, repeated ANOVA shows that the WSS95 performances across TI, AB, PT, and MH are not statistically different ($P=.079$).

Table 2. Single element WSS95 performance.

SR report topic	TI	AB	PT	AU	MH
ACE inhibitors	76.49	71.07	33.22	0	47.37
ADHD	80.26	65.10	22.56	0	47.00
Antihistamines	13.55	15.81	32.58	0	2.58
Atypical antipsychotics	17.23	20.54	19.64	0	9.46
Beta blockers	44.74	49.95	43.77	0	28.67
Calcium channel blockers	19.38	16.34	18.64	0	20.94
Estrogens	29.35	29.08	17.93	0	38.59
NSAIDS	63.36	66.67	58.27	0	33.84
Opioids	8.30	9.82	37.23	0	6.48
Oral hypoglycemics	11.73	12.13	22.27	0	7.55
Proton pump inhibitors	43.74	15.60	35.48	0	20.56
Skeletal muscle relaxants	0	36.03	74.68	0	42.85
Statins	25.52	30.17	13.31	0	13.68
Triptans	45.60	42.47	28.17	0	33.23
Urinary incontinence	30.89	18.65	43.43	0	26.91
Average WSS95	34.01	33.30	33.41	0	25.31

[Table 3]() shows the highest F_1 performance and corresponding timing during the recommendation process. When performance is good, the highest F_1 usually occurs during the early stage (discussed in the Methods section). We found that AB and PT gain the best F_1 (0.3683 and 0.3437, respectively); MH and TI have lower F_1 scores (0.3116 and 0.3039, respectively). Again, AU gets the worst F_1, only 0.1365. We also examined the corresponding timing of the highest F_1. We observed that the best F_1 value appears when 5% to 20% of articles are screened, which is at the early stage of recommendation. MEDLINE

elements with higher F_1 scores and lower percentages of articles screened indicate high accuracy performance during the early stage of recommendation (eg, AB). We concluded that AB and PT bring the best early stage performance; in other words, the recommendation accuracy of AB and PT in the beginning is better than that of the other elements. However, repeated ANOVA shows that the F_1 performances across TI, AB, PT, and MH are not statistically different ($P=.073$). Pairwise comparison only finds significant difference between TI and AB (AU is not considered due to its inferior performance).

XSL•FO
RenderX

Table 3. Single element F_1 performance; percentage of articles screened at F_1.

SR report topic	TI	AB	PT	AU	MH
	F_1 (%)	F_1 (%)	F_1 (%)	F_1 (%)	F_1 (%)
ACE inhibitors	0.3444 (4)	0.3121 (4)	0.2182 (<1)	0.1872 (6)	0.2368 (1)
ADHD	0.2885 (10)	0.3824 (6)	0.2963 (<1)	0.0909 (<1)	0.5556 (4)
Antihistamines	0.2593 (12)	0.4000 (3)	0.2759 (<1)	0.1111 (<1)	0.3333 (3)
Atypical antipsychotics	0.3447 (26)	0.4248 (14)	0.4363 (5, 12[a])	0.0135 (<1)	0.3113 (40)
Beta blockers	0.1972 (1)	0.2710 (5)	0.2105 (<1)	0.0417 (<1)	0.0957 (19)
Calcium channel blockers	0.2026 (10)	0.2672 (11)	0.2662 (15)	0.1261 (9)	0.2579 (2)
Estrogens	0.5140 (36)	0.5612 (29)	0.4937 (18)	0.0244 (<1)	0.5536 (39)
NSAIDS	0.4368 (34)	0.5870 (13)	0.6761 (8)	0.4853 (24)	0.3650 (24)
Opioids	0.2727 (<1)	0.1429 (<1)	0.2222 (<1)	0.1111 (<1)	0.2500 (<1)
Oral hypoglycemics	0.4509 (88)	0.4603 (76)	0.5019 (78)	0.0145 (<1)	0.4527 (53)
Proton pump inhibitors	0.3333 (1)	0.3860 (5)	0.1299 (42)	0.0377 (<1)	0.1775 (25)
Skeletal muscle relaxants	0.1429 (<1)	0.1981 (<1)	0.2286 (2)	0.1429 (<1)	0.2222 (<1)
Statins	0.2278 (6)	0.2479 (1)	0.4019 (4)	0.1484 (12)	0.1563 (1)
Triptans	0.1739 (10)	0.360 (4)	0.2569 (13)	0.0690 (<1)	0.2750 (8)
Urinary incontinence	0.3697 (24)	0.5243 (19)	0.5405 (10)	0.4444 (13)	0.4317 (30)
Average[b]	0.3039 (18)	0.3683 (13)	0.3437 (14)	0.1365 (5)	0.3116 (17)

[a] Both 5% and 12% have $F_1 = 0.4363$. The average of 5% and 12% (8.5%) is taken to calculate the average value on the last row of the table.

[b] <1% is considered as 1% for calculating the average percentage.

Composited Elements Performance

Different MEDLINE elements play different roles in the systematic review process, and their corresponding performance varied greatly as described above. To further explore their predictive abilities, we examined their collaborative performances. In total we examined 22 combinations and chose the top WSS95 performance of 6 combinations (see Table 4). Each of the 6 combination performances has an average of more than 36% WSS95. Table 5 shows the F_1 performance of the 6 combinations.

We also conducted statistical analysis with repeated ANOVA for the composited elements performance. For WSS95, the results show that there is no statistical difference in WSS95 performance across the 6 combinations ($P=.332$). For F_1 performance, there is also no statistical significant difference across the 6 combinations ($P=.069$).

In summary, we found that the predictive ability of MEDLINE elements varies according to systematic review topics. Overall, TI and PT have better WSS95 performance on average but are not statistically different. AB has the best average F_1 scores and is statistically better than TI.

XSL·FO

RenderX

Table 4. WSS95 of the top 6 combinations.

SR report topic	TI+AB	TI+AB +MH	TI+AB +PT	TI+AB +AU	TI+AB +PT+AU	TI+AB+MH +PT+AU
ACE inhibitors	76.38	76.85	74.29	75.79	73.70	75.08
ADHD	80.38	79.79	67.92	80.14	67.92	56.17
Antihistamines	16.13	10.65	24.52	16.13	24.52	18.39
Atypical antipsychotics	20.89	14.20	17.95	20.63	17.77	14.38
Beta blockers	60.14	60.09	65.01	60.96	64.72	65.21
Calcium channel blockers	18.23	18.64	17.32	18.39	17.49	22.82
Estrogens	33.42	36.14	22.55	33.97	22.55	29.08
NSAIDS	72.26	75.57	77.35	70.48	76.34	77.86
Opioids	6.01	11.75	8.98	5.95	8.98	12.17
Oral hypoglycemics	11.33	13.12	13.52	11.13	13.52	12.72
Proton pump inhibitors	19.20	21.31	19.65	19.05	19.65	20.11
Skeletal muscle relaxants	41.94	46.44	58.55	41.87	58.49	60.01
Statins	29.10	27.11	27.80	30.96	27.71	26.07
Triptans	48.29	51.71	39.64	50.52	39.79	40.98
Urinary incontinence	12.84	11.01	20.80	12.84	20.80	14.37
Average	36.44	36.96	37.06	36.59	36.93	36.35

Table 5. F_1 of the top 6 combinations.

SR report topic	TI+AB F_1 (%)	TI+AB+MH F_1 (%)	TI+AB+PT F_1 (%)	TI+AB+AU F_1 (%)	TI+AB+PT+AU F_1 (%)	TI+AB+MH+PT+AU F_1 (%)
ACE inhibitors	0.4156 (1)	0.4000 (2)	0.4051 (1)	0.3902 (2)	0.3971 (4)	0.3774 (3)
ADHD	0.4000 (3)	0.4688 (5)	0.5455 (4)	0.4286 (6)	0.5306 (3)	0.5818 (4)
Antihistamines	0.3226 (5)	0.3333 (10)	0.2903 (15)	0.3226 (5)	0.2903 (15)	0.2813 (15)
Atypical antipsychotics	0.4364 (16)	0.4241 (15)	0.4887 (15)	0.4411 (17)	0.4856 (15)	0.4606 (15)
Beta blockers	0.2800 (3)	0.3043 (2)	0.3590 (2)	0.2667 (3)	0.3596 (2)	0.3333 (3)
Calcium channel blockers	0.2335 (8)	0.2620 (11)	0.2804 (9)	0.2323 (8)	0.2816 (9)	0.2995 (9)
Estrogens	0.6000 (30)	0.6237 (29)	0.6047 (25)	0.5979 (31)	0.6118 (24)	0.6171 (26)
NSAIDS	0.6667 (16)	0.6154 (16)	0.6966 (12)	0.6471 (16)	0.6809 (13)	0.6667 (15)
Opioids	0.3000 (0)	0.3158 (<1)	0.3000 (<1)	0.3000 (<1)	0.3000 (<1)	0.3158 (<1)
Oral hypoglycemics	0.4497 (90)	0.4541 (88)	0.4553 (86)	0.4489 (92)	0.4561 (75)	0.4635 (82)
Proton pump inhibitors	0.4384 (7)	0.4737 (5)	0.5172 (5)	0.4552 (7)	0.5455 (5)	0.5079 (6)
Skeletal muscle relaxants	0.2222 (1)	0.2353 (<1)	0.2500 (<1)	0.2222 (1)	0.2500 (<1)	0.2667 (<1)
Statins	0.2994 (2)	0.3281 (1)	0.3382 (1)	0.2959 (2)	0.3358 (2)	0.3465 (1)
Triptans	0.3636 (3)	0.3913 (3)	0.3556 (3)	0.3556 (3)	0.3529 (4)	0.3913 (3)
Urinary incontinence	0.5063 (12)	0.5347 (19)	0.5505 (21)	0.5263 (11)	0.5507 (9)	0.5843 (15)
Average[a]	0.3956 (13)	0.4110 (14)	0.4291 (14)	0.3954 (14)	0.4286 (12)	0.4329 (13)

[a]<1% is considered as 1% for calculating the average percentage.

Performance Comparison With Existing Literature

Here we also compared our WSS95 performance with existing machine learning model training approaches (we were not able to compare the F_1 performances as they were not provided). Since TI+AB+PT has the simplest combination and its performance is equivalent or better than others, we chose TI+AB+PT (weight setting = 1:1:1) to compare against existing machine learning model training approaches, including voting perceptron-based automated citation classification system (VP), factorized complement naïve Bayes with weight engineering (FCNB/WE) and support vector machine (SVM) (Table 6).

Table 6. WSS95 comparison with the Cohen and Matwin systems across 15 SR topics.

SR report topic	Cohen 2006 [7] (VP[a])	Cohen 2008 [12] (SVM[b])	Matwin 2010 [15] (FCNB/WE[c])	Our study (TI+AB+PT)
ACE inhibitors	56.61	73.30	52.30	74.29
ADHD	67.95	52.60	62.20	67.92
Antihistamines	0	23.60	14.90	24.52
Atypical antipsychotics	14.11	17.00	20.60	17.95
Beta blockers	28.44	46.50	36.70	65.01
Calcium channel blockers	12.21	43.00	23.40	17.32
Estrogens	18.34	41.40	37.50	22.55
NSAIDS	49.67	67.20	52.80	77.35
Opioids	13.32	36.40	55.40	8.98
Oral hypoglycemics	8.96	13.60	8.50	13.52
Proton pump inhibitors	27.68	32.80	22.90	19.65
Skeletal muscle relaxants	0	37.40	26.50	58.55
Statins	24.71	49.10	31.50	27.80
Triptans	3.37	34.60	27.40	39.64
Urinary incontinence	26.14	43.20	29.60	20.80
Average	23.43	40.80	33.50	37.06

[a]VP: voting Perceptron-based automated citation classification system

[b]FCNB/WE: factorized complement naïve Bayes with weight engineering

[c]SVM: support vector machine

The repeated ANOVA test shows significant different across four studies (P=.005). The pairwise comparison with Hommel adjustment (Table 7) shows that there is no significant difference between our study and either Cohen 2008 [12] or Matwin 2010 [15] (P=.4979, .4979) but is significantly better than Cohen 2006 [7] (P=.0475). In summary, our methods provide competitive results to traditional machine learning model training approaches.

We were not able to compare side by side with the Wallace group [13,14] because they used different systematic reviews.

Their performance is by far the best among machine learning model training approaches (nearly 50% work reduction without missing any relevant articles) as they incorporate active learning with user interaction, which accepts feedback from users (similar to our Step D in Figure 3) [13,14]. This outcome is predictable as machine learning uses training data to model the classifier. With a large amount of training data, the classifier can perform almost perfectly. However, it is encouraging to us that without using algorithms to formulate a classification model, we are currently able to perform similarly to the model training approaches.

Table 7. The P values of pairwise comparison of four studies.

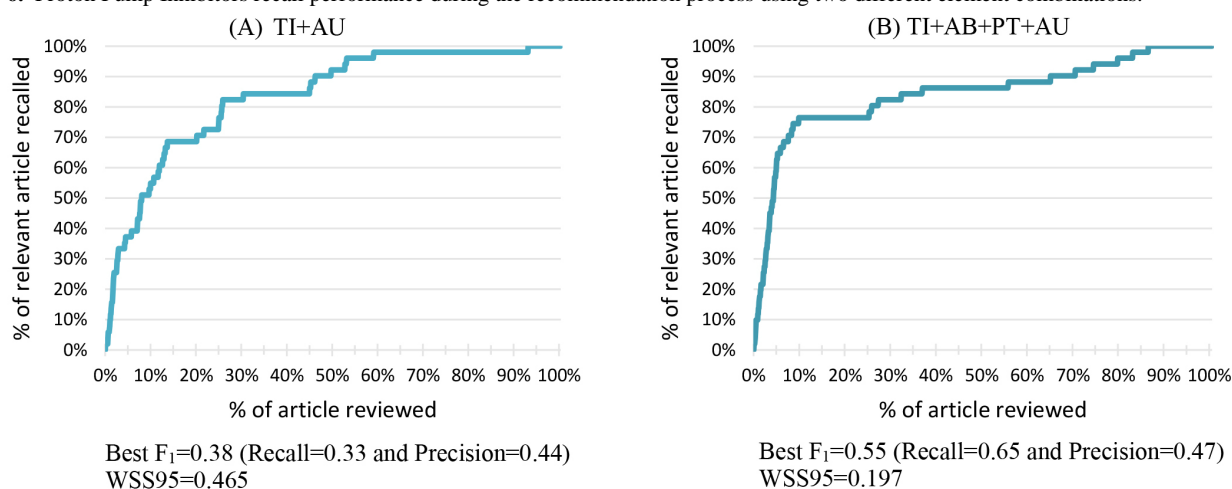|	Cohen 2006	Cohen 2008	Matwin 2010	Our study (TI+AB+PT)
Cohen 2006	—	0.0012	0.0433	0.0475
Cohen 2008	0.0012	—	0.0649	0.4979
Matwin 2010	0.0433	0.0649	—	0.4979
Our study	0.0475	0.4979	0.4979	—

Performance With Weight Parameters

Since different systematic reviews have diverse scopes (for example, one may require sufficient study information from an AB while another may have strict criteria on PT), we were interested in whether different weight parameters would alter the performance. We conducted experiments on different weight settings (eg, TI:PT:AU=3:1:2, TI:PT:AU=2:2:1, TI:PT:AU=3:2:1). The results revealed that when one element's weight was increased to achieve a higher performance for some reports, some other reports would have performance degradation. Overall, we could not find a universal weight setting that benefited all reports. This may be explained in part by the diverse scopes captured in different systematic reviews. In addition, although some weighted combinations bring better global performance (ie, average WSS95 among 15 reports), the enhancement from the baseline (elements in the combination are equally weighted) is limited. For example, consider the combination of TI+PT+AU, the baseline performance (TI:PT:AU=1:1:1) evaluated by average WSS95 is 35.45%, while the performance of its weighted one (TI:PT:AU=3:1:2) (37.30%) gains less than 2%. There is not much improvement with weighted parameters.

Interpreting the Inconsistency of WSS95 and F_1

During our experiments, we also discovered inconsistencies in performance with respect to WSS95 and F_1. For example, some combinations had high F_1 performance with low WSS95 and vice versa. We examined the recall performance during the entire recommendation process. Figure 6 presents the performance curves of the Proton Pump Inhibitors systematic review with two different element combinations, TI+AU (Figure 6A) and TI+AB+PT+AU (Figure 6B) (all equally weighted). The x-axis represents the percentage of articles screened (or recommended); the y-axis represents the recall rate. From the Figure 6, we see that in the early screening stage (5% of articles screened), curve B (recall of 70%) is steeper than curve A (recall of 40%). This also shows in their F_1 scores: the highest F_1 scores of curve A and B are 0.3778 and 0.5455, respectively, during the early screening stage. However, at the later stage, curve A reaches the recall of 100% faster than curve B after screening 60% of articles (WSS95 scores of curve A and B are 46.51% and 19.65%, respectively). In summary, current performance measures using WSS95, area under the curve, precision, and recall could not reflect the performance over time. Some elements may accelerate the performance in the beginning of the recommendation (screening) process. Using multiple performance measures and especially including the highest F_1 at a certain time point can better help us recognize the strength and weakness of different elements during the entire screening process

Figure 6. Proton Pump Inhibitors recall performance during the recommendation process using two different element combinations.

(A) TI+AU

Best F_1=0.38 (Recall=0.33 and Precision=0.44)
WSS95=0.465

(B) TI+AB+PT+AU

Best F_1=0.55 (Recall=0.65 and Precision=0.47)
WSS95=0.197

Discussion

Customizable Weight Parameters May Enhance Performance More Efficiently

Due to the fact that different systematic reviews have different review scopes, we could not identify one universal weight setting which could be successfully applied to every systematic review. A similar idea was mentioned in Matwin's work [15], where weight parameters (or weight multipliers) were tunable and being modified with regards to different systematic reviews. While different systematic reviews should have different weight multiplier values, we also agree that the process of computing such a value for every systematic review would be very time consuming [15]. Therefore, instead of finding the best weight parameters for each systematic review, flexible, customizable weight parameters for human reviewers based on their systematic review scopes and screening priority would be more useful and practical. Without adjusting weight parameters, our average performance is higher than the FCNB/WE approach [15] (Table 6). It is likely that we could improve even further when adjustable weight parameters are provided to human reviewers.

Moving Toward an Efficient and Generalizable Approach

Currently the work of biomedical text classification for the purpose of reducing systematic review workload has mainly used machine learning model training approaches. Naïve Bayes and SVM are two widely applied machine learning algorithms.

Although these machine learning approaches provide excellent performance in text classification on specific systematic review topics, it is a challenge to apply the existing machine learning algorithms to other new systematic review topics. It could be time consuming to construct training models as well [15]. In addition, the implementation of machine learning approaches usually requires an understanding of the algorithm. For example, operators need to choose a kernel or tune the setting of parameters for the SVM algorithm. Thus, it is difficult to apply the approach to a new systematic review topic without a well-trained classification model or without significant machine learning knowledge. Other approaches, such as text mining or statistical approaches, were also studied to facilitate the systematic review process [29,31], but they also rely heavily on prior decisions to find key terms to differentiate between the relevant and irrelevant classes, which is very similar to supervised machine learning.

Overcoming the limitations mentioned above, we provided a generalizable approach which can be easily deployed to facilitate any systematic review. Also, because we established an article network providing similarity relationship between articles, the iterative interactive recommendation process takes almost no time. Currently, our processing time to construct an article network takes from several seconds to several minutes for 300 to 3500 articles, but the recommendation step is real-time. This processing time is reasonable considering the non-trivial steps of building article networks. To be specific, this is polynomial time processing, not linear time processing. In our study, the backend programs for the computation of similarity matrixes are written in C/C++, which is the most efficient approach from the perspective of computer architecture and compiler. We also plan to improve the time responses for larger systematic reviews that may contain ten thousand articles or more. Most importantly, our approach can be applied to any systematic review topic and nontechnical human reviewers can use it with ease.

Study Limitation

This study only uses 15 DERP reports for evaluation. Although it is our assumption that our approach will be applicable globally, datasets from other systematic review teams are needed to further demonstrate our hypothesis. Our future plans include collaborating with other systematic review teams.

Future Direction

As we have discussed in the Methods section, different article elements have different predictive abilities regarding the evaluation scheme of WSS95 or F_1 score. With a better F_1 score and a lower WSS95, combinations containing AB or MH are more likely to elicit good performance in the beginning but have difficulty reaching 100% recall. On the other hand, although the combination of TI, PT, and AU can reach a better overall workload saved, the recall rises slowly (low accuracy) in the beginning of the recommendation process. This inspires us to utilize multiple types of weight settings and take advantage of different article element strengths during different recommendation phases (early-, mid-, and late-phases). We plan to implement automatic detection and adjustment when information from elements has been exhausted, which indicates the time to alter the combination of elements and weight parameters. For instance, when a series of N recommended articles is classified as excluded by human reviewers, we take it as a signal for adjustment as the current setting can no longer provide a good recommendation. Another example is to first apply the combination of AB and MH, as they provide high accuracy in the early recommendation stage, and then automatically adjust to the combination of TI, PT, and AU in the subsequent phase. Further research is also necessary to investigate proper adjustments of weight parameters under different conditions.

In the near future, we will also provide visualized article networks where relationships between articles could be intuitively represented and comprehended by humans. Network-based analysis will be conducted and network metrics like graph diameter, centrality, and module classes (by communication detection) will be reported. Such visualizations have the potential to enable the identification of clusters of articles and knowledge gaps in a targeted area. Lower density in such visualizations of the network could also indicate fewer related articles published or vice versa.

Conclusions

We demonstrated a new approach to assist the systematic review article screening process. We established article networks based on article similarity that facilitate the process of interactive article recommendation. We calculated article similarities using MEDLINE elements and examined the predictive ability of the MEDLINE element(s). We found that TI and PT have the best WSS95 performance, and AB and PT provide the best F_1 scores during the early stage of the recommendation process. However, no statistical difference was found.

Using our approach, we are able to achieve an average of 37% WSS95 with equally weighted combination of TI, AB, and PT. The statistical analysis also demonstrated that it is competitive with existing approaches. Based on findings and lessons learned from this study, we are currently deploying the approach into a prototype public online system, ArticleNet, to assist the article screening process.

XSL•FO

RenderX

Conflicts of Interest

None declared.

References

1. Yuan Y, Hunt RH. Systematic reviews: the good, the bad, and the ugly. Am J Gastroenterol 2009 May;104(5):1086-1092. [doi: 10.1038/ajg.2009.118] [Medline: 19417748]

2. Akobeng AK. Understanding systematic reviews and meta-analysis. Arch Dis Child 2005 Aug;90(8):845-848 [FREE Full text] [doi: 10.1136/adc.2004.058230] [Medline: 16040886]

3. Mulrow CD. Rationale for systematic reviews. BMJ 1994 Sep 3;309(6954):597-599 [FREE Full text] [Medline: 8086953]

4. Manchikanti L. Evidence-based medicine, systematic reviews, and guidelines in interventional pain management, part I: introduction and general considerations. Pain Physician 2008;11(2):161-186 [FREE Full text] [Medline: 18354710]

5. Lavis JN. How can we support the use of systematic reviews in policymaking? PLoS Med 2009 Nov;6(11):e1000141 [FREE Full text] [doi: 10.1371/journal.pmed.1000141] [Medline: 19936223]

6. Phillips B, Ball C, Sackett D. Oxford Centre for Evidence-Based Medicine (CEBM). 2009. Oxford Centre for Evidence-Based Medicine: Levels of Evidence URL: http://www.cebm.net/index.aspx?o=1025 [accessed 2012-09-10] [WebCite Cache ID 6AZnaE7iG]

7. Cohen AM, Hersh WR, Peterson K, Yen PY. Reducing workload in systematic review preparation using automated citation classification. J Am Med Inform Assoc 2006;13(2):206-219 [FREE Full text] [doi: 10.1197/jamia.M1929] [Medline: 16357352]

8. Glasziou PP. Information overload: what's behind it, what's beyond it? Med J Aust 2008 Jul 21;189(2):84-85. [Medline: 18637773]

9. Clarke M. The Cochrane Collaboration and the Cochrane Library. Otolaryngol Head Neck Surg 2007 Oct;137(4 Suppl):S52-S54. [doi: 10.1016/j.otohns.2007.05.050] [Medline: 17894947]

10. Shojania KG, Sampson M, Ansari MT, Ji J, Doucette S, Moher D. How quickly do systematic reviews go out of date? A survival analysis. Ann Intern Med 2007 Aug 21;147(4):224-233. [Medline: 17638714]

11. Tsafnat G, Glasziou P, Choong MK, Dunn A, Galgani F, Coiera E. Systematic review automation technologies. Syst Rev 2014;3:74 [FREE Full text] [doi: 10.1186/2046-4053-3-74] [Medline: 25005128]

12. Cohen AM. Optimizing feature representation for automated systematic review work prioritization. 2008 Presented at: Annual Symposium Proceedings/AMIA Symposium; 2008; Washington, DC p. 121-125.

13. Wallace BC, Trikalinos TA, Lau J, Brodley C, Schmid CH. Semi-automated screening of biomedical citations for systematic reviews. BMC Bioinformatics 2010;11:55 [FREE Full text] [doi: 10.1186/1471-2105-11-55] [Medline: 20102628]

14. Wallace BC, Small K, Brodley CE, Lau J, Trikalinos TA. Deploying an interactive machine learning system in an evidence-based practice center: abstract. 2012 Presented at: 2nd ACM SIGHIT International Health Informatics Symposium; 2012; Miami, Florida.

15. Matwin S, Kouznetsov A, Inkpen D, Frunza O, O'Blenis P. A new algorithm for reducing the workload of experts in performing systematic reviews. J Am Med Inform Assoc 2010;17(4):446-453 [FREE Full text] [doi: 10.1136/jamia.2010.004325] [Medline: 20595313]

16. Salton G, Buckley C. Improving retrieval performance by relevance feedback. In: Jones KS, Willett P, editors. Readings in Information Retrieval. Burlington, MA: Morgan Kaufmann Publishers Inc; 1997:355-364.

17. Ruthven I, Lalmas M. A survey on the use of relevance feedback for information access systems. Knowl Eng Rev 2003;18(2):95-145.

18. Garfield E, Pudovkin A, Istomin V. Algorithmic citation‐linked historiography: Mapping the literature of science. In: Proceedings of the American Society for Information Science and Technology. 2002 Presented at: ASIS&T; 2002; Philadelphia, PA.

19. MEDLINE/PubMed Data Element (Field) Descriptions.: NIH US National Library of Medicine URL: http://www.nlm.nih.gov/bsd/mms/medlineelements.html [accessed 2014-03-08] [WebCite Cache ID 6axCSptUP]

20. OHSU Center for Evidence-Based Policy. Drug Effectiveness Review Project (DERP) Systematic Drug Class Review Gold Standard Data URL: http://skynet.ohsu.edu/~cohenaa/systematic-drug-class-review-data.html [accessed 2015-08-21] [WebCite Cache ID 6axCmYzFa]

21. Kim S, Choi J. Improving the performance of text categorization models used for the selection of high quality articles. Healthc Inform Res 2012 Mar;18(1):18-28 [FREE Full text] [doi: 10.4258/hir.2012.18.1.18] [Medline: 22509470]

22. Tan PN, Steinbach M, Kumar V. Introduction to Data Mining. Boston, MA: Pearson Addison Wesley; 2006.

23. PubMed Help. PubMed Stopword List URL: http://www.ncbi.nlm.nih.gov/books/NBK3827/ [accessed 2015-08-21] [WebCite Cache ID 6axDPDpEt]

24. Porter M. An algorithm for suffix stripping. Program 1980;14(3):130-137.

25. Bekhuis T, Tseytlin E, Mitchell KJ, Demner-Fushman D. Feature engineering and a proposed decision-support system for systematic reviewers of medical evidence. PLoS One 2014;9(1):e86277 [FREE Full text] [doi: 10.1371/journal.pone.0086277] [Medline: 24475099]

26.   Wright SP. Adjusted P-values for simultaneous inference. Biometrics 1992;48(4):1005-1013.

27.   Blakesley RE, Mazumdar S, Dew MA, Houck PR, Tang G, Reynolds CF, et al. Comparisons of methods for multiple hypothesis testing in neuropsychological research. Neuropsychology 2009 Mar;23(2):255-264 [FREE Full text] [doi: 10.1037/a0012850] [Medline: 19254098]

28.   Bekhuis T, Demner-Fushman D. Towards automating the initial screening phase of a systematic review. Stud Health Technol Inform 2010;160(Pt 1):146-150. [Medline: 20841667]

29.   Kastrin A, Peterlin B, Hristovski D. Chi-square-based scoring function for categorization of MEDLINE citations. Methods Inf Med 2010;49(4):371-378. [doi: 10.3414/ME09-01-0009] [Medline: 20091016]

30.   Frunza O, Inkpen D, Matwin S. Building systematic reviews using automatic text classification techniques. 2010 Presented at: Paper presented at: Proceedings of the 23rd International Conference on Computational Linguistics: Posters; 2010; Beijing, China.

31.   Shemilt I, Simon A, Hollands GJ, Marteau TM, Ogilvie D, O'Mara-Eves A, et al. Pinpointing needles in giant haystacks: use of text mining to reduce impractical screening workload in extremely large scoping reviews. Res Synth Methods 2014 Mar;5(1):31-49. [doi: 10.1002/jrsm.1093] [Medline: 26054024]

Abbreviations

AB: abstract
ANOVA: analysis over variance
AU: author
DERP: Drug Effectiveness Review Project
EBM: evidence-based medicine
EX: excluded
F_1: F-measure
FCNB/WE: factorized complement naïve Bayes with weight engineering
IN: included
MH: MeSH, or medical subject heading
PMID: PubMed Indentifier
PT: publication type
SVM: support vector machine
TI: title
VP: voting perceptron-based automated citation classification system
WSS95: worked saved over sampling at 95% recall

Original Paper

A Web-Based, Hospital-Wide Health Care-Associated Bloodstream Infection Surveillance and Classification System: Development and Evaluation

Yi-Ju Tseng[1,2], PhD; Jung-Hsuan Wu[3], MS; Hui-Chi Lin[4], BS; Ming-Yuan Chen[5], MS; Xiao-Ou Ping[6], PhD; Chun-Chuan Sun[4], BS; Rung-Ji Shang[5], PhD; Wang-Huei Sheng[4,7], MD, PhD; Yee-Chun Chen[4,7,8], MD, PhD; Feipei Lai[1,3,6], PhD; Shan-Chwen Chang[7], MD, PhD

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

[2]Computational Health Informatics Program, Boston Children's Hospital, Boston, MA, United States

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

[4]Center for Infection Control, National Taiwan University Hospital, Taipei, Taiwan

[5]Information Systems Office, National Taiwan University Hospital, Taipei, Taiwan

[6]Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan

[7]Department of Internal Medicine, National Taiwan University Hospital and College of Medicine, Taipei, Taiwan

[8]National Institute of Infectious Diseases and Vaccinology, National Health Research Institutes, Miaoli, Taiwan

Corresponding Author:
Yee-Chun Chen, MD, PhD
Department of Internal Medicine
National Taiwan University Hospital and College of Medicine
No. 7 Chung-Shan South Road
Taipei, 100
Taiwan
Phone: 886 223123456 ext 65054
Fax: 886 223971412
Email: yeechunchen@gmail.com

Abstract

Background: Surveillance of health care-associated infections is an essential component of infection prevention programs, but conventional systems are labor intensive and performance dependent.

Objective: To develop an automatic surveillance and classification system for health care-associated bloodstream infection (HABSI), and to evaluate its performance by comparing it with a conventional infection control personnel (ICP)-based surveillance system.

Methods: We developed a Web-based system that was integrated into the medical information system of a 2200-bed teaching hospital in Taiwan. The system automatically detects and classifies HABSIs.

Results: In this study, the number of computer-detected HABSIs correlated closely with the number of HABSIs detected by ICP by department (n=20; r=.999 P<.001) and by time (n=14; r=.941; P<.001). Compared with reference standards, this system performed excellently with regard to sensitivity (98.16%), specificity (99.96%), positive predictive value (95.81%), and negative predictive value (99.98%). The system enabled decreasing the delay in confirmation of HABSI cases, on average, by 29 days.

Conclusions: This system provides reliable and objective HABSI data for quality indicators, improving the delay caused by a conventional surveillance system.

XSL•FO
RenderX

Introduction

Background

Health care-associated infections (HAIs), adverse events related to health care, excess mortality and morbidity, and resource use are responsible for augmenting antimicrobial resistance [1,2]. Surveillance of HAIs is an essential component of infection control programs in health care settings. The goals of surveillance are to assess the disease incidence, identify the niche and opportunity for improvement, monitor the efficacy of interventions, and support the rationale behind changes in policies [3]. Previous studies have reported that HAIs have decreased ly in hospitals that adopted surveillance programs in the 1980s in the United States [4,5]. In 1981, a hospital-wide HAI surveillance program was initiated at the National Taiwan University Hospital (NTUH). Factors such as periodic feedback to the departments or wards and intensified interventions resulted in a decrease in surgical-site and respiratory tract infections in the 1980s [6]. Following the upgrading of the infection prevention and control program in 2004, there was a significant reduction in bloodstream infections, HAIs in intensive care units, and HAIs caused by multidrug-resistant organisms (MDROs) during the period from 2004 to 2007 [7].

Attention to HAIs has increased partially because of legislative mandates for reporting and reimbursement policies [8]. However, conventional HAI surveillance systems require considerable human involvement in integrating and interpreting data and are labor intensive, performance dependent, and tend to divert resources that are necessary for implementing control measures and prevention activities [9]. Relying on employees in institutions in an environment where reporting HAIs can be associated with punitive consequences is suboptimal [10]. Furthermore, the decision rules (ie, the case definitions) of HAIs are complicated when the complexity of the current health care in tertiary care hospitals is considered. Utilization of hospital discharge registry data delays the detection of HAI, eventually resulting in ineffective identification of problems [11]. Recent studies have identified interinstitutional variability in surveillance techniques, an inconsistency that affects the validity of publicly reported HAI data [12]. Developing reliable and objective HAI definitions and automated processes for infection detection is crucial; however, transformation into automated surveillance system remains challenging [8,10].

Study Objective

In continuation with our previous efforts in developing a Web-based MDRO surveillance system that automatically identifies and accurately detects suspicious outbreaks of MDROs [13], implementation of which could save 1 person-day daily, we conducted this study with the aim of developing a Web-based automatic surveillance and classification system for health care-associated bloodstream infection (HABSI), the most common HAIs at NTUH [6]. In addition, performance of the system was evaluated by comparing the proposed system with a conventional infection control personnel (ICP)-based surveillance system.

Methods

Hospital Setting and Study Population

The study was conducted at NTUH, a 2200-bed teaching hospital that provides primary and tertiary care for the adult and the pediatric population in Taiwan. This study was approved by the Institutional Ethics Review Board of NTUH (NTUH-200904014R). In 2011, NTUH served 2,309,108 outpatients, received 106,950 emergency visits, and discharged 104,899 patients (723,505 patient-days).

Two sets of blood samples from separate venipuncture sites for bacterial culture were routinely collected from patients who were newly diagnosed with sepsis. An additional sample was collected after 45-60 minutes to define continuous bloodstream infection (BSI). Only 1 blood sample for a follow-up culture was collected to confirm the clearance of BSI. Of the 80,327 blood specimens that were sent for isolation and identification of pathogens, 991 (1.23%) were obtained through a single blood draw. A total of 8745 samples grew 1 or more pathogens (10.88%); of these, 1908 exhibited HABSIs. The pooled mean of HABSI incidence was 14.7 episodes/1000 patients (range 0.2-112.7/1000 patients by department) and 2.13 episodes/1000 patient-days.

Conventional ICP-Based HAI Surveillance System

A prospective, hospital-wide on-site surveillance of HAIs, initiated in 1981 [6], was conducted by ICPs who reviewed microbiological data daily and visited inpatient units weekly to identify patients with HAIs according to definitions of the Centers for Disease Control and Prevention (CDC) [14] before we implemented the Web-based surveillance system. If required, the ICPs consulted physicians, particularly infectious disease physicians, to interpret the medical information of patients who have received complicated interventions. Data are collected on standardized data-collection forms and input into the computer database manually. In addition, the ICPs monitor culture results from the clinical microbiology laboratory daily to identify MDROs. The unit-specific incidences of HAIs, including overall, unit-based, and site-specific infection rates, are analyzed monthly and compared with historical data. Feedback is provided to each unit to stimulate intervention measures.

The key data are collected by systematically reviewing hospital information systems (HISs), laboratory information systems (LISs), and handwritten charts. Because of advances in medical information system, data of HISs and LISs are currently stored as electronic medical records (EMRs). However, data generated in the medical information system are scattered in numerous databases, and data access is hindered by several interfaces. In addition, data must be integrated, interpreted, and transformed into meaningful information.

Web-Based HAI Surveillance and Classification System

We established a rule-based HABSI surveillance and classification system (the system), which was implemented on October 1, 2010. The current version was revised on September 20, 2012 (Figure 1). The system screens HIS and LIS data daily to detect HABSI candidates according to the well-defined detection rules. The system detects and classifies HABSIs

automatically and reserves professional autonomy by requiring further confirmation. Figure 1 shows the general architecture of the system, including the user layer, the database layer, and the 3-part system layer (data collection, candidate detection, and HAI management). The system adheres to service-oriented architecture (SOA) and Health Level Seven (HL7) standards and can be adapted in other information systems [15].

This data-collection subsystem collects relevant data from HIS and LIS using HL7 standards, which was extended stepwise from the previous version [13]. For example, data related to age, sex, ward transfer, admission date, and discharge diagnosis are obtained from the hospital administrative system. Re-admissions within 2 days of discharge are linked to the previous hospital stay and considered to be a single hospitalization episode. Procedure codes are obtained from the hospital billing system. Data on body temperature, heart rate, respiratory rate, and presence of device including indwelling catheters are obtained from the nursing system. The use of antimicrobial agents is obtained from the pharmacology prescription system. The laboratory data consist of specimen information and microbiological data. We used the specimen log-in time as a proxy for the time of specimen collection at bedside and the infection time.

To develop the detection rules in the candidate-detection subsystem, the ICPs have adapted objective components of the National Health Care Safety Network (NHSN) definitions established by the CDC [14] and modified them according to the consensus between Taiwan Center for Disease Control and NTUH (see Multimedia Appendix 1 for detection rules and

Multimedia Appendices 2 and 3 for list of devices and signs and symptoms, respectively). Computer engineers have established HABSI detection rules accordingly (Figure 2). Thus, there are differences between the detection rules and NHSN definition. Although the secondary HABSI was removed from the NHSN definition in 2008, we used the data in the NTUH infection control system to maintain data consistency for time trend analysis [6,7]. The primary HABSI was divided into 3 subtypes for quality-improvement purpose, including central line-related BSI (CRBSI), central line-associated BSI (other than CRBSI), and other primary HABSI. Detection rules did not include free text data that were not available in the EMRs, such as chills and apnea, during the study period.

The HAI management subsystem, established on July 1, 2007 [13], and upgraded periodically (refer to the screenshot in Multimedia Appendix 4), consists of data integration, display, and a data-modification user interface for facilitating the surveillance of HAIs. The HAI management subsystem has a single entry point for HIS through the browser [16] and comprises a storage information component relevant to HAI. The information for each event of HAI includes patient demographic data, diagnoses, procedures, medications, and microbiology reports to facilitate the confirmation of HAI by ICPs. If required, the ICPs can modify these HABSI cases or add additional HABSIs that were undetectable by the subsystem. The latter might occur for new units, new pathogens, new procedures, or other elements, which are not yet included in the current database. Furthermore, the system provides data analysis and process-control charts [13].

Figure 1. General architecture of the Web-based health care-associated infection (HAI) surveillance and classification system. DB: database; HL7: Health Level Seven; SOA: service-oriented architecture.
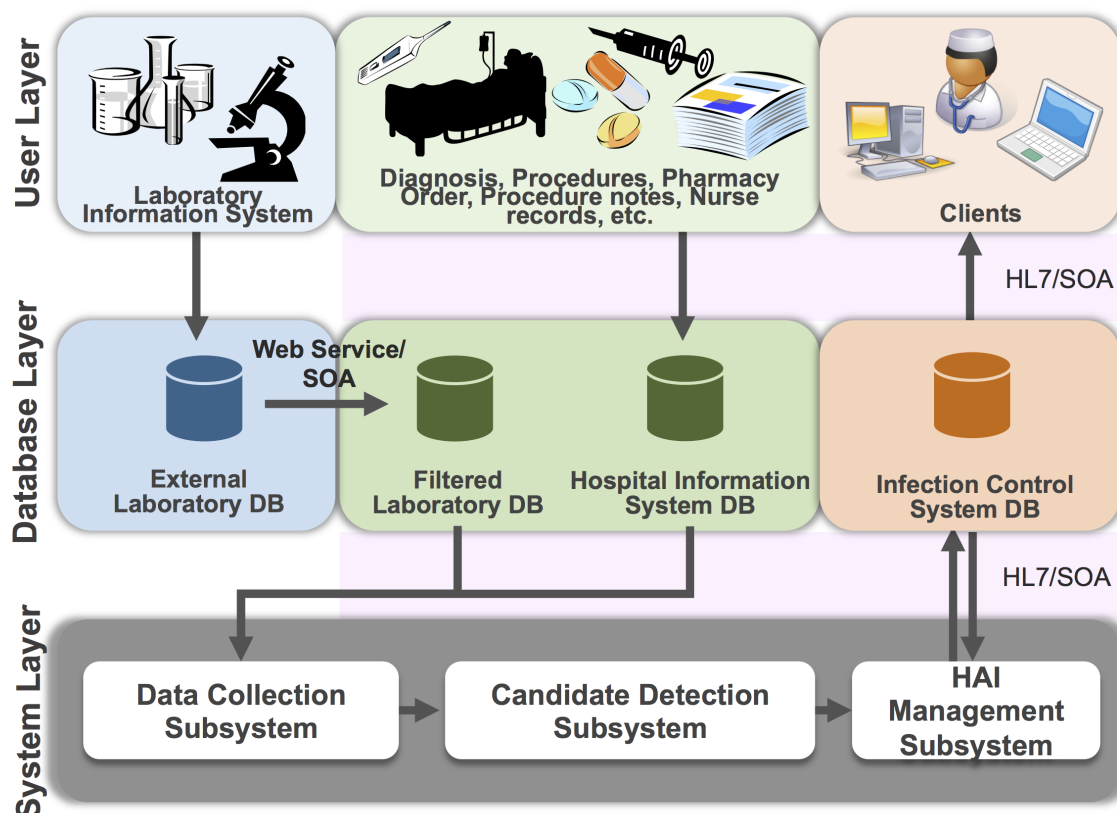
Figure 2. Computer algorithms to detect health care-associated bloodstream infection (HABSI) by active daily screening of data from hospital information system and laboratory information system. HABSIs are classified into primary HABSI (PRIM), secondary HABSI (SEC), and clinical sepsis (CSEP) as described in Multimedia Appendix 1. Polymicrobial and persistent BSI criterion here are to eliminate false signals due to duplicate counting, etc.

XSL•FO
RenderX
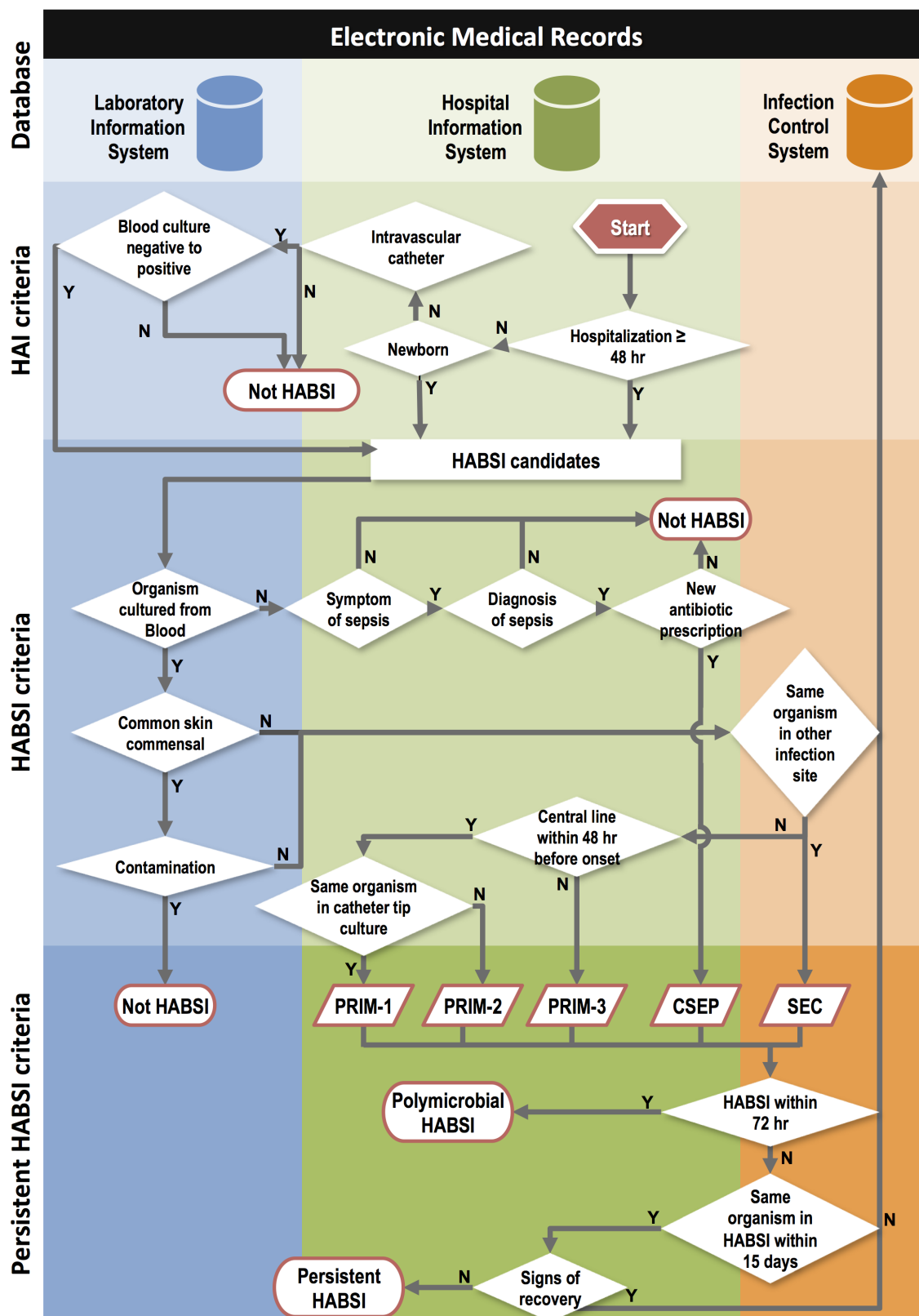
Evaluation of System Performance and Statistical Analyses

Clinically useful tests must be valid and reliable and have a reasonable turnaround time. Thus, we conducted a 3-aspect evaluation, including accuracy, reliability, and efficiency, of the system. Figure 3 summarizes the objectives, methods, and evaluation periods. Reliability of the system was evaluated before implementation of the first version of the system in 2010; accuracy of HABSI rules was evaluated in October 2012. Furthermore, we evaluated, and continue to evaluate, the stepwise improvement in efficiency after implementation of the HAI management system and the HABSI surveillance and classification system.

We first evaluated the performance of the system during the developmental phase (ie, before implementation of the system) regarding its potential to provide data for quality indicators. Computer-detected HABSIs were compared with ICP-detected HABSIs as a routine practice between July 1, 2010, and September 30, 2010. The correlation between these 2 data sources was analyzed according to department distribution and time trend of HABSIs.

On the basis of inconsistent and varied performances of the conventional ICP-based surveillance system, we further evaluated the performance after implementing the system using ICP-defined reference standards. To generate high-quality reference standards, 11 ICPs performed a retrospective review of all medical data of patients who were admitted between October 1, 2012, and October 31, 2012, to identify HABSI cases based on NTUH detection rules, and one of the authors (H-CL) validated the results. The sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV) of the system, and Cohen kappa coefficient [17] were calculated in this evaluation method. All the performance indicators were calculated based on whether a patient had HABSI or not.

We then compared the delay in HABSI confirmation (as a proxy for practice efficiency) before (October 2007-September 2010) and after (October 2010-September 2013) the system implementation. The delay in confirmation was defined as the intervals between the HABSI confirmation dates (complete data entry and confirmed by ICPs in the HAI management subsystem) and log-in dates of the first blood specimen with positive results and was calculated by averaging the number of delay days in each month. The study periods were selected taking into account the seasonal variation of HAI rates demonstrated previously [7].

Figure 3. Timeline of development and performance evaluation of the health care-associated bloodstream infections (HABSIs) surveillance and classification system. ICP: infection control personnel.
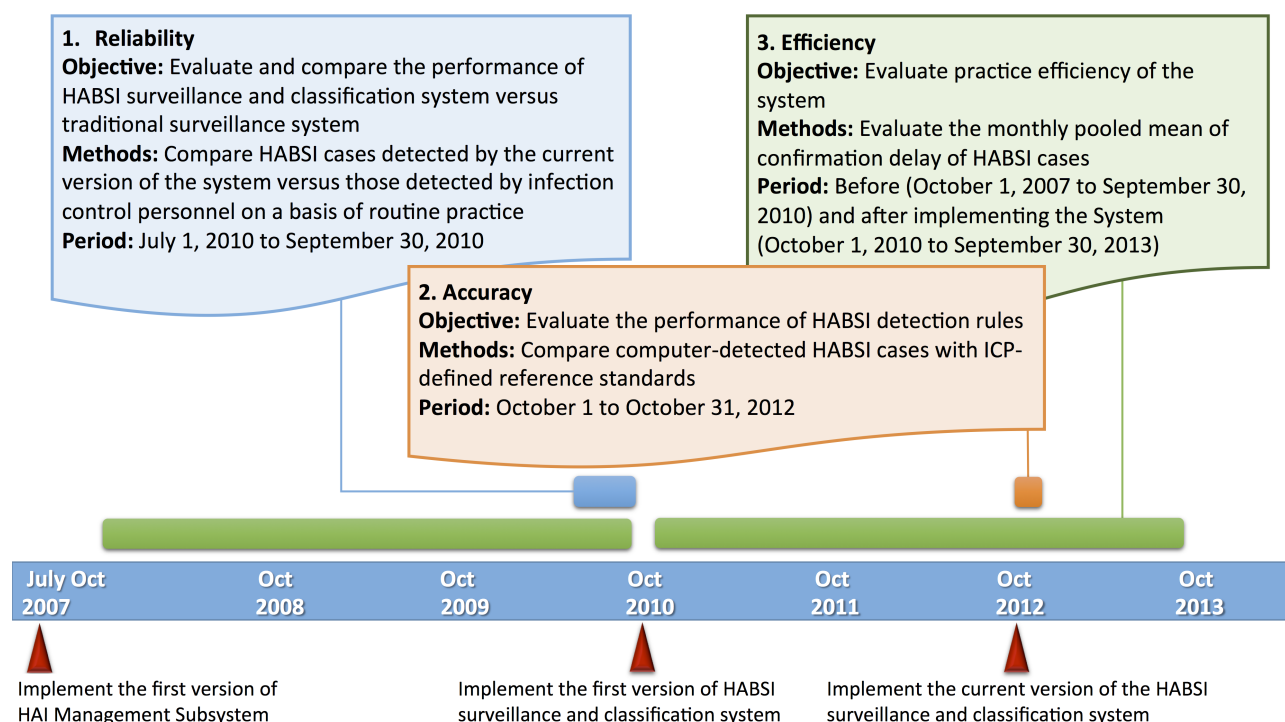

Results

Performance of the System as a Provider of Quality Indicator

During the 14-week study period (July 1, 2010-September 30, 2010), 501 episodes of ICP-detected HABSIs and 479 episodes of computer-detected HABSIs were found throughout the 20 departments. The data were highly correlated by place and time (Figure 4). These results indicated that the data provided by the system are suitable quality indicators. Thus, we implemented the system on October 1, 2010.

Figure 4. Correlation of 501 episodes of infection control personnel (ICP)-detected health care-associated bloodstream infection (HABSI) and 479 episodes of computer-detected HABSI from 20 departments during the 14-week study period. (A) Perfect agreement of HABSI episodes by department (n=20, Pearson correlation, $r<.999$, $P<.001$). (B) Perfect agreement of HABSI episodes by time (n=14, Pearson correlation, $r=.941$, $P<.001$).
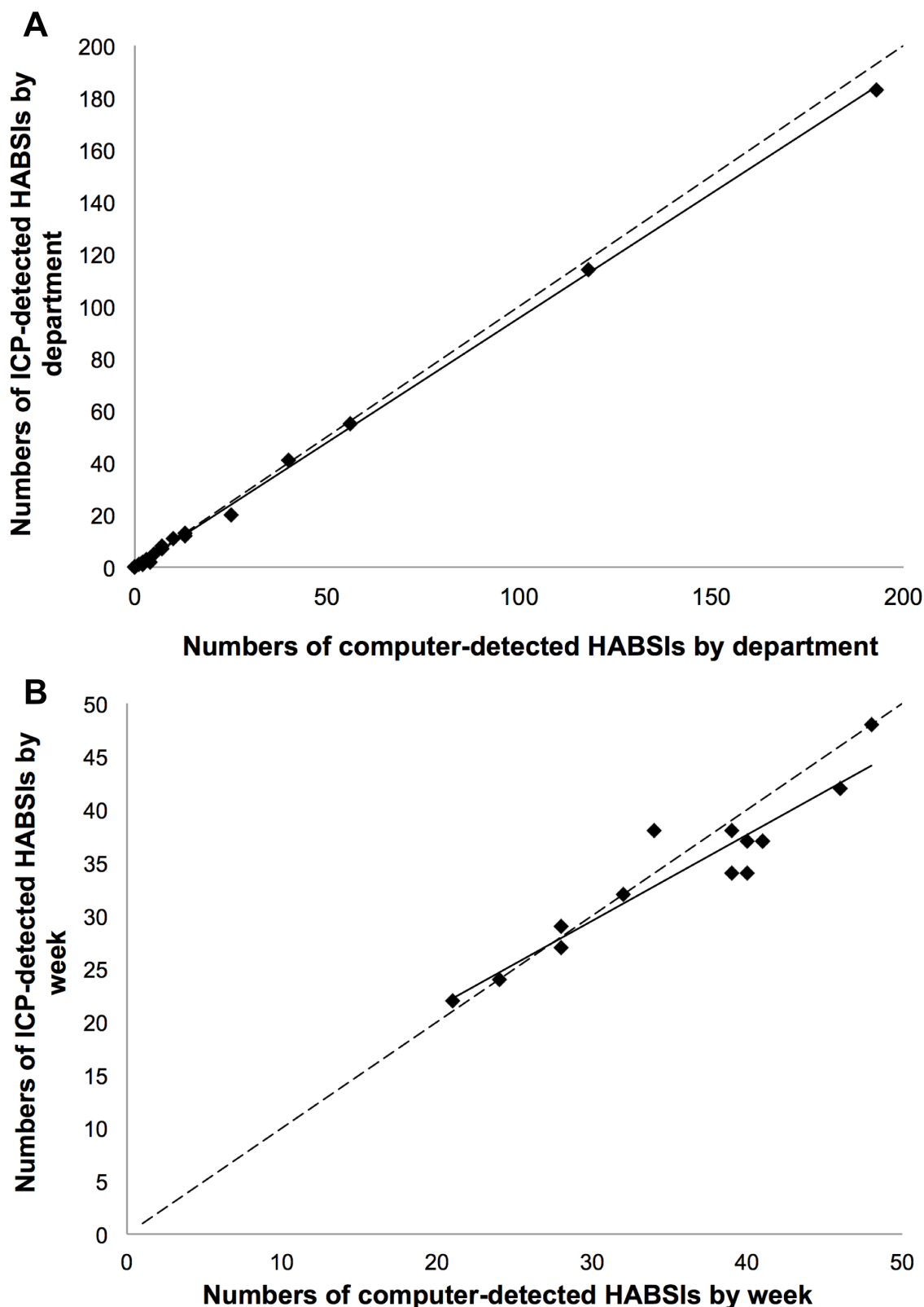
Accuracy of the Detection Rule

In October 2012, the system identified 167 episodes of HABSIs (Figure 5), including 160 of 163 reference standard episodes in

31 days (Table 1). The sensitivity and specificity of the HABSI classification system were 98.16% (95% CI 94.29-99.52) and 99.96% (95% CI 99.91-99.98), respectively. The PPV and NPV

XSL•FO
RenderX

were 95.81% (95% CI 91.22-98.15) and 99.98% (95% CI 99.95-100.00), respectively. Moreover, the agreement between the computer-detected HABSIs and the reference standard was nearly perfect (Cohen kappa coefficient .97; 95% CI 0.95-0.99).

The performance of the system for detecting central line-associated HABSI was also excellent (sensitivity 97.14%, specificity 99.94%, PPV 91.07%, NPV 99.94%, and Cohen kappa coefficient .94).

Table 1. Comparison of the case detection results of the health care-associated bloodstream infection surveillance and classification system with infection control personnel reference standard between the periods October 1 and October 31, 2012.

Infection control personnel reference standard	Automated surveillance classification		
	HABSI	Not HABSI	Total
HABSI	160	3[a]	163
Not HABSI	7[b]	17,824	17,831
Total	167	17,827	17,994

[a]Retrospective review by 2 investigators independently confirmed that these 3 episodes of HABSI due to common skin commensals were missed due to fever criteria (temperature > 38°C): 1 patient received antipyretic agents, 1 with a and sustained increase in temperature (>1°C) from baseline but less than 38°C, and in the other patient fever was documented only in the progress note and was missed by using this fever criteria.

[b]Four false-positive cases due to revision of final laboratory reports after "recall day." One episode of community-acquired BSI was detected as HABSI due to delay in transportation of specimen to microbiology laboratory. Two were cases of persistent bloodstream infection.

Figure 5. Computer algorithms identifying 167 events of health care-associated bloodstream infection among 8862 inpatients and 9132 patients in the emergency department between October 1 and October 31, 2012 (31 days).
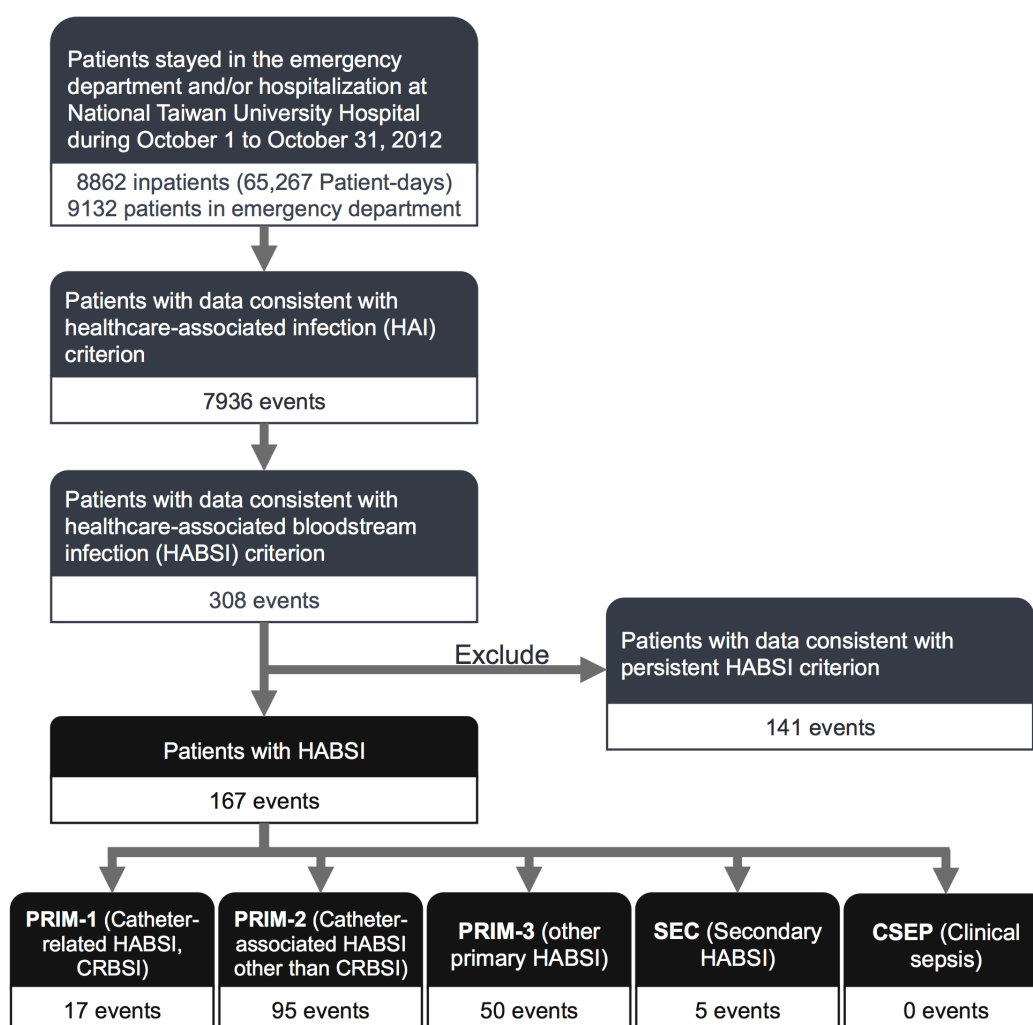
Decrease in the Delay of HABSI Confirmation

The delay in HABSI confirmation was reduced from 43.58 (SD 15.57) days before the system implementation (October 2007-September 2010, 1096 days) to 14.58 (SD 4.64) days after the implementation (October 2010-September 2013, 1096 days; $P<.001$). Figure 6 shows that the time trend of the delay in

HABSI confirmation, which was as high as 90 days in July 2007, decreased after the implementation of the HAI management subsystem in July 2007, and further improved after automating the system in October 2010. The delay in the second half of 2013 was only 5.78 (SD 0.91) days. Conversely, without the system, the delay increased during H1N1 influenza pandemics and when preparing for international accreditation.

Figure 6. The detection delay of health care-associated bloodstream infection (HABSI) decreased gradually from July 2007 to December 2013. The first version of health care-associated infection (HAI) management subsystem has been developed to facilitate infection control personnel-based surveillance program since July 2007. This was revised stepwise and has been operation as an automatic system since October, 2010. In March 2009, this hospital initiated preparedness for international hospital accreditation, which was scheduled 1 year later. Influenza pandemic occurred in April 2009.
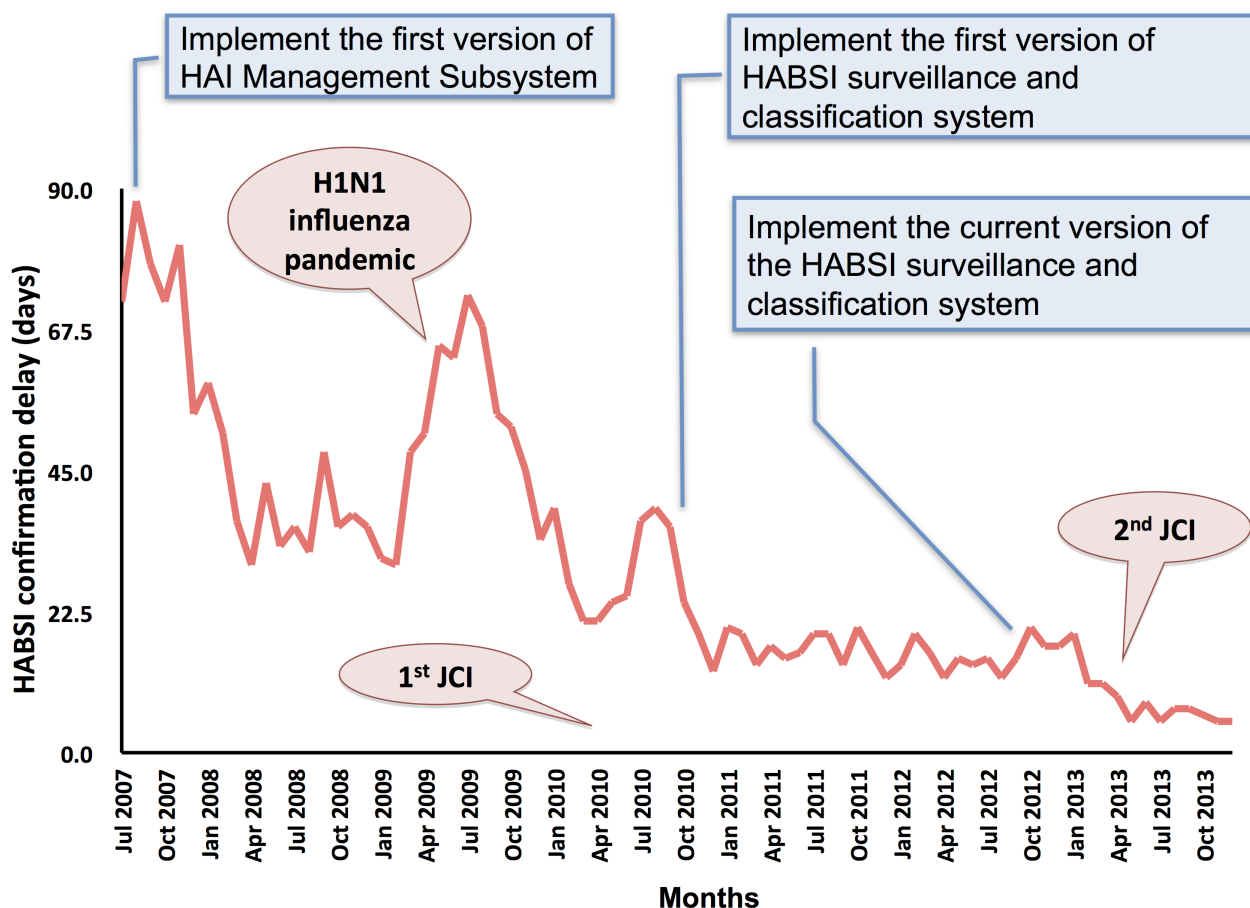
Discussion

Preliminary Findings

This Web-based, automated HABSI surveillance and classification system used discrete data elements obtained from HISs, and LISs provided data highly correlated with conventional ICP surveillance system. The performance was excellent regarding sensitivity, specificity, PPV, and NPV, and was in agreement with reference standards; the system reduced the delay in confirmation, on average, by 29 days. The system improves practice efficiency, enabling ICPs to intensify intervention and further reduce HAI rates.

Computer-assisted HAI surveillance and classification systems are widely implemented [11,18-35]. Studies have demonstrated using various algorithms for detecting HAIs, although with

varied outcomes (summary in Multimedia Appendix 5) [25-35]. The most critical performance characteristics of these kinds of surveillance systems are sensitivity and NPV; the efficiency of the system can be assessed according to the PPV [36]. Compared with reference standards, the current version of detection rules and computer algorithms performed excellently with regard to these 3 parameters (sensitivity 98.16%, NPV 99.98%, and PPV 95.81%) because of the following reasons: Through cross-talk among ICPs, infectious disease physicians, and engineers, we integrated clinical know-how and translated international case definitions to define detection rules and construct computer algorithms. The system was established and revised through a plan-do-check-act cycle in a general hospital, which serves a varied patient population and offers numerous procedures. Furthermore, we evaluated the clinical utility of the system,

comparing it with the prospective, hospital-wide, conventional surveillance system and reference standards.

In this study, the HABSI detection rules (see Multimedia Appendix 1) were clearly defined, and computer algorithms (Figure 2) provided excellent results. We adapted the US CDC definition of HAI, adding rules related to re-admission within 48 hours and neonates, and a "hospital-acquired" rule, defined as a positive blood culture that was obtained 48 hours or more after admission. These rules included rules for clinical sepsis, and the system actively screened the heart rate, respiratory rate, and body temperature for infection-related symptoms and signs. These efforts facilitated ameliorating the potential underestimation of HAI when only laboratory data were used. The HABSI detection rules included polymicrobial and persistent BSI criterion to eliminate the majority of false signals (eg, duplicate counting). In addition, classification of HABSIs is flexible to addressing local policy and ICP requests to compare them with the NTUH historical data.

The system detects and classifies HABSIs automatically and ensures professional autonomy by requiring further confirmation. Each episode of HAI requires confirmation by ICPs. The system presents detailed information about each HABSI candidate systematically to support decision making. The main reason for this design is because the system is imperfect (see the "Limitations" section). Furthermore, because HAIs are rare in hospitalized patients, the system aims to select potential HAI candidates and exclude patients who do not have an HAI and hence do not require review by ICPs.

This study verifies the potential of the system to provide data for quality indicators. The system enabled sustainable surveillance, generating data that were correlated with conventional surveillance systems by department and time. In addition, the delay in HABSI confirmation decreased to 5.78 (SD 0.91) days in the second half of 2013. Because of the reduced length of hospital stay and the increased threat of emerging infectious diseases, early detection of HAIs can enable identifying the reservoir or index case and providing early intervention before pathogens spread further. Currently, the delay in HABSI confirmation is caused by the time required to identify the positive blood cultures and microorganisms; the system detects and analyzes results of blood cultures to prevent false alerts. Furthermore, the major challenge encountered when sharing automated HABSI surveillance systems between hospitals is different HIS settings [22]; the Web-based system, which adheres to SOA and HL7 standards, can be easily extended to and adapted for use with other medical information systems.

Limitations

Although our results suggested that the system performs well, this study had several limitations. First, data integrity and instantaneity substantially affect performance, because this system uses EMRs from many sources. In addition, not all data required for HAI surveillance [37] are available on the Web, because EMRs were not fully implemented at NTUH until January 1, 2014. Second, the quality of source data, which is related to the performance of clinical practice and EMRs, affected the results. As much as 1.2% (991/80,327) of the blood specimens collected in 2011 were obtained through a single blood draw (reasons described in the "Methods" section), and this affected the identification of common skin commensals and resulted in a false-positive HABSI case (Table 1). We reviewed the medical records and determined that the false-positive result was caused by the delay in specimen delivery and log-in time. Third, the system updates laboratory data from LIS within a fixed period (recall day); however, data are not repeated during a subsequent hospital stay. This resulted in 4 false-positive HABSI cases, because the laboratory reports were revised in the LIS after recall day and included common skin commensal contaminants that were not updated in the system.

Fourth, the agreement regarding the place of onset (responsible ward) was not evaluated in this study, because patients are frequently transferred to different wards and electronic clinical data regarding symptoms and signs of infection were unavailable during the study period. Fifth, the case definition of HABSI is complicated when clinical scenarios are taken into consideration. For disagreements and received revaluations (Table 1), all the false-negative results were due to the case definition of fever (>38°C). Furthermore, we did not evaluate the reduction of person-hours after implementing the system, as described in previous studies [13,38,39], or the subsequent effect on the reduction of HABSI by reallocating ICP time and effort from collecting data to improving program quality, as described in a previous research of a hospital-wide hand-hygiene promotion program [7] and care bundles for device-associated infection to prevent HABSI.

Conclusions

This fully automated system that can be integrated in medical information systems detects and classifies HABSI within 5.78 (SD 0.91) days after occurrence, enabling the opportunity for early intervention. Currently, the system and other components of the infection control system [13,39] operate well and have become indispensable tools for infection control programs. Future studies using clinical data from complete EMRs and refining classification algorithms or adopting multivariable prediction models are warranted [36]. According to the results of this pilot study for HABSI automated surveillance, further efforts for other HAI surveillance are underway.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Detection rules for the health care-associated bloodstream infection (HABSI) Surveillance and Classification System at National Taiwan University Hospital (NTUH) and corresponding Centers for Disease Control and Prevention (CDC) National Healthcare Safety Network (NHSN) definition of health care-associated infection (HAI) [14].

[PDF File (Adobe PDF File), 53KB - medinform_v3i3e31_app1.pdf]

Multimedia Appendix 2

The list of central lines and intravascular devices.

[PDF File (Adobe PDF File), 29KB - medinform_v3i3e31_app2.pdf]

Multimedia Appendix 3

The definition of symptoms/signs of sepsis in pediatric patients.

[PDF File (Adobe PDF File), 20KB - medinform_v3i3e31_app3.pdf]

Multimedia Appendix 4

Summary of characteristics and performances of computer-assisted health care-associated infection surveillance systems in the literature.

[PDF File (Adobe PDF File), 40KB - medinform_v3i3e31_app4.pdf]

Multimedia Appendix 5

User interface of HAI Management System. This system integrated all the information which was needed for HAI decision.

[PDF File (Adobe PDF File), 210KB - medinform_v3i3e31_app5.pdf]

References

1. European Centre for Disease Prevention and Control. European Centre for Disease Prevention and Control: Annual Epidemiological Report on Communicable Diseases in Europe 2008. Stockholm, Sweden; 2008. URL: http://ecdc.europa.eu/en/publications/Publications/0812_SUR_Annual_Epidemiological_Report_2008.pdf [accessed 2014-12-23] [WebCite Cache ID 6V2YbKfyw]
2. Klevens RM, Edwards JR, Richards CL, Horan TC, Gaynes RP, Pollock DA, et al. Estimating health care-associated infections and deaths in U.S. hospitals, 2002. Public Health Rep 2007;122(2):160-166 [FREE Full text] [Medline: 17357358]
3. McKibben L, Horan TC, Tokars JI, Fowler G, Cardo DM, Pearson ML, Healthcare Infection Control Practices Advisory Committee. Guidance on public reporting of healthcare-associated infections: Recommendations of the Healthcare Infection Control Practices Advisory Committee. Infect Control Hosp Epidemiol 2005 Jun;26(6):580-587. [doi: 10.1086/502585] [Medline: 16018435]
4. Gaynes R, Richards C, Edwards J, Emori TG, Horan T, Alonso-Echanove J, et al. Feeding back surveillance data to prevent hospital-acquired infections. Emerg Infect Dis 2001;7(2):295-298 [FREE Full text] [doi: 10.3201/eid0702.700295] [Medline: 11294727]
5. Haley RW, Culver DH, White JW, Morgan WM, Emori TG, Munn VP, et al. The efficacy of infection surveillance and control programs in preventing nosocomial infections in US hospitals. Am J Epidemiol 1985 Feb;121(2):182-205. [Medline: 4014115]
6. Chuang YC, Chen YC, Chang SC, Sun CC, Chang YY, Chen ML, et al. Secular trends of healthcare-associated infections at a teaching hospital in Taiwan, 1981-2007. J Hosp Infect 2010 Oct;76(2):143-149. [doi: 10.1016/j.jhin.2010.05.001] [Medline: 20663585]
7. Chen YC, Sheng WH, Wang JT, Chang SC, Lin HC, Tien KL, et al. Effectiveness and limitations of hand hygiene promotion on decreasing healthcare-associated infections. PLoS One 2011;6(11):e27163 [FREE Full text] [doi: 10.1371/journal.pone.0027163] [Medline: 22110610]

8. Talbot TR, Bratzler DW, Carrico RM, Diekema DJ, Hayden MK, Huang SS, Healthcare Infection Control Practices Advisory Committee. Public reporting of health care-associated surveillance data: Recommendations from the healthcare infection control practices advisory committee. Ann Intern Med 2013 Nov 5;159(9):631-635. [doi: 10.7326/0003-4819-159-9-201311050-00011] [Medline: 24189597]

9. Emori TG, Edwards JR, Culver DH, Sartor C, Stroud LA, Gaunt EE, et al. Accuracy of reporting nosocomial infections in intensive-care-unit patients to the National Nosocomial Infections Surveillance System: A pilot study. Infect Control Hosp Epidemiol 1998 May;19(5):308-316. [Medline: 9613690]

10. Trick WE. Decision making during healthcare-associated infection surveillance: A rationale for automation. Clin Infect Dis 2013 Aug;57(3):434-440 [FREE Full text] [doi: 10.1093/cid/cit249] [Medline: 23592826]

11. Goto M, Ohl ME, Schweizer ML, Perencevich EN. Accuracy of administrative code data for the surveillance of healthcare-associated infections: A systematic review and meta-analysis. Clin Infect Dis 2014 Mar;58(5):688-696 [FREE Full text] [doi: 10.1093/cid/cit737] [Medline: 24218103]

12. Hebden JN. Rationale for accuracy and consistency in applying standardized definitions for surveillance of health care-associated infections. Am J Infect Control 2012 Jun;40(5 Suppl):S29-S31. [doi: 10.1016/j.ajic.2012.03.009] [Medline: 22626460]

13. Tseng YJ, Wu JH, Ping XO, Lin HC, Chen YY, Shang RJ, et al. A Web-based multidrug-resistant organisms surveillance and outbreak detection system with rule-based classification and clustering. J Med Internet Res 2012;14(5):e131 [FREE Full text] [doi: 10.2196/jmir.2056] [Medline: 23195868]

14. Horan TC, Andrus M, Dudeck MA. CDC/NHSN surveillance definition of health care-associated infection and criteria for specific types of infections in the acute care setting. Am J Infect Control 2008 Jun;36(5):309-332. [doi: 10.1016/j.ajic.2008.03.002] [Medline: 18538699]

15. Health Level Seven. HL7 Standard v2.5. 2003. URL: http://www.hl7.org/ [accessed 2014-12-23] [WebCite Cache ID 6V2ZCmgpr]

16. Hsieh SH, Hsieh SL, Cheng PH, Lai F. E-health and healthcare enterprise information system leveraging service-oriented architecture. Telemed J E Health 2012 Apr;18(3):205-212. [doi: 10.1089/tmj.2011.0100] [Medline: 22480301]

17. Carletta J. Assessing agreement on classification tasks: The kappa statistic. Comput Linguist 1996 Jun 1;22(2):249-254.

18. Steurbaut K, Van HS, Colpaert K, Lamont K, Taveirne K, Depuydt P, et al. Use of web services for computerized medical decision support, including infection control and antibiotic management, in the intensive care unit. J Telemed Telecare 2010;16(1):25-29. [doi: 10.1258/jtt.2009.001008] [Medline: 20086264]

19. Bellini C, Petignat C, Francioli P, Wenger A, Bille J, Klopotov A, et al. Comparison of automated strategies for surveillance of nosocomial bacteremia. Infect Control Hosp Epidemiol 2007 Sep;28(9):1030-1035. [doi: 10.1086/519861] [Medline: 17932822]

20. Chalfine A, Cauet D, Lin WC, Gonot J, Calvo-Verjat N, Dazza F, et al. Highly sensitive and efficient computer-assisted system for routine surveillance for surgical site infection. Infect Control Hosp Epidemiol 2006 Aug;27(8):794-801. [doi: 10.1086/506393] [Medline: 16874638]

21. Apte M, Landers T, Furuya Y, Hyman S, Larson E. Comparison of two computer algorithms to identify surgical site infections. Surg Infect (Larchmt) 2011 Dec;12(6):459-464 [FREE Full text] [doi: 10.1089/sur.2010.109] [Medline: 22136489]

22. Hota B, Lin M, Doherty JA, Borlawsky T, Woeltje K, Stevenson K, CDC Prevention Epicenter Program. Formulation of a model for automating infection surveillance: Algorithmic detection of central-line associated bloodstream infection. J Am Med Inform Assoc 2010;17(1):42-48 [FREE Full text] [doi: 10.1197/jamia.M3196] [Medline: 20064800]

23. Klompas M, Yokoe DS. Automated surveillance of health care-associated infections. Clin Infect Dis 2009 May 1;48(9):1268-1275 [FREE Full text] [doi: 10.1086/597591] [Medline: 19335166]

24. Brossette SE, Hacek DM, Gavin PJ, Kamdar MA, Gadbois KD, Fisher AG, et al. A laboratory-based, hospital-wide, electronic marker for nosocomial infection: The future of infection control surveillance? Am J Clin Pathol 2006 Jan;125(1):34-39 [FREE Full text] [Medline: 16482989]

25. Woeltje KF, McMullen KM, Butler AM, Goris AJ, Doherty JA. Electronic surveillance for healthcare-associated central line-associated bloodstream infections outside the intensive care unit. Infect Control Hosp Epidemiol 2011 Nov;32(11):1086-1090. [doi: 10.1086/662181] [Medline: 22011535]

26. Choudhuri JA, Pergamit RF, Chan JD, Schreuder AB, McNamara E, Lynch JB, et al. An electronic catheter-associated urinary tract infection surveillance tool. Infect Control Hosp Epidemiol 2011 Aug;32(8):757-762. [doi: 10.1086/661103] [Medline: 21768758]

27. McBryde ES, Brett J, Russo PL, Worth LJ, Bull AL, Richards MJ. Validation of statewide surveillance system data on central line-associated bloodstream infection in intensive care units in Australia. Infect Control Hosp Epidemiol 2009 Nov;30(11):1045-1049. [doi: 10.1086/606168] [Medline: 19803720]

28. Pokorny L, Rovira A, Martín-Baranera M, Gimeno C, Alonso-Tarrés C, Vilarasau J. Automatic detection of patients with nosocomial infection by a computer-based surveillance system: A validation study in a general hospital. Infect Control Hosp Epidemiol 2006 May;27(5):500-503. [doi: 10.1086/502685] [Medline: 16671032]

29. Leth RA, Møller JK. Surveillance of hospital-acquired infections based on electronic hospital registries. J Hosp Infect 2006 Jan;62(1):71-79. [doi: 10.1016/j.jhin.2005.04.002] [Medline: 16099539]

30. Trick WE, Zagorski BM, Tokars JI, Vernon MO, Welbel SF, Wisniewski MF, et al. Computer algorithms to detect bloodstream infections. Emerg Infect Dis 2004 Sep;10(9):1612-1620 [FREE Full text] [doi: 10.3201/eid1009.030978] [Medline: 15498164]

31. de Bruin JS, Adlassnig K, Blacky A, Mandl H, Fehre K, Koller W. Effectiveness of an automated surveillance system for intensive care unit-acquired infections. J Am Med Inform Assoc 2013;20(2):369-372 [FREE Full text] [doi: 10.1136/amiajnl-2012-000898] [Medline: 22871398]

32. Bouam S, Girou E, Brun-Buisson C, Karadimas H, Lepage E. An intranet-based automated system for the surveillance of nosocomial infections: Prospective validation compared with physicians' self-reports. Infect Control Hosp Epidemiol 2003 Jan;24(1):51-55. [doi: 10.1086/502115] [Medline: 12558236]

33. van Mourik MS, Moons KG, van Solinge WW, Berkelbach-van der Sprenkel JW, Regli L, Troelstra A, et al. Automated detection of healthcare associated infections: External validation and updating of a model for surveillance of drain-related meningitis. PLoS One 2012;7(12):e51509 [FREE Full text] [doi: 10.1371/journal.pone.0051509] [Medline: 23236510]

34. Du M, Xing Y, Suo J, Liu B, Jia N, Huo R, et al. Real-time automatic hospital-wide surveillance of nosocomial infections and outbreaks in a large Chinese tertiary hospital. BMC Med Inform Decis Mak 2014;14:9 [FREE Full text] [doi: 10.1186/1472-6947-14-9] [Medline: 24475790]

35. Woeltje KF, Butler AM, Goris AJ, Tutlam NT, Doherty JA, Westover MB, et al. Automated surveillance for central line-associated bloodstream infection in intensive care units. Infect Control Hosp Epidemiol 2008 Sep;29(9):842-846. [doi: 10.1086/590261] [Medline: 18713052]

36. van Mourik MS, Troelstra A, van Solinge WW, Moons KG, Bonten MJ. Automated surveillance for healthcare-associated infections: Opportunities for improvement. Clin Infect Dis 2013 Jul;57(1):85-93 [FREE Full text] [doi: 10.1093/cid/cit185] [Medline: 23532476]

37. Woeltje KF, Lin MY, Klompas M, Wright MO, Zuccotti G, Trick WE. Data requirements for electronic surveillance of healthcare-associated infections. Infect Control Hosp Epidemiol 2014 Sep;35(9):1083-1091. [doi: 10.1086/677623] [Medline: 25111915]

38. Hajdu A, Eriksen HM, Sorknes NK, Hauge SH, Loewer HL, Iversen BG, et al. Evaluation of the national surveillance system for point-prevalence of healthcare-associated infections in hospitals and in long-term care facilities for elderly in Norway, 2002-2008. BMC Public Health 2011;11:923 [FREE Full text] [doi: 10.1186/1471-2458-11-923] [Medline: 22165849]

39. Tseng YJ, Chen YC, Lin HC, Wu JH, Chen MY, Lai F. A web-based hospital-acquired infection surveillance information system. New York: IEEE; 2010 Oct Presented at: 10th IEEE International Conference on Information Technology and Applications in Biomedicine (ITAB); Nov 3-5, 2010; Corfu, Greece p. 1-4. [doi: 10.1109/ITAB.2010.5687808]

Abbreviations

CABSI: central line-associated bloodstream infection
CDC: Disease Control and Prevention
CRBSI: central line-related bloodstream infection
EMRs: electronic medical records
HABSI: health care-associated bloodstream infection
HAIs: health care-associated infections
HIS: hospital information systems
HL7: Health Level Seven
ICP: infection control personnel
LISs: laboratory information systems
MDROs: multidrug-resistant organism
NHSN: National Health Care Safety Network
NPV: negative predictive value
NTUH: National Taiwan University Hospital
PPV: positive predictive value
SOA: service-oriented architecture

XSL•FO
RenderX

Original Paper

Meaningful Use of Electronic Health Records: Experiences From the Field and Future Opportunities

Sarah Patricia Slight[1,2], MPharm, PhD, PGDip; Eta S Berner[3], EdD; William Galanter[4], MS, MD, PhD; Stanley Huff[5,6], MD; Bruce L Lambert[7], PhD; Carole Lannon[8,9], MD, MPH; Christoph U Lehmann[10], MD, FAAP, FACMI; Brian J McCourt[11], BA; Michael McNamara[12], MD; Nir Menachemi[13], MPH, PhD; Thomas H Payne[14], MD; S Andrew Spooner[15], MD, MS, FAAP; Gordon D Schiff[2,16], MD; Tracy Y Wang[11], MD, MHS, MSc; Ayse Akincigil[17], PhD; Stephen Crystal[17], PhD; Stephen P Fortmann[18], MD; Meredith L Vandermeer[18], MPH; David W Bates[2,16], MD, MSc

[1]Division of Pharmacy, School of Medicine Pharmacy and Health, Durham University, Durham, United Kingdom

[2]The Centre for Patient Safety Research and Practice, Division of General Internal Medicine, Brigham and Womens Hospital, Boston, MA, United States

[3]Center for Health Informatics for Patient Safety/Quality, Department of Health Services Administration, University of Alabama at Birmingham, Birmingham, AL, United States

[4]University of Illinois at Chicago, Chicago, IL, United States

[5]Department of Biomedical Informatics, University of Utah, Salt Lake City, UT, United States

[6]Intermountain Healthcare, Salt Lake City, UT, United States

[7]Center for Communication and Health, Department of Communication Studies, Northwestern University, Chicago, IL, United States

[8]James M Anderson Center for Health Systems Excellence, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, United States

[9]Department of Pediatrics, University of Cincinnati, Cincinnati, OH, United States

[10]Pediatrics and Biomedical Informatics, Vanderbilt University, Nashville, TN, United States

[11]Duke Clinical Research Institute, Duke University, Durham, NC, United States

[12]Departments of Pediatrics and Medical Informatics (DMI), Northwest Permanente, Portland, OR, United States

[13]Richard M Fairbanks School of Public Health, Indiana University, Indianapolis, IN, United States

[14]Department of Medicine, University of Washington, Seattle, WA, United States

[15]Division of Biomedical Informatics, Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, OH, United States

[16]Harvard Medical School, Boston, MA, United States

[17]Institute for Health, Health Care Policy, and Aging Research, Rutgers, The State University of New Jersey, New Brunswick, NJ, United States

[18]Center for Health Research, Kaiser Permanente Northwest, Portland, OR, United States

Corresponding Author:
David W Bates, MD, MSc
The Centre for Patient Safety Research and Practice
Division of General Internal Medicine
Brigham and Womens Hospital
1620 Tremont Street
Boston, MA, MA 02120-1613
United States
Phone: 1 617 732 5650
Fax: 1 617 732 7072
Email: dbates@partners.org

Related Article:

This is a corrected version. See correction statement: http://medinform.jmir.org/2015/3/e32/

Abstract

Background: With the aim of improving health care processes through health information technology (HIT), the US government has promulgated requirements for "meaningful use" (MU) of electronic health records (EHRs) as a condition for providers

XSL•FO
RenderX

receiving financial incentives for the adoption and use of these systems. Considerable uncertainty remains about the impact of these requirements on the effective application of EHR systems.

Objective: The Agency for Healthcare Research and Quality (AHRQ)-sponsored Centers for Education and Research in Therapeutics (CERTs) critically examined the impact of the MU policy relating to the use of medications and jointly developed recommendations to help inform future HIT policy.

Methods: We gathered perspectives from a wide range of stakeholders (N=35) who had experience with MU requirements, including academicians, practitioners, and policy makers from different health care organizations including and beyond the CERTs. Specific issues and recommendations were discussed and agreed on as a group.

Results: Stakeholders' knowledge and experiences from implementing MU requirements fell into 6 domains: (1) accuracy of medication lists and medication reconciliation, (2) problem list accuracy and the shift in HIT priorities, (3) accuracy of allergy lists and allergy-related standards development, (4) support of safer and effective prescribing for children, (5) considerations for rural communities, and (6) general issues with achieving MU. Standards are needed to better facilitate the exchange of data elements between health care settings. Several organizations felt that their preoccupation with fulfilling MU requirements stifled innovation. Greater emphasis should be placed on local HIT configurations that better address population health care needs.

Conclusions: Although MU has stimulated adoption of EHRs, its effects on quality and safety remain uncertain. Stakeholders felt that MU requirements should be more flexible and recognize that integrated models may achieve information-sharing goals in alternate ways. Future certification rules and requirements should enhance EHR functionalities critical for safer prescribing of medications in children.

Introduction

The Health Information Technology for Economic and Clinical Health (HITECH) Act was signed into law on February 17, 2009 with the commitment of substantial financial resources to expand the use of electronic health records (EHRs) and great hopes of promoting improvements in the efficiency of health care for all Americans. This effort is being led by the Centers for Medicare & Medicaid Services (CMS) and the Office of the National Coordinator for Health Information Technology (ONC) at the US Department of Health and Human Services (HHS) [1,2]. As a condition for clinicians and hospitals to receive incentive payments, they needed to use certified EHRs in a meaningful manner (ie, "meaningful use" [MU]). More specifically, this involved using EHRs to improve quality, safety, and efficiency; reduce health disparities; engage patients and family in their health; improve care coordination and population and public health; and maintain privacy and security of patient health information [3].

The CMS EHR incentive programs have included 3 stages to date, each with its own specific objectives, measures, and standards. The final rules for MU Stage 1, which specify the criteria that eligible professionals and hospitals need to meet to qualify for incentives, went into effect on September 26, 2010. The rules defined 15 core and 10 menu-set objectives that focused on providers capturing and sharing patient data. The onset of MU Stage 2 criteria was delayed until 2014 and concentrated on advanced clinical processes and more rigorous health information exchange (HIE). Specific to the Stage 2 objectives was the expectation that patients would be provided with secure online access to their health information. The Health Information Technology Policy Committee (HITPC), which advises the government on its EHR incentive program, submitted

its preliminary recommendations for MU Stage 3 to the ONC in early 2013. As part of the federal rule-making process, these preliminary Stage 3 recommendations were released for public comment and generated a high volume of responses [4]. These responses play a key role in informing the future direction of MU and related health information technology (HIT) policies, with Stage 2 now extended through to 2016 and Stage 3 scheduled to begin in 2017. At this time, relatively little has been published about professionals' experiences with implementing the Stage 2 requirements. A large number of these core measures are associated with the entering, recording, or ordering of medicines. Our goals were to examine critically the impact of MU to date, both experiences with Stage 2 and reactions to Stage 3 recommendations, with a particular focus on medication requirements along with related broader policy and implementation issues. We used this information to develop a set of recommendations to help inform future policies.

Methods

We gathered the perspectives of a wide range of professionals (N=35) representing academicians, practitioners, policy makers, and senior management officials identified through the CERTs, henceforth referred to as "stakeholders." Stakeholders initially met in June 2014 as part of the national CERT steering committee meeting to discuss the purpose and content of this document and included representatives from different health care and academic organizations including: Agency for Healthcare Research and Quality (AHRQ) (n=5), Kaiser Permanente (n=4), Brigham and Women's Hospital (n=3), Cincinnati Children's Hospital Medical Center (n=2), Food and Drug Administration (FDA) (n=2), Duke University (n=3), Rutgers University (n=2), University of Alabama at Birmingham (n=2), Intermountain Healthcare (n=1), University of Illinois

at Chicago (n=1), Northwestern University (n=1), University of Washington (n=1), University of Maryland (n=1), Baylor Scott and White Health (n=1), Baylor College of Medicine (n=1), Blue Cross Blue Shield Association (n=1), and a variety of others (n=4). A number of stakeholders occupied roles such as Chief Medical Information Officer or Chief Medical Informatics Officer in their respective health care organizations. A number of open-ended questions were posed to the group including:

- What were your experiences of implementing Stage 2 MU requirements?
- What key challenges did you face?
- How were these challenges overcome (or could they be overcome in the future)?
- What are your thoughts on the proposed Stage 3 recommendations?
- Do you think there were any important areas omitted in the proposed Stage 3 recommendations?

Specific issues and recommendations were presented, discussed, and agreed on as a group. Some of these issues that were agreed on by the stakeholders have been documented and supported by relevant literature. We used the principles of consensus decision making; all stakeholders were (1) involved in the group discussions (inclusive), (2) encouraged to contribute opinions and suggestions (participatory), (3) given the opportunity to build on one another's suggestions (collaborative), (4) afforded equal input into the process (egalitarian), and (5) allowed to voice any particular concerns that they may have so that the group could incorporate them into the emerging domains (cooperative). These include, for example, how organizational differences in the delivery of health care could impact stakeholders' ability to achieve MU requirements, challenges and opportunities for rural communities, and how EHRs could be improved to support safer and more effective prescribing for children. The public commentary available on the government website was reviewed to help inform these discussions [4]. A summary of the key findings were presented to the group as an oral presentation (via a webinar) in January 2015 and a concerted attempt was made to reach full agreement on the key domains (principle of agreement seeking). All stakeholders had the opportunity to provide feedback both verbally and electronically, and all feedback was incorporated. The stakeholders were convened for a second face-to-face meeting at the start of March 2015 and gave their final approval to the manuscript's content and recommendations. All authors listed on this manuscript participated in these meetings. In the sections that follow, we discuss these 6 domains, which include some of the key objective(s) on which the HITPC requested comment and the HITPC identification number to facilitate cross-referencing.

Results

Accurate Medication Lists and Medication Reconciliation

When a patient is transferred from one health care setting or provider to another, it is essential that accurate and up-to-date information about the patient's medications be provided. This enables health care professionals responsible for the patient's care to identify any medication changes or discrepancies between the prior and current medication lists. This process of medication reconciliation helps health care providers make informed decisions and safely monitor their patients' care [5]. A Stage 2 core measure recommended that medication reconciliation be performed for more than 50% of patients transitioning into the care of the eligible provider or admitted to the eligible hospital's or Critical Access Hospital's (CAHs) inpatient or emergency department (SGRP 302). However, the consensus of the stakeholder group was that this process of medicine reconciliation is very important and requires attention. The quality and accuracy of these medication lists are often poor and providing patients with medication lists that are of dubious quality (due to missing, duplicated, or inaccurate prescription information) can pose a risk to patient safety. Medication lists can also fall short, for example, by excluding important information critical to pediatric dosing, such as the intended weight-based dose, adjustments made based on gestational age, and dose rounding. As part of the medication reconciliation process, prescribers and nonprescribers (eg, medical assistants) are now entering medical information about patient medications such as a report that the patient is not taking a drug. This "not taking" data element fails to capture whether the drug has or has not been prescribed or discontinued, or whether the patient is choosing not to take the medication. The ambiguity in the meaning of the data element "not taking" introduces considerable variation in how individuals handle this information in the EHR system and raises questions on how the quality of this process would be measured or monitored.

Better electronic tools are needed to assist with this medication reconciliation process [6]. For example, 3 stakeholders highlighted how Partners Healthcare developed an electronic postdischarge tool that presents the ambulatory EHR medication list (preadmission) alongside the discharge medication list on the same screen with all differences in dose or frequency highlighted [7]. Medications can then be efficiently added to, updated, or deleted from the EHR medication list. The primary care provider could also "verify" that a medication was up-to-date, thus helping other clinicians judge the accuracy of medication information. This electronic tool is one example of automated approaches that could more actively involve the primary care provider and improve patient safety at the transition from hospital to primary care.

A Stage 3 recommendation was that EHR systems should provide functionality to help maintain an up-to-date accurate medication list (SGRP 106); therefore, the incorporation of external data, such as pharmacy dispense status notifications, into vendor EHR systems was proposed for a future stage of MU (SGRP 125). These data could better inform users as to whether a patient had their prescription(s) filled, was taking 2 kinds of the same drug (including detection of abuse), or was using multiple drugs whose indications overlap. All stakeholders agreed that such needed interoperability poses additional challenges related to data validity, reliability, and integrity, and concerns about the willingness, timing, and ability of pharmacies to make these data available electronically.

One specific recommendation from the stakeholder group was that medication cancelations should be transmitted to pharmacies. This is often done in the inpatient setting, but it is not done in the outpatient setting, although a standard does exist. If this were done, it could help resolve many discrepancies in medication reconciliation.

Accurate Problem Lists and the Shift in Health Information Technology Priorities

An accurate list of a patient's problems and allergies represents a key component of the patient's EHR. Problem lists contain a list of patients' problems or diagnoses and may be used by clinicians to familiarize themselves with the needs of a patient and orient caregivers to the reasons why a patient may be on a particular medication or regimen. If a problem is properly documented in a patient's EHR, their clinician can receive appropriate alerts and reminders to guide care. The problem list also helps primary care practices to correctly identify disease-specific populations and create patient registries ensuring that all patients benefit from the most up-to-date evidence-based care.

The MU Stage 3 recommendations expanded the scope of reconciliations to include those of medication allergies and problems (SGRP 302). Many stakeholders recognized the importance of obtaining patients' input on the accuracy of problem lists (SGRP 105) in the process of reconciliation. However, concerns were raised by the stakeholder group about whether and how patients should contribute to the same up-to-date problem list as clinicians and, if so, whether this may confuse and possibly interfere with the credibility of the list [4]. Some patients do not actually have any active problems and stakeholders grappled with the need to distinguish the explicit absence of a problem from the situation in which a problem may exist but was not entered (or does not fit the criteria that CMS has determined for what constitutes a problem). For example, one stakeholder highlighted how Intermountain Healthcare had asked their physicians to enter problems or "no problems" in the chart to comply with MU, but in actual use, many items on the problem list were not "problems" according to CMS rules and so "no CMS problems" was entered instead. This proved confusing for clinicians to interpret.

One Stage 2 core measure recommended maintaining an up-to-date problem list of current and active diagnoses (SGRP 105) and a medication allergy list (SGRP 107). Stage 3 recommendations expand on these basic requirements proposing that EHR systems should also provide functionality to help keep both problem and allergy lists accurate and up-to-date. One stakeholder explained how the University of Washington has developed new functionality using natural language processing to help achieve this objective for EHR problem lists. However, because of the burden of complying with MU requirements, other work that was not directly tied to MU incentives was postponed or halted. For example, before the launch of the MU incentive program, there were active clinical decision support (CDS) initiatives on-going at the University of Washington for the early detection of sepsis, identifying non-ICU patients at risk of clinical deterioration, complying with guidelines to reduce ventilator-associated pneumonia, venous thromboembolism, and other complications of ICU care—all leading causes of patient harm. However, to meet MU requirements, work on these projects was deferred and the clinical analysts, engineers, and senior programming staff were redirected to work on implementing MU requirements. One stakeholder reported a similar stifling of innovation at Intermountain Healthcare, where the implementation of MU capabilities delayed other EHR development projects, such as the replacement of legacy system functionality in labor and delivery, electronic consent handling, clinical HIE workflow integration, and replacement/enhancement of inpatient computerized provider order entry (CPOE) functionality. The consensus of the stakeholder group was that this might represent an opportunity cost for innovation. Institutions understandably may place priority on innovations that will bring known rewards, even if the innovations would not be as high a priority if there were no incentives. These unintended consequences of the MU incentives can be instructive to consider as other "pay for performance" programs are initiated.

The definition of CPOE by CMS is "a provider's use of computer assistance to directly enter medical orders (eg, medications) from a computer or mobile device" [8]. The Stage 3 MU measure recommended 60% of medication orders and 60% of laboratory and radiology orders (as opposed to 30% in Stage 2 MU) be recorded by the eligible or authorized provider using CPOE. Stakeholders supported the inclusion of drug-drug interaction (DDI) checking in CPOE systems for "never" combinations (SGRP 101)—combinations that have the potential for severe adverse effects if prescribed together. Questions frequently arose about who would create and maintain such an externally vetted list of DDI alerts for "never" combinations. Two stakeholders suggested that the creation of a national knowledge base, which is managed centrally, might be one possible option to consider so that each organization does not have to individually reinvent the wheel. However, most stakeholders felt the overall utility of DDI alerts was mixed because of a plethora of what clinicians perceived were "nuisance alerts" that they mostly ignored. All alerts need to be implemented thoughtfully with careful attention paid to the balance between sensitivity and specificity, how the alerts are delivered to providers, how intrusive they are to provider workflow, and the provider's clinical specialty and patient population. Stakeholders felt that organizations should be allowed flexibility in managing DDI alert implementation to ensure that it does not result in too many false-positive warnings and inaccurate or trivial information, resulting in alert fatigue [4].

Accurate Allergy Lists and Allergy-Related Standards Development

Overview

Stage 2 and 3 recommendations require EHRs to maintain active medication allergy lists (SGRP 107). This includes functionality that codes medication allergies and links them to related drug family and code-related reactions. It known that an allergy that is entered as free text in the EHR is neither interoperable across clinical information systems nor easily usable for CDS applications, such as drug-allergy interaction checking.

However, the US government has not yet specified which standard terminologies should be used to structure and encode allergy information. The consensus of the stakeholder group was that defining allergy standards will be essential to facilitate both documentation and the exchange of information between health care settings [4]. One stakeholder highlighted how Goss et al [9] defined a set of desirable characteristics to assess allergy standards and terminologies, and conducted an analysis to examine the content coverage of each existing standard terminology within specific domains. Systemized Nomenclature of Medical Clinical Terms (SNOMED CT) was found to fulfill the greatest number of desirable characteristics, whereas RxNorm provided the most comprehensive coverage for representing drug allergens, followed by Unique Ingredient Identifier (UNII) and SNOMED CT. Unfortunately, no single terminology was found to be, by itself, a complete solution. SNOMED CT was the only terminology to contain concepts to represent "no known allergies." Failure to document positive findings may result in compliance issues and can potentially jeopardize patient safety [9]. There is a lack of validated outcome measures or service accreditation standards, which would allow improved measurement of the quality of allergy services provided [10]. The stakeholder group agreed that further work is needed to develop a common terminology model, which will reconcile overlapping concepts and terms.

Supporting Safer and More Effective Prescribing for Children

Stakeholders, especially those from the Cincinnati CERT that specializes in pediatric medication use, expressed concern about the lack of attention paid to pediatric prescribing in the MU criteria. Although Stage 2 and Stage 3 MU objectives required CPOE systems to be used for 60% of medication orders (SGRP 101), EHR functionalities to assist with the prescribing of medications for children have not been specifically mentioned or recommended. This is despite the fact that prescribing medicines for children is reported in the literature to carry disproportionately higher safety risks and be more error prone compared to prescribing for adults [11]. A child's continuously changing physiology [12] and limited ability to tolerate errors [13,14] require consideration of gestational age, actual age, weight, length, body surface area, and body mass index when prescribing drugs [15]. With almost one-quarter of the US population being children [16], it stands to reason that EHR functionalities should be developed and widely implemented to promote safer pediatric prescribing.

The American Academy of Pediatrics (AAP), AHRQ, and Health Level 7 (HL7) International, have described desirable functionalities for EHRs in pediatric populations. Major areas include immunization management, growth tracking, medication dosing, data norms, and privacy in special pediatric populations [17]. For safe prescribing, pediatric drug dosages are usually best calculated on the basis of body weight [18,19]. Stakeholders pointed out how it is possible for an EHR system to use this value to suggest doses or indeed request that the weight be updated or entered in the system if absent. EHR systems could also help minimize errors in computing of a volume of liquid for a particular dose and round it to a convenient volume to be administered by a caregiver. Because data norms and values (eg, body measurements and vital signs) change continuously with age, EHRs can also assist with the calculation and flagging of abnormal values. Furthermore, they can generate instructions to the pharmacy to dispense the drug in a particular way [17]. Textbox 1 lists EHR functionalities that stakeholders considered important in prescribing for children.

Textbox 1. Electronic health record functionalities that stakeholders considered important in prescribing for children.

- Weight-based/body surface-based dose calculations and range checks [14]

- Ability to detect erroneously entered weights [14,20,21]

- Display of patient specific units of measure (eg, grams) along with the data values [22]

- Rounding of medication doses to appropriate decimal precision with special consideration of the low-weight patients [23,24]

- Display of data that influenced the final dose and amount in the prescription, particularly to dispensing pharmacists [25]

- Display of normal pediatric dose ranges and advice when no pediatric references exist [26]

- Use pediatric dose ranges for alerts using patient weight/age with soft-stops for adult dose [27]

- Appropriate alerts for age correction for preterm infants, neonates, and low-weight patients [28]

- Recommendation of optimized dispensing format (liquid, tablet, etc) or concentration for the patient [22,29]

- Adolescent patients require a level of confidential care, especially when prescribing medications for reproductive or mental health issues [30,31]

Stage 3 recommendations propose a new measure that would require health care providers to generate and transmit discharge prescriptions electronically (SGRP 103). Although this objective may improve workflow for pediatric providers and reduce the risk of illegible handwriting and transcription errors, the stakeholder group felt that it does not focus on the decision support required to generate correct prescriptions and may simply enable faster generation and transmission of potentially erroneous orders. Current formats for electronic prescription messages do not include body weight or any details about the calculations that yielded the dose [32]. Thus, the consensus of the stakeholder group was that few of the Stage 2 requirements were aligned sufficiently with the functionalities considered critical for the accurate prescribing of medications in children and it was key that this issue be addressed in the development of future recommendations.

Challenges and Opportunities for Rural Communities

Awards totaling US $10 million were collectively granted to 5 domestic institutions to support HIT curriculum development in April 2010 to the University of Alabama at Birmingham, Johns Hopkins University, Columbia University, Duke University, and Oregon Health and Science University. Each of these Curriculum Development Centers was given responsibility to develop, revise, and share curriculum components covering a specific set of HIT content areas. The ultimate aim was to prepare future professionals to meet emerging workforce needs. Despite the initial HITECH funding for training, stakeholders felt that the needs of the HIT workforce in rural areas across the country have not been met yet. Rural communities are more likely to have smaller practices, which have been among the last to embrace electronic medical records [33]. They have fewer resources to both purchase EHRs and to hire and retain HIT support staff. The overall IT infrastructure in many of these areas (as in some low-resource urban areas) is poor, which makes it even more challenging to participate in the electronic information exchange. Thus, patients with complex conditions in rural communities may not benefit from the quality improvements that the MU incentives are designed to deliver.

According to the stakeholders, especially those from the University of Alabama at Birmingham CERT that specializes in workforce training, several steps have been taken to address these issues. In addition to the workforce training programs, 62 Regional Extension Centers (RECs) have been established with US $677 million in funding from the ONC to provide on-the-ground assistance to smaller rural practices. In 2011, the ONC announced an additional US $12 million in new technical support assistance to help CAHs and rural hospitals adopt and become meaningful users of certified HIT. This funding was in addition to the $20 million provided to RECs in September 2010 to provide technical assistance to the CAHs and rural hospitals [34]. In addition, University of Alabama at Birmingham and Columbia University collaborated with representatives from several of the other RECs to adapt the original training curriculum so that it would be better suited to the needs of rural and low-resource urban practices. In 2013, the Health Services and Resource Administration (HRSA) funded rural networks in 15 states to develop rural HIT workforce development programs to provide education, apprenticeships, and job placements in rural practices [35]. HRSA, AHRQ, and ONC have also developed resources, checklists, and toolkits to help sites unable to afford expensive outside consultation [36].

One stakeholder pointed out that as more hospitals and practices begin to meet the MU criteria, some of the traditional boundaries that have separated rural primary care practices from tertiary care centers in large urban areas may begin to disappear. Primary care practices may have more access to information about their patients' hospitals stays. Tertiary care hospitals are likely to have a substantial number of patients from surrounding rural areas who can benefit from patient portals or similar mechanisms to promote patient engagement (SGRP 204A). However, patient engagement is likely to be another challenge going forward with rural residents, considering unreliable Internet connections, low health literacy, and lack of resources. Although MU requirements currently set a low percentage of patients who are expected to use the portals, the consensus of the stakeholder group was that systems must be scalable if more patients are to benefit, which will likely entail use of novel technologies such as mobile devices.

Achieving Meaningful Use: Easier for Some Than for Others?

Many different stakeholders supported the MU general goal that providers should have appropriate information about patients transitioning into their care (SGRP 303). Stage 3 recommendations expanded on this Stage 2 objective by specifying the types of information that should be included in the summary care record, such as a concise narrative section, goals, instructions, and care team members. The consensus of the stakeholder group was that some organizations, such as Kaiser Permanente or Intermountain Healthcare, might find it easier to achieve this objective than others. Such well-established integrated delivery systems have organized, coordinated, and collaborative networks that bring together various health care providers to deliver coordinated care to a defined patient population [37]. They include primary and specialty outpatient care, as well as community and tertiary hospital services. The effective use of HIT is a key attribute of successful integrated delivery systems [37,38]. For example, in the case of Kaiser Permanente or the Veterans Affairs systems, the same longitudinal EHR is accessible and shared by both primary care physicians and specialists, thus facilitating the tracking of patients across the continuum [38]. Kaiser Permanente also has an integrated pharmacy system that is used for most patient prescriptions. One stakeholder highlighted how, for the past 20 years, Kaiser Permanente has had a bidirectional electronic HL7-based interface in place in their pharmacy systems, which has ensured that the information presented to their patients was consistent, whether they were engaged with clinical operations, outpatient pharmacy locations, or mail order pharmacy services. It also meant that the Stage 2 recommendation to generate and transmit permissible discharge prescriptions electronically (SGRP 103) was easily achievable for all eligible providers. However, this stakeholder also explained how other measures, such as Summary of Care documentation at time of transitions with external organizations, have required substantial resources to fund technical and operational change that has impacted less than 2% of Kaiser Permanente's patient population. Care should be taken to avoid MU requirements that are unnecessarily burdensome to mature, typically staff model systems that have historically been the leaders in integrated use of clinical information.

Another issue raised by a different stakeholder related to whether organizations are using existing functionality (eg, Surescripts) or have chosen to develop their own. Kaiser Permanente and other integrated delivery systems lacked the functionality to bring medication information from external pharmacies into their EHR system and were swayed by the MU incentives to add this to their systems. However, the value of this functionality within staff model systems such as Kaiser Permanente is likely to be low in light of the fact that Kaiser Permanente patients obtain nearly all their medications from Kaiser Permanente. Stakeholders agreed in principle that external interoperability

functionality can help maintain accurate medication and problem lists, although they felt that implementation should be flexibly based on the organizational-specific contexts. They also felt that many of the specific criteria should be postponed until the technological, operational, and legal issues are more fully evolved and the quality and accuracy of tools are sufficiently tested.

Finally, Stage 3 recommendations propose a new measure that would require health care providers to use CPOE for referrals/transition of care orders (SGRP 130). One stakeholder highlighted how some organizations, including Intermountain Healthcare, already use extensive CPOE/CDS capabilities and other advanced functionality and questioned the value of spending considerable resources to develop functionality that they believed would add little to their existing systems simply to meet MU requirements. For example, for the successful attestation of Stage 1, Intermountain Healthcare estimated that its 696 eligible professionals and 22 hospitals were eligible for approximately US $46.3 million. The high degree of coordination already inherent in their delivery model and IT systems meant that total costs for the implementation of Stage 1 recommendations were considerably lower than for others at an estimated US $17.3 million, resulting in a net revenue benefit of US $29 million. Although this financial benefit may seem substantial, another stakeholder pointed out how these total implementation costs may not reflect the "true" cost because they did not include the development of the system's computer network (in their case, this was already in existence) or the disruption caused by HIT implementations and upgrades. Therefore, the consensus of the stakeholder group was that it is important to understand the current structural advantages of existing integrated delivery systems in the achievement of MU objectives and to recognize the need for future MU requirements to be applied and interpreted more flexibly. Textbox 2 lists a summary of the key issues for each domain.

Textbox 2. A summary of the key issues in each domain.

---

1. Accurate Medication Lists and Medication Reconciliation

- The quality and accuracy of these medication lists is often poor and providing patients with medication lists that are of dubious quality can pose a risk to patient safety.

- Better electronic tools are needed to assist with this medication reconciliation process.

- The incorporation of external data, such as pharmacy dispense status notifications, into vendor EHR systems could better inform providers about a patient's medicines usage.

2. Accurate Problem Lists and the Shift in HIT Priorities

- EHR systems should also provide functionality to help keep both problem and allergy lists accurate and up-to-date.

- Institutions understandably may place priority on innovations that will bring known rewards, even if the innovations would not be as high a priority if there were no incentives.

- All CDS alerts need to be implemented thoughtfully with careful attention paid to the balance between sensitivity and specificity, and how the alerts are delivered to providers.

3. Accurate Allergy Lists and Allergy-Related Standards Development

- Defining allergy standards will be essential to facilitate both documentation and the exchange of information between health care settings.

4. Supporting Safer and More Effective Prescribing for Children

- Data norms and values change continuously with age and EHRs can assist with the calculation and flagging of abnormal values.

- Few Stage 2 requirements were aligned sufficiently with the functionalities considered critical for the accurate prescribing of medications in children and it was key that this issue be addressed in the development of future recommendations.

5. Challenges and Opportunities for Rural Communities

- Despite the initial HITECH funding for training, the needs of the HIT workforce in rural areas across the country have not been met.

- Patient engagement is likely to be challenge going forward with rural residents, considering unreliable Internet connections, low health literacy, and lack of resources.

- Although MU requirements currently set a low percentage of patients who are expected to use the portals, systems must be scalable if more patients are to benefit, which will likely entail use of novel technologies such as mobile devices.

6. Achieving MU: Easier for Some Than for Others?

- Some MU measures have been easily achievable for integrated delivery systems, whereas other measures have required substantial resources to fund and have impacted only a small portion of their patient population.

- Future MU requirements need to be applied and interpreted more flexibly.

Discussion

We assessed stakeholders' learning and experiences from the implementation of MU requirements over the past 4 years, with a particular focus on medication requirements and attempted to identify problem areas where midcourse corrections might be helpful. Six specific issues were highlighted, all of which present opportunities for improvement. The implementation of MU capabilities was reported to have stifled innovation at some organizations. This appears to run counter to the ONC's goal of encouraging innovation and creating "an environment of testing, learning, and improving, thereby fostering breakthroughs that quickly and radically transform health care" [39]. The challenge in many organizations was that resources were largely focused on implementing basic MU criteria and diverted away from addressing other meaningful local problems and creating innovative solutions.

Likewise, although the EHR incentive program was viewed as a valuable opportunity to encourage provider-level clinical quality measure (CQM) innovation and perform provider-level CQM testing, some stakeholders felt it distracted them at least temporarily from their efforts to develop and implement such quality measurement and improvement systems. The HITPC also raised the possibility of allowing health care organizations to submit a locally developed CQM as a menu item, in lieu of one of the existing measures specified in the MU program [3]. Health care organizations may find this difficult to achieve, especially if their clinical analysts, engineers, and senior programming staff are focused on achieving MU requirements rather than other EHR development projects. Furthermore, the Food and Drug Administration Safety and Innovation Act (FDASIA) working group was clear that any new HIT regulatory framework should promote innovation rather than stifle it [40]. This FDASIA working group recommended more local HIT configuration and integration, as well as more control and accountability for outcomes of use. A greater emphasis should be placed on local HIT configuration that addresses population health needs. Thus, MU requirements will need to change and evolve over the next few years to achieve this broader and more flexible orientation. The concerns we have identified have spurred the following recommendations:

1. Definitions of transitions in care should enable and support shared patient record systems. Better tools and interoperability with external data are needed for effective and efficient medication reconciliation. On the other hand, measures should not drive unnecessary or unreliable data transmission.
2. Development of a common terminology model is needed to facilitate documentation and encoding of key data elements, notably patient allergies.
3. Future MU certification rules and requirements should consider EHR functionalities that are critical, but often lacking, for the accurate prescribing of medications in children.
4. Future MU requirements should put more emphasis on flexibly understanding, incorporating, and supporting local HIT configurations that address population health needs.

5. The MU objectives should acknowledge the diversity of health care systems. For example, integrated delivery systems are more likely to achieve the goal of information sharing because of their integrated structure, greater functionality, and improved interoperability. From the policy perspective, this could be handled by offering exceptions or alternate routes for qualification.

The sampling strategy used in this study ensured that the perspectives of highly knowledgeable informants from the 5 AHRQ-sponsored CERTs were captured. Our sample included those directly involved in the implementation of MU criteria (eg, Chief Medical Information Officer or Chief Medical Informatics Officer) and those who were knowledgeable about, but not directly involved in, the day-to-day implementation work (eg, academicians, practitioners, policy makers). Participants were free to raise any issues that they felt were relevant to the topic under discussion. Consequently, we believe that the information gathered was reflective of genuine concerns and views. All stakeholders were given an opportunity to provide feedback on the key domains, ensuring that the conclusions accurately reflected the opinions and views collected. A limitation of this study was that it was performed in the US context and, therefore, could be viewed as less applicable to other countries. However, we believe that the implementation and adoption of EHRs is highly heterogeneous across health care systems and countries, and will be of interest.

The future course that the federal government will take with respect to HIT and policy measures is uncertain. It is not clear whether there will be a fourth stage of MU, although that currently seems unlikely. Taking stock of the important ways MU has been successful in achieving many of its objectives—such as dramatically increasing the number of medications ordered electronically—as well as where it encountered predicted and unanticipated problems, will be critical to mapping the next steps. Overall, the incentives and specific MU criteria will almost certainly be less important than they have been in the future as information systems more broadly improve their functionality and many of the challenges that we face today become embedded as the standard of care. It does appear that certification will continue to be important, although providers have recently called for separating MU from certification [41]. The ONC will likely continue to (appropriately) maintain its "bully pulpit" role in helping to encourage and accelerate the development of standards and interoperability among other needs. Finally, it appears likely that a national Center for HIT Safety will be established, a development many of the CERT stakeholders welcomed, especially given the valuable role CERTs have historically played in the coordination of national medication improvement efforts [39].

Regardless, we believe it will be important for the federal government to address some of the issues we have identified in this paper, including problems with how medication reconciliation is being promoted, the issues around accurate problem lists and the shift in HIT priorities, supporting safer and effective prescribing for children and rural communities, and making achieving MU more likely to result in the care improvement desired by all stakeholders. Any new policy will

introduce new problems and it is essential for the federal government and others to consider how best to address these issues and others through the MU incentive program.

Authors' Contributions

DWB conceived this manuscript. SPS reviewed the public commentary available on the government website and led the writing of this manuscript with TYW, SAS, CUL, NM, ESB, BG, BLL, THP, SH, and MM contributing. All coauthors commented on subsequent drafts and gave their approval for the final version to be published. SPS and DWB acts as guarantors.

Conflicts of Interest

Dr Bates and Dr Lehmann have served as members of the HIT Policy Committee. Dr Huff has served as a member of the HIT Standards Committee.

References

1. HealthIT.gov. 2015. Health IT regulations: meaningful use regulations URL: http://www.healthit.gov/policy-researchers-implementers/meaningful-use-regulations [accessed 2015-05-24] [WebCite Cache ID 6Ylk9mfv2]
2. Blumenthal D, Tavenner M. The "meaningful use" regulation for electronic health records. N Engl J Med 2010 Aug 5;363(6):501-504. [doi: 10.1056/NEJMp1006114] [Medline: 20647183]
3. Department of Health and Human Services, Office of the National Coordinator for Health Information Technology. Request for Comment Regarding the Stage 3 Definition of Meaningful Use of Electronic Health Records (EHRs). 2013 Jan 14. URL: http://www.healthit.gov/sites/default/files/hitpc_stage3_rfc_final.pdf [accessed 2015-09-13] [WebCite Cache ID 6bWMjeeEY]
4. regulations.gov. Your voice in federal decision-making URL: http://www.regulations.gov/ [accessed 2015-04-08] [WebCite Cache ID 6XdQbcd1z]
5. EHR Incentive Program. Eligible Professional Meaningful Use Menu Set Measures Measure 6 of 9. 2014 May. Medication reconciliation URL: http://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/downloads/7_Medication_Reconciliation.pdf [accessed 2015-04-08] [WebCite Cache ID 6XdRfFrnu]
6. Schnipper JL, Hamann C, Ndumele CD, Liang CL, Carty MG, Karson AS, et al. Effect of an electronic medication reconciliation application and process redesign on potential adverse drug events: a cluster-randomized trial. Arch Intern Med 2009 Apr 27;169(8):771-780. [doi: 10.1001/archinternmed.2009.51] [Medline: 19398689]
7. Schnipper JL, Liang CL, Hamann C, Karson AS, Palchuk MB, McCarthy PC, et al. Development of a tool within the electronic medical record to facilitate medication reconciliation after hospital discharge. J Am Med Inform Assoc 2011 May 1;18(3):309-313 [FREE Full text] [doi: 10.1136/amiajnl-2010-000040] [Medline: 21486889]
8. The Centers for Medicare & Medicaid Services. Stage 2: Eligible Hospital and Critical Access Hospital Meaningful Use Core Measures Measure 1 of 16. 2012 Oct. URL: http://www.cms.gov/site-search/search-results.html?q=Eligible%20Hospital%20and%20Critical%20Access%20Hospital%20Meaningful%20Use%20Core%20Measures%20Measure%201%20of%2016.%20Stage%202 [accessed 2015-04-08] [WebCite Cache ID 6XdSm5XDp]
9. Goss FR, Zhou L, Plasek JM, Broverman C, Robinson G, Middleton B, et al. Evaluating standard terminologies for encoding allergy information. J Am Med Inform Assoc 2013;20(5):969-979 [FREE Full text] [doi: 10.1136/amiajnl-2012-000816] [Medline: 23396542]
10. Royal College of Physicians. 2010. Allergy services: still not meeting the unmet need URL: https://www.rcplondon.ac.uk/sites/default/files/documents/allergy-services-still-not-meeting-the-unmet-need.pdf [accessed 2015-05-24] [WebCite Cache ID 6Yllb4yT1]
11. Gerstle RS, Lehmann CU, American Academy of Pediatrics Council on Clinical Information Technology. Electronic prescribing systems in pediatrics: the rationale and functionality requirements. Pediatrics 2007 Jun;119(6):e1413-e1422. [doi: 10.1542/peds.2007-0889] [Medline: 17545368]
12. Hughes RG, Edgerton EA. Reducing pediatric medication errors: children are especially at risk for medication errors. Am J Nurs 2005 May;105(5):79-80, 82, 85 passim. [Medline: 15867545]
13. Stavroudis TA, Miller MR, Lehmann CU. Medication errors in neonates. Clin Perinatol 2008 Mar;35(1):141-61, ix. [doi: 10.1016/j.clp.2007.11.010] [Medline: 18280880]
14. Kim GR, Lehmann CU, Council on Clinical Information Technology. Pediatric aspects of inpatient health information technology systems. Pediatrics 2008 Dec;122(6):e1287-e1296. [doi: 10.1542/peds.2008-2963] [Medline: 19047228]
15. Lowry S, Quinn M, Ramaiah M, Brick D, Patterson E, Zhang J, et al. A Human Factors Guide to Enhance EHR Usability of Critical User Interactions When Supporting Pediatric Patient Care. 2012 Jun. URL: http://www.nist.gov/healthcare/usability/upload/NIST-IR-7865.pdf [accessed 2015-09-13] [WebCite Cache ID 6bWNQboed]

XSL·FO
RenderX

16.   KIDS COUNT data center. 2014. Total population by child and adult populations URL: http://datacenter.kidscount.org/data/tables/99-total-population-by-child-and-adult [accessed 2015-04-08] [WebCite Cache ID 6XdSuHso0]

17.   Spooner SA, Council on Clinical Information Technology, American Academy of Pediatrics. Special requirements of electronic health record systems in pediatrics. Pediatrics 2007 Mar;119(3):631-637. [doi: 10.1542/peds.2006-3527] [Medline: 17332220]

18.   Maat B, Au YS, Bollen CW, van Vught AJ, Egberts TC, Rademaker CM. Clinical pharmacy interventions in paediatric electronic prescriptions. Arch Dis Child 2013 Mar;98(3):222-227. [doi: 10.1136/archdischild-2012-302817] [Medline: 23184352]

19.   Ginzburg R, Barr WB, Harris M, Munshi S. Effect of a weight-based prescribing method within an electronic health record on prescribing errors. Am J Health Syst Pharm 2009 Nov 15;66(22):2037-2041. [doi: 10.2146/ajhp080331] [Medline: 19890088]

20.   Selbst SM, Fein JA, Osterhoudt K, Ho W. Medication errors in a pediatric emergency department. Pediatr Emerg Care 1999 Feb;15(1):1-4. [Medline: 10069301]

21.   Condren M, Studebaker IJ, John BM. Prescribing errors in a pediatric clinic. Clin Pediatr (Phila) 2010 Jan;49(1):49-53. [doi: 10.1177/0009922809342459] [Medline: 19643978]

22.   Croskerry P, Shapiro M, Campbell S, LeBlanc C, Sinclair D, Wren P, et al. Profiles in patient safety: medication errors in the emergency department. Acad Emerg Med 2004 Mar;11(3):289-299 [FREE Full text] [Medline: 15001413]

23.   Patterson ES, Zhang J, Abbott P, Gibbons MC, Lowry SZ, Quinn MT, et al. Enhancing electronic health record usability in pediatric patient care: a scenario-based approach. Jt Comm J Qual Patient Saf 2013 Mar;39(3):129-135. [Medline: 23516763]

24.   Johnson KB, Lee CK, Spooner SA, Davison CL, Helmke JS, Weinberg ST. Automated dose-rounding recommendations for pediatric medications. Pediatrics 2011 Aug;128(2):e422-e428 [FREE Full text] [doi: 10.1542/peds.2011-0760] [Medline: 21788218]

25.   Johnson KB, Ho Y, Cala CM, Davison C. Showing your work: impact of annotating electronic prescriptions with decision support results. J Biomed Inform 2010 Apr;43(2):321-325 [FREE Full text] [doi: 10.1016/j.jbi.2009.11.008] [Medline: 19995617]

26.   Kirkendall ES, Spooner SA, Logan JR. Evaluating the accuracy of electronic pediatric drug dosing rules. J Am Med Inform Assoc 2014 Feb;21(e1):e43-e49 [FREE Full text] [doi: 10.1136/amiajnl-2013-001793] [Medline: 23813541]

27.   Scharnweber C, Lau BD, Mollenkopf N, Thiemann DR, Veltri MA, Lehmann CU. Evaluation of medication dose alerts in pediatric inpatients. Int J Med Inform 2013 Aug;82(8):676-683. [doi: 10.1016/j.ijmedinf.2013.04.002] [Medline: 23643148]

28.   D'Agostino J, Gerdes M, Hoffman C, Manning M, Phalen A, Bernbaum J. Provider use of corrected age during health supervision visits for premature infants. J Pediatr Health Care 2013;27(3):172-179.

29.   Zimmer KP, Miller MR, Lee BH, Miller RE, Lehmann CU. Electronic narcotic prescription writer: use in medical error reduction. J Patient Saf 2008;4(2):98-105.

30.   Anoshiravani A, Gaskin GL, Groshek MR, Kuelbs C, Longhurst CA. Special requirements for electronic medical records in adolescent medicine. J Adolesc Health 2012 Nov;51(5):409-414. [doi: 10.1016/j.jadohealth.2012.08.003] [Medline: 23084160]

31.   Britto MT, Tivorsak TL, Slap GB. Adolescents' needs for health care privacy. Pediatrics 2010 Dec;126(6):e1469-e1476 [FREE Full text] [doi: 10.1542/peds.2010-0389] [Medline: 21098148]

32.   National Council for Prescription Drug Programs. 2014. Standards information URL: http://www.ncpdp.org/Standards [accessed 2015-05-24] [WebCite Cache ID 6YlnxjggO]

33.   National Rural Health Association. The future of rural health URL: http://www.ruralhealthweb.org/index.cfm?objectid=EAB2AE78-3048-651A-FE4CBFB6C083F34F., [WebCite Cache ID 6XdS5vIBh]

34.   HealthcareITNews. 2011 Feb 08. ONC awards additonal $12M to 48 RECs for EHR adoption URL: http://www.ruralhealthweb.org/index.cfm?objectid=EAB2AE78-3048-651A-FE4CBFB6C083F34F., [accessed 2015-05-24] [WebCite Cache ID 6YlmrBrbT]

35.   PR Newswire. 2013 Sep 12. HHS awards $5.3 million to boost health IT for rural areas URL: http://www.prnewswire.com/news-releases/hhs-awards-53-million-to-boost-health-it-for-rural-areas-223471081.html [accessed 2015-04-08] [WebCite Cache ID 6XdUzXxKY]

36.   Agency for Healthcare Research Quality. Appendix B: Getting Ready: A Planning Checklist for Rural and Community Hospitals Considering Implementing Health IT. 2012 Oct. URL: https://healthit.ahrq.gov/sites/default/files/docs/page/AppendixB.pdf [accessed 2015-09-13] [WebCite Cache ID 6bWOHGN6q]

37.   Enthoven AC. Integrated delivery systems: the cure for fragmentation. Am J Manag Care 2009 Dec;15(10 Suppl):S284-S290 [FREE Full text] [Medline: 20088632]

38.   Maeda J, Lee K, Horberg M. Comparative health systems research among Kaiser Permanente and other integrated delivery systems: a systematic literature review. Perm J 2014;2014(3):18.

39.   Office of the National Coordinator for Health Information Technology. Federal Health Information Technology Strategic Plan 2011-2015. 2015. URL: http://www.healthit.gov/sites/default/files/utility/final-federal-health-it-strategic-plan-0911.pdf [accessed 2015-05-24] [WebCite Cache ID 6Ylnl5DgR]

40.   Slight S, Bates D. A risk-based regulatory framework for health IT: recommendations of the FDASIA working group. J
       Am Med Inform Assoc 2014:1-4.
41.   California Medical Association. 2015 Jan 23. AMA asks feds to decouple EHR certification from meaningful use URL:
       http://www.cmanet.org/news/detail/?article=ama-asks-feds-to-decouple-ehr-certification [accessed 2015-05-24] [WebCite
       Cache ID 6YlnZzcTX]

Abbreviations

CAH: Critical Access Hospital
CDS: clinical decision support
CERT: Centers for Education and Research in Therapeutics
CPOE: computerized provider order entry
CQM: clinical quality measure
DDI: drug-drug interaction
EHR: electronic health records
FDASIA: Food and Drug Administration Safety and Innovation Act
HIE: health information exchange
HIT: health information technology
HITECH: Health Information Technology for Economic and Clinical Health
HITPC: Health Information Technology Policy Committee
HRSA: Health Services and Resource Administration
MU: meaningful use
REC: Regional Extension Centers
UNII: Unique Ingredient Identifier

<u>Original Paper</u>

Building Data-Driven Pathways From Routinely Collected Hospital Data: A Case Study on Prostate Cancer

Joao H Bettencourt-Silva[1,2]; Jeremy Clark[3]; Colin S Cooper[3]; Robert Mills[2]; Victor J Rayward-Smith[1]; Beatriz de la Iglesia[1]

[1]School of Computing Sciences, University of East Anglia, Norwich, United Kingdom

[2]Norfolk & Norwich University Hospital NHS Foundation Trust, Norwich, United Kingdom

[3]School of Biological Sciences, University of East Anglia, Norwich, United Kingdom

Corresponding Author:
Joao H Bettencourt-Silva
School of Computing Sciences
University of East Anglia
Norwich Research Park
Earlham Road
Norwich, NR4 7TJ
United Kingdom
Phone: 44 1603592300
Fax: 44 1603593345
Email: jhbs@cmp.uea.ac.uk

Abstract

Background: Routinely collected data in hospitals is complex, typically heterogeneous, and scattered across multiple Hospital Information Systems (HIS). This big data, created as a byproduct of health care activities, has the potential to provide a better understanding of diseases, unearth hidden patterns, and improve services and cost. The extent and uses of such data rely on its quality, which is not consistently checked, nor fully understood. Nevertheless, using routine data for the construction of data-driven clinical pathways, describing processes and trends, is a key topic receiving increasing attention in the literature. Traditional algorithms do not cope well with unstructured processes or data, and do not produce clinically meaningful visualizations. Supporting systems that provide additional information, context, and quality assurance inspection are needed.

Objective: The objective of the study is to explore how routine hospital data can be used to develop data-driven pathways that describe the journeys that patients take through care, and their potential uses in biomedical research; it proposes a framework for the construction, quality assessment, and visualization of patient pathways for clinical studies and decision support using a case study on prostate cancer.

Methods: Data pertaining to prostate cancer patients were extracted from a large UK hospital from eight different HIS, validated, and complemented with information from the local cancer registry. Data-driven pathways were built for each of the 1904 patients and an expert knowledge base, containing rules on the prostate cancer biomarker, was used to assess the completeness and utility of the pathways for a specific clinical study. Software components were built to provide meaningful visualizations for the constructed pathways.

Results: The proposed framework and pathway formalism enable the summarization, visualization, and querying of complex patient-centric clinical information, as well as the computation of quality indicators and dimensions. A novel graphical representation of the pathways allows the synthesis of such information.

Conclusions: Clinical pathways built from routinely collected hospital data can unearth information about patients and diseases that may otherwise be unavailable or overlooked in hospitals. Data-driven clinical pathways allow for heterogeneous data (ie, semistructured and unstructured data) to be collated over a unified data model and for data quality dimensions to be assessed. This work has enabled further research on prostate cancer and its biomarkers, and on the development and application of methods to mine, compare, analyze, and visualize pathways constructed from routine data. This is an important development for the reuse of big data in hospitals.

XSL•FO
RenderX

Introduction

Clinical Pathways

Clinical pathways, also known as care or critical pathways, have been introduced in health care systems to improve the efficiency of care, while maintaining or improving its quality [1]. In 1995, Pearson et al [2] described critical pathways as a management plan that "displays goals for patients and provides the sequence and timing of actions necessary to achieve these goals with optimal efficiency". More recently they have been described as a concept for making patient centered care operational, and for "supporting the modelling of patient groups with different levels of predictability" [1]. Clinical pathways are developed by multidisciplinary teams and rely on evidence from the literature, operational research, and patient involvement methodologies [1].

Over the years, pathways evolved from paper-based to computerized pathways [3,4], and there have been efforts to integrate them with electronic health records [4,5]. The support for guidelines and pathways is one of the most promising fields for knowledge-based systems in health care [6]. The standard functions of pathways have been proposed in [4], and a strong emphasis is given to the statistics function to implement automated methods for checking the occurrence of variance (ie, discrepancies between planned and observed events).

There are several definitions of clinical pathways in the literature, but in this paper they are defined as an ordered set of patient-centric events and information relevant to a particular clinical condition. In this paper, a clinical pathway is not described in the context of an intervention, but in the context of the description, analysis, and evaluation of clinical parameters for a specific condition over time. The pathways are also data driven and allow the inspection of routine hospital data that would otherwise be overlooked. Furthermore, we place particular importance on the use of clinical biomarkers and other indicators (such as blood readings) in pathways, as they enable a thorough inspection of data quality, as well as further clinical studies observing trends over time.

Analysis of Clinical Pathways

The analysis of clinical pathways is a topic receiving increasing attention in medical informatics, but techniques often require extensive clinical expert knowledge and can be laborious. Huang and Duan [7] used process mining techniques to measure clinical behavior derived from clinical workflow logs and to help identify novel process patterns. According to them, clinical pathway analysis has been defined as the process of discovering knowledge about clinical activities in patients' care journeys. Ultimately the goal is to utilize the discovered knowledge for pathway (re)design, optimization, decision support, audit, or management, and one of the major challenges reported was the derivation of compact, yet high quality, patterns that cover the most useful medical behaviors in clinical practice.

Process mining techniques are promising analysis techniques in the context of clinical pathways. However, it has been reported that traditional process mining algorithms do not cope well with unstructured processes like those commonly found in a hospital environment [8,9], and that they may not produce clinically meaningful visualizations. The heterogeneity and incompleteness of the data are major obstacles in achieving meaningful models, yet an application to stroke has proved fruitful [10]. An aim of this paper is to produce pathways that may be suitable for process mining. For this, data quality is key, but consensus and definitions are lacking [11,12], and intelligent agents that explore quality issues are needed [12].

The use of routine data or workflow logs in the construction of clinical pathways is a key topic receiving increasing attention in the literature [7-9]. In hospitals, such efforts rely heavily on the hospital information systems (HIS) and electronic health records (EHR), and the availability and quality of the information conveyed in them. Indeed, hospitals often opt for implementing several commercial departmental systems, creating "islands" of information across various departments [13,14]. This can significantly hinder the process of extraction and collation of detailed patient-centric information to create clinical pathways. The methods presented in this paper attempt to overcome some of these difficulties.

Data Quality in Electronic Health Records

A review on data quality in EHR [11] identified five data quality dimensions described in the literature: (1) completeness, (2) correctness, (3) concordance, (4) plausibility, and (5) currency. However, the authors identified that not all dimensions are commonly or consistently assessed, and further work is needed toward the adoption of systematic, statistically based methods of data quality assessment. The work presented in this paper enables the inspection of data quality dimensions with a particular emphasis on assessing the completeness of pathway information using biomarker expert rules.

Overall, this paper describes a framework for building and visualizing prostate cancer pathways using routinely collected data from a large United Kingdom National Health Service (NHS) hospital. This approach does not involve workflow logs produced by HIS or EHR, but rather, the patient-centric data conveyed in them. Our previous work on methods for the collection of patient-centric data from multiple HIS [14] has underpinned this research.

Prostate Cancer

The latest estimates of global incidence indicate that prostate cancer has become the second most common cancer in men [15]. In the United Kingdom, it is the most common male cancer, accounting for 25% of all malignancies [16]. In recent years, there has been a generalized increase in reported incidence, but, despite this, the mortality rates have been on the decline [16-18]. Nevertheless, the economic burden of prostate cancer will continue to rise due to increased diagnosis, diagnosis

XSL•FO

RenderX

at an earlier stage, and prolonged survival [18]. It has been reported that new strategies need to be devised to increase the efficiency of health care provision for this type of cancer in order to tackle the increasing burden [18]. Prostate Specific Antigen (PSA), a biochemical marker used clinically for prostate cancer detection and prognosis, is associated with substantial overdiagnosis and excessive treatment [19], which makes its utility as a screening test controversial, and warrants the need for further studies.

The National Institute for Clinical Excellence (NICE) in the United Kingdom publishes clinical guidelines and has recently developed the NICE pathways, a tool that visually represents the recommendations and guidelines on a specific clinical or health topic [20]. Following the NICE pathway, patients with suspected prostate cancer are directed through from referral, to assessment, diagnosis, and communication; their needs are then often discussed at a multidisciplinary team meeting; admission and treatment options are selected as appropriate, and ultimately patients are followed up, and outcomes assessed. During each step of the pathway, relevant patient-centric data are produced and often stored in a variety of different HIS. Clinicians wishing to investigate prostate cancer, say to establish the merits of alternative treatment and management options, would have a powerful tool if access to the integrated data was facilitated in an electronic and canonical form. However, as is often the case with HIS, database systems and their data are heterogeneous, and data quality, accessibility, and interface vary considerably.

Objectives

The aims of the work presented here can be divided into two: (1) to generate individual data-driven patient-centric pathways from routinely collected hospital data for prostate cancer, and (2) to evaluate the completeness and utility of the generated pathways for investigating biomarker trends. The latter allows for the selection of high quality data for clinical studies and decision making, which, in turn, enables the (re)design, management, and optimization of pathways. We focus on a definition of a pathway as a data structure that synthesises knowledge, and facilitates the development of methods for the computation of variance and other statistics. The framework presented in this paper, together with their formalisms, should allow and encourage other tools and techniques, such as process mining or ad-hoc algorithms to be used.

Methods

Prostate Cancer Case Study

A case study on prostate cancer was carried out at the Norfolk & Norwich University Hospital (NNUH) NHS Foundation Trust with data from this hospital only. Appropriate credentials were obtained from the National Research Ethics Service (Norfolk) and NNUH research governance committees, and no patient consent was required. The data were anonymized and no patient sensitive information such as names or addresses was used.

This section first summarizes the methods for data collection from multiple hospital sources under the subheading "The Operational Data Store", and it is followed by the definition of a pathway under the subheading "Extraction of the Study Datasets". Descriptions of the methods to build a pathway dictionary and to generate the pathways are given under the subheading "Building the Pathway Dictionary and Database". The subsection "Visualizing Pathways and Overall System Architecture" introduces the system to integrate, visualize, and analyze the pathways, as well as a novel graphical representation, and the subsection "Assessing Completeness Using Biomarker Information" describes a method for assessing the quality of the pathways.

The Operational Data Store

Electronic patient data in hospitals are usually complex and heterogeneous [21,22], scattered through several information sources or HIS, and its retrieval methods are often ad-hoc and poorly described in the literature [14,23]. A previously proposed data extraction process [14] was used to collect patient-centric data from HIS, and it is summarized in this section.

The process involves liaising with domain experts (or subject matter experts) to identify data sources where information related to prostate cancer patients is likely to be stored (eg, radiology). In this case study, the team of experts included a urology consultant, prostate cancer geneticists, a consultant oncologist, a histopathologist, and a chemical pathologist. For each data source identified (a EHR or HIS), the data extraction process [14] was followed. The process consists of four key steps and Figure 1 shows this: (1) system understanding, where each data source is investigated and details about the system are gathered; (2) data understanding, where data familiarization, selection, and building the data dictionary occurs; (3) extraction preparation, where data extraction methods are prepared or reviewed; and (4) extraction and evaluation, where data are extracted, validated, and the process is evaluated.

An example of an input data source is the laboratory information system (LAB), where information on the PSA and other blood tests are stored. Following the data extraction process in Figure 1, a thorough inspection of the system is carried out first (system understanding step). This required the involvement of domain experts (clinical and administrative), obtaining relevant access credentials and previewing the system, and resulted in an understanding of the way in which blood tests are requested and how that information flows in and out of the LAB system. The next step deals with understanding the data. In this example, data on PSA were explored, including details on how it had been recorded over time, data field semantics, and available patient and blood test identifiers that, for example, allow the retrieval of unique blood tests for each patient. Once both systems and data were investigated with respect to the required information (in this example, the PSA), then a suitable data extraction strategy is devised. Finally, the selected methods are tested to ensure that they produce the same desired results. In this example, the LAB system offered an on-line analytical processing interface, where additional training and input from domain experts was required in order to produce database queries that retrieved the PSA test data along with dates, times, and identifiers for data linkage purposes. Sample datasets are extracted in a suitable format, and subsequently they are evaluated. The evaluation consists of cross-checks against the LAB system and patient notes, and a careful examination for

missing or erroneous values (for example, nonnumeric values were identified in some of the PSA test results: <0.1 ng/ml). Erroneous values are corrected when possible (for example, <0.1 was reformatted to 0.05) or their records are eliminated. Finally, a study dataset is produced for the LAB system containing the PSA tests. A second output is metadata (about the source, its tables, attributes, and values) that is generated at each step of the process, and allows it to be repeated and documented over time.

The process is repeated for every data source where information on prostate cancers is likely to exist, and this ultimately generates an operational data store (ODS), which is similar to a data warehouse from where specific data marts can be extracted. The ODS contains relevant metadata and detailed, routinely collected information on the selected case study. By enabling the inspection, linkage, and compilation of cohorts, it helps to overcome the types of heterogeneity commonly found between HIS such as technical differences, syntactic, and semantic heterogeneity. This process is also suitable for, and greatly facilitated in, less heterogeneous environments where data sharing standards exist.

Overall, the data extraction process enables the use of routinely collected data to build a repository containing all interactions of the patient with the hospital. This process can be repeated so that the ODS continues to be populated with new records. The methods of extraction are reviewed and revised over time. The costs associated with this process depend on the functionalities

of the HIS, particularly with respect to the retrieval of cohorts of patients, as well as documentation and support. The process may be time consuming in systems where no querying tools are available, and alternative methods are required. Overall, the most time consuming step of the process, given our experience in this case study, was system understanding, where a substantial amount of time was spent liaising with hospital information technology managers and other staff, and the second most time consuming step was extraction preparation. However, after the first iteration, the process becomes streamlined and only minor adjustments may be required even in heterogeneous environments. The process is also applicable to more structured environments where reduced costs are to be expected. Different problem domains are not expected to require other costs, as these are mostly dependent on the HIS rather than on particular data elements.

The data available in the ODS may be more than required for a particular clinical study, as the retrieval process is based on minimum use of constraints. However, this provides a holistic representation of the patients, including their demographics, comorbidities, test results, or other information, and is limited by the availability of electronic information in the HIS. The selection of specific data elements from the ODS that will form a pathway is performed later ("Building the Pathway Dictionary and Database") in consultation with the domain experts. A summary of the data retrieved from the ODS for creating pathways is given in the Results section.

Figure 1. Simplification of the data extraction process [14].
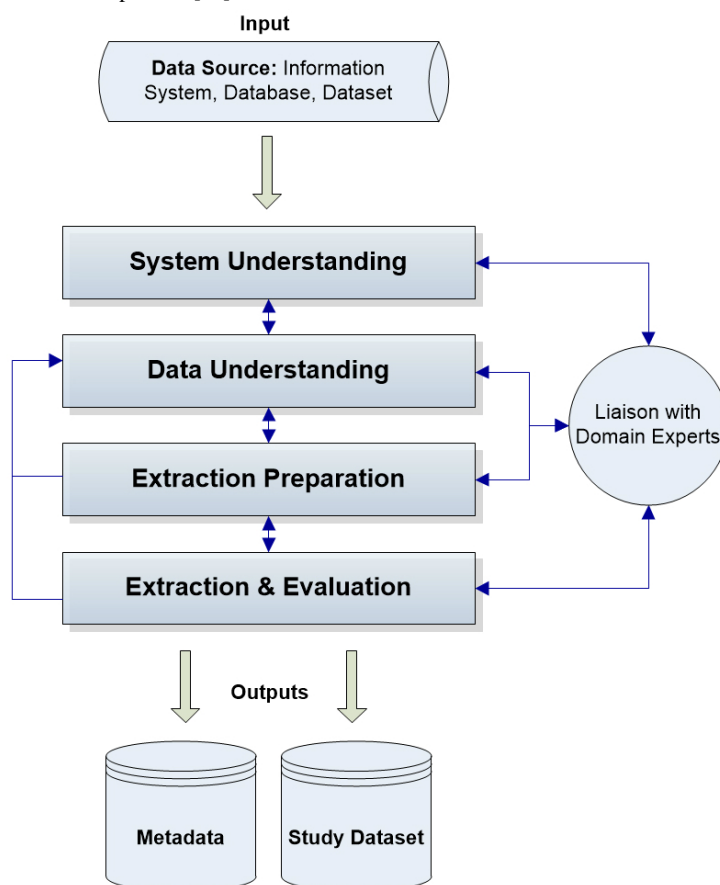
Extraction of the Study Datasets

In the case study on prostate cancer, the ODS contains information from the following systems: administration, cancer waiting times, histopathology, radiology, biochemistry, operating theater, orthopaedics, oncology, and radiotherapy. However, not all sources are used in the pathways presented in this paper, as later explained in "Building the Pathway Dictionary and Database". Table 1 shows the data sources used in the development of the pathways. Retrieving diagnosis codes from the administration and histopathology systems first identified the prostate cancer cohort, and it was later validated with information from the local cancer registry.

Table 1. Data sources used for the development of the pathways.

Data source (abbreviation)	Description of selected data
Administration (ADM)	Patient episodic information, comorbidities, and clinical coding.
Histopathology (HIST)	Histopathology reports and extracted Gleason grades.
Radiology (RAD)	Radiological imaging limited by reports where the word prostate occurs.
Biochemistry (LAB)	PSA tests. However, other blood tests can be added.
Operating theater (OT)	Operating theater procedures and coding.
Radiotherapy (RT)	Radiotherapy treatments dates and number of sessions.
Cancer registry datasets (CR)	The cancer registry dataset includes some of the above data, which can be used for quality checking purposes, and additional data such as cause of death.

The Prostate Cancer Cohort

For the prostate study, a cohort of 1904 patients diagnosed with prostate cancer (average age 72, SD 9) between 2004 and 2010 was selected for retrieval from the ODS. This represents a subset of the total number of prostate cancers, where it was possible to accurately ascertain both diagnosis and treatment dates. Ascertainment of nearly 20% of the original cohort was not possible due to the information not being consistently recorded, to changes in systems and the way they are used, and to data quality not being consistently inspected prior to 2008. Data from 2003 were collected and used as potential "screening" and from 2011 as follow-up. This time window delimits patient pathways. Date and cause of death were collected from the cancer registry early in 2012. All patients in the cohort have a diagnosis date and have been offered treatment as per the UK guidelines. The UK national cancer waiting time guidelines stipulate that all suspected cancers in the NHS should be offered treatment (including active surveillance) within 31 or 62 days, according to the national cancer waiting times guidelines. As per the cancer waiting times guidelines, all patients were followed up after diagnosis and, in this cohort, 2.21% (42/1904) did not agree on any form of treatment. This differs from active surveillance, in that the latter requires the patient and clinician to agree to monitor tumor growth.

Additional information not consistently recorded in HIS (eg, tumor staging) was retrieved from the local cancer registry (CR) using deterministic record linkage on national health identifiers and dates of birth. The registry served as a source of validation for the collected data as most of the critical data elements often used in prostate cancer studies will be present in the local CR. However, additional hospital data that were routinely collected, but not present in national audit reports or cancer registries (such as biomarker trends or imaging) increases the value and completeness of the pathways. In particular, the value of the biomarker in determining the quality of the pathways is discussed later in this paper.

Defining a Pathway

In order to create pathways, data elements are selected from the ODS and its sources (Table 1). A formal definition of a pathway is given in the "Definitions" subsection, and further details on the selection of data elements and their inclusion in a pathway data dictionary are given in the "Building the Pathway Dictionary and Database" subsection. The developed software environment, data flows, and visualizations are described in the subsection "Visualizing Pathways and Overall System Architecture", and the proposed methods to compute completeness based on biomarker elements within a pathway are given in the subsection "Assessing Completeness Using Biomarker Information".

Definitions

Let D represent the pathway dictionary, where the i-th entry has a code $c_i(1 \leq i \leq n)$ in a total of n possible codes described in detail in Table 2. C_E is the subset of codes containing timed events, and C_I the subset containing informational elements, such as demographics. By associating a zero time with informational elements, all events in the pathway can be viewed as timed events.

A *pathway activity* A is then defined as four-tuple $A=(r,t,c,v)$ where,

- r is the patient identifier
- $c \in C$ is an event code
- t is the time in days before or since the day of diagnosis recorded for patient r
- v is a value, numerical or categorical, associated with dictionary code c

A *pathway* for patient, r, is represented as a chronological sequence of activities, $P=(A_1,A_2,...,A_m)$, where

1. A_i is of the form (r, t_i, c_i, v_i) for $1 \leq i \leq m$,
2. $t_i \leq t_i+1$ for $1 \leq i \leq m$,
3. any A_i with $c \in C_I$ has $t_i=0$,

4. if $A_i = (r, t_i, c_i, v_i)$ and $A_{i+1} = (r, t_{i+1}, c_{i+1}, v_{i+1})$, then there is no activity $A = (r, t, c, v)$ where

5. $t_i < t < t_{i+1}$, and

6. all relevant activities involving patient r appear in P.

Note that when $t_i = t_{i+1}$ for $1 \leq i \leq m-1$, the corresponding activities A_i and A_{i+1} are concurrent.

A simple pathway for patient $r=1$ might be $P = \langle A_1 = (1,-28,P,45)$, $A_2 = (1,0,D,2)$, $A_3 = (1,1,G,"4+3")$, $A_4 = (1,1,H,"Cyproterone Acetate")$, $A_5 = (1,151,R,"37")$, $A_6 = (1,260,P,0.2)$, $A_7 = (1,340,P,0.05)$, $A_8 = (1,539,P,0.05) \rangle$.

In this patient's pathway, the first PSA test was elevated at 45 ng/ml, and this led to the diagnosis of stage 2 prostate cancer, with a Gleason grade of 4+3. Note that the biopsy was performed as an outpatient event, and hence, it is unavailable in this pathway, however, the histopathological findings of that biopsy are present. The patient then agreed to undergo hormone therapy (cyproterone acetate) and a subsequent 37 sessions of radiotherapy. The number of radiotherapy sessions is recorded as value of element code R. Information on specific sessions was not consistently available at the time, and was therefore not used. The radiotherapy sessions were then followed by PSA readings of 0.2 ng/ml and two readings <0.1 ng/ml, which indicate a good response to treatment.

The above model of expressing pathway activities is similar to the entity-attribute-value (EAV) data model [24], where concepts are described in an attribute in a row. Later, the i2b2 data model [25] expanded on the EAV model to account for time (start and end dates for each observation). This, together with a star schema, has been described as an extremely efficient way of querying data, as a large index can be built to encompass all patients' data in the master table [25]. The proposed pathways model expands the EAV model in that every row has an associated time, and this is important because pathways are ordered sets of events. With regards to the i2b2 model, the proposed pathways include fewer elements in the master table, and focus on a sequential representation and processing of pathway activities. In addition, activities and their pathways can also be linked to other tables (and dimensions) that store other types of information, similarly to what is accomplished by the star schema in the i2b2 model. The proposed pathways model is part of an overall framework environment that is described in detail in the subsection "Visualizing Pathways and Overall System Architecture".

Table 2. Pathway dictionary for prostate cancer.

Class	Code	Name	Type	Data source	Frequency, n (%)
Demographics	Q	Deprivation score	Information	CR[a]	1904/1904 (100.00)
Demographics	A	Age at diagnosis	Information	CR[a]	1904/1904 (100.00)
Demographics	Z	Death	Event	CR[a]+ODS (ADM[b])	402/1904 (21.11)
Demographics	L	Clinical trial	Information	ODS (ADM[b])	22/1904 (1.16)
Demographics	X	Other cancers	Event	CR[a]	406/1904 (21.32)
Diagnostics	D	Diagnosis and staging	Event	CR[a]+ODS (HIST[c]+ADM[b])	1904/1904 (100.00)
Diagnostics	G	Histology grade	Event	CR[a]+ODS (HIST[c])	1609/1904 (84.51)
Diagnostics	I	Imaging	Event	ODS (RAD[d])	291/1904 (15.28)
Diagnostics	P	PSA test	Event	ODS (LAB[g])	1814/1904 (95.27)
Treatment	S	Surgery	Event	CR[a]+ODS (OT[e])	640/1904 (33.61)
Treatment	R	Radiotherapy	Event	CR[a]+ODS (RT[f])	395/1904 (20.75)
Treatment	C	Chemotherapy	Event	CR[a]+ODS (ADM[b])	8/1904 (0.42)
Treatment	O	Orchidectomy	Event	CR[a]+ODS (OT[e])	2/1904 (0.11)
Treatment	H	Hormone	Event	CR[a]+ODS (ADM[b])	960/1904 (50.42)
Treatment	W	Active surveillance	Event	CR[a]+ODS (ADM[b])	422/1904 (22.16)
Treatment	N	No treatment	Information	CR[a]+ODS (ADM[b])	42/1904 (2.21)

[a]CR = Cancer Registry datasets

[b]ADM = administration

[c]HIST = histopathology

[d]RAD = radiology

[e]OT = operating theater

[f]RT = radiotherapy

[g]LAB = biochemistry

Building the Pathway Dictionary and Database

The selection of key informational requirements for the pathways is facilitated by the patient-centric approach to data collection [14]. The ODS contains data from the retrieved hospital sources and metadata, which allows for the inspection, linkage, and integration of semantic and syntactically different data. Nevertheless, the ODS may contain information outside the domain of a specific pathway. Therefore, in order to build a pathway dictionary, it is crucial to identify, select, and retrieve key data elements. Figure 2 illustrates the process of building a pathway dictionary from the data in the ODS, and is inspired by the similar data warehousing technique of extract-transform-load [26]. The pathway dictionary can be regarded as a simple ontological knowledge base, built by a bottom-up process, from available data to concepts. Temporal ontologies have been developed [27], yet for the definition of

pathways, the above time-oriented data structure together with a pathway dictionary was sufficient to enable temporal abstractions.

The dictionary building process, based on input from domain experts, literature survey, and current prostate cancer guidelines, involves gathering relevant data elements and applying transformations to either create new features or strip out irrelevant elements (eg, hospital events that are neither exclusive nor relevant to the treatment of prostate cancer). At the end of this process, and for each data element, a flat file with the data corresponding to that element is created in the four-tuple transactional format described in the subsection "Definitions". The steps involved in this process are described in detail below. At present, the system has not used multimedia or other large files, but plans are underway to ensure that such files can be encrypted and stored locally. In such cases a pointer to the file would be included in the relevant data element.

Figure 2. Methodology to build pathways' dictionary and database (DB).
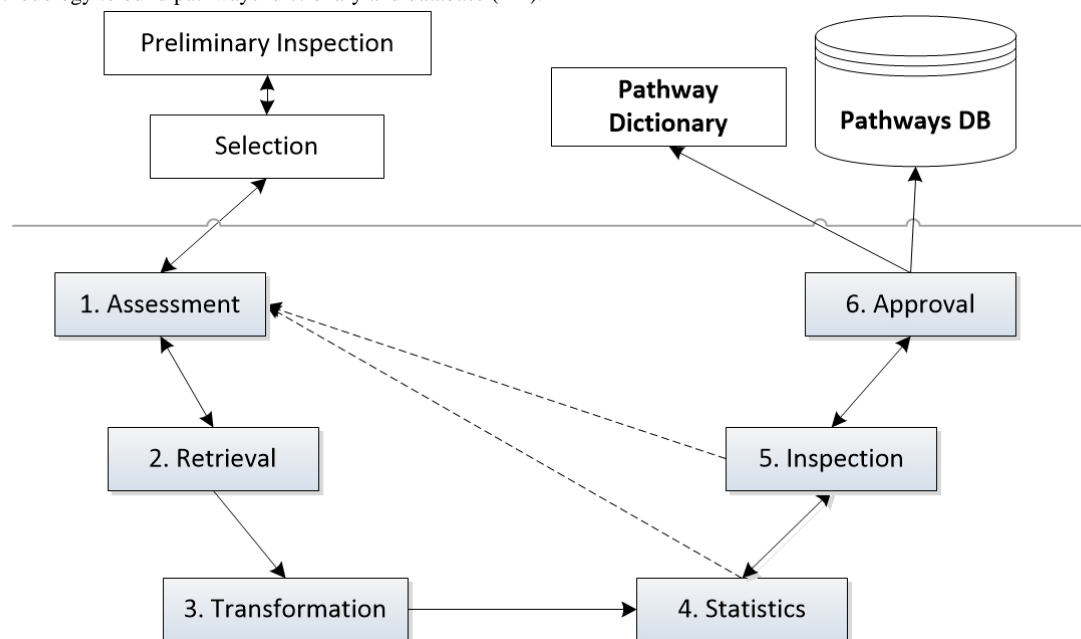
Preliminary Inspection and Selection

The domain experts collaborate on a first inspection of the available data in the ODS to help with the identification of key data elements to be included in the pathways. This involves examining summary statistics (such as frequencies of biochemistry tests or other descriptive statistics) and metadata (such as attributes descriptions, semantics, or expected outliers) from the ODS, and is important as it sets the granularity of the pathways and the extent to which they can be meaningful for a particular disease. This process was mostly ad-hoc, as each data element required different statistics, and was also based on contributions by the domain experts. Data updates, however, may be processed automatically once the pathway dictionary is built.

For the prostate cancer study, we defined three classes of information: demographics, diagnostics (including investigations), and treatment. Hence, the selected elements in this step have an associated class. Further to this, each element type can either be a timed event, describing a particular activity at a given time, or auxiliary information, such as demographic data or other nonevent data, such as a patient's participation in a clinical trial. Both class and type are two properties common to all elements of the pathway, and can be determined a priori or throughout the process of building the pathways as explained below.

For each selected data element, the following six steps are carried out to create a complete pathway dataset and dictionary. Throughout the following steps, we will use the example of the biomarker test for prostate cancer (PSA test) as a data element.

Assessment

The first step is to inspect the element's values as well as its semantics, syntax, and data type, and any potential limitations that may interfere with the consistency of the data element. Additional mapping, linkage, and transformations may be necessary to enforce a consistent format and these should be

identified here. An example arising from the PSA test was the need for the removal of values that include symbols, such as "<1", meaning the PSA test value is less than one. In this case, such values were replaced by 0.5. A first classification of the element is also given by assigning a dictionary code and the element type (informational or timed event); in this case the code for the timed event PSA test is "P".

Retrieval

The set of attributes and values for the data element are retrieved from the ODS. In the case of the PSA test, the attributes in the ODS include dates of test authorization, date of entry, value, comments, clinical history, fasting, blood reading thresholds, and the patient identifiers. More complex data elements, such as social determinants can also be created from the information available in the ODS, but they might require additional or specific preprocessing. For example, in this case study, deprivation score was included in the pathways (code Q in Table 2), and regarded as an informational element. Because of the way in which the deprivation index was recorded, the data element was in this case time-independent, and handled differently by the pathways framework described in the next section.

When retrieving information to create or update a data element, rule-based deterministic record linkage can be used to enforce constraints. In the course of the case study, the retrieval step was used to select data within the study time period as well as validating data from the hospital sources against the cancer register datasets, where possible, in terms of completeness, correctness, and concordance. The retrieved attributes must have the information required by pathway definition. The data for the particular element are then stored, and in this case study a comma separated file (CSV) is created to this effect. For PSA tests, the attributes selected from the ODS to be included in the pathway were date of entry (date when the sample was taken from the patient within the selected time period), the value, and the identifier that allows linkage. Working with CSVs can

introduce additional technical challenges, in particular when different database or spreadsheet systems are used. In this case study, the data available in the ODS were extracted in a format that is compatible with CSV. However, additional checks using raw text editors and spreadsheet software may need to be performed after the data are extracted, so as to inspect and ensure that the exported data meet the required CSV constraints.

Transformation

The retrieved data file is converted into the pathway data structure, with attributes *Identifier, Code, Date* (instead of time)*, and *Value*, where *Code* is a constant. *Date* is used here, but it will later be converted into time, *t*, zeroed at diagnosis date. The latter, by removing full dates, allows an additional layer of anonymity to the pathways, as well as a basis for comparison among patients. Any necessary transformations and formatting changes identified in the previous steps are undertaken here.

Summary Statistics

Summary statistics are produced in this step. These include distributions of the *Value* attribute, which can help to detect potential bias, together with overall support (ie, total number of patients); value-specific support (ie, number of patients on each value category); and extremes. Such statistics may help to detect and correct quality issues by assessing completeness (missing data), correctness, and plausibility. Additionally, other statistics may be produced, such as the number of values within a range; this is particularly useful for producing a summary of abnormal blood readings, such as raised PSA tests.

Inspection

Together with the domain experts, the retrieved data and descriptive statistics are inspected. The values of the attributes are also checked for format consistency and the quality dimensions described above. At this stage, a decision regarding the data element is reached. The element may be:

- Kept as is, should it contain sufficient information and adequate support;
- Rejected, because there may not be enough information or support, or because the formats or data types no longer match those previously collected. The latter may lead to a reevaluation of the methods used to extract the data. However, this is not expected to happen when the process in Figure 1 is followed, and consistent metadata is also collected; and
- Subject to decomposition, into two or more elements, should the values of the element vary qualitatively, creating a source of ambiguous information, or should the requirements of a particular study involve inspecting a particular quantitative range, such as the abnormal range of a blood reading.

In the example of the PSA test, the data element was kept after the values were set to a canonical form.

An example of an element that was rejected in the case study is *biopsy*, because of insufficient support (this is further discussed below). A further example of an element that was split was *surgery*, into *orchidectomy* and *surgery* (prostatectomy). Another example of an element that was split

was *radiotherapy*, where, for the analysis of the trend of PSA, only *radical radiotherapy* was interesting to investigate, as it affects the PSA.

Approval and Update

Upon inspection, a decision is made regarding the data element and its values. When the decision is favorable, an update is carried out. The update is concerned with the technical work of merging the table containing the data element and its values with the pathways database master table. Further transformations are also carried out to sort the master table by date and patient identifier, and to compute time *t* zeroed at diagnosis date. This can be achieved by either creating an informational element providing the date of diagnosis or by programmatically isolating the specific date from an existing element and subsequently setting *t* for all activities in a pathway. The pathways dictionary is then updated with summary information.

The process of building the pathways data dictionary can be revisited to accommodate new data or to change the way in which informational elements are modelled. For example, should informational elements later be provided with a time-point, these can be remodelled as timed events and instructions can be added so that the software framework handles them differently. The latter is described in more detail below.

Visualizing Pathways and Overall System Architecture

A system responsible for the integration, visualization, and analysis of pathways and related data was developed. Figure 3 illustrates the overall environment of the developed carcinoma of the prostate visualization and interpretation system (CaP VIS), the ODS, and the previous method of building the pathway dictionary. Figure 3 also shows the ways in which the data flow from the sources, and the steps involved in bringing detailed pathways into the visualization and interpretation system, the analysis, or query engines. The steps of the two main processes that feed data into the CaP VIS system start from the ODS and are enumerated. Secondary processes are highlighted with dotted lines.

The main process responsible for producing the pathways starts from the ODS and follows steps 1a to 5a in Figure 3. Datasets were extracted from the ODS in the pathway format defined in the subsection "Defining a Pathway", and used to build the pathway dictionary (as described in the subsection "Building the Pathway Dictionary and Database"), and the raw pathways database (following steps 1a and 2a). The pathways engine, which works with the information stored in the raw pathways database (step 3a), is responsible for the segmentation, summarization, cleansing, and indexing of the raw pathways. Such operations together allow for the mapping, selection, and retrieval of individual or groups of similar paths using regular expressions or ad-hoc algorithms. The detailed pathways are organized by patient identifier and stored as "plots" (following step 4a) that allow an interpreter and the visualization software (CaP VIS) to produce a detailed graphical representation (step 5a). The interpreter will parse each activity from a pathway and, based on the dictionary and a set of rules determined for each element code, plot the corresponding graphical representation. An important feature of the visualization system is to integrate

the pathways with histopathological or further clinical information. A coding lookup table was added in order to translate and present diagnosis (International Classification of Diseases, ICD) and procedures (Office of Population Censuses and Surveys) codes (highlighted by the dotted lines in Figure 3). Because the time length of different pathways can vary considerably, it was important for the plot to be interactive, allowing zoom and rescale, as well as mechanisms for graphical conflict resolution (ie, avoiding overlapping elements). Figures 4 and 5 show sample output from the visualization software and a patient pathway and related information, including the pathway data format. The analysis engine can be used by the CaP VIS software to compute statistics for the pathways, but it can also be used on its own to develop algorithms that work with the pathways data. The subsection "Assessing Completeness Using Biomarker Information" demonstrates the use of the analysis engine in computing completeness scores for the PSA values in pathways. The analysis engine consists of a set of functions and libraries that are built in to main software, written in Python. In order to access the engine and perform operations, Python scripts can be written to access relevant functions that read information from pathways, generate graphical representations, compute PSA kinetics, or other statistics.

The CaP VIS system is also fed additional data that can be linked with pathway information. This process follows steps 1b to 3b in Figure 3, and produces a database of other clinical information not included in the pathways dictionary, such as full histopathological text reports. The latter could still be included in the pathways, but in our case study, this information was not part of the desired graphical representation of a pathway, and so it was more efficient for it to be accessed differently. Furthermore, this enables the system to use additional data that are not part of the pathway. The CaP VIS system integrates this information and shows a novel graphical representation of the patients' pathways. Figure 4 shows the left side of the CaP VIS screen where the graphical representation of a pathway is visible, and Figure 5 shows the right side of the screen with additional information pertaining to that pathway. Together, the two figures show the full screen of the system.

The way in which pathway elements are plotted in CaP VIS depends upon their code, type, and value. In the pathway plot seen in Figure 4, the x-axis represents the time in days, zeroed on diagnosis of prostate cancer, and the y-axis represents the biochemical marker, PSA. Other data elements are plotted either as vertical lines dividing the pathway into segments, or as further information captions along the x-axis or y-axis, as needed. The plot illustrates a total of 32 events and informational elements. Vertical lines pertaining to treatments or diagnostics are accompanied by the respective element code from the dictionary on top. There are three types of vertical lines that are plotted: diagnosis (code, D, solid line), treatment (codes, H for hormone therapy and S for surgery, dashed), and death (bold). The latter is accompanied, along the x-axis, by ICD coding for the causes of death as well as age at death, whereas the diagnosis line is accompanied by age at diagnosis, tumor staging, and Gleason grade. Treatments are plotted as dashed lines and further biopsies, dotted lines. The lines may overlap; however, color coding and scaling are available to further investigate smaller segments of the pathway when necessary.

The main CaP VIS system screen contains three areas on the right side of the screen (Figure 5) to enable the inclusion of the histopathology text reports, pathway details, and annotations. The histopathology box can be toggled between a summary of the pathway statistics including PSA kinetics (measurements of change over time, widely used to assess recurrence) [28] or the histopathology text report. The pathway details box includes the pathway data in the format described in the subsection "Defining a Pathway", and further information computed based on that data. Other screens include the detailed PSA kinetics regression line (seen in Figure 5 above histopathology text report, including doubling time and velocity), and a further screen (not shown here) summarizing the details of all 1904 pathways, which can be used to query the cohort. Overall, the CaP VIS system allows a thorough inspection of biomarker trends and other electronically available data on patients with prostate cancer by clinicians and researchers.

Figure 3. Data flow diagram illustrating the relationship between the operational data store (ODS, in bold), the pathway and analysis engine, the carcinoma of the prostate visualization and interpretation system (CaP VIS), and other interactions including lookup databases (DB) for International Classification of Diseases (ICD) and OPCS (Office of Population Censuses and Surveys) coding. The two main processes that feed information into the CaP VIS system are enumerated. The pathways data follows steps 1a to 5a, while other information follows steps 1b to 3b. Secondary processes are highlighted with dotted lines.
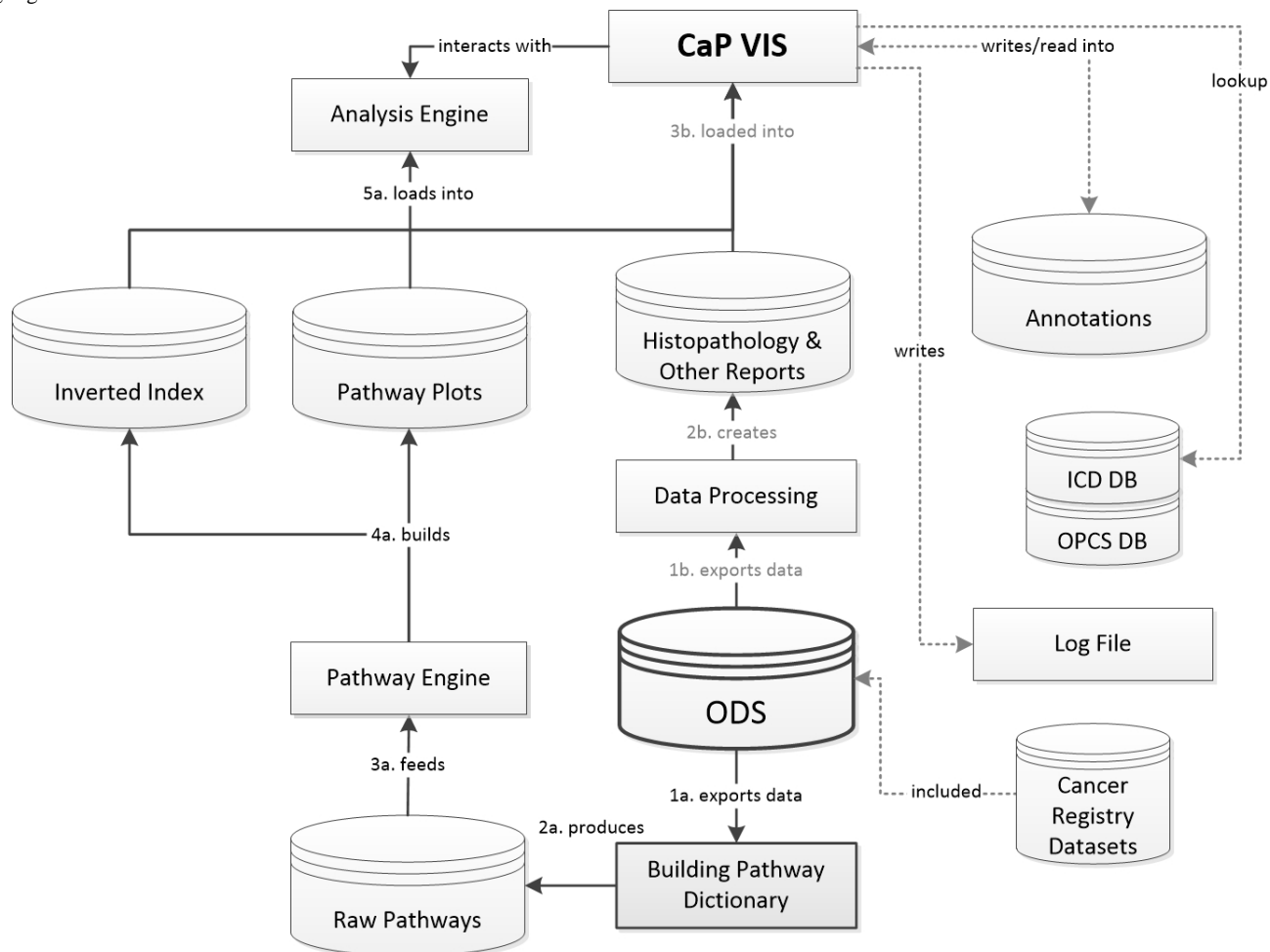
Figure 4. The CaP VIS system showing the left side of the screen with the graphical representation of a castration resistant patient pathway. The patient was first treated with hormone therapy and had a subsequent palliative prostatic resection. The plotted pathway shows the trend of the PSA biomarker together with diagnosis line and treatments.

XSL•FO
RenderX
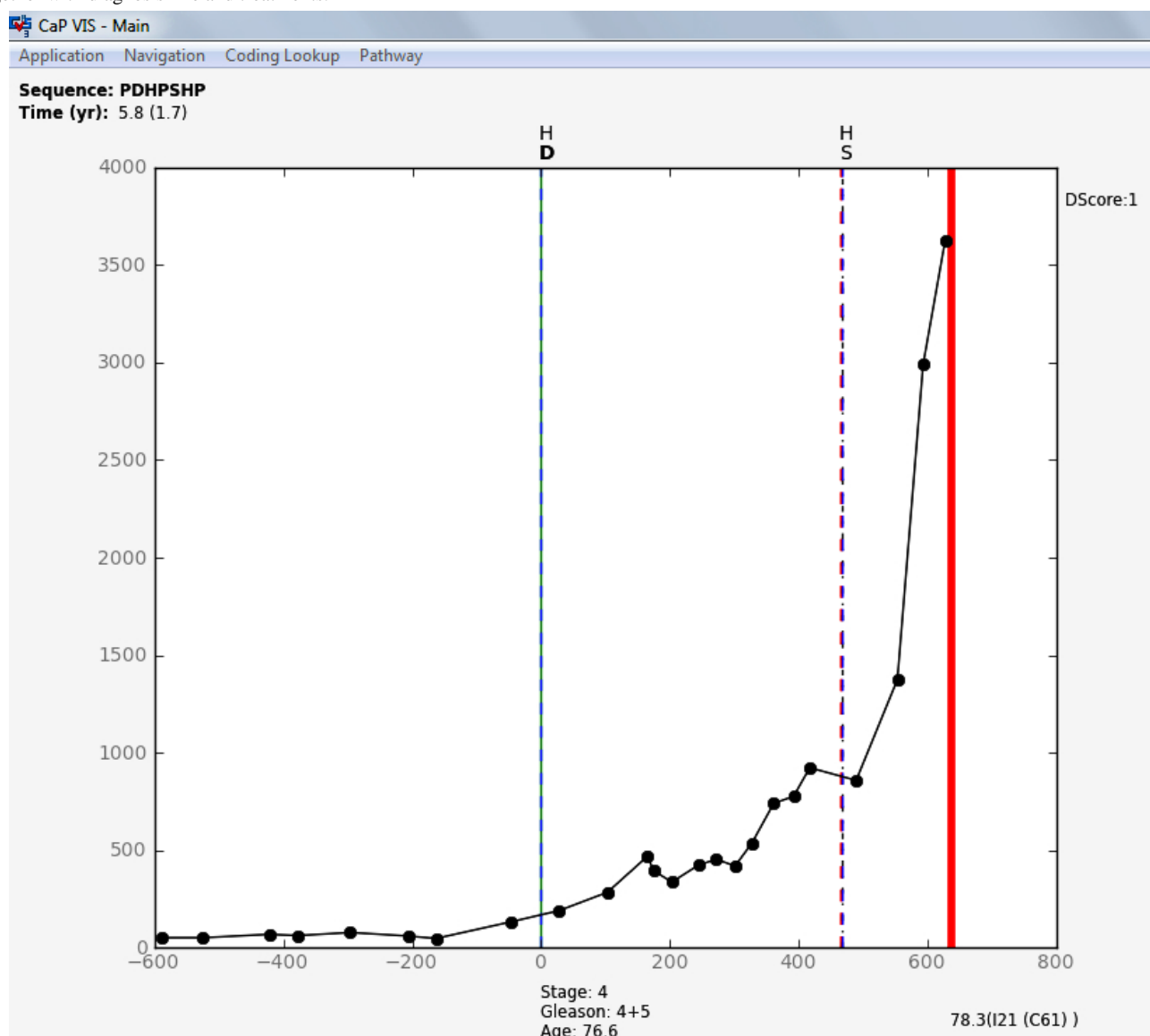
Figure 5. The carcinoma of the prostate visualization and interpretation system (CaP VIS) system showing the right side of the screen where additional information about the patient pathway is seen. There are three sections that show the histopathology report, the pathway details in the format presented in this paper, the annotations section, and a fourth section (hidden) contains pathway statistics. An overlapping window shows in detail the prostate specific antigen (PSA) kinetics before treatment, computed for this pathway. The toolbar above the histopathology section allows the user to zoom and pan the plotted pathway (in Figure 4), as well as to save the plotted figure to file, search for particular patient pathway, or navigate to the previous or next patient pathway. Doubling time (DT) and velocity (V).
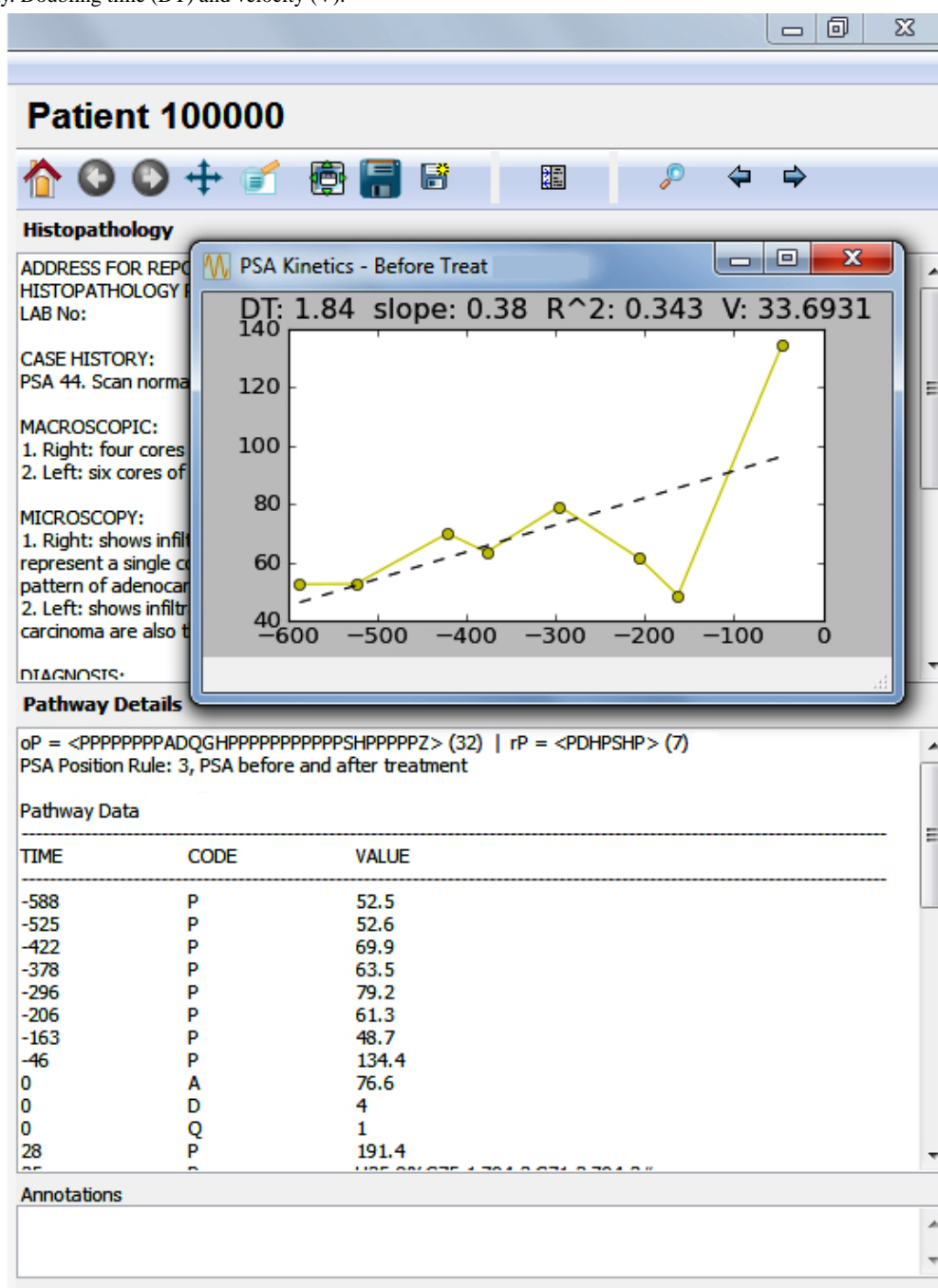
Assessing Completeness Using Biomarker Information

Routinely collected data can vary in quality, and it is important to assert the quality of the elements in the pathway so that they can be selected or discarded for clinical analysis. The above sections dealt with the definition and building of pathways from routine data, and this section introduces a method for inspecting their quality with the aim of selecting pathways for clinical studies. We already discussed that upon extraction from the ODS, data elements were previously cross-validated against a trusted source, namely, the CR. However, the biomarker information available in hospitals and included in this study enabled additional quality assurance. To this effect, we investigate methods of computing the completeness of pathways from the biomarker information. For this, rule-based scores were computed. The data elements used to assess pathway quality were the PSA and all radical treatments available in the pathway dictionary (ie, treatments that have an impact on the biomarker). The reason for choosing these elements is their interest for the analysis of prostate cancer, and hence, their ability to indicate the quality of the data for that specific purpose. Clinicians are also interested in biomarker trends and

in comparing patients under different treatment regimes. Similar methods may be applied to other variables or domains, and should allow the assessment of data quality for different clinical investigations.

Rule-Based Scores

Given the defined dictionary and its underlying format, it is possible to create a knowledge base of rules to aid the process of computing completeness scores for particular elements of the pathway. It is often difficult to convey and analyze a biomarkers' information in pathways, but here it was possible to compute their trends and to allow those computations to inform on the quality of the pathway. In the particular case of prostate cancer, the trend of PSA readings across the pathway is of interest. We identified, guided by domain experts, two major sets of rules in which the biomarker can be used to assess the completeness of a pathway with respect the clinical domain. The first set of rules relies on the position of biomarker readings in the pathway, whereas the second relies on identifying clinical interventions that justify abrupt changes in biomarker values. Rules can be applied programmatically by running Python

scripts in the analysis engine. In this case study, the rules were used to help determine the quality of the pathways for future research. However, similar rules can be built to assess adherence to guidelines, or to perform complex data queries.

Positioning of Biomarker Readings

As some of the intended clinical investigations pertain to PSA trends and associated treatments, it is important to have complete PSA trends within a pathway. In this context, a pathway should include biomarker readings before and after treatment so that the effect of treatment on the biomarker can be elucidated in posterior analyses. We can therefore compute a partial score of a pathway as a result of a set of rules on the occurrence of PSA readings. The rules are presented in Table 3 with their respective score and the percentage of pathways where the rule applied. The computation of the positioning score involves iterating through pathways' codes and flagging occurrences of the PSA and their position with respect to treatments (excluding active surveillance). The most informative pathways should have one or more readings before and after treatment and the least informative have no PSA readings.

Table 3. PSA availability and positioning rules with respective scores and coverage.

Positioning score	Rule description	Coverage, n (%)
0	No PSA readings found.	90/1904 (4.73)
1	One or more readings found before treatment (or no treatment) and none after treatment.	77/1904 (4.04)
2	One or more readings found after treatment and none before treatment.	158/1904 (8.30)
3	One or more readings found before and after treatment.	1579/1904 (82.93)

Substantiation of Biomarker Variation

Further rules can be devised to ascertain quality. For example, biological variations, in this case expressed by the PSA, should often be accompanied by evidence of some clinical intervention or other relevant factor. In the case of prostate cancer, an analysis of the PSA curve can be undertaken to identify major changes in PSA readings. In this case, the most significant drop in PSA should be associated with treatment to the prostate. A complete pathway for our purposes should attempt to provide explanations for such drops in the form of some clinical intervention. In this case, the computation of a score involves looking at every pair of PSA readings and then identifying the maximum absolute drop. Searching between the pair of readings to identify an element of substantiation, which in this case study was set to be any radical treatment, follows this. The result of this rule is a Boolean value, stating whether substantiation of a

large change in the biomarker trend was detected. Although this rule may in most cases provide relevant insights on data quality for patients with prostate cancer, the use of other biomarker variations to inform on quality should be carried out with caution, as other potential factors could introduce bias. This has been discussed in detail in [29,30].

Overall Score

An overall score for completeness can then be computed based on both positioning of biomarker readings and substantiation of major variation. It is worth noting that pathways that receive a positioning score of 0 or 1 could not have substantiation by definition, as no PSA values appear after treatment. The overall score is an ordered set of values in which the highest score is awarded to the pathways with the highest positioning scores that are substantiated. The overall scores are exemplified in Table 4 (see Multimedia Appendix 1).

Table 4. Completeness scoring system for PSA trends in prostate cancer pathways.

Overall score	Biomarker			Frequency, n (%)	Average number of unique elements
	Positioning	Substantiation	Description		
S0	0	N/A[a]	No readings found.	90/1904 (4.73)	3.26 (SD .64)
S1	1	N/A[a]	One or more readings found before treatment (or no treatment), and no readings after.	77/1904 (4.04)	4.72 (SD 1.02)
S2	2	N/A[a]	One or more readings found after treatment, and no readings before.	102/1904 (5.36)	4.71 (SD 1.03)
S3	3	No	One or more readings found before and after treatment.	393/1904 (20.64)	4.56 (SD .99)
S4	2	Yes	One or more readings found after treatment and major biomarker variation explained.	56/1904 (2.94)	4.70 (SD .88)
S5	3	Yes	One or more readings found before and after treatment and major biomarker variation explained.	1186/1904 (62.29)	4.80 (SD .92)

[a]N/A=not applicable

Results

Building Pathways

The development of a framework to build, analyze, and visualize pathways from routinely collected hospital data made it possible to create individual patient pathways for 1904 patients, while integrating clinical information from several HIS.

The developed data dictionary contains 16 elements, described in Table 2. The data sources specify whether the elements were collected from the ODS (hospital systems, together with an abbreviation of the respective system) or the CR. Elements present on both sources have been cross-validated so their quality is assured. The quality and accuracy of the data elements present in the pathways was ensured in the process of building the pathway dictionary. Quality checks were also performed when building the ODS, and additional clerical review was undertaken manually.

Table 2 shows the element's frequency, and indicates the percentage of pathways in which that particular element is present. Table 2 also gives the percentage of patients who died in this cohort during the time of observation (ie, pathways including a death event, 21.11%, 402/1904). These deaths are not exclusive to prostate cancer, and the percentage should not be used to determine a measure of survival from prostate cancer. It will be possible, however, to undertake survival analyses in future studies.

Regarding biopsies, they are only coded if performed as an inpatient event, and hence, only extensive biopsies were retrieved. As a result, biopsy events were removed from the

dictionary and are not used in the current study, but can be kept for future studies. The frequency of imaging events was low (only captured imaging events on 15.28%, 291/1904, of all pathways), and it reflects the nature of the retrieval methods from radiology, which are based on a text search of the word "prostate". Further data elements that have not been added to the dictionary here, but will be added in future studies, include further biochemistry tests as well as comorbidities and hospital stays, which may or may not be related to prostate cancer.

The analysis engine computed descriptive statistics, such as the various frequencies of the elements of the dictionary. A summary of the pathway statistics for all pathways is given in Table 5. Descriptive statistics are important as they convey information about the pathways. They can also give rise to quality indicators, but we found these methods alone not to be sufficient to determine quality.

The use of routinely collected hospital data for timed events indicates with certainty that a particular activity occurred; however, its absence may not indicate the opposite. Existing data may be used in validity checks for the completeness of the data, for example, the PSA biomarker can act as an alert for potential missing activities at particular time intervals. The pathways' data structure and analysis engine enabled the computation of completeness scores for the purpose of selecting pathways with similar data points to analyze the biomarker trend. The analysis engine allows other rules to be implemented, including measuring the time between PSA measurements, for example. The following sections show the results of the application of the rules and their impact on quality assessment for research purposes.

Table 5. Summary of pathway statistics.

Statistic	Value
Average number of unique activities	4.66 (SD 1.03)
Average pathway length	1795 days (SD 1724)
Average pathway length from diagnosis	1017 days (SD 653)
Most common activity code, n (%)	P 1723/1904 (90.49)
Five most common start codes, n (%)	P 1388/1904 (72.89)
	X 222/1904 (11.65)
	D 141/1904 (7.40)
	G 79/1904 (4.14)
	L 22/1904 (1.15)
Five most common terminal codes, n (%)	P 1394/1904 (73.21)
	Z 399/1904 (20.95)
	G 59/1904 (3.09)
	W 12/1904 (0.63)
	R 10/1904 (0.52)
Total number of unique pathway sequences	694
Most common pathways' sequence (repetitions truncated), n (%)	{P,D,G,H,P} 135/1904 (7.09)
	{P,D,G,W,P} 130/1904 (6.82)
	H[a] 907/1904 (47.63)
	S[b] 518/1904 (27.20)
	W[c] 318/1904 (16.70)
Most common treatment regimes (where first and second treatment modality are within 92 days of each other), n (%)	SW[d] 59/1904 (3.09)
	SH[e] 22/1904 (1.15)

[a]H=hormone therapy alone [b]S=surgery alone [c]W=watchful waiting alone [d]SW = surgery and watchful waiting within 92 days [e]SH = surgery and hormone therapy within 92 days

Inspection of the Positioning of Readings

The application of rules on the positioning of the PSA biomarker allowed the identification of (82.93%) 1579/1904 pathways, where it was possible to plot the trend of the biomarker through treatment (scores S3+S5 in Table 4). The framework presented above made possible the inspection of data elements in relation to other events plotted chronologically. It is also possible to compute the proximity between elements. For example, treatment elements within 90 days were grouped together to form treatment packages. The type of rules proposed here allow for the assessment of the timeliness and completeness dimensions of data quality.

Inspection of the Substantiation Rule

Overall, it was possible to ascertain the biomarker variation substantiation rule for 61.08% (1163/1904) of the pathways. We also identified that 4.14% (79/1904) of pathways with two or more PSA readings had a constant or always rising PSA trend. These were merged with the overall substantiation number, making 65.23% (1242/1904) the total number of pathways with a positive substantiation rule (scores S4+S5 in Table 4). Substantiation does not occur when a treatment element is not present in the biomarker interval of interest, or if the treatment date is inaccurate. This may indicate missing information in the case of prostate cancer. The substantiation rule allows for the elimination of pathways with insufficiently accurate information to study the biomarker trend. This rule enables the assessment of completeness and timeliness dimensions of data quality. However, it should be used with caution, in particular in other domains, as other factors may also explain the variations in the biomarker trends.

A Hybrid Scoring System

A hybrid scoring system for the completeness of the pathways combines both biomarker rules described above (positioning and substantiation), and it is given in Table 4 and examples are given in the Appendix (see Multimedia Appendix 1). The overall score ranges from least complete (score S0) to most complete (score S5), and were automatically computed based on the criteria set in the rules above. Example pathways for each computed score are given in the Appendix (see Multimedia Appendix 1). This particular set of rules aims to identify the completeness of the pathways based on the prostate cancer biomarker. It is also possible to extend the framework presented in this paper to create other quality scenarios involving more robust and detailed rules based on biomarkers or other aspects of the pathways. Examples of pathway plots automatically drawn by the CaP VIS system are available in the Appendix (see Multimedia Appendix 1) and illustrate each of the five completeness scores.

Further Analysis of Data Quality

A further analysis on surveillance regimes made possible the observation that 7.3% (25/342) of those on surveillance (as first treatment) had a subsequent treatment within at least a year, and therefore left surveillance. For those that did not have a subsequent treatment (92.6%, 317/342), it was possible to investigate any substantial drops in PSA, which may be indicative of unrecorded treatments. By establishing a drop ratio calculated as the maximum PSA drop divided by the PSA at diagnosis, we noted that 30.6% (97/317) of pathways on surveillance regimes show a drop over a 0.5 ratio, whereas 15.7% (50/317) had a drop >1. This analysis is only preliminary, but it may indicate that patients received treatment, yet these have not been recorded or carried out at this hospital. Such pathways could be excluded from analyses or be further explored to seek plausible reasons for the unexplained variation in biomarker trend. Again, this is an example of the type of analysis enabled by the framework and the pathways' data structure.

The analyses on quality also led to improvements in the data collection process. It was possible, for example, to identify patients that only had PSA readings after treatment, as well as those without PSA readings before diagnosis. This process yielded a small number of pathways (2.00%, 38/1904) where there had been earlier PSA readings, but these were not linked to the patient's main hospital number in the hospital data warehouses, and hence, were missed on retrieval (not present in the ODS). Such cases are not expected to occur frequently, and do not affect any of the hospital administration or clinical operations. However, they can diminish the amount of information available for the use of routinely collected hospital data for analysis. In this instance, as only a small number of cases were affected, they were manually fixed. The exercise, however, uncovered the need for further checks by the hospital on the data warehouse to ensure consistency of recordings.

Framework, Developed Software, and Visualization

The developed framework and visualization software enabled the visualization of all 1904 patient pathways with their corresponding biomarker trends. This gives clinicians access to trends that may have been previously much harder to observe. Furthermore, the system is flexible and extensible to include other data elements such as blood readings. For example, Figure 6 shows the PSA values and the haemoglobin (Hb) readings. The shaded area is the normal range for Hb. In this case, the drop in Hb on the day of surgery reveals perioperative bleeding. This information, when computed for all patients, would enable a study of the length of time that patients take to recover after surgery. This illustrates the flexibility of the combined framework and visualization tool, and provides access to a number of studies with data that was otherwise not readily available or contextualized. Furthermore, by plotting this, clinicians are able to see the full profile of the patient with respect to diagnosis, treatments, and how these affect the biomarker and other blood values. The pathways dictionary can continue to be developed to introduce additional information to this graphical representation.

This work contrasts with other established summarization and visualization systems, such as LifeLines [31], HARVEST [32], and others [33,34], in that it provides a succinct graphical and temporal representation that enables clinicians to promptly read a large number of data points and their interactions for a given patient in a single graph. However, this approach was developed to work with a single clinical domain of interest, while other systems may cope with multiple or overlapping domains and more complex data interactions, thereby summarizing larger amounts of information from EMRs. Nevertheless, the overall software and framework are also capable of handling the temporal complexity of constantly changing variables and producing unique meaningful visualizations for clinicians and other scientists.

Additionally, the framework presented made possible the inspection of data quality dimensions similar to those described in [11], including those that are least often assessed. The inspection of some of the dimensions, however, depends on the availability of the data elements in the sources. Currency (or timeliness) has been considered a fundamental dimension, yet it is often not assessed and only measured using a single approach [11]. The pathway data structure presented here includes time as one of the key variables, hence, it allows for the examination of currency; pathways are arranged chronologically and allow for concurrent elements. For example, in the case study, treatments within 90 days were considered as a treatment package. Another example of currency evaluation is the identification and discarding of data elements not relevant at particular intervals, as exemplified by the positioning rules. Furthermore, the plausibility and concordance dimensions were assessed with respect to PSA using the substantiation rule, the completeness dimension using the positioning rule, and correctness and completeness dimensions assessed by cross-referencing against the CR. The methods used correspond to log review (currency); validity checks (plausibility and correctness); element presence (completeness) and agreement (concordance); data source agreement; and gold standard (completeness and correctness).

The proposed framework and developed software should also allow for the selection and extraction of particular datasets with complete data for process mining and other analysis. It has been reported that the evaluation of the quality of process mining event logs relies on trustworthiness (recorded events actually happened), completeness, and well defined semantics [35]. These can be achieved by selecting pathways with required data points using the proposed framework. Furthermore, the visualization system allows for the close inspection and contextualization of pathways, illustrating particular paths with similar features, such as the ones exemplified in the Appendix (see Multimedia Appendix 1). In summary, the proposed framework, when used in hospitals, would facilitate the retrieval, selection, and inspection of patient pathways, and also the further steps of data mining analysis using appropriate methodologies.

XSL•FO

RenderX

Figure 6. Pathway plot showing the prostate specific antigen (PSA) (round markers) and haemoglobin (Hb) readings (star markers) together. As a result of the prostatectomy event (S) the PSA dropped and Hb also dropped due to normal perioperative bleeding. The shaded area denotes the normal range for Hb.
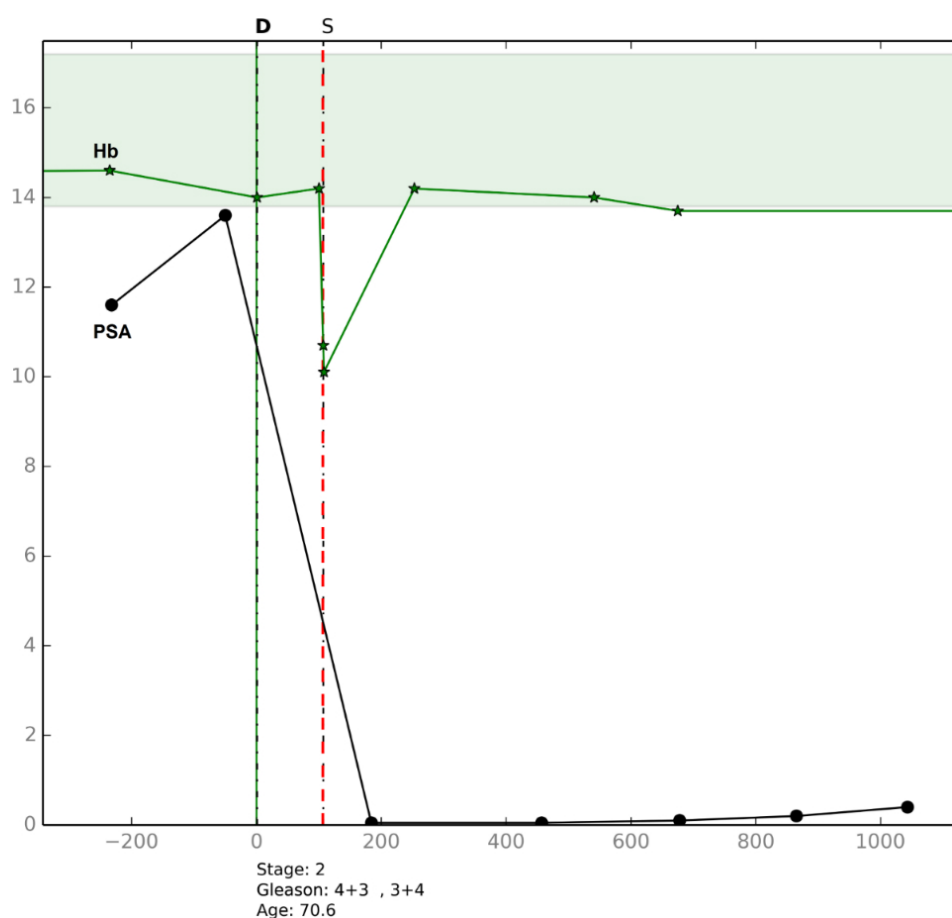
Discussion

Principal Findings

Based on the prostate cancer case study carried out at a large regional NHS hospital, a framework, which enables the secondary uses of routinely collected hospital data, was developed and presented in this paper. The main components of this framework (Figure 3) are the ODS containing patient-centric data, used to build the pathways based on the methodology presented in Figure 2; the pathways engine; analysis engine; and the visualization software. The underlying pathway data structure, in some aspects similar to the EAV data model, retains some degree of patient privacy and together with the dictionary provides a simplified, yet flexible and powerful, platform for the complex querying and analysis of patient information and disease pathways. It enables the summarization and extension of pathways, as well as the aggregation of similar sequences. It is also possible to capture and plot pathways with concurrent elements, and to develop algorithms to further explore the data and investigate quality issues. Furthermore, the methodology used to build the pathway dictionary, as well as the formalisms presented here, can be transported to other domains and settings. This is particularly true because the pathways dictionary can be remodelled to accommodate other

data elements and research interests. Likewise, the framework is capable of plotting other continuous or categorical variables. The software has also been developed in a way that accommodates changes, as it focuses on the pathway data model (subsection "Defining a Pathway") that is not designed for a specific disease. Nevertheless, in this paper, the pathways were constructed using a case study on prostate cancer, and further work is underway to apply these methods to other domains, where the emphasis is on different clinical parameters.

The process of integrating routinely collected electronic data may produce pathways that may not be informative or complete. A topic, which, to our knowledge, has received little attention in the literature, is the computation of quality indicators for data-driven pathways. Such indicators are important to enable the selection of study-relevant high quality data for clinical investigation. The methods developed in this paper enable us to discard pathways that, because of the nature of electronically routinely collected hospital data, fail to provide enough or sufficiently accurate information to be used in clinical analyses.

We have shown that methods for pathway quality measurement can rely on biological marker trends, as they are often the response to some parallel process. In the case of the PSA, a sharp decline in the average readings would most likely indicate treatment to the prostate, which suppressed the production of

the antigen. This allows us to ascertain whether treatment records are missing. Similar approaches, however, should be used carefully so as to take into account any possible confounding factors. Algorithms were written to compute completeness based on prostate cancer biomarker rules, creating an overall scoring system (Table 4). Once researchers are satisfied that the PSA trends have sufficient data points and are substantiated (ie, they receive a high completeness score), they can investigate those PSA trends as predictors of prognosis in the disease. Such research is seldom undertaken due to the unavailability of data, but may lead to improved outcomes for patients and health services.

We investigated the cohort of 1904 patients, automatically built their respective pathways, and computed completeness scores with regards to the prostate cancer biomarker. Overall, 65.23% (1242/1904) of pathways attained the two highest scores, while 82.93% (1579/1904) attained the highest PSA positioning rule. Hence, these pathways contain sufficient biomarker information to aid clinical investigations on the biomarker trends. We have shown that routinely collected data can be transformed and prepared for clinical research, decision making, and decision support.

The flexibility of the data structure allows the insertion and removal of dictionary elements, and work is underway to include additional blood tests and comorbidities to the pathways, as depicted in Figure 6. The work presented here has also enabled future research into pathway adherence and variance metrics, particularly with respect to the UK NICE guidelines. This work is possible in the first instance by analyzing similar pathway sequences, and then by programmatically accessing detailed pathway information using the analysis engine. This paper describes methods for data collection, presentation, and quality assessment that can be applicable to build other disease pathways in other settings. We are also motivated by further work on mining pathways, in particular, the computation of similarity of biomarker trends, and the application of clustering algorithms and survival analysis in the context of pathways.

Limitations

The framework and pathways were built using a case study on prostate cancer where there was a particular clinical interest on the biomarker trends. This specific working domain may introduce some limitations to the reproducibility of this work;

however, further research is underway to apply the approach to other domains, specifically in the construction of pathways for acute stroke.

The number of data elements used in the pathway data dictionary was also a limiting factor, however, they were sufficient to study the PSA trends and to select cohorts with similar baseline features for further research. The pathways data structure presented in this paper has coped with the addition of new data elements, but further work is required to assess the quality and availability of other routinely collected data. Further work on the methods for evaluating quality is also needed, and it is hoped that the adoption of systematic methods, such as those presented in this paper, encourages further research in this area.

With regards to privacy, the pathways data structure includes an anonymized patient identification number, replaces specific dates with time zeroed at diagnosis, and suppresses patient names, addresses, and postcodes. These have been sufficient to ensure the anonymity of the patients. However, it may be possible to utilize specific information to attempt to identify individuals, particularly as new data elements containing specific information are added. Further work may be required to anonymize additional information, such as histopathological text reports, and to ensure that the system is fully resistant to privacy attacks.

The timeliness of the process of retrieving routine data and feeding them into the pathways database depends on the availability of the data in the ODS, and it can be a limitation. Similarly, the process of building the pathways dictionary and liaising with domain experts may introduce delays. However, once the dictionary is agreed and the data and their sources are fully understood and accessible, creating individual pathways in near real-time is possible. In this case study, the process of transforming data from the ODS into the pathways database for a new data element could be achieved in a few hours, however, the retrieval of the data from the sources onto the ODS and liaison with domain experts and other hospital staff could introduce significant delays up to several weeks. This case study was also undertaken in a single large hospital and, although the challenges are reportedly similar elsewhere, it is expected that the time and effort to feed new routine data can vary considerably.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Examples of pathway plots drawn by the developed CaP VIS system for each of the six possible completeness scores.

[TIF File, 335KB - medinform_v3i3e26_app1.tif]

References

1. Vanhaecht K, Panella M, van Zelm R, Sermeus R. An overview on the history and concept of care pathways as complex interventions. International Journal of Care Pathways 2010 Sep 01;14(3):117-123. [doi: 10.1258/jicp.2010.010019]

2. Pearson SD, Goulart-Fisher D, Lee TH. Critical pathways as a strategy for improving care: Problems and potential. Ann Intern Med 1995 Dec 15;123(12):941-948. [Medline: 7486490]

3. Luc K, Todd J. e-pathways: Computers and the patient's journey through care. In: E-pathways: Computers And the Patient's Journey Through Care. UK: Radcliffe Medical Press LTD; 2003.

4. Wakamiya S, Yamauchi K. What are the standard functions of electronic clinical pathways? Int J Med Inform 2009 Aug;78(8):543-550. [doi: 10.1016/j.ijmedinf.2009.03.003] [Medline: 19345608]

5. Veselý A, Zvárová J, Peleska J, Buchtela D, Anger Z. Medical guidelines presentation and comparing with electronic health record. Int J Med Inform 2006;75(3-4):240-245. [doi: 10.1016/j.ijmedinf.2005.07.016] [Medline: 16165396]

6. Spreckelsen C, Spitzer K, Honekamp W. Present situation and prospect of medical knowledge based systems in German-speaking countries: Results of an online survey. Methods Inf Med 2012;51(4):281-294. [doi: 10.3414/ME11-01-0084] [Medline: 22526065]

7. Huang Z, Lu X, Duan H. On mining clinical pathway patterns from medical behaviors. Artif Intell Med 2012 Sep;56(1):35-50. [doi: 10.1016/j.artmed.2012.06.002] [Medline: 22809825]

8. Lang M, Bürkle T, Laumann S, Prokosch HU. Process mining for clinical workflows: Challenges and current limitations. Stud Health Technol Inform 2008;136:229-234. [Medline: 18487736]

9. Mans R, Schonenberg M, Song M, Bakker P. Application of process mining in healthcare - A case study in a Dutch hospital. In: Fred A, Filipe J, Gamboa H, editors. Biomedical Engineering Systems and Technologies. Berlin: Springer Berlin Heidelberg; 2009:425-438.

10. Mans R, Schonenberg H, Leonardi G, Panzarasa S, Cavallini A, Quaglini S, van der Aalst Wil. Process mining techniques: An application to stroke care. Stud Health Technol Inform 2008;136:573-578. [Medline: 18487792]

11. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: Enabling reuse for clinical research. J Am Med Inform Assoc 2013 Jan 1;20(1):144-151 [FREE Full text] [doi: 10.1136/amiajnl-2011-000681] [Medline: 22733976]

12. Liaw ST, Rahimi A, Ray P, Taggart J, Dennis S, de LS, Yeo A E T, et al. Towards an ontology for data quality in integrated chronic disease management: A realist review of the literature. Int J Med Inform 2013 Jan;82(1):10-24. [doi: 10.1016/j.ijmedinf.2012.10.001] [Medline: 23122633]

13. Sujansky W. Heterogeneous database integration in biomedicine. J Biomed Inform 2001 Aug;34(4):285-298 [FREE Full text] [doi: 10.1006/jbin.2001.1024] [Medline: 11977810]

14. Bettencourt-Silva J, De La Iglesia B, Donell S, Rayward-Smith V. On creating a patient-centric database from multiple Hospital Information Systems. Methods Inf Med 2012;51(3):210-220. [doi: 10.3414/ME10-01-0069] [Medline: 21818520]

15. Globocan, prostate cancer incidence, mortality and prevalence worldwide in 2008, GLOBOCAN, 2008. 10th, 2015 URL: http://globocan.iarc.fr/ [accessed 2015-01-10] [WebCite Cache ID 6VTxujNe1]

16. Cancer Research UK, Prostate cancer statistics. 2015. URL: http://www.cancerresearchuk.org/cancer-info/cancerstats/types/prostate/ [accessed 2015-01-10] [WebCite Cache ID 6VTxUF5AS]

17. Collin SM, Martin RM, Metcalfe C, Gunnell D, Albertsen PC, Neal D, et al. Prostate-cancer mortality in the USA and UK in 1975-2004: An ecological study. Lancet Oncol 2008 May;9(5):445-452 [FREE Full text] [doi: 10.1016/S1470-2045(08)70104-9] [Medline: 18424233]

18. Roehrborn CG, Black LK. The economic burden of prostate cancer. BJU Int 2011 Sep;108(6):806-813. [doi: 10.1111/j.1464-410X.2011.10365.x] [Medline: 21884356]

19. Holmström B, Johansson M, Bergh A, Stenman UH, Hallmans G, Stattin P. Prostate specific antigen for early detection of prostate cancer: Longitudinal study. BMJ 2009;339:b3537 [FREE Full text] [Medline: 19778969]

20. National institute for health and clinical excellence. Prostate cancer overview URL: http://pathways.nice.org.uk/pathways/prostate-cancer [accessed 2015-06-22] [WebCite Cache ID 6ZU18HCyb]

21. Cios KJ, Moore W. Uniqueness of medical data mining. Artif Intell Med 2002;26(1-2):1-24. [Medline: 12234714]

22. Safran C, Perreault L. Management of information in integrated delivery networks. In: Shortliffe EH, Perreault LE, Wiederhold G, Fagan LM, editors. Medical Informatics: Computer Applications in Health Care and Biomedicine. New York: Springer; 2003:359-396.

23. de Keizer N F, Ammenwerth E. The quality of evidence in health informatics: How did the quality of healthcare IT evaluation publications develop from 1982 to 2005? Int J Med Inform 2008 Jan;77(1):41-49. [doi: 10.1016/j.ijmedinf.2006.11.009] [Medline: 17208040]

24. Friedman C, Hripcsak G, Johnson S, Cimino J, Clayton P. NCBI. 1990. A generalized relational schema for an integrated clinical patient database URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2245527/ [accessed 2015-07-05] [WebCite Cache ID 6ZnzCwhP0]

25. Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc 2010;17(2):124-130 [FREE Full text] [doi: 10.1136/jamia.2009.000893] [Medline: 20190053]

XSL•FO
RenderX

26.   Brazhnik O, Jones JF. Anatomy of data integration. J Biomed Inform 2007 Jun;40(3):252-269 [FREE Full text] [doi: 10.1016/j.jbi.2006.09.001] [Medline: 17071142]

27.   O'Connor MJ, Shankar R, Parrish DB, Das AK. Knowledge-data integration for temporal reasoning in a clinical trial system. Int J Med Inform 2009 Apr;78 Suppl 1:S77-S85. [doi: 10.1016/j.ijmedinf.2008.07.013] [Medline: 18789876]

28.   Thompson IM, Carroll PR, Carducci MA. Recommendations for defining and treating high risk localized prostate cancer. J Urol 2006 Dec;176(6 Pt 2):S6-S10; quiz S3. [doi: 10.1016/j.juro.2006.09.016] [Medline: 17084170]

29.   Pivovarov R, Albers DJ, Sepulveda JL, Elhadad N. Identifying and mitigating biases in EHR laboratory tests. J Biomed Inform 2014 Oct;51:24-34. [doi: 10.1016/j.jbi.2014.03.016] [Medline: 24727481]

30.   Pivovarov R, Albers DJ, Hripcsak G, Sepulveda JL, Elhadad N. Temporal trends of hemoglobin A1c testing. J Am Med Inform Assoc 2014;21(6):1038-1044. [doi: 10.1136/amiajnl-2013-002592] [Medline: 24928176]

31.   Plaisant C, Milash B, Rose A. LifeLines: Visualizing personal histories. New York, NY, USA: ACM; 1996 Presented at: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems; 1996; Vancouver p. 221-227. [doi: 10.1145/238386.238493]

32.   Hirsch JS, Tanenbaum JS, Lipsky GS, Liu C, Schmitz E, Hashorva D, et al. HARVEST, a longitudinal patient record summarizer. J Am Med Inform Assoc 2015 Mar;22(2):263-274 [FREE Full text] [doi: 10.1136/amiajnl-2014-002945] [Medline: 25352564]

33.   Pivovarov R, Elhadad N. Automated methods for the summarization of electronic health records. J Am Med Inform Assoc 2015 Apr 15 [FREE Full text] [doi: 10.1093/jamia/ocv032] [Medline: 25882031]

34.   West VL, Borland D, Hammond WE. Innovative information visualization of electronic health record data: A systematic review. J Am Med Inform Assoc 2015 Mar;22(2):330-339 [FREE Full text] [doi: 10.1136/amiajnl-2014-002955] [Medline: 25336597]

35.   van der Aalst W, Adriansyah A, de Medeiros AKA. Process mining manifesto. In: Business process management workshops. Berlin: Springer Berlin Heidelberg; 2012:169-194.

Abbreviations

CaP VIS: carcinoma of the prostate visualization and interpretation system
CR: cancer registry
CSV: comma separated file
EAV: entity-attribute-value
EHR: electronic health records
Hb: Haemoglobin
HIS: hospital information systems
ICD: International Classification of Diseases
LAB: laboratory information system
NHS: National Health Service
NNUH: Norfolk & Norwich University Hospital
NICE: National Institute for Clinical Excellence
ODS: operational data store
PSA: prostate specific antigen

XSL•FO
RenderX

Original Paper

Analysis of PubMed User Sessions Using a Full-Day PubMed Query Log: A Comparison of Experienced and Nonexperienced PubMed Users

Illhoi Yoo[1,2*], PhD; Abu Saleh Mohammad Mosa[2,3*], M. S.

[1]Department of Health Management and Informatics, School of Medicine, University of Missouri, Columbia, MO, United States

[2]Informatics Institute, University of Missouri, Columbia, MO, United States

[3]Institute for Clinical and Translational Science, School of Medicine, University of Missouri, Columbia, MO, United States

[*]all authors contributed equally

Corresponding Author:
Illhoi Yoo, PhD
Department of Health Management and Informatics
School of Medicine
University of Missouri
Five Hospital Dr.
CE718 Clinical Support and Education Building (DC006.00)
Columbia, MO, 65212
United States
Phone: 1 573 882 7642
Fax: 1 573 882 6158
Email: yooil@health.missouri.edu

Abstract

Background: PubMed is the largest biomedical bibliographic information source on the Internet. PubMed has been considered one of the most important and reliable sources of up-to-date health care evidence. Previous studies examined the effects of domain expertise/knowledge on search performance using PubMed. However, very little is known about PubMed users' knowledge of information retrieval (IR) functions and their usage in query formulation.

Objective: The purpose of this study was to shed light on how experienced/nonexperienced PubMed users perform their search queries by analyzing a full-day query log. Our hypotheses were that (1) experienced PubMed users who use system functions quickly retrieve relevant documents and (2) nonexperienced PubMed users who do not use them have longer search sessions than experienced users.

Methods: To test these hypotheses, we analyzed PubMed query log data containing nearly 3 million queries. User sessions were divided into two categories: experienced and nonexperienced. We compared experienced and nonexperienced users per number of sessions, and experienced and nonexperienced user sessions per session length, with a focus on how fast they completed their sessions.

Results: To test our hypotheses, we measured how successful information retrieval was (at retrieving relevant documents), represented as the decrease rates of experienced and nonexperienced users from a session length of 1 to 2, 3, 4, and 5. The decrease rate (from a session length of 1 to 2) of the experienced users was significantly larger than that of the nonexperienced groups.

Conclusions: Experienced PubMed users retrieve relevant documents more quickly than nonexperienced PubMed users in terms of session length.

XSL•FO
RenderX

Introduction

Background

Methods of information seeking have become much easier, faster, and inexpensive since the 1990s with the advent of information technologies (ITs) including the Internet, digital libraries (eg, electronic full-text databases), and online search software/services such as Google Scholar and PubMed. [1-3]. Since then, immense change in scientific-information-seeking behavior has been observed, including among professionals, scholars, and scientists in the area of biomedical and health sciences [3-6]. There is unprecedented growth of biomedical information, which has been doubling every 5 years [7,8]. This large amount of scientific information from multiple sources (eg, journals) is currently integrated in electronic bibliographic databases and accessible through online search software [3,9]. For example, PubMed, which is maintained by the United States National Library of Medicine (NLM), is one of the largest and most authoritative online biomedical bibliographic databases in the world [10-12]. As of June 2015, PubMed contained more than 24 million citations and abstracts from approximately 5600 biomedicine and health-related journals. Health care professionals consider PubMed to be one of the most important and reliable sources of up-to-date health care evidence [13,14]. PubMed also plays a very important role in the process of literature-based discovery [15].

Recent years have seen a rising trend in biomedical information seeking from PubMed [16,17]. About two-thirds of PubMed users are domain experts (eg, health care professionals) and one-third are lay people [18]. Previous studies have examined the effects of domain expertise/knowledge on search performance using PubMed [6,19-21]. However, very little is known about PubMed users' knowledge of information retrieval (IR) functions and their usage in query formulation.

The goal of this study was to shed light on how PubMed users perform their search queries by analyzing a full-day query log. The hypotheses of this study were that (1) experienced PubMed users who use system functions such as Medical Subject Heading (MeSH) terms and search field tags quickly retrieve relevant documents and (2) nonexperienced PubMed users who do not use them have longer search sessions than experienced users, because they identify their information needs through subsequent queries by narrowing and/or broadening their queries. In order to test the hypotheses, we analyzed a full day of PubMed log data. We assumed that if a session was closed within a few queries, the session was successful (meaning that relevant documents were retrieved), even if a session close did not always mean successful IR.

In this study, experienced PubMed users were defined as users who used advanced PubMed IR functions for query formulation. The proper use of IR functions (described in the next section) is key for efficient and effective PubMed searches [6,8,22-27] because, unlike Google, PubMed does not sort search results by relevance. Studies have shown that experienced users are more likely to use IR functions than novice users. Xie and Joo (2012) [28] performed a study on factors affecting the selection of search tactics and demonstrated that expert participants were

more willing to use advanced IR functions. The study [28] used the definition of expert IR users from Holscher and Strube (2000) [29], in which expert users were defined as users having the "knowledge and skills" necessary to utilize information-seeking systems successfully. Holscher and Strube (2000) [29] also recognized that "expert users use advanced IR functions much more than average users." Earlier studies also demonstrated that experienced searchers are more knowledgeable of the content and structure of the IR system and more likely to interact with the system [30,31]. Penniman (1981) [32] defined experienced PubMed users based on the frequency of PubMed searches and concluded that experienced searchers use more search functions than nonexperienced searchers. In addition, many studies have demonstrated that experienced users use more advanced IR functions and show better IR performance than novices [33-37].

PubMed System Functions

PubMed system functions include search field tags, MeSH terms (used for indexing PubMed articles), truncation, and combining searches using search history. In PubMed, bibliographic information is stored in a structured database with 65 fields including title, abstract, author, journal or proceeding, publication type, and publication date. PubMed provides 48 search field tags in order to facilitate searching in its various database fields; a description for each search field is available at the NLM website [38] (last revised and updated November 2012). Thus, PubMed is a field-oriented search system in which search terms are tagged with search field tags and appended using Boolean operators (ie, AND, OR, and NOT). Using search field tags, PubMed users can limit the search to a specific field for each search keyword. A search field tag is attached to a search term by enclosing the search field name in square brackets (eg, "myocardial infarction" [Title]). The NLM indexes PubMed documents using the MeSH vocabulary after indexers read full papers (not just abstracts). Usually, 5-10 MeSH terms are assigned to a PubMed document. Truncation is used to search for the first 600 variations of a truncated word in PubMed. However, PubMed allows an asterisk (*) at the end of a word only; "?" is not used in PubMed. For example, the search term nutrition* will search for nutritional and nutritionists. Finally, the combining search function using search history enables PubMed users to readily use and combine previous search results using Boolean operators and search history indexes. For example, after a PubMed search for *diabetes mellitus*, the search result can be readily combined with one using a new search keyword *hypertension*: #1 AND *hypertension* (#1 indicates *diabetes mellitus*).

Related Studies

The study of information-seeking behavior is very important for the user-centric design of online IR systems including digital libraries. Individuals' knowledge and skills related to information seeking are the primary determinants of their online IR performance. According to Marchionini (1995) [39], there are four types of expertise that determine information-seeking performance: general cognitive abilities, domain knowledge, overall experience of online information seeking, and experience or knowledge of the functions of the IR system. Most intellectual

activities like the information-seeking process involve planning (eg, query term selection), progress monitoring (eg, the number of returned documents), decision making (eg, when to continue or stop the search), and reflecting on past activities (eg, refining the search query for a better search result). Marchionini (1995) [39] stated that people's perceptual and cognitive processes (known as cognitive abilities) are used in completing these tasks. As a common expectation, a person with higher cognitive abilities should perform better at information seeking than someone with lower cognitive abilities. However, few studies have investigated which cognitive abilities are linked to information seeking performance [1,29,39-42]. Hersh et al (2002) [24] assessed three cognitive factors (spatial visualization, logical reasoning, and verbal reasoning) that were found to affect IR performance, and found that PubMed/MEDLINE search experience and spatial visualization were the main factors in successful PubMed searches.

The second major area of expertise is the knowledge of information seekers in their area of interest (known as domain knowledge). The NLM reported that almost two-thirds of PubMed users are health care professionals and scientists (ie, domain experts), whereas the remainder are the general public [18]. Studies have demonstrated that methods of conducting information seeking tasks by domain experts are different from those of novice users [1,5]. In addition, overall IR performance of domain experts is better than that of novice users in various IR systems such as web and hypertext searches [29,42-46], and online bibliographic database searches [33,42,47]. A similar result has also been observed for PubMed searches [20,48]. PubMed search studies demonstrated that PubMed users with domain knowledge usually spent less time and retrieved more information than PubMed users with less domain knowledge. On the other hand, some studies measured user-searching performance (in terms of recall and precision) and concluded that domain knowledge did not significantly affect information-seeking performance. These studies were performed with the DIALOG database [49], an online library catalog [50], and the MEDLINE search system [19-21].

The other two determinants of search performance (ie, overall experience using online information seeking and experience or knowledge of the functions of the IR system) can be considered together as procedural knowledge for using the IR system [6]. Previous studies have demonstrated that such experience improves IR performance for various search systems such as web, hypertexts, file collections, and bibliographic DBs including PubMed [21,24,35,42,44,45,51]. Egan (1988) [52], Hölscher and Strube (2000) [29], and Jenkins et al (2003) [44] found that domain knowledge helped to improve search performance only if users had sufficient procedural knowledge including experience with online searching and search software/systems. In their literature review, Vibert et al (2009) [6] mainly compared the effects of domain knowledge on PubMed searches between expert and novice groups, and demonstrated that domain knowledge does not help to improve search performance if users do not have procedural knowledge. In addition, the study [6] suggested that knowledge in a broad scientific field can compensate for a lack of knowledge in a specific domain, and that the main determinant of bibliographic

search performance is individual cognitive abilities. Thus, people with basic domain knowledge in their area of interest, higher cognitive abilities, and sufficient procedural knowledge regarding the bibliographic search system should efficiently perform information-seeking tasks (eg, query selection and decisions about search discontinuation). Some recent studies found that most academic researchers and health care professionals including physicians do not use advanced IR functions but only natural language for PubMed searches [6,51,53-55]. Another very recent study of PubMed by Macedo-Route et al (2012) [56] concluded that the way researchers use PubMed is nearly the same as the way IR novices do ("mostly typing a few keywords and scanning the titles retrieved by the tool"). Several studies have shown that medical librarians (considered experienced users in the study) use more IR functions for PubMed searches and their IR performance is better than regular users [20,36,57,58].

In this study, our goal was to compare experienced versus nonexperienced users' searching behavior in terms of session length (ie, the number of queries per session). We used a full-day PubMed query log for that purpose. There are a number of approaches for studying user-searching behavior such as eye tracking, surveys, and search log analysis. Search log analysis has become a viable solution for many applications including search engines [16,17,59-63]. One major advantage of search log analysis over other methods is that actual searches by a large number of real users can be analyzed, while other methods usually examine searches from only tens up to hundreds of users. A search engine stores users' query text along with other information including user IP addresses in query log files.

Silverstein et al (1999) [59] and Jansen et al (2000) [60] analyzed a query log from the AltaVista and Excite web search engines, respectively. Silverstein et al (1999) [59] reported three important facts: (1) users rarely navigate beyond the first page of search results, (2) they rarely resubmit a refined query (similar to Jansen et al (2000)'s [60] finding), and (3) most queries are short in length. Herskovic et al (2007) [16] carried out a similar study with a PubMed log and reported statistical information on PubMed usage (including the number of users, queries per user, sessions per user, and frequently used search terms and search field tags). The PubMed log data were used for segmenting query sessions [64], evaluating the PubMed Automatic Term Mapping (ATM) [65], and annotating PubMed queries using the Unified Medical Language System (UMLS) [66]. NLM researchers used month-long PubMed log data for categorizing PubMed queries [17,66], creating a query suggestion database [67], and identifying related journals for user queries [68]. Both of the full-day-long and month-long datasets are publicly available. However, the month-long dataset does not contain actual user queries. For this reason, we used the full-day-long PubMed log data.

The focus of this study is different from that of the eight studies that used PubMed log data [16,17,63-68]. We focused on comparing experienced versus nonexperienced users' searching behavior in terms of session length (the number of queries in a session). To the best of our knowledge, there is no study with this focus.

Methods

Data Cleaning and Preprocessing

The dataset used in this study is a plain text file containing a full-day's query log of PubMed that was obtained from the NLM FTP site (Refer to [69] to access the data). There are nearly 3 million queries issued by 626,554 distinct users.

The data cleaning and preprocessing steps are presented in Figure 1. We found 1146 records with empty user IDs, 76 records with unusual user IDs (we believe they were errors), and 77,923 records with no user-query text. These records (79,145/2,996,301, 2.64%) were eliminated from the dataset.

Figure 1. Data cleaning and preprocessing.
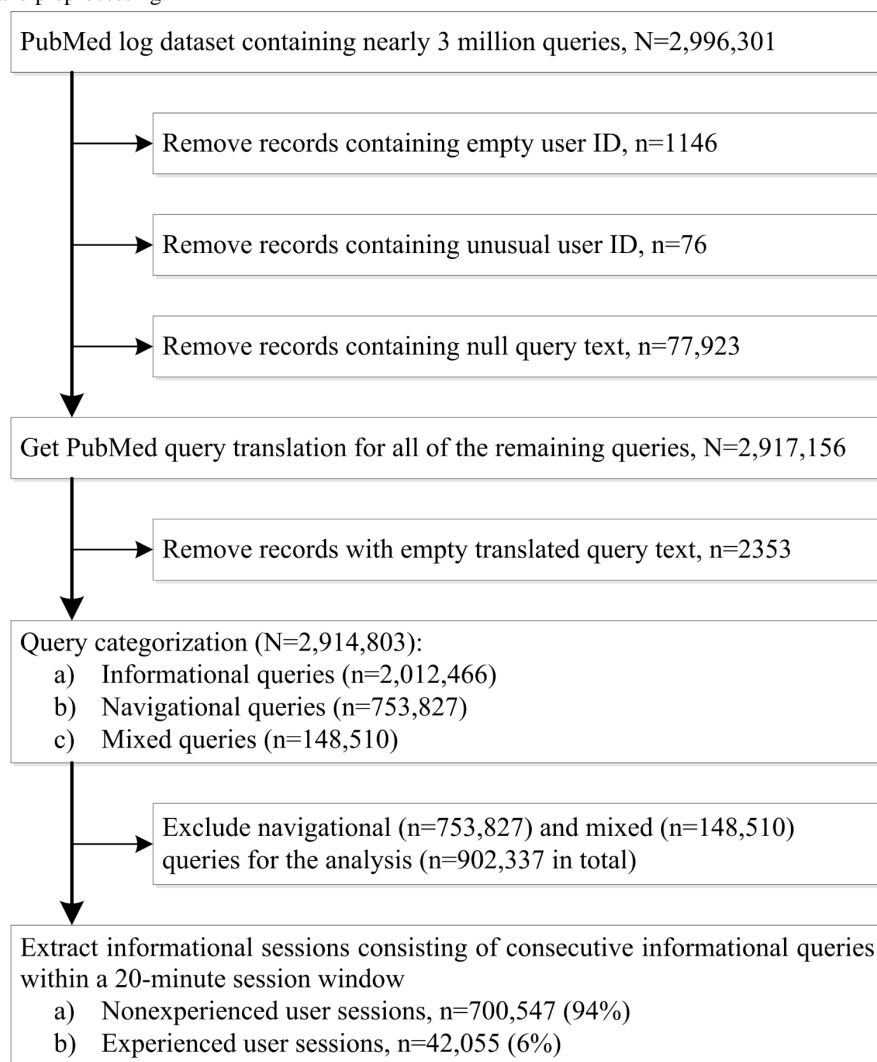
Query Categorization

The user queries in the PubMed log file are categorized as informational, navigational, or mixed according to the purpose of the search expressed in the query. Informational queries are intended to fulfill end users' information needs (eg, "diabetes mellitus" [MeSH]) and navigational queries are intended to retrieve specific documents (eg, Yoo [author] AND Mosa [author]). Mixed queries have both intentions (eg, searching for a specific topic within a specific journal). Refer to Broder (2002) [70] and Herskovic et al (2007) [16] for details of web search types and PubMed search types, respectively.

In order to identify the purpose of user queries for query categorization, we used PubMed's ATM. Every PubMed user query is automatically translated by ATM to improve overall IR performance and the translated query is actually used for the PubMed search; if a query contains double quotation marks or search tags, those parts (words or terms) are not translated. The ATM translation identifies each term in a query and adds an appropriate search tag to the term. We categorized PubMed queries using ATM-added tags as well as user-added tags after ATM translations. PubMed provides 48 search tags (refer to the PubMed Help website [71] for details), which are classified into informational and navigational tags [69]. Queries containing only informational tags are identified as informational queries. Navigational queries are queries containing navigational or citation-related tags. Queries containing both informational and navigational tags are identified as mixed queries, unless the original query contains an indication of a navigational query. Figure 2 presents a flow diagram for query categorization. A total of 2353 queries resulted in empty query translation. These

were removed from the analysis. The translated query texts were then parsed to extract the search tags.

The search tag extraction process involved a semiautomatic approach consisting of two steps: the semiautomatic construction of a list of search tags and their variations, and the automatic extraction of the search tags including their variations from the queries using the search tag list. A total of 963 unique substrings were extracted from the queries in the first step. The first step (a partial manual step) was required for two reasons: (1) for each search tag there are several variations that are not fully documented even though they are correctly recognized by the PubMed system; for example, [Author Name], [Author], [AU Name], [Auth], and [AU] represent the same search tag header but only [Author Name] and [AU] are documented in the PubMed Help web page, and (2) incorrect search tags (eg, typos like [Atuhor]) used in PubMed queries are not recognized by the PubMed system but a domain expert could correctly recognize and read those intentions. The extracted search tags from the translated queries were then analyzed to identify query types. Since navigational search tags are mainly used to retrieve specific documents rather than to fulfill information needs, we excluded navigational and mixed queries from the analysis, assuming informational search tags are primarily used for information needs.

Figure 2. Query categorization.
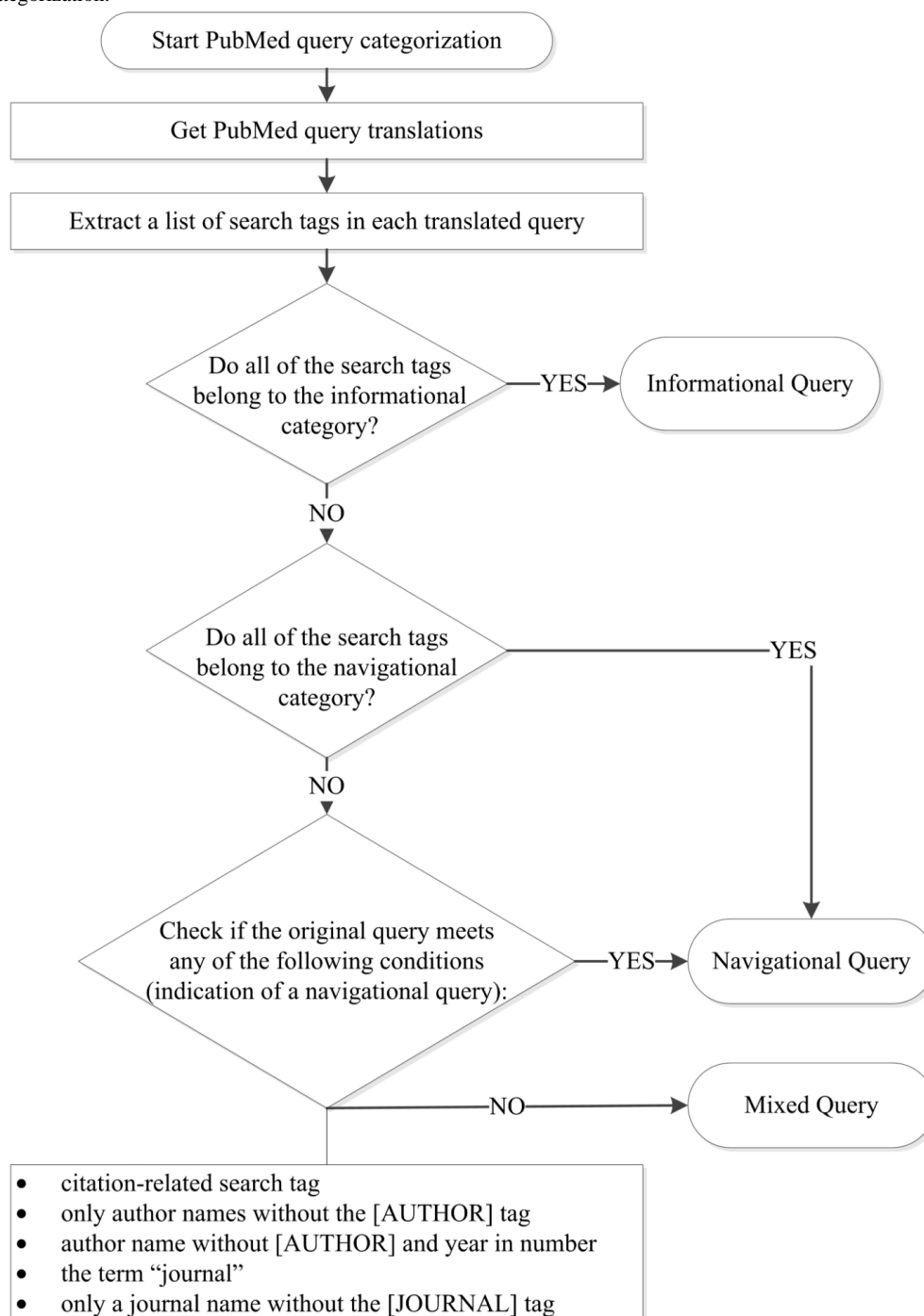
Session Segmentation

Information seeking is defined as "the process of repeatedly searching over time in relation to a specific, but possibly an evolving information problem" [72]. Swanson et al (1977) [73] defined information seeking as a trial-and-error process, in which the initial search query is refined at every step, based on the search results in the previous queries. IR users often perform multiple queries in a row for the same information problem. The IR community has coined the term session in this regard. Silverstein et al (1999) [59] defined a session as "a series of queries by a single user made within a small range of time; a session is meant to capture a single user's attempt to fill a single information need." In order to segment queries by a user into sessions, most studies utilized temporal clues such as temporal threshold (ie, time cutoff) between two consecutive queries [59,74-78] or temporal constraint [79] (Refer to a recent survey article by Gayo-Avello (2009) [80] for details). This process (ie, session segmentation) provides valuable insights into users' search behavior and interactions with the IR system.

In this study, we employed both the session-shift and temporal-constraint-based sliding window for session segmentation. This is because several studies reported the average duration of user sessions for query log analysis (meaning that the maximum length of session window can be chosen based on those results for session segmentation) [81-83]. In our study, we set the maximum length of the sliding window to be 20 minutes. The choice of a 20-minute session window was based on two biomedical IR studies. The first was a qualitative study with human subjects that showed most PubMed users successfully completed their task within a 15-minute period, whereas many took more than 15 minutes [6]. The second was a randomized controlled trial on biomedical information retrieval demonstrating that the average time to solve a biomedical information problem ranges from 14 to 17 minutes [84]. In addition to temporal constraint, we used change of query types as session shift. As a result, a change from an informational query to a navigational query was considered a session boundary.

Using this method, we extracted 742,602 user sessions from more than 2 million informational queries. User sessions were divided into two categories: experienced and nonexperienced. Experienced sessions were those in which queries were formed using system functions such as MeSH terms and search field tags. Otherwise, a user session was considered nonexperienced. For example, while a query containing "hypertension [MeSH]" was considered experienced, a query with "high blood pressure" was considered nonexperienced, even though hypertension is a synonym of high blood pressure. This is because although for the query "high blood pressure," PubMed's ATM internally expands the query by adding the MeSH term *hypertension*, the MeSH term is ORed with the term high blood pressure (i.e., hypertension [MeSH] OR high blood pressure) and the lay term results in many irrelevant documents. Thus, the ATM is designed to increase recall at the cost of precision (refer to PubMed Help to understand how ATM works).

Results

First, we performed some basic statistical analysis on query and session data. The number of queries per user ranged from 1 to 8544 (an extreme outlier) with an average of 4.77 queries per user (SD 15.11, median 2). Figure 3 presents the proportion of users that submitted different numbers of queries and the proportion of queries submitted by the corresponding users. Many PubMed users submitted one query. About two-fifths (43%) of users submitted one query that represented around 9% of the total queries. The rest of the users (57%) performed multiple queries and those queries represented about 90% of the total queries. More than half of PubMed users performed one or at most two queries for their information needs. There was a gradual decrement in the proportion of users as the number of queries increased.

PubMed users may perform multiple IR sessions to fulfill their various information needs. In order to identify the purpose of each IR session, we categorized the queries in the log dataset as shown in Figure 2. Figure 4 presents the percentages of different query types. A total of 2,012,466 (69%) queries were identified as informational, 753,827 (26%) queries navigational, and 148,510 (5%) queries mixed. A total of 742,602 user sessions were identified from the informational queries. Because we compared experienced and nonexperienced search sessions, we further identified experienced and nonexperienced search sessions based on their system function usage from the user sessions (that are identified from the informational queries only, see Figure 4).

About 94% (=700,547/742,602) of the sessions were performed by nonexperienced-users and 6% (=42,055/742,602) of the sessions were performed by experienced users (see Figure 4). Some of the users (about 1.12%) performed both experienced and nonexperienced search sessions meaning that such sessions contain both experienced and nonexperienced queries. Since these users knew how to perform searches using advanced system functions, we considered them as experienced users. There are two possible explanations as to why they performed nonexperienced queries. First, they needed to express new concepts but there were no MeSH terms for the concepts. Thus, although they knew of advanced search functions such as MeSH terms, they could not avoid using natural language to describe concepts. Second, as Vibert et al (2009) [6] found, many PubMed users with search skills do not use search functions.

Figure 5 shows the histogram of the proportion of the experienced and nonexperienced users for the various session lengths (the number of queries in a session). Technically, the users in the figure indicate sessions. Because a user may have multiple sessions, a set of sessions that is performed by the same user cannot be matched with a specific (integer number of) session length, meaning that each session is independently treated in the analysis. For both of the groups, the proportion of users significantly decreased as the number of sessions increased. For experienced users, the session length ranged from 1 to 308 (an extreme outlier) with an average of 2.85 queries per session (SD 4.24, median 1). For nonexperienced users, session length ranged from 1 to 8522 (an extreme outlier) with

an average of 2.7 (SD 11.61, median 2). As the standard deviation values indicate, session length variation of nonexperienced sessions was higher than that of experienced sessions. Figure 5 clearly shows the difference between experienced users and nonexperienced users in terms of session length. While for users whose session length was 1 (ie, an ideal IR), the percentage of experienced users was higher than that of nonexperienced users (25,365/42,055, 60.31% vs 331,337/700,547, 47.30%), for users whose session length was 2 or 3, the percentage of the experienced group was lower than that of the nonexperienced group. This session length difference indicates that experienced users completed their searches earlier than nonexperienced users.

In addition, we measured user decrease rates of the experienced and nonexperienced users from the session length of 1 to 2, 3, 4, and 5. Because the ideal session length is 1 (meaning that a user fulfills his or her information need with only one query), the baseline session length should be 1 (the ideal session). Decrease rates from the baseline indicate the success of the IR session (at retrieving relevant documents). Figure 6 compares decrease rates from the baseline of the two user groups. The decrease rate of the experienced users at the session length of 2 was significantly higher than that of the nonexperienced group (the formula to calculate the rate of the experienced users at the session length of 2 is: $1 - \#$ of experienced sessions at the session length of 2/# of experienced sessions at the session length of 1, or $1 - 3969/25,365 = 84.30\%$). The decrease rates of the two groups indicated that most experienced PubMed user sessions were closed within only one query (note the median of the session lengths was 1) (in other words, the initial or first query satisfied the users' information needs) and nonexperienced user sessions (median of 2) were longer than those of the experienced group.

Figure 3. Percentage of users and queries per number of queries.
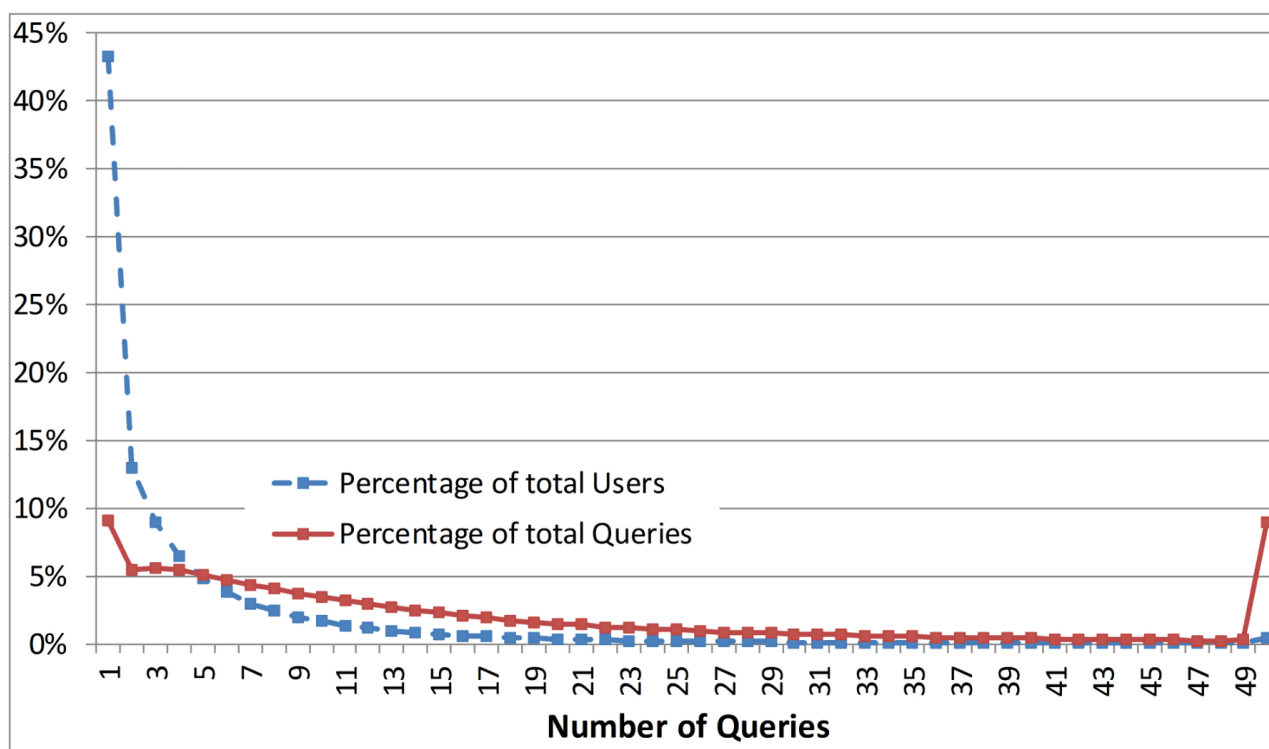
Figure 4. Query types and session types.
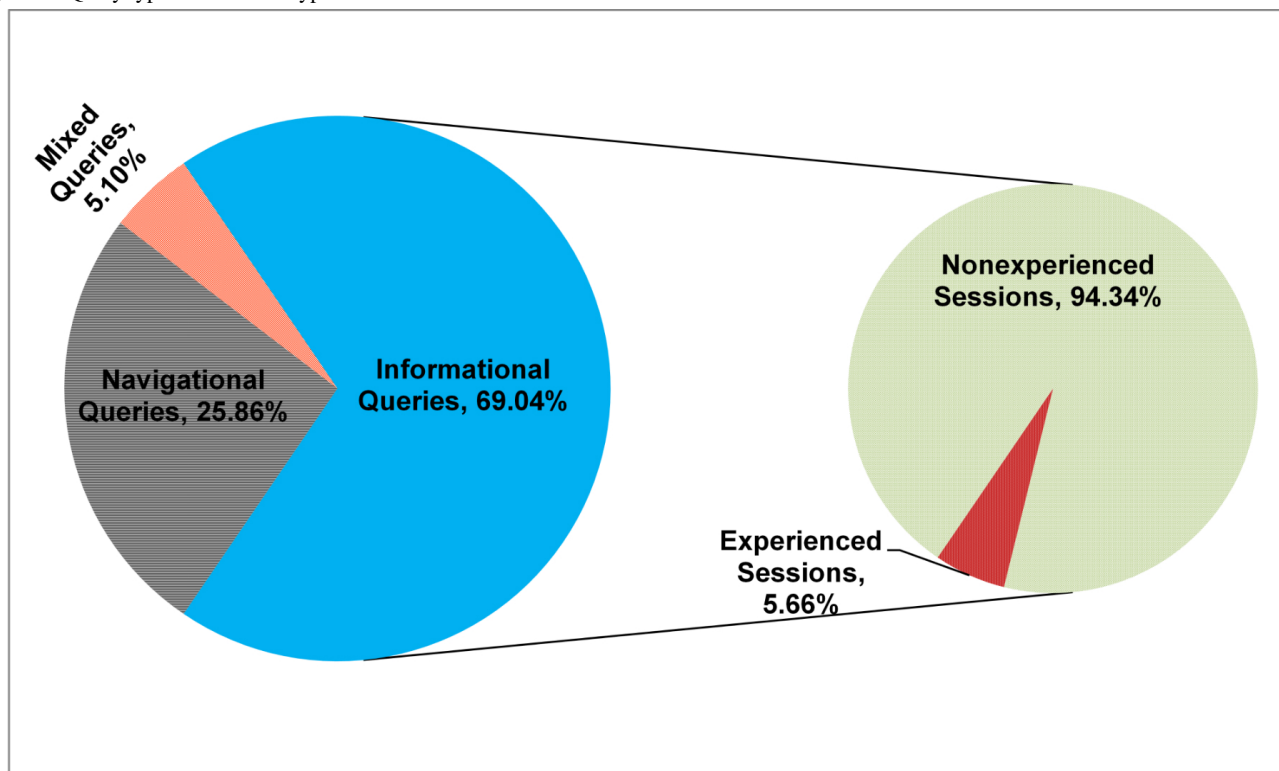
Figure 5. Percentages of experienced and nonexperienced users per session length (# of queries per session).
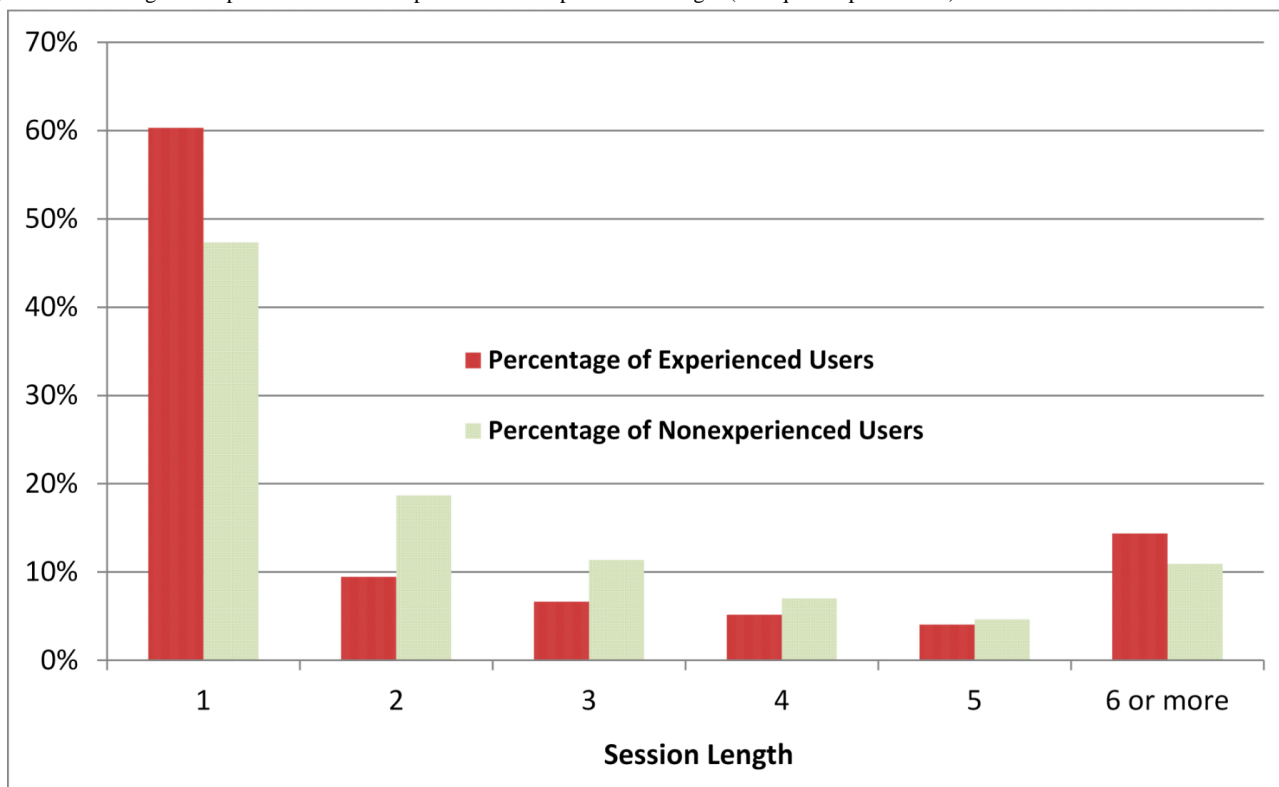
Figure 6. Decrease rates of experienced and nonexperienced users by session length (# of queries per session).
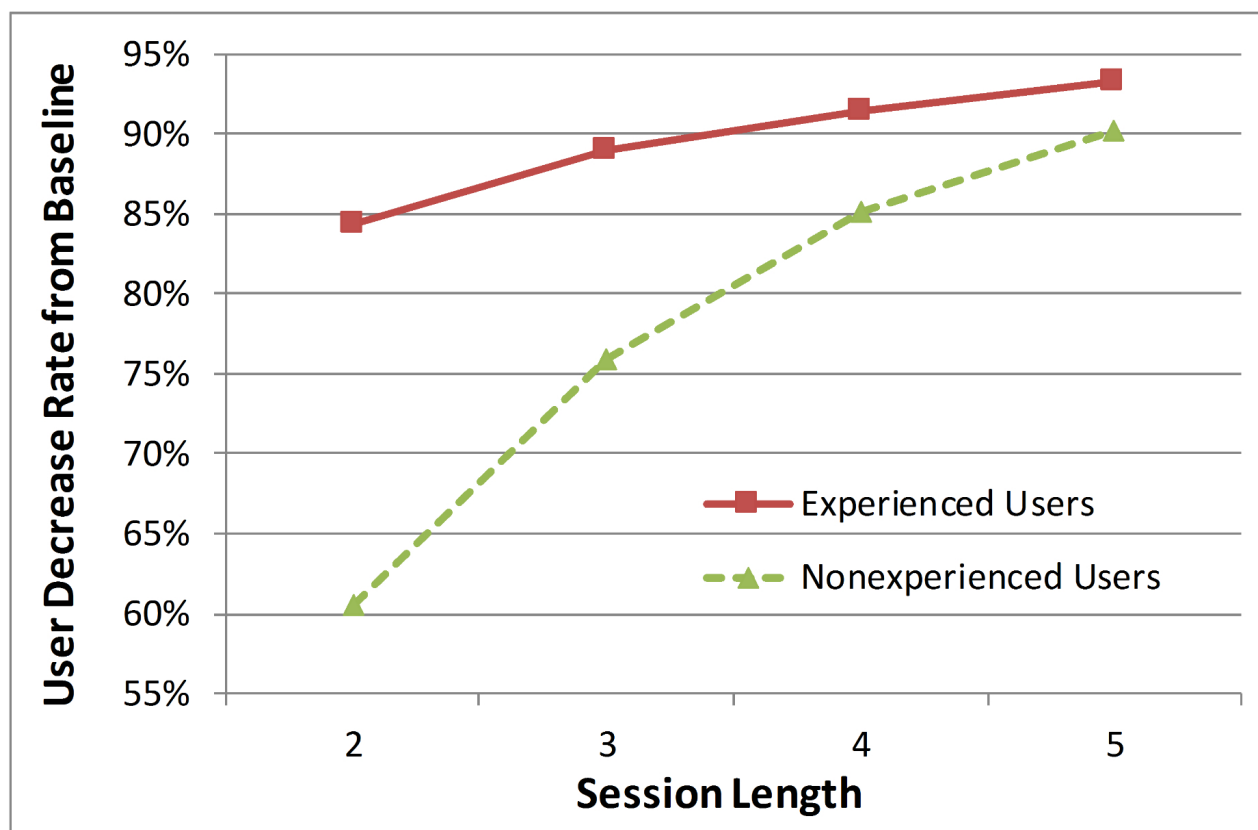
Discussion

Principal Findings

In bibliographic searches like PubMed searches, procedural knowledge is an important factor to improve the overall performance of information retrieval. Procedural knowledge includes experience using online search systems and their search functions. Earlier studies demonstrated that PubMed users perform searches with higher recall and precision if PubMed search functions are used [25,26,85-89]. These studies used at most tens of human subjects for their experiments. In this study, to check the effect of IR functions on PubMed searches, we performed an analysis on a very large scale. The full-day PubMed log data we used contained nearly 3 million user queries issued by more than 0.6 million users. To our knowledge, this study is the first in the field of biomedical and health informatics to use log data containing nearly 3 million queries to compare search performance and behavior of experienced and nonexperienced users. For the analysis, we first categorized queries into informational or navigational based on their underlying intentions, and then identified 0.7 million informational query sessions from more than 2 million informational queries. An informational query session consisted of one or many informational queries in a row within a 20-minute session window. Sessions were further categorized into experienced and nonexperienced user sessions. To test our hypotheses, we compared experienced and nonexperienced users, and found that experienced PubMed users quickly retrieved relevant documents and nonexperienced PubMed users had longer search sessions than experienced users.

Limitations

There are some limitations of this study. First, the PubMed query log data used in this study could have been biased in terms of IR function usage because the data contained search queries for one day only. Second, we used a predetermined time cutoff (20 minutes) for determining search sessions since the log data did not contain any session-related information. It is possible for a PubMed user to perform more than one session in 20 minutes. However, according to recent studies [6,84], most users complete their search session within 20 minutes. At the same time, it is not common that PubMed users spend more than 20 minutes on a search session; more than 65% of PubMed users perform one to three queries per session (see Figure 3). Third, the classification of users based on the use of search tags is not always correct. In other words, the user classification names (ie, experienced and nonexperienced user groups) do not always necessarily indicate that, for example, all the users in the nonexperienced user group are PubMed novice users. At the same time, we believe the group included some experienced users. There are two reasons why experienced users sometimes do not use search functions: first, in order to find "recently published" articles one must use natural language (nonMeSH terms) because those articles are not indexed yet (indexing lag); second, using MeSH terms requires one to search the MeSH database first before conducting PubMed searches (this is an additional step).

Fourth, we assumed if a session was closed within a few queries, the session was successful (meaning that their information needs were fulfilled) even if a session close does not always mean successful IR. This assumption is based on the fact that nearly

77% of users had only 1 to 3 queries in a session. We believe that most searches are successful. If most searches were unsuccessful, one would expect that most users would not use PubMed again. However, according to the NLM, the number of PubMed users has been increasing. In fact, there is no way to know if a session has been successful using the log data; using web log information is the only solution to this problem but this information is not available. We believe that some sessions that are closed within a few queries are unsuccessful. However, the gaps between the decrease rates of the experienced and nonexperienced users (especially at the session length of 2, see Figures 5 & 6) clearly indicate that most sessions that are closed within a few queries are successful. In fact, these limitations are related to the use of log data, rather than direct data from human subjects, for the analysis. In other words, the limitations are simply drawbacks of using log data that we cannot readily overcome.

Current Applicability of the Log Data Analysis to PubMed

It is unknown when the PubMed query data were collected, for confidentiality reasons. However, they are at least 9 years old. One might argue that this study based on old log data is still currently applicable, because the NLM has added many features to improve the performance and user interface of PubMed. Some examples are related citations, automatic term mapping, and PubMed Clinical Queries. PubMed is significantly different from how it was 9 years ago, in terms of the user interface and internal processes for better information retrieval. However, it is imperative to ascertain whether the new features and user interface retrieve documents that are more relevant or lead to better PubMed searches. Studies have found that most PubMed users *still* have difficulty finding relevant documents for patient care in PubMed and do not want to use PubMed for their information needs (instead they want to use UpToDate and/or Google).

There are many recent studies (published in 2010 or later) that found that physicians prefer UpToDate and/or Google to PubMed, and that UpToDate and/or Google provide more answers to clinical questions. Thiele and colleagues (2010) [90] evaluated four search tools (Google, Ovid, PubMed, and UpToDate) widely used to answer clinical questions. They found that Google was the most frequently used search engine for patient care, and Google and UpToDate were faster and brought more clinical answers than PubMed and Ovid. Shariff and colleagues (2013) [91] compared the performance of searches in PubMed and Google Scholar by evaluating the recall and precision of the searches (the first 40 search result records were analyzed) to determine how well search engines answered nephrological questions. The recall of Google Scholar was two times higher than that of PubMed (indicating documents twice as relevant) while the precision of Google Scholar was slightly higher than that of PubMed (indicating less irrelevant documents in the search result). Another advantage of Google Scholar was that it provided nearly three times more links to full-text documents than PubMed. Duran-Nelson and colleagues (2013) [92] carried out a survey to uncover how internal medicine residents use resources (such as UpToDate, Google/Google Scholar, and PubMed) for point-of-care (POC) clinical decision

making. The top two resources the residents used daily at the POC were UpToDate and Google. Of interest, although the residents thought both UpToDate and PubMed provided trustworthy information for patient care, only 20 residents used PubMed daily while nearly 140 residents used UpToDate daily. In addition, the biggest barrier to using PubMed was speed (it took more time to find clinical answers with PubMed). Cook and colleagues (2013) [93] performed a study similar to Duran-Nelson's (Duran-Nelson et al, 2013) [92]. This focus group study (based on a brief survey) showed that physicians used UpToDate two times as much as PubMed, and physicians regarded PubMed as less useful in POC learning due to the time required to find relevant information through PubMed searches. Sayyah Ensan and colleagues (2011) [94] compared PubMed Clinical Queries and UpToDate to determine their ability to answer clinical questions and the time required to find answers. Their findings were that (a) physicians obtain more answers using UpToDate (76%) than PubMed Clinical Queries (43%), and (b) the median times spent retrieving answers using UpToDate and PubMed Clinical Queries were 17 minutes and 29 minutes, respectively. Nourbakhsh and colleagues (2012) [95] evaluated PubMed and Google Scholar with four clinical questions. The first 20 citations/results were analyzed and classified into three relevance groups (clearly relevant, possibly relevant, and not relevant). They found Google Scholar retrieved more relevant documents than PubMed (80% vs 67.6%). Thiele and colleagues (2010) [96] conducted a survey of medical students, residents, and attending physicians on computer use and four search engines widely used to answer clinical questions (Google, Ovid, PubMed, and UpToDate), and compared the search engines in terms of accuracy, speed, and user confidence. Results showed that 33% and 32% of physicians used UpToDate and Google, respectively, for answering their clinical questions, while only 13% of physicians used PubMed. The authors found that Google and UpToDate answered more clinical questions correctly and more quickly than PubMed.

In sum, the findings of these recent studies indicate that the information retrieval features of PubMed are inferior to other electronic resources or search engines such as UpToDate and Google. In other words, most PubMed users still have considerable difficulty obtaining relevant documents/information despite its many new features. As a result, physicians spend more time finding relevant information with PubMed. This problem is critical for PubMed because recent studies still show that the main barrier to POC learning is lack of time [90] [91] [92] [93] [97] [98]. We believe, based on these recent studies that virtually nothing has changed in terms of information-seeking behavior and PubMed from the user's perspective.

Conclusions

The PubMed log analysis indicated that experienced PubMed users quickly retrieved relevant documents in terms of session length and nonexperienced PubMed users had longer search sessions than experienced users. We believe there are a few potential solutions to this problem. First, the NLM could design and provide a novel PubMed user interface for nonexperienced users so that they can readily utilize advanced search functions without special training in PubMed. Second, because it is

imperative for health professionals (especially physicians) to learn the system functions and MeSH vocabulary for better PubMed searches, the NLM could award grant funding only to institutes that regularly train health professionals in PubMed search skills. Third, the NLM could develop a sophisticated relevance-sorting algorithm similar to Google's, so that PubMed users can quickly find relevant documents. Currently, PubMed provides a relevance sorting option. However, it is not the default sorting option as of 17 June 2015 and we believe there should be a significant improvement to the sorting algorithm. This PubMed search problem is not just an information retrieval issue but also a health care practice matter, because health professionals, especially physicians, could significantly improve the quality of patient care and effectively educate chronic patients using clinical and medical information and knowledge obtained from PubMed searches.

Acknowledgments

Conflicts of Interest

None declared.

References

1.   Downing R, Moore J, Brown S. The effects and interaction of spatial visualization and domain expertise on information seeking. Comput Human Behav 2005;21(2):195-209.
2.   Rouet J. The Skills of Document Use: From Text Comprehension to Web-Based Learning. Mahwah, NJ: Lawrence Erlbaum Associates, Inc; 2006.
3.   Tenopir C, King D, Boyce P, Grayson M, Zhang Y, Ebuen M. Patterns of journal use by scientists through three evolutionary phases. D-Lib Magazine 2003 May;9(5) [FREE Full text] [doi: 10.1045/may2003-contents]
4.   De Groote SL, Dorsch JL. Measuring use patterns of online journals and databases. J Med Libr Assoc 2003 Apr;91(2):231-240 [FREE Full text] [doi: 10.1043/0025-7338(2003)091<0231:MUPOOJ>2.0.CO;2] [Medline: 12883574]
5.   Vibert N, Rouet J, Ros C, Ramond M, Deshoullieres B. The use of online electronic information resources in scientific research: The case of neuroscience. Libr Inf Sci Res 2007 Dec;29(4):508-532. [doi: 10.1016/j.lisr.2007.05.001]
6.   Vibert N, Ros C, Bigot L, Ramond M, Gatefin J, Rouet J. Effects of domain knowledge on reference search with the PubMed database: An experimental study. J Am Soc Inf Sci 2009 Jul;60(7):1423-1447. [doi: 10.1002/asi.21078]
7.   Mattox DE. Welcome to ARCHIVES CME. Arch Otolaryngol Head Neck Surg 2000 Jul;126(7):914. [Medline: 10889012]
8.   Muin M, Fontelo P, Liu F, Ackerman M. SLIM: an alternative Web interface for MEDLINE/PubMed searches - a preliminary study. BMC Med Inform Decis Mak 2005;5:37 [FREE Full text] [doi: 10.1186/1472-6947-5-37] [Medline: 16321145]
9.   Bar-Ilan J, Fink N. Preference for electronic format of scientific journals—A case study of the Science Library users at the Hebrew University. Libr Inf Sci Res 2005 Jun;27(3):363-376. [doi: 10.1016/j.lisr.2005.04.011]
10. Sood A, Ghosh A. Literature search using PubMed: an essential tool for practicing evidence- based medicine. J Assoc Physicians India 2006 Apr;54:303-308. [Medline: 16944614]
11. Bahaadinbeigy K, Yogesan K, Wootton R. MEDLINE versus EMBASE and CINAHL for telemedicine searches. Telemed J E Health 2010 Oct;16(8):916-919. [doi: 10.1089/tmj.2010.0046] [Medline: 20925566]
12. Yoo I, Marinov M. Recent research for MEDLINE/PubMed: short review. USA: ACM; 2010 Presented at: Proceedings of the ACM fourth international workshop on Datatext mining in biomedical informatics; October 26 - 30, 2010; Toronto, ON, Canada p. 69-70. [doi: 10.1145/1871871.1871887]
13. Haux R, Grothe W, Runkel M, Schackert HK, Windeler HJ, Winter A, et al. Knowledge retrieval as one type of knowledge-based decision support in medicine: results of an evaluation study. Int J Biomed Comput 1996 Apr;41(2):69-85. [Medline: 8803668]
14. Nankivell C, Wallis P, Mynott G. Networked information and clinical decision making: the experience of Birmingham Heartlands and Solihull National Health Service Trust (Teaching). Med Educ 2001 Feb;35(2):167-172. [Medline: 11169091]
15. Baker NC, Hemminger BM. Mining connections between chemicals, proteins, and diseases extracted from Medline annotations. J Biomed Inform 2010 Aug;43(4):510-519 [FREE Full text] [doi: 10.1016/j.jbi.2010.03.008] [Medline: 20348023]
16. Herskovic JR, Tanaka LY, Hersh W, Bernstam EV. A day in the life of PubMed: analysis of a typical day's query log. J Am Med Inform Assoc 2007;14(2):212-220 [FREE Full text] [doi: 10.1197/jamia.M2191] [Medline: 17213501]
17. Islamaj DR, Murray G, Névéol A, Lu Z. Understanding PubMed user search behavior through log analysis. Database (Oxford) 2009;2009:bap018 [FREE Full text] [doi: 10.1093/database/bap018] [Medline: 20157491]
18. Lacroix E, Mehnert R. The US National Library of Medicine in the 21st century: expanding collections, nontraditional formats, new audiences. Health Info Libr J 2002 Sep;19(3):126-132. [Medline: 12390234]

19. Hersh W, Hickam D. Use of a multi-application computer workstation in a clinical setting. Bull Med Libr Assoc 1994 Oct;82(4):382-389 [FREE Full text] [Medline: 7841907]

20. McKibbon KA, Haynes RB, Dilks CJ, Ramsden MF, Ryan NC, Baker L, et al. How good are clinical MEDLINE searches? A comparative study of clinical end-user and librarian searches. Comput Biomed Res 1990 Dec;23(6):583-593. [Medline: 2276266]

21. Pao ML, Grefsheim SF, Barclay ML, Woolliscroft JO, McQuillan M, Shipman BL. Factors affecting students' use of MEDLINE. Comput Biomed Res 1993 Dec;26(6):541-555. [Medline: 8112055]

22. Gallagher PE, Allen TY, Wyer PC. How to find evidence when you need it, part 2: a clinician's guide to MEDLINE: the basics. Ann Emerg Med 2002 Apr;39(4):436-440. [Medline: 11919532]

23. Chapman D. Advanced search features of PubMed. J Can Acad Child Adolesc Psychiatry 2009 Feb;18(1):58-59 [FREE Full text] [Medline: 19270851]

24. Hersh WR, Crabtree MK, Hickam DH, Sacherek L, Friedman CP, Tidmarsh P, et al. Factors associated with success in searching MEDLINE and applying evidence to answer clinical questions. J Am Med Inform Assoc 2002;9(3):283-293 [FREE Full text] [Medline: 11971889]

25. Richter RR, Austin TM. Using MeSH (medical subject headings) to enhance PubMed search strategies for evidence-based practice in physical therapy. Phys Ther 2012 Jan;92(1):124-132 [FREE Full text] [doi: 10.2522/ptj.20100178] [Medline: 21979271]

26. Darmoni SJ, Soualmia LF, Letord C, Jaulent M, Griffon N, Thirion B, et al. Improving information retrieval using Medical Subject Headings Concepts: a test case on rare and chronic diseases. J Med Libr Assoc 2012 Jul;100(3):176-183 [FREE Full text] [doi: 10.3163/1536-5050.100.3.007] [Medline: 22879806]

27. Hubert G, Cabanac G, Sallaberry C, Palacio D. Query operators shown beneficial for improving search results. In: Gradmann S, editor. Research and Advanced Technology for Digital Libraries: International Conference on Theory and Practice of Digital Libraries. Berlin: Springer Berlin Heidelberg; 2011:118-129.

28. Xie I, Joo S. Factors affecting the selection of search tactics: Tasks, knowledge, process, and systems. Inf Process Manag 2012 Mar;48(2):254-270. [doi: 10.1016/j.ipm.2011.08.009]

29. Hölscher C, Strube G. Web search behavior of Internet experts and newbies. Comput Netw 2000 Jun;33(1-6):337-346. [doi: 10.1016/S1389-1286(00)00031-1]

30. Oldroyd B. Study of strategies used in online searching 5: differences between the experienced and the inexperienced searcher. Online Rev 1984 Mar;8(3):233-244. [doi: 10.1108/eb024156]

31. Harter S. Online searching styles: An exploratory study. Coll Res Libr 1984 Jul 01;45(4):249-258. [doi: 10.5860/crl_45_04_249]

32. Penniman D. Modeling and Evaluation of On-Line User Behavior. Final Report to the National Library of Medicine. Dublin, OH: OCCL; Sep 1981.

33. Hsieh-Yee I. Effects of search experience and subject knowledge on the search tactics of novice and experienced searchers. J Am Soc Inf Sci 1993 Apr;44(3):161-174 [FREE Full text] [doi: 10.1002/(SICI)1097-4571]

34. Drabenstott K. Do nondomain experts enlist the strategies of domain experts? J Am Soc Inf Sci Technol 2003 Jul;54(9):836-854. [doi: 10.1002/asi.10281]

35. Tabatabai D, Shore B. How experts and novices search the Web. Libr Inf Sci Res 2005 Mar;27(2):222-248. [doi: 10.1016/j.lisr.2005.01.005]

36. Gardois P, Calabrese R, Colombi N, Deplano A, Lingua C, Longo F, et al. Effectiveness of bibliographic searches performed by paediatric residents and interns assisted by librarians. A randomised controlled trial. Health Info Libr J 2011 Dec;28(4):273-284. [doi: 10.1111/j.1471-1842.2011.00957.x] [Medline: 22051126]

37. Fields B, Keith S, Blandford A. Designing for expert information finding strategies. In: Fincher S, editor. People and Computers XVIII - Design for Life: Proceedings of HCI 2004 (Pt. 18). London, UK: Springer London; 2005:89-102.

38. US National Library of Medicine. National Institutes of Health. 2005 May 17. MEDLINE/PubMed Data Element (Field) Descriptions URL: http://www.nlm.nih.gov/bsd/mms/medlineelements.html [accessed 2014-07-30] [WebCite Cache ID 6RRV0Kbq8]

39. Marchionini G. Information Seeking in Electronic Environments. Cambridge: Cambridge University Press; 1995.

40. Chen C, Rada R. Interacting with hypertext: A meta-analysis of experimental studies. Hum Comput Interact 1996 Jun 1;11(2):125-156. [doi: 10.1207/s15327051hci1102_2]

41. Palmquist R, Kim K. Cognitive style and on-line database search experience as predictors of Web search performance. J Am Soc Inf Sci 2000;51(6):558-566. [doi: 10.1002/(SICI)1097-4571(2000)51:6<558::AID-ASI7>3.0.CO;2-9]

42. Marchionini G, Dwiggins S, Katz A, Lin X. Information seeking in full-text end-user-oriented search systems: The roles of domain and search expertise. Libr Inf Sci Res 1993;15(1):35-69.

43. Hembrooke H, Granka L, Gay G, Liddy E. The effects of expertise and feedback on search term selection and subsequent learning. J Am Soc Inf Sci 2005 Jun;56(8):861-871. [doi: 10.1002/asi.20180]

44. Jenkins C, Corritore C, Wiedenbeck S. Patterns of information seeking on the Web: A qualitative study of domain expertise and Web expertise. Inf Technol Soc 2003;1(3):64-89.

45. Ju B. Does domain knowledge matter: Mapping users' expertise to their information interactions. J Am Soc Inf Sci 2007 Nov;58(13):2007-2020. [doi: 10.1002/asi.20680]

46. Patel S, Drury C, Shalin V. Effectiveness of expert semantic knowledge as a navigational aid within hypertext. Behav Inf Tech 1998 Jan;17(6):313-324. [doi: 10.1080/014492998119274]

47. Wildemuth B, de Bliek R, Friedman C, File D. Medical students' personal knowledge, searching proficiency, and database use in problem solving. J Am Soc Inf Sci 1999;46(8):590-607. [doi: 10.1002/(SICI)1097-4571(199509)46:8<590::AID-ASI7>3.0.CO;2-#]

48. Hersh WR, Hickam DH. How well do physicians use electronic information retrieval systems? A framework for investigation and systematic review. JAMA 1998 Oct 21;280(15):1347-1352. [Medline: 9794316]

49. Saracevic T, Kantor P. A study of information seeking and retrieving. II. Users, questions, and effectiveness. J Am Soc Inf Sci 1988 May;39(3):177-196. [doi: 10.1002/(SICI)1097-4571(198805)39:3<177::AID-ASI3>3.0.CO;2-F]

50. Allen B. Topic knowledge and online catalog search formulation. Libr Q 1991 Apr;61(2):188-213.

51. Aula A, Nordhausen K. Modeling successful performance in Web searching. J Am Soc Inf Sci 2006 Oct 1;57(12):1678-1693. [doi: 10.1002/asi.20340]

52. Egan D. Individual differences in human-computer interaction. In: Handbook of Human-Computer Interaction. Amsterdam, Netherlands: Elsevier; 1988:543-568.

53. Cullen RJ. In search of evidence: family practitioners' use of the Internet for clinical information. J Med Libr Assoc 2002 Oct;90(4):370-379 [FREE Full text] [Medline: 12398243]

54. Davies KS. Physicians and their use of information: a survey comparison between the United States, Canada, and the United Kingdom. J Med Libr Assoc 2011 Jan;99(1):88-91 [FREE Full text] [doi: 10.3163/1536-5050.99.1.015] [Medline: 21243061]

55. Markey K. Twenty-five years of end-user searching, Part 1: Research findings. J Am Soc Inf Sci 2007 Jun;58(8):1071-1081. [doi: 10.1002/asi.20462]

56. Macedo-Rouet M, Rouet J, Ros C, Vibert N. How do scientists select articles in the PubMed database? An empirical study of criteria and strategies. Eur Rev Appl Psychol 2012 Apr;62(2):63-72. [doi: 10.1016/j.erap.2012.01.003]

57. Lasserre K. Expert searching in health librarianship: a literature review to identify international issues and Australian concerns. Health Info Libr J 2012 Mar;29(1):3-15. [doi: 10.1111/j.1471-1842.2011.00974.x] [Medline: 22335285]

58. McKibbon KA, Haynes RB, Johnston ME, Walker CJ. A study to enhance clinical end-user MEDLINE search skills: design and baseline findings. Proc Annu Symp Comput Appl Med Care 1991:73-77 [FREE Full text] [Medline: 1807700]

59. Silverstein C, Marais H, Henzinger M, Moricz M. Analysis of a very large web search engine query log. SIGIR Forum 1999 Sep 01;33(1):6-12. [doi: 10.1145/331403.331405]

60. Jansen B, Spink A, Saracevic T. Real life, real users, and real needs: a study and analysis of user queries on the web. Information Processing & Management 2000 Mar;36(2):207-227. [doi: 10.1016/S0306-4573(99)00056-4]

61. Facca F, Lanzi P. Mining interesting knowledge from weblogs: a survey. Data Knowl Eng 2005 Jun;53(3):225-241. [doi: 10.1016/j.datak.2004.08.001]

62. Murray G, Teevan J. Query log analysis: social and technological challenges. SIGIR Forum 2007 Dec 01;41(2):112-120. [doi: 10.1145/1328964.1328985]

63. Doğan R, Murray G, Névéol A, Lu Z. Characterizing user search behavior in PubMed. : American Medical Informatics Association; 2010 Presented at: AMIA 2010 Annual Symposium; 2015 Nov 16; Washington, DC p. 1094.

64. Lu Z, Wilbur W. Improving accuracy for identifying related PubMed queries by an integrated approach. J Biomed Inform 2009 Oct;42(5):831-838 [FREE Full text] [doi: 10.1016/j.jbi.2008.12.006] [Medline: 19162232]

65. Lu Z, Kim W, Wilbur WJ. Evaluation of query expansion using MeSH in PubMed. Inf Retr Boston 2009;12(1):69-80 [FREE Full text] [doi: 10.1007/s10791-008-9074-8] [Medline: 19774223]

66. Névéol A, Islamaj DR, Lu Z. Semi-automatic semantic annotation of PubMed queries: a study on quality, efficiency, satisfaction. J Biomed Inform 2011 Apr;44(2):310-318 [FREE Full text] [doi: 10.1016/j.jbi.2010.11.001] [Medline: 21094696]

67. Lu Z, Wilbur WJ, McEntyre JR, Iskhakov A, Szilagyi L. Finding query suggestions for PubMed. 2009 Nov 14 Presented at: AMIA Annu Symp Proc; 2009; San Francisco, USA p. 396-400.

68. Lu Z, Xie N, Wilbur WJ. Identifying related journals through log analysis. Bioinformatics 2009 Nov 15;25(22):3038-3039 [FREE Full text] [doi: 10.1093/bioinformatics/btp529] [Medline: 19734155]

69. Mosa ASM, Yoo I. A study on PubMed search tag usage pattern: association rule mining of a full-day PubMed query log. BMC Med Inform Decis Mak 2013;13:8 [FREE Full text] [doi: 10.1186/1472-6947-13-8] [Medline: 23302604]

70. Broder A. A taxonomy of web search. SIGIR Forum 2002 Sep 01;36(2):3-10. [doi: 10.1145/792550.792552]

71. PubMed. U.S. National Library of Medicine, National Institute of Health. PubMed Help URL: http://www.ncbi.nlm.nih.gov/books/NBK3827/pdf/Bookshelf_NBK3827.pdf [accessed 2015-06-18] [WebCite Cache ID 6ZOCemft9]

72. Spink A, Bateman J, Jansen B. Searching heterogeneous collections on the Web: behaviour of Excite users. Inf Res 1998 Oct;4(2):12-27.

73. Swanson D. Information retrieval as a trial-and-error process. Libr Q 1977;47(2):128-148.

74. Downey D, Dumais S, Horvitz E. Models of searching and browsing: languages, studies, and applications. In: Proceedings of the 20th international joint conference on Artifical intelligence. San Francisco, CA: Morgan Kaufmann Publishers Inc; 2007 Presented at: IJCAI; 2007 Jan 6-12; Hyderabad, India p. 2740-2747.

75. Huang C, Chien L, Oyang Y. Relevant term suggestion in interactive web search based on contextual information in query session logs. J Am Soc Inf Sci 2003 May;54(7):638-649. [doi: 10.1002/asi.10256]

76. Buzikashvili N, Jansen B. Limits of the web log analysis artifacts. 2006 Presented at: Workshop on Logging Traces of Web Activity: The Mechanics of Data Collection; 2006; Edinburgh, Scotland.

77. He D, Göker A. Detecting session boundaries from web user logs. 2000 Apr 05 Presented at: Proceedings of the 22nd Annual Colloquium on Information Retrieval Research; 2000; Cambridge p. 57-66.

78. Radlinski F, Joachims T. Query chains: learning to rank from implicit feedback. New York, NY: ACM Press; 2005 Presented at: Proceedings of the eleventh ACM SIGKDD international conference on knowledge discovery in data mining; 2005 Aug 21-24; Chicago, IL, USA p. 239-248. [doi: 10.1145/1081870.1081899]

79. Shi X, Yang C. Mining related queries from search engine query logs. New York, NY, USA: ACM Press; 2006 Presented at: Proceedings of the 15th international conference on World Wide Web - WWW; May 23-26, 2006; Edinburgh, Scotland p. 943-944. [doi: 10.1145/1135777.1135956]

80. Gayo-Avello D. A survey on session detection methods in query logs and a proposal for future evaluation. Inf Sci 2009 May 30;179(12):1822-1843. [doi: 10.1016/j.ins.2009.01.026]

81. Catledge L, Pitkow J. Characterizing browsing strategies in the World-Wide Web. Proc Int World Wide Web Conf 1995 Apr;27(6):1065-1073. [doi: 10.1016/0169-7552(95)00043-7]

82. Jansen B, Spink A. An analysis of Web information seeking and use: Documents retrieved versus documents viewed. 2003 Presented at: Proceedings of the 4th International Conference on Internet Computing; 2003 Jun 23-26; Las Vegas, Nevada, USA p. 65-69.

83. Jansen B, Spink A, Blakely C, Koshman S. Defining a session on Web search engines. J Am Soc Inf Sci 2007 Apr;58(6):862-871. [doi: 10.1002/asi.20564]

84. Ahmadi S, Faghankhani M, Javanbakht A, Akbarshahi M, Mirghorbani M, Safarnejad B, et al. A comparison of answer retrieval through four evidence-based textbooks (ACP PIER, Essential Evidence Plus, First Consult, and UpToDate): a randomized controlled trial. Med Teach 2011;33(9):724-730. [doi: 10.3109/0142159X.2010.531155] [Medline: 21854150]

85. Haynes RB, McKibbon KA, Walker CJ, Ryan N, Fitzgerald D, Ramsden MF. Online access to MEDLINE in clinical settings. A study of use and usefulness. Ann Intern Med 1990 Jan 1;112(1):78-84. [Medline: 2403476]

86. Bernstam E, Kamvar S, Meric F, Dugan J, Chizek S, Stave C. Oncology patient interface to MEDLINE. 2001 Presented at: Proceedings of the American Society of Clinical Oncology; 2001; USA p. 244a.

87. Agoritsas T, Merglen A, Courvoisier DS, Combescure C, Garin N, Perrier A, et al. Sensitivity and predictive value of 15 PubMed search strategies to answer clinical questions rated against full systematic reviews. J Med Internet Res 2012;14(3):e85 [FREE Full text] [doi: 10.2196/jmir.2021] [Medline: 22693047]

88. Shariff SZ, Sontrop JM, Haynes RB, Iansavichus AV, McKibbon KA, Wilczynski NL, et al. Impact of PubMed search filters on the retrieval of evidence by physicians. CMAJ 2012 Feb 21;184(3):E184-E190 [FREE Full text] [doi: 10.1503/cmaj.101661] [Medline: 22249990]

89. Ugolini D, Neri M, Casilli C, Bonassi S. Development of search filters for retrieval of literature on the molecular epidemiology of cancer. Mutat Res 2010 Aug 30;701(2):107-110. [doi: 10.1016/j.mrgentox.2010.05.016] [Medline: 20685344]

90. Thiele RH, Poiro NC, Scalzo DC, Nemergut EC. Speed, accuracy, and confidence in Google, Ovid, PubMed, and UpToDate: results of a randomised trial. Postgrad Med J 2010 Aug;86(1018):459-465. [doi: 10.1136/pgmj.2010.098053] [Medline: 20709767]

91. Shariff SZ, Bejaimal SA, Sontrop JM, Iansavichus AV, Haynes RB, Weir MA, et al. Retrieving clinical evidence: a comparison of PubMed and Google Scholar for quick clinical searches. J Med Internet Res 2013;15(8):e164 [FREE Full text] [doi: 10.2196/jmir.2624] [Medline: 23948488]

92. Duran-Nelson A, Gladding S, Beattie J, Nixon LJ. Should we Google it? Resource use by internal medicine residents for point-of-care clinical decision making. Acad Med 2013 Jun;88(6):788-794. [doi: 10.1097/ACM.0b013e31828ffdb7] [Medline: 23619072]

93. Cook DA, Sorensen KJ, Hersh W, Berger RA, Wilkinson JM. Features of effective medical knowledge resources to support point of care learning: a focus group study. PLoS One 2013;8(11):e80318 [FREE Full text] [doi: 10.1371/journal.pone.0080318] [Medline: 24282535]

94. Sayyah EL, Faghankhani M, Javanbakht A, Ahmadi S, Baradaran HR. To compare PubMed Clinical Queries and UpToDate in teaching information mastery to clinical residents: a crossover randomized controlled trial. PLoS One 2011;6(8):e23487 [FREE Full text] [doi: 10.1371/journal.pone.0023487] [Medline: 21858142]

95. Nourbakhsh E, Nugent R, Wang H, Cevik C, Nugent K. Medical literature searches: a comparison of PubMed and Google Scholar. Health Info Libr J 2012 Sep;29(3):214-222. [doi: 10.1111/j.1471-1842.2012.00992.x] [Medline: 22925384]

96. Thiele RH, Poiro NC, Scalzo DC, Nemergut EC. Speed, accuracy, and confidence in Google, Ovid, PubMed, and UpToDate: results of a randomised trial. Postgrad Med J 2010 Aug;86(1018):459-465. [doi: 10.1136/pgmj.2010.098053] [Medline: 20709767]

97. Cook DA, Sorensen KJ, Hersh W, Berger RA, Wilkinson JM. Features of effective medical knowledge resources to support point of care learning: a focus group study. PLoS One 2013;8(11):e80318 [FREE Full text] [doi: 10.1371/journal.pone.0080318] [Medline: 24282535]

98. van Dijk N, Hooft L, Wieringa-de WM. What are the barriers to residents' practicing evidence-based medicine? A systematic review. Acad Med 2010 Jul;85(7):1163-1170. [doi: 10.1097/ACM.0b013e3181d4152f] [Medline: 20186032]

Abbreviations

ATM: Automatic Term Mapping
IR: information retrieval
MeSH: Medical Subject Heading
NLM: National Library of Medicine

XSL•FO
RenderX

Corrigenda and Addenda

Metadata Correction: 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 Cassale[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:

Correction of: http://medinform.jmir.org/2015/2/e23/

Nick Peterson, PhD (Lumiata, Inc, San Mateo, CA, United States) was inadvertently omitted from the list of authors in "A Web-Based Tool for Patient Triage in Emergency Department Settings: Validation Using the Emergency Severity Index" (JMIR Med Inform 2015;3(2):e23). The author Nick Peterson should have been added after Kim Branson in the original published manuscript. This error has been corrected in the online version of the paper on the JMIR Med Inform website on August 11, 2015, together with publishing this correction notice. This was done after submission to PubMed Central and other full-text repositories. This correction notice has been submitted to PubMed, the original paper resubmitted to PubMed Central, and the metadata has been resubmitted to CrossRef with publishing this correction notice.

Corrigenda and Addenda

Metadata Correction: Meaningful Use of Electronic Health Records: Experiences From the Field and Future Opportunities

Sarah Patricia Slight[1,2], MPharm, PhD, PGDip; Eta S Berner[3], EdD; William Galanter[4], MS, MD, PhD; Stanley Huff[5,6], MD; Bruce L Lambert[7], PhD; Carole Lannon[8,9], MD, MPH; Christoph U Lehmann[10], MD, FAAP, FACMI; Brian J McCourt[11], BA; Michael McNamara[12], MD; Nir Menachemi[13], MPH, PhD; Thomas H Payne[14], MD; S Andrew Spooner[15], MD, MS, FAAP; Gordon D Schiff[2,16], MD; Tracy Y Wang[11], MD, MHS, MSc; Ayse Akincigil[17], PhD; Stephen Crystal[17], PhD; Stephen P Fortmann[18], MD; Meredith L Vandermeer[18], MPH; David W Bates[2,16], MD, MSc

[1]Division of Pharmacy, School of Medicine Pharmacy and Health, Durham University, Durham, United Kingdom

[2]The Centre for Patient Safety Research and Practice, Division of General Internal Medicine, Brigham and Womens Hospital, Boston, MA, United States

[3]Center for Health Informatics for Patient Safety/Quality, Department of Health Services Administration, University of Alabama at Birmingham, Birmingham, AL, United States

[4]University of Illinois at Chicago, Chicago, IL, United States

[5]Department of Biomedical Informatics, University of Utah, Salt Lake City, UT, United States

[6]Intermountain Healthcare, Salt Lake City, UT, United States

[7]Center for Communication and Health, Department of Communication Studies, Northwestern University, Chicago, IL, United States

[8]James M Anderson Center for Health Systems Excellence, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, United States

[9]Department of Pediatrics, University of Cincinnati, Cincinnati, OH, United States

[10]Pediatrics and Biomedical Informatics, Vanderbilt University, Nashville, TN, United States

[11]Duke Clinical Research Institute, Duke University, Durham, NC, United States

[12]Departments of Pediatrics and Medical Informatics (DMI), Northwest Permanente, Portland, OR, United States

[13]Richard M Fairbanks School of Public Health, Indiana University, Indianapolis, IN, United States

[14]Department of Medicine, University of Washington, Seattle, WA, United States

[15]Division of Biomedical Informatics, Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, OH, United States

[16]Harvard Medical School, Boston, MA, United States

[17]Institute for Health, Health Care Policy, and Aging Research, Rutgers, The State University of New Jersey, New Brunswick, NJ, United States

[18]Center for Health Research, Kaiser Permanente Northwest, Portland, OR, United States

Corresponding Author:
David W Bates, MD, MSc
The Centre for Patient Safety Research and Practice
Division of General Internal Medicine
Brigham and Womens Hospital
1620 Tremont Street
Boston, MA, MA 02120-1613
United States
Phone: 1 617 732 5650
Fax: 1 617 732 7072
Email: dbates@partners.org

Related Article:

Correction of: http://medinform.jmir.org/2015/3/e30/

During article production, the co-author Meredith Vandermeer, MPH (Center for Health Research, Kaiser Permanente Northwest, Portland, OR, United States) was omitted from the list of authors in the paper "Meaningful Use of Electronic Health Records: Experiences From the Field and Future Opportunities" (JMIR Med Inform 2015;3(3):e30). The author Vandermeer should have been added after Stephen P Fortmann in the original published manuscript. This error has been corrected in the online version of the paper on the JMIR Med Inform website on Sept 25, 2015, together with publishing this correction notice. This

was done before submission to PubMed Central and other full-text repositories. This correction notice has been submitted to PubMed and the metadata has been resubmitted to CrossRef with publishing this correction notice.

XSL•FO

RenderX