Automated detection of harm in healthcare with information technology: a systematic review

Malavika Govindan,1 Aricca D Van Citters,2 Eugene C Nelson,1 Jane Kelly-Cummings,3 Gautham Suresh4

ABSTRACT
Context To improve patient safety, healthcare facilities are focussing on reducing patient harm. Automated harm-detection methods using information technology show promise for efficiently measuring harm. However, there have been few systematic reviews of their effectiveness.
Objective To perform a systematic literature review to identify, describe and evaluate effectiveness of automated inpatient harm-detection methods.
Methods Data sources included MEDLINE and CINAHL databases indexed through August 2008, extended by bibliographic review and search of citing articles. The authors included articles reporting effectiveness of automated inpatient harm-detection methods, as compared with other detection methods. Two independent reviewers used a standardised abstraction sheet to extract data about automated and comparison harm-detection methods, patient samples and events identified. Differences were resolved by discussion.
Results From 176 articles, 43 articles met inclusion criteria: 39 describing field-defined methods, two using natural language processing and two using both methods. Twenty-one studies used automated methods to detect adverse drug events, 10 detected general adverse events, eight detected nosocomial infections, and four detected other specific adverse events. Compared with gold standard chart review, sensitivity and specificity of automated harm-detection methods ranged from 0.10 to 0.94 and 0.23 to 0.98, respectively. Studies used heterogeneous methods that often were flawed.
Conclusion Automated methods of harm detection are feasible and some can potentially detect patient harm efficiently. However, effectiveness varied widely, and most studies had methodological weaknesses. More work is needed to develop and assess these tools before they can yield accurate estimates of harm that can be reliably interpreted and compared.

INTRODUCTION
It is widely recognised that harm caused by the healthcare system is a major source of morbidity and mortality in hospitalised patients.1 An estimated 15 million instances of medical harm occur in the USA every year.2 However, the lack of simple, practical and accurate methods to identify adverse events in hospitals has hampered efforts to develop routine monitoring systems, assess the impact of interventions to prevent harm and compare interhospital performance.

Detecting incidence and types of patient harm are prerequisites for implementing strategies to prevent harm. Manual, comprehensive chart review by trained professionals has been used in key studies and can be considered the gold-standard harm-detection method.3–6 However, this approach requires time and trained abstractors, thereby decreasing its feasibility as a pragmatic method for routine measurement of adverse events.

Several organisations are currently using the Institute for Healthcare Improvement’s Global Trigger Tool, which is based on manual chart review, and allows targeted chart review to identify harm more efficiently than comprehensive chart review and more extensively than voluntary reporting of harm.

Automated strategies of harm detection that use computerised methods to scan patient records may require fewer time and personnel resources than traditional methods, and can potentially provide real-time surveillance alerts. We performed this review to: (1) identify types of automated methods of inpatient harm detection described in published literature, (2) describe types of events identified by these methods and (3) evaluate accuracy of these methods in identifying harm. We also independently evaluated the quality and validity of key studies.

METHODS
Definitions
In this review, we used the terms harm, automated harm detection and gold standard chart review as defined in Box 1.

Data sources/study selection
We (MG and AVC) identified articles for this review through a literature search of MEDLINE (start date 1950) and CINAHL (start date 1982) using the following search terms: (harm OR adverse event OR adverse drug event OR nosocomial infection) AND (automated OR computerised OR electronic) AND (identify OR detect OR detection OR recognise OR recognition). We identified additional articles using bibliographic review of key articles, the ‘related articles’ feature of Medline, and the ‘find similar’ and ‘find citing articles’ feature of CINAHL. We reviewed the title and abstract of each article, and obtained the full text of relevant articles. We limited our search to English language articles indexed through 31 August 2008.

We included studies that: (a) occurred in an inpatient setting, (b) described an automated harm-detection method, (c) measured actual harm and (d) compared the automated method to an alternative method of harm detection.

Data extraction and analysis
We developed and tested a standardised data form and extracted the following variables from included articles: details of patient sample, methodology used for automated harm detection, nature of events identified, description of alternative method of harm detection and comparisons of events detected by automated and alternative methods. Data were extracted by MG and AVC, with uncertainties resolved by discussion and consensus.

We critically appraised each study that compared the automated method of harm detection to a gold standard chart review, using published criteria for validity of diagnostic test studies.7 We assessed each study for: (a) independent, blind comparison of the automated method with a gold standard method, (b) performance of the gold standard assessment regardless of the automated method’s results and (c) validation of the assessment in a second, independent set of patients.

If studies provided adequate data, we independently calculated the sensitivity, specificity and positive and negative predictive values of the automated harm-detection method.

RESULTS
Selection of articles
One hundred and seventy-six articles were reviewed for potential inclusion, of which 43 provided information on validity of automated methods of harm detection.8–50 The remaining articles were excluded because they: were review articles on harm-detection methodologies (n=9)51–59; did not focus on detection of harm (n=26) or automated methods (n=22); did not include a comparison group (n=17); were not limited to inpatients (n=13); were descriptive papers of a program, incident reporting system, algorithm or computer simulation (n=55); were commentaries or editorials (n=11); or were repeat publications (n=2).

The methodologies and results from the 43 included studies are described in online appendix 1. Of these, 14 studies compared the automated harm-detection methodology to a gold standard chart review, and their methods and results are summarised in tables 1, 2.

As shown in online appendix 1, 20 studies were conducted among adult populations, three in paediatric patients, two among all age groups, one in geriatric patients, one among Medicare beneficiaries and one among patients 14 years and older. The most common hospital settings were general medical units (n=14), followed by general surgical units (n=8), medical, surgical or general intensive care units (n=8), medical subspecialties (n=5), neonatal and paediatric intensive care units (n=5) and obstetric units (n=2). The target population and setting were unstated in 15 studies.

Data sources for automated harm-detection methods
Automated harm-detection methods were classified into field-defined and natural language-processing systems. Field-defined systems relied on computerised detection using pre-existing numeric or coded data stored in medical records. Natural language processing relied on computerised analysis of free text within a medical record to detect language indicative of harm. Field-defined and natural language-processing systems are described in table 5.

Forty-one of 43 studies used field-defined systems for automated harm detection. The nature of the programs, databases used, data fields used and types of harm detected within this category were source-specific. Typical sources of data for field-defined programs included laboratory, radiology, microbiology, pharmacy, and administrative and billing databases. Five of 43 studies used natural language-processing systems. The most common source of data was discharge summaries. Radiology reports, chart text, daily progress notes, consultation notes, nursing records, and procedure or operative reports also were used.

Degree of automation
Twenty-five studies (58%) reported on detection tools that were partially automated,8–14 21–25 31 32 34–38 40 45–48 50 and 14 studies (33%) described fully automated tools15–17 19 26–30 33 41 42 44 49 and one study (2%) reported both fully and partially automated systems.20 The degree of automation was unclear in three reports (7%).18 39 45

Types of events identified
Automated methods for detecting harm predominantly focused on identification of adverse drug events (ADEs) (n=21, 49%).11 12 18 21–26 29–32 35–38 43 45 50 Ten automated methods (23%) focused on general adverse events,8–10 19 33 34 40 46–48 eight (19%) focused on nosocomial infection,14 20 28 39 41 42 44 49 and four (9%) focused on other specific adverse events (eg, decubitus ulcers, surgical complications).13 15 17 27

Accuracy of automated harm-detection methods
Only 14 studies15 17 18 20 22 23 26 30 32–34 44 47 48 compared an automated harm-detection method with ‘gold-standard’ adverse event detection and were eligible for critical appraisal of validity (table 2). Methodologies used to evaluate these automated systems were heterogeneous. Seven studies (50%) applied the gold standard using independent, blind evaluators. Eight studies (57%) applied the gold standard independently of the outcome from the automated method. One study (7%) validated the results of the automated method in an independent, second set of patients.

Table 4 shows the sensitivity, specificity, and positive and negative predictive values of the automated methods that were
### Table 1: Summary of studies comparing automated harm-detection methods with gold standard chart review

<table>
<thead>
<tr>
<th>Reference</th>
<th>Patient sample and time frame</th>
<th>Sampling strategy</th>
<th>Specialty</th>
<th>Events identified</th>
<th>Automated event dataset sample size</th>
<th>Comparison event dataset sample size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nebeker et al.</td>
<td>18 Adults 2001 and 2003</td>
<td>Random</td>
<td>General Surgery</td>
<td>Adverse drug events</td>
<td>397 admissions</td>
<td>397 admissions</td>
</tr>
<tr>
<td>Zhan et al.</td>
<td>17 Medicare benefic. 2002 to 2004</td>
<td>Random, Flagged sample</td>
<td>General Surgery</td>
<td>Adverse drug events, specifically postoperative deep venous thrombosis and/or pulmonary embolism</td>
<td>20,868 hospital discharges identified as surgical patients</td>
<td>20,868 hospital discharges identified as surgical patients</td>
</tr>
<tr>
<td>Brossette et al.</td>
<td>44 Unknown 1e 3 Dec 2003 and 2e 29 Apr 2004</td>
<td>Sequential</td>
<td>Unknown</td>
<td>Infection</td>
<td>907 admissions</td>
<td>907 admissions</td>
</tr>
<tr>
<td>Hougland et al.</td>
<td>30 Adults 1 Jan 2001 to 31 Dec 2001</td>
<td>Random, Flagged sample (from records with at least one flagged adverse drug event code)</td>
<td>Unknown</td>
<td>Unknown</td>
<td>1961 random, 1142 flagged</td>
<td>1961 random, 1142 flagged</td>
</tr>
<tr>
<td>Polancich et al.</td>
<td>15 Unknown Unknown Unknown</td>
<td>Unknown</td>
<td>Unknown</td>
<td>Hospital acquired decubitus ulcers</td>
<td>Unknown</td>
<td>Unknown</td>
</tr>
<tr>
<td>Dormann et al.</td>
<td>26 Adults 1 Sept 2000 to 28 Feb 2001</td>
<td>Sequential</td>
<td>Gastroenterology</td>
<td>Adverse drug events</td>
<td>316 patient records</td>
<td>40 patients records (10 very low probability, 10 high probability)</td>
</tr>
<tr>
<td>Levy et al.</td>
<td>23 All age groups 1 Apr 1997 to 31 May 1997</td>
<td>Sequential</td>
<td>General Medical</td>
<td>Adverse drug events</td>
<td>1000 charts</td>
<td>1000 charts</td>
</tr>
<tr>
<td>Azaz-Livshits et al.</td>
<td>22 All age groups 1 Apr 1995 to 31 May 1995</td>
<td>Sequential</td>
<td>General Medical</td>
<td>Adverse drug events</td>
<td>153 admissions</td>
<td>153 admissions</td>
</tr>
<tr>
<td>Jha et al.</td>
<td>32 Adults 1 Oct 1994 to 31 May 1995</td>
<td>Sequential</td>
<td>MICU, SICU, General Medical, General Surgical</td>
<td>Adverse drug events, adverse events, diagnostic errors, operative complications, falls</td>
<td>21,964 patient-days</td>
<td>21,964 patient-days</td>
</tr>
<tr>
<td>NLP</td>
<td>47 Adults 1 Jun 1999 to 31 Dec 2004</td>
<td>Sequential</td>
<td>MICU, SICU and other (placement of CVC)</td>
<td>Adverse drug events related to central venous catheter, CVC complications, falls</td>
<td>40 patients records (10 very low probability, 10 high probability)</td>
<td>40 patients records (10 very low probability, 10 high probability)</td>
</tr>
<tr>
<td>Forster et al.</td>
<td>34 Adults FY 2002</td>
<td>Random</td>
<td>General Medical, General Surgical</td>
<td>Adverse drug events</td>
<td>245 patients</td>
<td>245 patients</td>
</tr>
<tr>
<td>Melton and Hripcsak</td>
<td>48 Unknown 1996 to 2000</td>
<td>Random (charts), Sequential (electronic discharge summaries)</td>
<td>Unknown</td>
<td>Adverse drug events, adverse events, diagnostic errors, operative complications, falls</td>
<td>1000 charts</td>
<td>1000 charts</td>
</tr>
<tr>
<td>Murff et al.</td>
<td>33 Adults 1 Jan 2000 to 30 Jun 2000</td>
<td>Random (Cohort 1), Sequential (Cohort 2)</td>
<td>General Medical, Medicine subspecialties</td>
<td>Adverse drug events, adverse events, diagnostic errors, operative complications, falls</td>
<td>Cohort 1: 21,964 patient-days Cohort 2: 21,964 patient-days</td>
<td>Cohort 1: 21,964 patient-days Cohort 2: 21,964 patient-days</td>
</tr>
</tbody>
</table>

*Sampling strategy refers to the method by which charts were chosen to be screened by the automated tool. Unless specifically noted, the same sampling strategy also applies to the gold standard method. A scoring system was developed by the authors to reflect the probability of the adverse event in question relating to the central venous catheter placement. This system is described in the text and in table 1 of the paper.*

CVC, central venous catheter; MICU, Medical Intensive Care Unit; NYPORTS, New York Patient Occurrence Reporting and Tracking System; PSI, patient safety indicators; SICU, Surgical Intensive Care Unit.
Table 2  Evaluation of validity of studies comparing automated method to gold standard chart review

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Field-defined</td>
<td>Nebeker et al(^{18})  Computer algorithms  Chart review for study, unclear if strategy aims to be Full or Partial</td>
<td>ICD-9 CM codes</td>
<td>Medical record</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Study used Hougland et al(^{30})'s methodology to specifically apply HOCTA (hierarchically optimal classification tree analysis) to administrative data to develop surveillance rules for the identification of ADEs manifesting as either bleeding or delirium. Requires expert computer programming. DVT/PE events flagged by ICD-9 CM codes were compared with those discovered by gold standard chart review. The sample studied was a random sample abstracted by the Medicare Patient Safety Monitoring System.</td>
<td></td>
</tr>
<tr>
<td>Zhan et al(^{17})  Patient Safety Indicators  Full</td>
<td>ICD-9 CM codes</td>
<td>Medical record</td>
<td>Unknown</td>
<td>Unknown</td>
<td>No</td>
<td>Nosocomial Infection Marker (NIM) program by Med Mined, Birmingham, Alabama. Took about 10 min/week to maintain. Total time for NIM: 2 h/10 000 admissions, compared with medical record review at 1.5 full time employees per 10 000 admissions.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brossette et al(^{44})  Nosocomial Infection Marker  Full</td>
<td>Medical record and Lab database</td>
<td>Medical record</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Designed to test validity of Agency for Healthcare Research and Quality (AHRQ) PSIs for detecting hospital acquired decubitus ulcers. Only a sample of cases was manually reviewed.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hougland et al(^{30})  Automated ICD-9 code strategy  Full: Review of flagged charts here for study purposes</td>
<td>ICD-9 CM codes</td>
<td>Medical record</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Expert panel identified 416 ICD-9 CM codes to represent ADEs (flagged ADEs). Then chart review performed to ascertain codes' ability to detect/identify ADE.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Polancich et al(^{15})  Patient Safety Indicators  Full</td>
<td>Administrative data, Billing data, ICD-9 CM diagnosis and procedure codes</td>
<td>Medical Record</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Designed to test validity of Agency for Healthcare Research and Quality (AHRQ) PSIs for detecting hospital acquired decubitus ulcers. Only a sample of cases was manually reviewed.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dormann et al(^{26})  Automated laboratory signal detection  Full</td>
<td>Demographics, History, Lab findings, Drugs, &amp; Diagnosis</td>
<td>Medical record</td>
<td>Unknown</td>
<td>Unknown</td>
<td>No</td>
<td>Used automated lab signals (ALS) and changes in ALS to identify ADEs. Automated system used to flag potential ADEs, which were then sent as an alert to physicians. Use of delta ALS (change) resulted in improvement over Dormann et al(^{25})'s methodology.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trick et al(^{20})  Computer algorithm  Full and Partial</td>
<td>Medical record; Lab, pharmacy, &amp; radiology database; Microbiology</td>
<td>Medical record; Yes Lab, pharmacy, &amp; radiology database; Microbiology</td>
<td>Yes</td>
<td>No</td>
<td>Comparison of manual and computer assisted bloodstream central venous catheter infection surveillance using data from two hospitals. Different computer algorithms developed for full or partial automation were tested.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Levy et al(^{23})  Automated laboratory signal detection  Partial</td>
<td>Lab database</td>
<td>Lab database and clinical data</td>
<td>Unknown</td>
<td>Yes</td>
<td>No</td>
<td>Implementation of the pilot program described in Azaz-Livshits et al(^{22}) Computerised lab data monitored to detect ADEs using the same signals as the pilot study.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Azaz-Livshits et al(^{22})  Automated laboratory signal detection  Partial</td>
<td>Lab database and clinical data</td>
<td>Lab database and clinical data</td>
<td>Unknown</td>
<td>Yes</td>
<td>No</td>
<td>Pilot program to develop and assess computerised laboratory data as a detection tool for ADE in 34-bed medical ward in Jerusalem, Israel. Lab signals generated by computer, then verified by team. Limited computerised patient data at this hospital; however lab data were fully electronic. Cost of this system reasonable compared with costs of ADEs.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Continued
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Jha et al 32</td>
<td>Automated triggers</td>
<td>Partial</td>
<td>Medical record</td>
<td>Medical record</td>
<td>Yes</td>
<td>Unknown</td>
<td>No</td>
<td>Study of computer-based ADE identification using modified Classen 1991 rules to create automated triggers with which the electronic record was screened. Rules modified during the study to increase PPV, and new rules created. Trained reviewer and physician were blinded to detection method. 11 person-hours per week for automated method versus 55 for chart review and 5 for voluntary reporting.</td>
</tr>
<tr>
<td>NLP</td>
<td>Computer algorithms &amp; Natural Language Processing</td>
<td>Partial</td>
<td>Text records: Daily progress notes; Consultation, Nursing, and Procedure notes; Operative reports; Discharge summaries</td>
<td>Text records: Daily progress notes; Consultation, Nursing, and Procedure notes; Operative reports; Discharge summaries</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Compared two methods for semiautomated review of text records within the VA database using NLP (MedLEE) and a phrase matching algorithm (PMA). Limited by incomplete or inaccurate documentation, incomplete coding, spelling errors, and sentence structure abbreviations. Time/technology intensive.</td>
</tr>
<tr>
<td>Forster et al 34</td>
<td>Computerised screen for trigger words in free text</td>
<td>Partial</td>
<td>Discharge summaries</td>
<td>Discharge summaries</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Automated adverse event lexicon made up of 104 terms used by Murff et al. Computerised search engine scanned discharge summaries (dsresearch desktop) to detect potential harm. Specificity higher for non-elective admissions and discharge summaries dictated by residents/staff versus medical students. Automated detection reduced physician time by one-fifth.</td>
</tr>
<tr>
<td>Melton and Hripcsak 48</td>
<td>Natural Language Processing</td>
<td>Partial</td>
<td>Discharge summaries</td>
<td>Full electronic chart and paper chart for a subset of 100 patients</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Natural Language Processing system (MedLEE) to identify 45 NY Patient Occurrence Reporting and Tracking System event types. Chart review by physician and independent informatician of random sample of 1000 charts to assess performance of NLP program. Results biased towards patients with electronic discharge summaries. This method is technologically intensive.</td>
</tr>
<tr>
<td>Murff et al 33</td>
<td>Computerised screen for trigger words in free text</td>
<td>Full (goal is a fully automated system, manual review of subsamples performed for study)</td>
<td>Discharge summaries</td>
<td>Medical record (not otherwise specified)</td>
<td>Yes†</td>
<td>Yes</td>
<td>Yes</td>
<td>Brigham and Women’s Hospital, using Brigham Integrated Computer system. Computerised screening tool searched free text discharge summaries for trigger words indicating possible adverse events. List of automated trigger words compiled using Harvard Medical Practice Study definitions as base. Electronic method alone versus electronic plus manual review compared for two cohorts. Reviewers blinded to whether screening tool had identified the admission.</td>
</tr>
</tbody>
</table>

ADE, adverse drug event; AE, adverse event; DVT, deep venous thrombosis; NLP, natural language processing; PE, pulmonary embolism; PPV, positive predictive value; PSI, patient safety indicators; VA, Veterans Administration.

*We define fully automated methods as those where the identification of harm was not followed by further chart review, and partially automated methods where patient records flagged by the automated detection of potential harm (eg, ‘trigger’) were manually reviewed to verify harm.

† Authors manually reviewed a random 25% sample of screened-negative charts, then used this random sample to estimate the number of adverse events occurring in entire set of screened-negative charts.
Table 3  Description and classification of field-defined and natural language processing systems for automated detection of harm*

<table>
<thead>
<tr>
<th>Automated method</th>
<th>Data source used</th>
<th>Events identified</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complications Screening Program (CSP)</td>
<td>ICD-9 CM codes</td>
<td>Adverse drug events, adverse surgical outcomes, infections, and miscellaneous complications such as falls</td>
<td>A computerised method for identifying potentially preventable complications of hospital care.</td>
</tr>
<tr>
<td>Health Evaluation through Logical Processing (HELP)</td>
<td>Electronic Medical Record: specifically including pharmacy, laboratory, radiology and surgery records</td>
<td>Adverse drug events, adverse medical device events, infection</td>
<td>Integrated electronic medical record of the LDS Hospital in Salt Lake City, Utah, which contains an interactive modular knowledge base that continually analyses information</td>
</tr>
<tr>
<td>Patient Safety Indicators (PSI)</td>
<td>Administrative data: billing information, ICD-9 CM diagnosis codes and procedure codes</td>
<td>Adverse events</td>
<td>A fully automated method developed by the Agency for Healthcare Research and Quality</td>
</tr>
<tr>
<td>Computer algorithms</td>
<td>Electronic Medical Record: components specific to the particular program: see online appendix 1</td>
<td>Adverse events, adverse drug events, infection</td>
<td>Specific, named computer programs</td>
</tr>
<tr>
<td>Lab signal detection tools</td>
<td>Laboratory Database</td>
<td>Adverse drug events</td>
<td>Automated tools search for key words or word combinations that signal potential or actual harm—for example, detection of elevated potassium levels</td>
</tr>
<tr>
<td>ICD-9 CM or billing code detection tools</td>
<td>Administrative data: ICD-9 CM or billing codes</td>
<td>Adverse drug events, infections, surgical complications</td>
<td>Automated tools scan for diagnosis, discharge, or billing codes that signal potential or actual harm—for example, evidence of antibiotic exposure following a postoperative infection</td>
</tr>
<tr>
<td>Tools using computerised triggers</td>
<td>Electronic Medical Record: multiple sources such as pharmacy, laboratory, and microbiology databases</td>
<td>Adverse events, adverse drug events, infection</td>
<td>Automated tools using multiple triggers to signal actual or potential harm—for example, detection of elevated potassium levels (laboratory database) combined with certain medication administration (pharmacy database). Among the various tools included in this category, there are four named systems: Dynamic Pharmaco-Monitoring System, Nosocomial Infection Marker, Event Detector, New York Antimicrobial Resistance Project.</td>
</tr>
<tr>
<td>Natural language processing systems</td>
<td>Free text in the Electronic Medical Record: discharge summaries, radiology reports, chart notes</td>
<td>Adverse events, infection</td>
<td>Sophisticated programs that 'read' free text via the application of computer logic</td>
</tr>
</tbody>
</table>

*Multiple detection strategies were used in several studies, including those that combined two or more field-defined systems,16–46 two natural language-processing systems,47 and both a field-defined and natural language-processing system.33–34
<table>
<thead>
<tr>
<th>Reference</th>
<th>Events identified by automated harm-detection method</th>
<th>Sensitivity (95% CI)</th>
<th>Specificity (95% CI)</th>
<th>Positive predictive value (95% CI)</th>
<th>Negative predictive value (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nebeker et al(^{18})</td>
<td>Calculated separately for bleeding/anticoagulation ADEs and delirium ADEs</td>
<td>Bleeding: 0.86(^{†})</td>
<td>NA</td>
<td>Bleeding: 0.12(^{†})</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Delirium: 0.89(^{†})</td>
<td>NA</td>
<td>Delirium: 0.03(^{†})</td>
<td>NA</td>
</tr>
<tr>
<td>Zhan et al(^{17})</td>
<td>DVT Cases</td>
<td>0.67 (0.58 to 0.76)</td>
<td>NA</td>
<td>0.31 (0.25 to 0.37)</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>PE cases</td>
<td>0.74 (0.59 to 0.90)</td>
<td>NA</td>
<td>0.24 (0.16 to 0.33)</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>DVT/PE Cases</td>
<td>0.68 (0.60 to 0.76)</td>
<td>NA</td>
<td>0.29 (0.24 to 0.34)</td>
<td>NA</td>
</tr>
<tr>
<td>Brossette et al(^{44})</td>
<td>Hospital-wide nosocomial infection</td>
<td>0.88 (0.82 to 0.94)</td>
<td>NA</td>
<td>0.78</td>
<td>NA</td>
</tr>
<tr>
<td>Hougland et al(^{29})</td>
<td>Codes for inpatient ADE</td>
<td>0.74 (0.59 to 0.90)</td>
<td>NA</td>
<td>0.52 (0.42 to 0.62)</td>
<td>NA</td>
</tr>
<tr>
<td>Polancich et al(^{25})</td>
<td>Patients with decubitus ulcers</td>
<td>0.91</td>
<td>0.23</td>
<td>0.10(^{†})</td>
<td>0.93</td>
</tr>
<tr>
<td>Dormann et al(^{26})</td>
<td>ADR positive admissions using NEW ALS</td>
<td>0.91</td>
<td>0.76</td>
<td>0.25(^{†})</td>
<td>0.87</td>
</tr>
<tr>
<td></td>
<td>ADR positive admissions using DELTA ALS</td>
<td>0.91</td>
<td>0.76</td>
<td>0.25(^{†})</td>
<td>0.87</td>
</tr>
<tr>
<td>Trick et al(^{20})</td>
<td>Hospital-acquired episodes of primary CVC associated bloodstream infections</td>
<td>0.81(^{†})</td>
<td>0.72(^{†})</td>
<td>0.62(^{†})</td>
<td>0.87(^{†})</td>
</tr>
<tr>
<td>Levy et al(^{29})</td>
<td>Admissions</td>
<td>0.63 (0.51 to 0.74)</td>
<td>0.42 (0.34 to 0.51)</td>
<td>0.34 (0.25 to 0.42)</td>
<td>0.70 (0.60 to 0.80)</td>
</tr>
<tr>
<td>Azar-Livshits et al(^{22})</td>
<td>Admissions</td>
<td>0.66 (0.51 to 0.81)</td>
<td>0.51 (0.42 to 0.60)</td>
<td>0.31 (0.21 to 0.41)</td>
<td>0.82 (0.73 to 0.91)</td>
</tr>
<tr>
<td>Jha et al(^{31})</td>
<td>ADE</td>
<td>NA(^{†})</td>
<td>0.16(^{†})</td>
<td>0.16(^{†})</td>
<td>NA(^{†})</td>
</tr>
<tr>
<td>NLP</td>
<td></td>
<td>0.80 (0.74 to 0.86)</td>
<td>0.90 (0.84 to 0.96)</td>
<td>0.88 (0.82 to 0.94)</td>
<td>0.96 (0.92 to 0.97)</td>
</tr>
<tr>
<td>Forster et al(^{14})</td>
<td>Cases</td>
<td>0.23 (0.11 to 0.35)</td>
<td>0.92 (0.86 to 0.96)</td>
<td>0.41 (0.22 to 0.59)</td>
<td>0.83 (0.78 to 0.88)</td>
</tr>
<tr>
<td>Melton et al(^{33})</td>
<td>ADE</td>
<td>0.23 (0.16 to 0.40)</td>
<td>0.98 (0.97 to 0.99)</td>
<td>0.47 (0.30 to 0.64)</td>
<td>0.96 (0.95 to 0.97)</td>
</tr>
<tr>
<td>Muff et al(^{33})</td>
<td>AE</td>
<td>0.69 (0.62 to 0.75)</td>
<td>0.48 (0.42 to 0.55)</td>
<td>0.52 (0.46 to 0.58)</td>
<td>0.65 (0.59 to 0.72)</td>
</tr>
<tr>
<td></td>
<td>Fully automated: 0.69 (0.62 to 0.75)</td>
<td>0.48 (0.42 to 0.55)</td>
<td>0.52 (0.46 to 0.58)</td>
<td>0.65 (0.59 to 0.72)</td>
<td>0.74 (0.69 to 0.79)</td>
</tr>
</tbody>
</table>

\(^{95\%}\) CIs for independently verified values reported in parentheses.

\(^{†}\) Denotes figures that we could not independently verify.

\(^{‡}\) Dormann et al\(^{26}\) defined the positive predictive value (PPV) as the number of alerts associated with adverse drug reactions (ADRs) out of the total number of alerts. Using this criteria, they found the following PPVs: New automatic laboratory signals (ALS) (574/2328) 25%; Delta ALS (189/580) 32%.

\(^{‡}\) Brossette et al\(^{44}\) reported a sensitivity of 0.86 and a specificity of 0.98. It is unclear how they identified true negative screens.

\(^{‡}\) Jha et al\(^{32}\) report a range of PPVs based on the first and final 8 weeks of data collection (0.16 and 0.23, respectively). We were able to independently verify the PPV for the first 8 weeks of the study only.

\(^{‡}\) Murff et al\(^{33}\) defined a PPV of 0.65 and a PPV of 0.64, respectively. We were able to independently verify the PPV for the first 8 weeks of the study only.

\(^{‡}\) Polancich et al\(^{25}\) reported a sensitivity of 0.86 and a specificity of 0.98. It is unclear how they identified true negative screens.

ADE, adverse drug event; AE, adverse event; CVC, central venous catheter; DVT, deep venous thrombosis; NLP, natural language processing; PE, pulmonary embolism; PMA, phrase matching algorithm.
compared against a gold standard chart review. Sensitivities of different methods ranged from 0.10 to 0.94, and specificities ranged from 0.23 to 0.98. Positive predictive values ranged from 0.08 to 0.84, and negative predictive values ranged from 0.70 to 0.96. Our independent assessment of validity allowed us to verify all published values for nine of the 14 studies that reported validity data.\textsuperscript{15} 17 22 23 30 33 34 48 Figure 1 displays the sensitivity and 1-specificity intersection points of methods used in these studies in a format similar to that of a receiver-operating characteristic curve.

DISCUSSION

Strategies to improve patient safety require efficient and accurate detection of patient harm. Automated methods of harm detection have been used for this purpose because they offer the potential to rapidly scan patient records with minimal human effort. This systematic review describes types of automated methods of harm detection used in inpatient settings, events identified by these methods and their accuracy.

We found two categories of automated harm detection described in the literature: field-defined systems (used in most studies) and natural language-processing systems. Most frequently laboratory, pharmacy and administrative databases were used to identify adverse drug events, general adverse events and nosocomial infections.

We found that the validity of studies describing automated harm-detection methods was variable. Of these studies, those attempting to identify ADEs\textsuperscript{18} 30 and nosocomial infections\textsuperscript{20} 44 using field-defined methods, and one attempting to identify multiple types of adverse events\textsuperscript{53} using natural language processing satisfied more validity criteria than others. We believe that automated harm-detection methods will have more validity if they attempt to identify events that are discrete, easily and reliably detected, and consistently documented in the chart, such as adverse drug events, nosocomial infections, pressure ulcers and postoperative complications.

Automated harm detection has the potential to positively impact clinical practice. While most automated methods retrospectively identified harm, eight were paired with real-time surveillance alerts that informed physicians or pharmacists of an adverse event. Such prospective surveillance systems can alert the clinical team of impending or ongoing harm, thus allowing early intervention to limit harm. Real-time alerts were present within methods for detecting adverse drug events,\textsuperscript{11} 21 23 26 35 45 general adverse events\textsuperscript{40} and nosocomial infection.\textsuperscript{14} Automated alerts were a component of the Health Evaluation through Logical Processing system\textsuperscript{11} 14 and were incorporated within methods using automated lab signal detection,\textsuperscript{23} 26 45 computer algorithms\textsuperscript{21} and other automated triggers.\textsuperscript{35} 40

Another potential benefit of automated detection is the reduction of person-hours required for harm surveillance. Few studies\textsuperscript{14} 21 22 24 32 34 38 40 44 provided information on financial or human resource requirements for implementing and maintaining automated detection tools. In general, the automated methods reviewed here require fewer person-hours than manual...
chart review. Field-defined strategies appear to be less technologically demanding than natural language-processing strategies. Sophisticated computer algorithms and natural language-processing programs require specialised subject knowledge, skill and time to develop, and require installation and instruction by experts. Whether costs to implement such programs are offset by savings from eliminating manual chart review and decreased patient harm is unknown and should be studied. Future studies also should quantify differences in time and personnel resources needed for the automated detection method, relative to other detection strategies.

To our knowledge, four of the 43 unique articles report on commercially available automated harm-detection systems (MedLEE, dtsearch desktop, Nosocomial Infection Marker (NIM), and Dynamic Pharmaco-Monitoring System). Other articles report on systems that employ data elements common across medical institutions (ie, ICD-9 codes used in the Complications Screening Program use software available to the VA or specific states (ie, RADARx, NY Antimicrobial Resistance Project) or are available through the Agency for Healthcare Research and Quality (ie, Patient Safety Indicators). The availability of the remaining detection systems is either institution-specific or not made clear by their developers.

While automated tools offer promise for efficient and accurate harm detection, there are important limitations that currently make them unsuitable for widespread application, particularly for interhospital comparisons. The reported sensitivity and specificity are variable and often low, suggesting that many episodes of harm may go undetected, and that many events identified will be false positives. Low accuracy may result from limited capability of the tool to detect events, or from flawed sources of data used for automated harm detection. For example, the reliability of field-defined systems can be affected by data entry errors or limited availability and accuracy of administrative codes, while natural language processing is sensitive to spelling and grammatical errors in free text. Both systems may include irrelevant or erroneous information, or exclude necessary information. For example, perhaps driven by medical-legal concerns, health professionals often do not include information about medical errors and resulting adverse events in their progress notes, problem lists and discharge summaries. Thus, an electronic medical record containing accurate, complete and easily accessible information can enhance the performance of an automated detection tool. Understanding these factors is important when evaluating the technological requirements, feasibility and inherent limitations of automated detection methods.

The variety of distinct automated methodologies makes comparisons between studies and between automated tools difficult and unreliable. Differences in the quality and content of data sources, as well as other unknowns such as accuracy of hospital documentation and coding practices, also complicate comparisons. The performance and methods of automated tools also may be institution-specific, making it difficult to generalise to other organisations or patient populations. For example, the Health Evaluation through Logical Processing system used by LDS Hospital in Salt Lake City, Utah relies on an advanced, highly integrated and dynamic information system that is not widely available.

We speculate that field-defined methods of automated harm detection will prove superior to natural language-processing methods, particularly if information about harm is accurately documented in electronic medical record systems in prespecified fields, thus allowing rapid and reliable detection of harm events. The methodological rigour of studies was variable. Only two-thirds of the 14 studies that compared an automated method with a gold standard chart review had verifiable validity results. Moreover, most studies compared automated harm-detection methods with other methods of data on patient harm (eg, voluntary reporting, unstandardised chart reviews, and prospective surveillance records). The validity of data from studies without chart review comparison is questionable given the absence of a defined denominator of events against which to measure the performance of the automated tool. The use of different methods, statistical analyses, denominator values and outcomes precludes a comparison of one automated method with another, as well as any attempt to statistically pool their results in a meta-analysis.

Other authors have summarised the literature on automated harm-detection methods, but most have focused on automated methods specific to a type of harm (ie, adverse drug events or nosocomial infections), patient population (ie, paediatrics), source of data (ie, administrative data or automated technology (ie, natural language processing)). Our systematic review included all types of automated methods, harm events and sources of data evaluated in an inpatient setting. Furthermore, we provide an additional level of critical appraisal compared with other systematic reviews. For example, while Bates et al address differences between study methodologies by noting the presence or absence of gold standard comparison, they do not assess validity of studies or independently verify reported data. To our knowledge, this is the first systematic review to critically assess methodological rigour and study validity.

While our review has several strengths, it also has limitations. First, the search strategy was limited to published English language articles. Second, we did not evaluate scientific meeting abstracts, nor did we contact investigators to identify unpublished studies. Third, publication bias must be considered in which studies with negative findings may not have reached dissemination venues. Fourth, most of the articles evaluated automated methods of harm detection among adults in general medical or surgical units, which may limit application to other populations and settings. Finally, our independent appraisal of the methodology and validity of key studies relied on information available within published articles. Our inability to verify the rigour and validity of all studies highlights the variation among even the most rigorous evaluations.

In conclusion, our review identified numerous automated methods of harm detection in two broad categories—field-defined methods and natural language processing—that identified a broad range of harm events, but particularly adverse drug events and nosocomial infections. Although many of these studies described the accuracy (sensitivity and specificity) of automated harm detection when compared with chart review, these results may not be valid due to methodological flaws in the conduct of many of these studies. Future studies assessing the performance of automated harm-detection methods should ensure that the gold-standard assessment (usually chart review) is performed by a blinded assessor, the gold-standard is applied independently of the results of the automated method (ie, charts not flagged by the automated method are reviewed for false negatives), and the automated method is tested in a set of patients that is independent of the set used to develop the automated method. Finally, efforts should be made to improve documentation of harm episodes in the patient record, in problem lists and when generating diagnosis codes, in order to
improve automated harm detection. Future research should also focus on developing methods for real-time harm detection. In this way, automated harm-detection tools will realize their potential to describe accurately the incidence of harm in hospitalized patients, monitor changes from preventive interventions, and compare institutions and individual health professionals. Establishing universal standards and guidelines for the development, testing and utilization of automated harm-detection methods, perhaps through a centralized agency, would allow data to be collected and compared in a rigorous, systematic fashion.

Summary
Automated methods of harm detection are feasible, allow rapid scanning of a large number of patient records with minimal effort and have the potential to identify events as they occur or soon thereafter. However, the heterogeneity of automated methodologies, the spectrum of study rigour and the widely varying accuracy data suggest that currently available automated methods poorly measure the true incidence of harm. These methods cannot replace chart review as the gold standard but can provide estimates of the frequency of harm that can allow hospitals to identify priorities for action, make decisions about safety interventions and potentially monitor change over time. As automated harm-detection tools and scientific methods to test them evolve, there exists a great potential to positively impact patient safety.

Acknowledgements
We are grateful for the administrative support provided by the Institute for Healthcare Improvement.

Funding
Funding for the literature review was provided by the Institute for Healthcare Improvement (IHI) to MG and ADVC. Subsequent data analysis and interpretation, as well as conceptualisation, preparation, and review of the manuscript were not financially supported.

Competing interests
JK-C was employed by Premier Inc. from 31 March 2007 to 2 July 2008. Premier has developed an automated event detection product. SafetySurveillance. This study does not reference or endorse this product. No other authors disclosed any potential conflicts of interest.

Provenance and peer review
Not commissioned; externally peer reviewed.

REFERENCES


Automated detection of harm in healthcare with information technology: a systematic review

Malavika Govindan, Aricca D Van Citters, Eugene C Nelson, Jane Kelly-Cummings and Gautham Suresh

Qual Saf Health Care 2010 19: e11 originally published online July 29, 2010
doi: 10.1136/qshc.2009.033027

Updated information and services can be found at:
http://qualitysafety.bmj.com/content/19/5/e11

These include:

Supplementary Material
Supplementary material can be found at:
http://qualitysafety.bmj.com/content/suppl/2011/06/09/qshc.2009.033027.DC1

References
This article cites 52 articles, 13 of which you can access for free at:
http://qualitysafety.bmj.com/content/19/5/e11#BIBL

Email alerting service
Receive free email alerts when new articles cite this article. Sign up in the box at the top right corner of the online article.

Notes

To request permissions go to:
http://group.bmj.com/group/rights-licensing/permissions

To order reprints go to:
http://journals.bmj.com/cgi/reprintform

To subscribe to BMJ go to:
http://group.bmj.com/subscribe/