MINING COMPLEX CLINICAL DATA FOR PATIENT SAFETY RESEARCH: A FRAMEWORK FOR EVENT DISCOVERY

George Hripcsak, MD, MS[1],

Suzanne Bakken, RN, DNSc, FAAN[1,2],

Peter D. Stetson, MD, MA[1,3]

Vimla L. Patel, PhD, D.Sc.[1,4]

[1]Department of Biomedical Informatics

[2]School of Nursing

[3]Department of Medicine

[4]Department of Psychiatry

Columbia University

New York, NY

Corresponding author:

George Hripcsak, MD, MS

622 West 168th Street, VC-5

New York, NY 10032

hripcsak@columbia.edu

(212) 305-5712 voice

(212) 305-3302 fax

Abstract

Successfully addressing patient safety requires detecting medical events effectively. Given the volume of patients seen at medical centers, detecting events automatically from data that are already available electronically would greatly facilitate patient safety work. We have created a framework for electronic detection. Key steps include: selecting target events, assessing what information is available electronically, transforming raw data such as narrative notes into a coded format, querying the transformed data, verifying the accuracy of event detection, characterizing the events using systems and cognitive approaches, and using what is learned to improve detection.

Introduction

The Institute of Medicine's report on medical errors demonstrates that adverse events in hospitalized patients are common [1]. A study of 30,121 randomly selected records of hospitalized patients admitted to acute-care hospitals in New York State in 1984 [2] showed that 3.7% had adverse events; of those, 2.6% caused permanent disability, 13.6% caused death, and 28% were negligent. A second study of 15,000 discharges from hospitals in Utah and Colorado in 1992 [3] showed that 2.9% had adverse events; of those, 2.2% caused permanent disability, 6.6% caused death, and 27 to 32% were negligent. Several studies have attempted to clarify the epidemiology of adverse events [4,5].

As these studies suggest, there are many types of errors that occur, some of which can lead to adverse events and increased cost. Addressing these errors requires understanding how and when they occur. In order to implement error prevention goals, one must evaluate the types of systems failures that occur in the medical system under scrutiny.

The first and critical step in this process is identifying medical errors. As Zapt and Reason point out, if an error is not detected, it cannot be managed [6]. They assert that the error detection rate must be high, for errors that are not detected for a long time could have disastrous consequences. Yet error detection rates are generally low. The New York State Health Department's mandatory event reporting program, for example, estimated that only 16% of code 605 (death within 48 hours of an operating room procedure) are being reported, and these should be among the easier events to detect [7]. Furthermore, some hospitals were reporting virtually no errors. Kaplan et al. found gross underreporting of errors in a transfusion service [8]. The importance of the large chart review studies [2,3] lies in the fact that errors are grossly underreported. There are a number of approaches to improving detection rates, including no fault near miss reporting [8], stricter mandated reporting with sanctions against low reporting [7], and meticulous chart review [2,3].

Purpose

The purpose of this paper is to describe a framework for an iterative approach to event discovery in electronic medical records that is based upon a clinical data repository and natural language processing techniques. We summarize our definitions and assumptions and the literature describing capture of medical errors, and we discuss knowledge gaps and issues related to the approach.

Definitions and assumptions

For this paper, we will adopt the following error-related definitions. A *medical error* is defined as the failure of a planned action to be completed as intended or the use of a wrong plan to achieve an aim [1]. An *adverse event* is defined as an injury caused by medical management rather than by the underlying disease or condition of the patient [1]. An *adverse outcome* is defined as an undesirable and unintended outcome of care such as prolonged hospitalization, disability, or death at the time of discharge [2]. A *near miss* (as defined by the Agency for Healthcare Research and Quality's Center for Quality Improvement and Patient Safety) is an event in which the unwanted consequences were prevented because there was a recovery by identification and correction of the failure. The recovery might be *planned* or *unplanned*. Therefore, the following situations are possible:

- medical error accompanied by an adverse outcome (e.g., a drug rash due to prescribing a medication to which the patient is known to be allergic)
- near miss, which is a medical error from which there has been a recovery (e.g., if the pharmacist catches a prescription to a medication to which the patient is known to be allergic)
- medical error without a recovery but without an adverse outcome because of luck or the robustness of human physiology (e.g., a medication to which the patient is known to be allergic is prescribed, dispensed, and taken, but the patient has no reaction)
- adverse outcome without an error (e.g., an allergic reaction to a medication for which there was no known allergy)

In this paper, we will use the term *event* to refer to any medical error (with or without recovery or harm) or adverse outcome. Event is therefore the broadest term and includes all of the definitions above.

Our framework includes the following assumptions. There exists a repository of rich clinical information that we hypothesize contains useful data about patient safety. The data are obscured, however, due to the way they are recorded. Much is in narrative form and therefore not amenable to traditional statistical analysis, and even when coded, the data are stored in complex, nested structures that may be difficult to use. A set of informatics tools exists—natural language processing, machine learning, etc.—that are capable of extracting the useful patient safety information from the repository in an automated or partially automated fashion. Manual review of the paper and electronic medical record (and in a smaller sample, interviews of the relevant care providers) can be used to create a reference standard to judge the accuracy of the automated system. Errors have structure that can be described using a systems approach to errors and using a cognitive approach. We may then use these descriptions to learn how to improve the automated system.

Capture of Medical Events

Various methods have been used to capture medical errors. While physician chart review of all cases is generally considered a gold standard, this is expensive. The primary method has been to rely on self-reporting in combination with non-physician screens to develop a subset of cases where medical errors are more likely. Several approaches have been taken to generate screening criteria with better characteristics. Screening using non-physician human review generally is accomplished by setting established screening criteria to flag cases where medical errors are more likely to occur. For example, the

Harvard Medical Practice Study [2] used 18 screening criteria and used medical analysts trained in medical record coding, terminology, and management [9].

Self-reporting systems are also widely used. O'Neil [10] studied 3141 patients using record review and an aggressive physician self-reporting system found 133 cases where adverse events had occurred. The physician self-reporting identified 89 while the chart review methods identified 85. Unfortunately, other studies have demonstrated that self-reporting systems also miss a significant portion of errors identified by though other means, and their success has been limited to relatively few settings where anonymity can be assured and daily feedback on the importance of reporting is emphasized [10,11].

Automated rule systems that flag potential adverse events have shown promise because they are relatively inexpensive compared to manual review if the data are collected for other purposes. They may detect errors not recorded using other techniques, and they can identify errors retrospectively or prospectively with the potential to facilitate prevention. Often this is dramatic. For example, at the LDS hospital in Utah, Classen and his colleagues demonstrated a sixty-fold increase in adverse drug event detection [10] and subsequently a 65% reduction in severe adverse drug events compared to historical controls [12]. Similarly, Jha et al. used a computer-based monitor to detect adverse drug events and compared the result to other methods [13]. Evans et al. used a computer-based monitor to address hospital-acquired infections and antibiotic use [14,15].

Using the 133 cases of known adverse events collected by O'Neil [10], Bates estimated the probability of the event being detected or prevented by computer systems with different levels of sophistication [16]. He estimated that 53% of the cases could be identified with demographic information, diagnostic test results and a list of current medications. Adding physician order entry and problem lists raised the identification rate to 89%. Five percent of the cases were felt to be preventable using the less sophisticated event monitors. Twenty three percent were felt to be preventable using higher level systems. Bates later went on to implement many of these rules during the implementation of his physician order entry system. An evaluation of that system over 4 years demonstrated a significant reduction medicine related errors and adverse drug events [17,18]. Prior to the implementation of physician order entry, the non-missed dose medication rate by chart review was estimated to be 142 per 1000 patient days. With the implementation of physician order entry, and progressive of refined rule sets, this rate fell 81% to 26.6 per 1000 patient days.

Nevertheless, the detection rate of adverse events by computer remains challenging. Much of the attention in identifying and preventing adverse events has focused on medication errors because they are common and important, but also because they are more easily identified. Non-medication errors may be equally important, however, because those errors may point to systematic errors that might be amenable to prevention by systems corrections.

The challenge is that for non-medication related errors, rules are difficult to identify and may depend on the more complex narrative data. A study by Kossovsky et al. [19], found that distinguishing planned from unplanned readmissions required narrative data from discharge summaries and concluded that natural language processing would be necessary to separate these cases automatically. Roos et al. [20] used claims data from Manitoba to identify complications leading to readmission and found reasonable predictive value, but their similar attempts to identify whether or not a diagnosis represented an in-hospital complication of care based on claims data met with difficulties only resolved through narrative data (discharge abstracts) [21]. Similarly, Iezonni et al. used administrative data to screen for complications [22], but later reported on the lack of completeness and accuracy of ICD9-CM coding [23]. Honigman et al. [24] used a medical lexicon tool on outpatient notes to find adverse drug events. The results were promising but limited by lack of full natural language understanding. A recent IOM report stated that the automation and linking of data on services provided to patients in ambulatory and institutional settings would provide a rich source of information for quality measurement and improvement purposes [25]. This is only useful if selected documents are reviewed or if documents can be coded automatically.

In study that is described in greater detail below, Knirsch et al. used a clinical event monitor linked to a natural language processing system to detect patients at high risk for active tuberculosis who where placed in shared rooms [26]. They reduced missed isolation errors by about half. Lau has reported the use of an expert system in the detection of diagnostic errors [27]. An expert system (Iliad) was used to screen cases under concurrent review by the Utah Physician Review Organization (UPRO). One hundred cases were randomly selected from 242 identified by the UPRO as having diagnoses recognized by Iliad. The UPRO identified 28 cases of the 100 with quality problems—mostly treatment and documentation errors. The expert system also identified 28 cases where there were potential diagnostic errors. These were subsequently reviewed by the UPRO, who confirmed 17 of the flagged cases (68%) as having quality problems. Importantly, six of the cases were detected by both mechanisms, however the problems detected in these cases were different, and the cases identified by the expert system appeared more serious.

A Framework for Event Discovery

Discovering medical events in an electronic medical record has several challenges [28]. Many of the data, especially the administrative coded data, are collected for financial purposes and therefore may not reflect clinical reality [29]. The information contained in narrative reports requires natural language processing [30] or information retrieval techniques [31] to be made available to analysis. The data are complex with deeply nested attributes thwarting most simple approaches to querying databases [32]. Other challenges include (1) the sparseness of the data—each patient has a different set of reports; (2) incomplete data collection; (3) inaccurate data collection; and (4) that many of the fields contain categorical data with many levels (e.g., ICD-9 codes), which can be difficult to handle in machine learning programs.

As a result, querying for relevant medical events is not a simple matter. In our patient safety research we are taking an iterative approach to discovering events:

1. Target events—Pick the target events of interest (either an actual list of known errors or a conceptual type of error to look for).
2. Repository—Begin with the full clinical repository or a purposely-defined subset.
3. Natural language processing—Use natural language processing to parse the narrative data and to create a fully coded repository.

4. Queries—Generate queries that detect and classify errors. They may be generated manually or automatically.
5. Verification—Verify the accuracy of the detection and classification by manual review, thus calculating performance and adding to the database of known errors.
6. Error description—Use a systems approach or a cognitive approach to describe the newly detected errors.
7. Feedback—Based on the errors uncovered in step 5 and the information learned in step 6, improve the natural language processor (step 3) and the queries (step 4), and possibly steer the next selection of target errors (step 1).

These steps are covered in the following sections.

Target events

a. Clinically-focused targets. There are several different approaches to select target events for detection in the electronic medical record. One approach, which is used in most of the literature on detecting events electronically, is to target specific events (e.g., inpatient falls) or specific types of events (e.g., drug interactions).

The choice of events depends on the clinical goals and priorities of the institution. Events can be defined as important by virtue of being common (with at least a moderate effect or potential effect on the patient); or of having high impact (despite being relatively rare); or of being particularly instructive (perhaps uncommon and with minor effect, but pointing to important system errors that might help prevent other more clinically important errors). The choice of events also depends on the types of data that are available electronically.

For example, drug interactions cannot be targeted well if medication order entry, pharmacy, and medication list information are absent from the electronic record.

b. Targets for calibration: mandatory reporting and published event rates. Electronic detection of events is still not well characterized. Evans et al. found that electronic detection complements other forms of event reporting, such as manually reported events, and that there may be little overlap [33]. Therefore, if electronic detection will be used operationally in an institution, it should be calibrated by external measures. One approach is to compare the result to a manual reporting system. Some states, for example, mandate reporting of certain events, and institutions implement manual reporting systems. In New York State, the NYPORTS initiative lists a set of errors and adverse outcomes that must be reported. By targeting those events for detection, one can use the manually reported events to estimate the sensitivity of the electronic detection method. The electronic method can then be used to augment the manual reporting methods.

Another way to calibrate is to select events that have been studied in similarly sized institutions [13,14,34]. One can use existing publications of automated or manual error detection as a candidate set of events to detect. This will allow comparison of error rates and a gross assessment of whether the approach is uncovering a reasonable number of events.

c. Explicit voluntary reporting in the medical record. Another approach is to look for general indicators that an event may be present without specifying the exact type of error. This approach can be used to discover new types of events and broaden the search criteria on the more specific approaches. Although it may not happen often, care providers occasionally document events as being errors or adverse outcomes in the electronic medical record. That is, the provider may state not only that there was a medical condition, but also the fact that the condition represents an error or adverse outcome. For example, the following sentences document an event in an outpatient visit note: “He did not take the antibiotics because I mistakenly prescribed augmentin when he has a penicillin allergy. Luckily, his pharmacist caught the error.”

Certain phrases may be indicative of such events: “untoward,” “nosocomial,” “inadvertent,” “error,” “adverse,” “unexpected,” etc. A minority of the occurrences of the phrases may actually represent relevant medical events. For example, “trauma” is likely to represent trauma that occurred before the medical encounter and which is the likely reason for the encounter rather than a mishap during the encounter. The predictive value of “trauma” to detect interesting medical events is likely to be low. Phrases like “inadvertent,” however, may be more fruitful. The tradeoff, however, is that terms with high predictive value tend to occur less frequently in text. Either way, the approach requires a manual review after the initial electronic screening to document which instances represent events of interest.

The sample of errors found using this approach is likely to be biased. It remains unclear at this time what prompts providers to document occurrences as errors in the record, but given the likely low rate of reporting, the events found in this fashion cannot be considered a representative sample. The approach may be useful nevertheless to uncover new types of errors that are not being targeted routinely today.

d. Conflicts in the record. Another broad-based approach is to look for cases in which conflicting evidence may signify a medical event. For example, the occurrence of a myocardial infarction in a non-cardiac admission demonstrates an adverse outcome and may point to an error. There are several overlapping types of conflicts, including various kinds of mismatches of diagnoses and treatments. The generic screening criteria of Bates et al. [35] can be seen as a form of finding various conflicts in the record (trauma in a patient not admitted for trauma, myocardial infarction after a procedure, new neurological defect, etc.). While these are broad, they are a useful starting set to determine if more information can be derived from the charts (especially the narrative data) to further narrow the search for events.

This approach can be carried out at several levels of granularity. The most general would be to simply look for all cases where there is any kind of conflict, such as a change in primary diagnosis. This would most likely lead to too many false positive events. The most specific would be to look for specific conflicts, such as myocardial infarction in an asthma patient. While this is useful and may uncover useful events, it requires a great deal of manual coding of rules. An intermediate solution is to look for classes of conditions or interventions involved in a conflict. For example, one could look for any cardiac diagnosis that shows up in a non-cardiac patient. This would catch not just myocardial infarction, but also heart failures and arrhythmias.

Clinical data repository

Event discovery may use the full clinical repository or a purposely-defined subset. In many institutions what we generically refer to as the “repository” in this paper is really a pair of databases with duplicate data: 1) transaction-oriented repository; and 2) clinical data warehouse. For example, at Columbia-Presbyterian Medical Center, the transaction-oriented repository [36] serves as the basis of the online clinical information system. It is a relational database clustered by patient and optimized for retrieving all the data for single patients. A clinical event monitor [37] sits atop this transaction-oriented repository and applies Arden Syntax-based rules (the Health Level Seven standard for representing health knowledge) [38,39] to clinical transactions. The other copy of the data is contained in the clinical data warehouse (Sybase relational database on a Sun UNIX server), which is clustered by data type and is optimized for cross-patient queries. This database is well suited to retrospective exploration and is used for clinical research and administrative support. The clinical data warehouse is used for the retrospective analysis, and the transaction-oriented repository is used for testing real time event detection.

Natural language processing

The medical record serves to document the patient’s medical condition, the interventions applied to the patient, and the patient’s response to those interventions. Unfortunately, the portion of the medical record that is stored in coded electronic form at most institutions—administrative, financial, and selected ancillary information—is a mere shadow of the patient’s true state and progress [23,40]. Narrative documents such as discharge summaries, operative reports, progress notes, admission notes, signout notes, consult notes, nursing notes, radiology reports, pathology reports, other ancillary reports, and outpatient notes contain a much more detailed description of the patient. Many of these documents are captured in electronic form at health care institutions today, but their narrative format make them inaccessible to large scale or automated analysis.

Natural language processing [41–47] offers a solution. It converts machine-readable narrative text into a structured form. For example, a natural language processor might code this excerpt from a radiographic report, “Improved patchy opacity in the left lower lobe, no effusions seen,” as follows:

```
finding: opacity
    descriptor: patchy
    body location: left lower lobe of lung
    change: better
finding: pleural effusion
    certainty: no
```

This structured format allows the data to be used for clinical research—generating and testing hypotheses with large sample sizes and screening patients for studies on a large scale—and for clinical care via automatically generated alerts and reminders. Natural language processing can put vast stores of coded information at the fingertips of researchers. We believe that such coded information would be invaluable to patient safety research [28].

Simpler text searches that are based upon information retrieval techniques (similar to MEDLINE or Web-based search engines) can be a somewhat useful alternative [24,48–50]. Rather than attempt to convert all narrative text to coded form, these techniques look for documents likely to contain concepts of interest. For example, one could look for "aspiration pneumonia" and its lexical variants to find patients who may have suffered a complication of conscious sedation.

Nevertheless, it has been demonstrated that natural language processing can achieve higher accuracy than such search techniques [45,51,52]. Most of the difference reported in these studies was attributable to reduced specificity, as the search engine selected many reports in which a condition was actually being denied (e.g., "no evidence to suggest pneumonia or pneumothorax").

At least two independent groups have demonstrated that natural language processing can be as accurate as expert human coders for coding radiographic reports, as well as more accurate than simple text-based methods, such as searching for relevant phrases in the reports [45,51,52]. Demonstration of the use of natural language processing to code complex reports, such as admission notes and discharge summaries, is promising [46,47,53–58]. The potential of natural language processing to facilitate clinical research has been recognized [59,60] and demonstrated on a stroke database of 471 patient records [61].

There have been at least two studies on the use of natural language processing for patient safety. One study used a system based on a medical terminology lexicon to find adverse drug events in outpatient visit notes [24]. The results were promising, but the authors noted that limitations in the technique—it was purely lexical rather than a full natural language understanding system—led to many false positives. Problems included recognizing negative terms, recognizing medical differentiating terms, and understanding the context within which common pathologic conditions were being used.

The other study was part of the Applied Informatics health information network project [62], in which a series of Arden Syntax modules were generated to estimate a patient's risk for active tuberculosis and sent an alert to the hospital epidemiologist when a high risk patient was placed in a shared hospital room. The prospective study showed that the system reduced the number of patients with active tuberculosis in shared rooms from 13 to 7 (30 tuberculosis patients had been placed in isolation appropriately by the physician), and recommended re-isolation of a patient who had been taken off isolation too soon [26]. The majority of those patients were uncovered by natural language processing of chest radiographic reports using the Medical Language Extraction and Encoding System (MedLEE) [44,63] developed at Columbia University and Queens College, NY. This study was an example of detecting errors in real time (patients at high risk for tuberculosis placed in a shared room against hospital policy), following by electronic alerts to create a recovery (switch to respiratory isolation).

The natural language processor may need to be modified to accommodate text relevant to patient safety. The form this modification takes depends on the processor's approach. For example, MedLEE [44] uses a lexicon to categorize words and phrases into semantic classes and a semantic grammar to parse sentences and create an intermediate representation that is then used to generate a structured output. It has been found that when moving to a new clinical area, most of the work involves expanding the lexicon rather than modifying the grammar (see, for example, work on head computerized tomography and on discharge summaries [54,61]). Therefore, one can exploit a patient safety ontology (e.g., the ontology by Stetson et al., which focused on communication errors [64]) to expand the lexicon with minimal additional work on the grammar.

Assessing whether an error occurred (versus an averse outcome without error, for example) may require a complex analysis of multiple clinical variables obtained in the report, as well as knowledge of the sequence of events. One can exploit information such as the section of the document (e.g., history of present illness versus past medical history), the paragraph, the ordering of sentences, and temporal cues from those sentences. The temporal sequence of findings without explicit temporal information is inferred from adjacent findings. Refinement of this approach to discourse analysis may be necessary.

Generating queries to detect events

Two overall approaches to queries may be used: manually written queries and automatically generated queries via machine learning. Queries can be authored by a knowledge engineer in collaboration with a clinical expert (with expertise in the area of the target events). The query authors study the target events, looking for clinical cues that might alert one to the events. A subset of the repository is chosen that is used in the process of developing the query. When the performance of the final query is measured using the repository, this subset is left out to avoid bias associated with testing on the training set. Defining the queries is then an iterative process.

Based on this analysis, an initial query will contain inclusion and exclusion criteria over the repository. For example, the target may be inpatients (based on registration data) who are adults (based on date of birth) and who had an adverse drug reaction. The latter might be based on direct mention of an allergic or adverse drug reaction in a resident signout note, nursing note, or discharge summary, or might be based on mention of a symptom (e.g., rash) accompanied by a relevant medication in the signout note, discharge summary, or pharmacy system. The various documents are linked by the patient's medical record number, the time of the document, and for certain reports, a case number.

Based on an initial retrieval from the repository, the authors can uncover other candidate terms to search for. Much important information is recorded redundantly in the record, so finding a case one way may lead to the discovery of new cues to find other cases. The cues may be found in the same documents or in other documents for the same patient. If manually reported events are available, the records of a subset of these can be reviewed, also looking for cues to detect such events in the repository. The queries can then be modified to reflect the new cues and rerun on the repository. The results are again checked in an iterative process until no further progress is made. The final queries are then run against the rest of the repository and verified as described below.

Querying natural language processed data can be challenging because the information may be nested and concepts may not always be linked. For example, when the following sentence is parsed by MedLEE—"History of skin rash thought to be allergic reaction to diuretics," the skin rash and the allergic reaction to diuretics are parsed correctly (even capturing "thought to be" as being less than fully certain), but the two concepts are not explicitly linked. If the goal were to find allergic reactions that specifically include a rash, then the query author might look for the two comments in the same or adjacent sentences. A resulting query, which also checked the age of the patient as of the date of the skin rash and whether the patient was currently being hospitalized, might be as follows ("#" signifies a comment):

```
finding1 belongs to class <medication>
and finding1.reaction = <allergy>
and finding2 = <rash>
and finding2.bodyloc = <skin>
and absolute_value(finding1.sid – finding2.sid) <= 1  # same or adjacent sentences
and years(finding1.primarytime – birthdate) >= 18
and admission1.dischargedate >= finding1.primarytime  # for any admission1
and admission1.admitdate <= finding1.primarytime
```

Machine learning may also play a role. It may uncover clinical cues to events that would not have been thought of by expert query authors. To employ supervised learning, one requires a training set of known events. This can be obtained from manual event reporting and from events detected and verified via manually written queries. A number of techniques can be used: decision trees, association rules, nearest neighbor, naïve Bayes, neural networks, support vector machines, etc [65].

Our experience [66–70] and that of others [71] in generating queries against the electronic medical record has been that expert-authored queries are as good or (usually) better than those generated by machine learning. Nevertheless, the approach can be useful to find new cues for event detection.

Verifying events

To assess the performance of the system, detected events must be verified for accuracy. General definitions of medical error [1], recovery [72], and adverse outcome [1] are available, but specific definitions for error, recovery, and adverse outcome will necessarily flow from the target event. For example, if the goal of the query is to find cases in which a myocardial infarction occurred in a non-cardiac case, then the adverse outcome will be myocardial infarction and its complications, errors will be actions (or lack of action) that could have led to the myocardial infarction, and recovery will not be relevant as the goal in this example is to find an adverse outcome. If, on the other hand, the goal is broader, such as self-reporting of medical errors in the patient record, then error will be defined as whatever the provider was self reporting (as long as it fits within the broad definition of medical error), and adverse outcome and recovery will be defined as being related to that reported error. If the provider says, “I mistakenly gave Augmentin to a penicillin allergic patient and the pharmacist caught it,” then the error is the prescribing of a penicillin-containing medication to a patient with a known allergy, the recovery is the pharmacist’s catching it, and there is no actual adverse outcome.

There are three levels of evidence available: (1) the electronic chart; (2) the electronic and paper charts taken together; and (3) the charts and direct interviews of the involved providers. These levels are nested, with higher-numbered levels including all the data from the lower levels. Automated queries are restricted to information in the electronic chart. The levels allow one to assess the accuracy of several approaches to event detection:

- How well do automated queries perform compared to all the evidence available (compare automated queries on level 1 to human review on level 3)?
- How well can a retrospective review be expected to perform (compare human review on level 2 to level 3)?

- What is the maximum achievable performance of the automated system given the available evidence (compare human review on level 1 to level 3)?
- How well does an automated system perform, given the evidence available to it (compare automated queries on level 1 to human review on level 1)?

At any level, it will frequently be unknowable whether there was an actual error, a recovery, or an adverse outcome. Therefore, reviewers can be asked to supply information along three axes for each event report: (1) whether it is likely that a medical error occurred, unlikely that it occurred, or unknowable from the data, and a brief orienting description; (2) whether it is likely that a recovery occurred, unlikely that it occurred, or unknowable from the data, and a brief orienting description; (3) whether it is likely that an adverse outcome occurred, unlikely that it occurred, or unknowable from the data, and a brief orienting description.

On a given axis, the performance of the automated queries can be estimated given cases where the answer in the comparison group (e.g., human review of the electronic and paper charts) is knowable. One can also report the maximum and minimum performance, assuming that the answers marked as unknowable in the comparison group would have agreed or disagreed with the answer given by the automated queries.

One can quantify performance of the automated queries in terms of positive predictive value, sensitivity, and specificity in detecting an event or a property of that event (did an error occur, did a recovery occur, etc.). In addition, receiver operating characteristic (ROC) curve area and inter-rater reliability can be quantified.

One can also quantify the automated queries’ ability to separate the various types of events: errors with adverse outcome, near misses, errors without recovery or harm, and adverse outcomes without error based on reviewers’ coding along the three axes: was there an error, was there a recovery, and was there an adverse outcome.

If the approach to error detection is iterative—failures to detect events properly are fed back to improve the queries (and other components such as a natural language processor)—then work on one type of target event may affect the performance on the next type of target event. Therefore, it is important to avoid testing on the training set. That is, the system must not be tested on the same events that were used to improve the components of the system, or else performance will be overestimated. By holding back a set of events (for each target event type) for final testing, but iterating several rounds of training on a single set and using cross validation to avoid over fitting the data, one can use the data efficiently but end up with a rigorous measure of performance. Cross validation is difficult to employ for manual tasks such as knowledge engineering, but it was used successfully for a natural language processing evaluation [51].

Error characterization via a systems approach and a cognitive approach

By a systems approach, we mean the characterization of an error in the context of the health care system and its processes, including the role that the external procedures of delivering care play in the error. One can employ the medical error classification scheme applied by Kaplan, Battles, and coworkers [8,73] to the MERS-TM medical errors reporting and analysis system, which was used in transfusion medicine. They identified errors by what happened, where in the process it occurred, when it happened, and who was involved in the event. They then applied Van der Schaaf's Eindhoven Classification Model for errors [72], which categorizes the errors as being due to system failures, including technical and organizational problems, and human errors, including problems with knowledge-based, rule-based, or skill-based behaviors. This approach is consistent with the frameworks of Reason [74] and Rasmussen [75]. Subsequently, entries in the database were examined using root cause analysis techniques.

Given a representative sample of apparent errors and, based on manual review of the electronic record, one can determine how much of the error classification scheme can be filled in using the available data. This, in turn, can be used to estimate how much of the scheme can be filled in automatically. There may be enough information to fill in most of the identifying attributes, but accurate classification may not be possible without in depth analysis.

In addition to partial automated coding of the errors, one can use this analysis to improve error detection itself. The scheme provides a framework for searching the electronic record in an organized fashion for information related the errors. One can learn of new cues that might facilitate error detection. That is, in exhaustively searching the record for evidence of technical or organizational problems that may have contributed to the error, one can find cues that might not otherwise have been thought of.

By a cognitive approach, we mean understanding how errors occur and how a provider’s internal knowledge of that error might be documented in the electronic medical record. In the absence of electronic medical records that capture data in a standardized and structured way, the only record one has to explain the basis for a clinician’s decisions is the clinical note. These include histories and physical examination (usually repeated in the discharge summary), progress notes (including resident signout notes), operative reports, consultation reports, and other text documents that are frequently created by clinicians as they describe their observations and record the actions that they have decided to take. Some notes may not contain enough information to provide a clear basis for the decisions that the clinicians have made. In an effort to understand problems of this sort in clinical notes, and in recognition that text documents are still the primary record of most decisions made in patient-care settings, one can use the methods of cognitive analysis to gain insight into the limitations and characteristics of dictated materials.

In particular, methods of comprehension analysis [76–81], namely propositional analysis and semantic networking [82–84], provide a basis for understanding the nature of reasoning that is reflected in dictated notes and for identifying gaps in logic, circularity, or non-sequiturs. Using such methods, one can look for any disconnect between the actions to be taken (e.g., treatment plans) and the information on which such actions are evidently based (e.g., history, physical examination data, and diagnostic assessments) [85,86]. Using specific methods of scoring the summary documents, one can study the text representation, which varies greatly among readers depending on the prior content knowledge and level of expertise. One can identify and characterize the nature of inferences generated as physicians attempt to link the content of a text to their personal world knowledge. This gives a way to identify sources of incorrect inferences (e.g., inferences that are based upon ambiguous text or incomplete text information). These documents can also be examined for the patterns of reasoning and evaluation of evidence. The goal is to use the results of these analyses to inform the future use of text processing techniques for error detection in dictated notes as well as the design of future electronic medical records that are intended to capture the basis for decisions so that suitable decision support can be offered at the point of care.

Bias

Recognizing that errors are frequently not documented in the medical record, one must also assess the bias associated with information system surveillance. That is, what types of errors are better documented in the electronic medical record, and what types of errors are missing. For example, adverse drug events, an important area in patient safety research, are in fact amenable to detection via information systems, whereas others may be missing. One can compare the distribution of errors detected via this system to that of the events reported to responsible agencies, the raw quality assurance referral forms (with a broader range of events), near miss reports, and large published surveys of errors (e.g., [2,3]). One can assess the distribution with respect to error type, type of provider involved, severity of the outcome, service, type of visit (inpatient, outpatient), and various patient characteristics.

Extensions

Prevention and recovery are obvious extensions of automated detection. To prevent and recover from errors, the errors must be detected in a timely fashion and the errors must be preventable or recoverable. One can estimate the proportion of errors that would have been preventable (e.g., [16]) and, more specifically, which would have been preventable via automated alerts and reminders. The latter requires (1) that the data source be available in real time (e.g., not discharge summaries), (2) that the error or potential error could be detected before the adverse outcome is irreversible, and (3) that avoidance or recovery could be triggered by an alert or reminder.

The utility of the methods in the framework for patient safety can be increased through refinement of the various steps and through improvement in natural language processing systems. In addition, the adoption and deployment of data standards in several areas are required to support these techniques. Use of a clinical document architecture standard such as the Health Level 7 CDA [87] will provide additional context to narrative data and consequently improve the performance of natural language processing. Widespread implementation of the National Center for Vital and Health Statistics-recommended core patient medical record information terminologies will increase the amount of coded data available for use in automated event detection [88].

Conclusions

Electronic detection of medical events appears to be an important and feasible avenue of patient safety research. Automated event detection requires a careful selection of target events and an assessment of what data are available electronically. Raw data such as narrative notes must be converted to coded form using natural language processing or information retrieval techniques. Detected events must be verified, and they may be characterized using a systems or cognitive approach. Designing effective queries is usually an iterative process. Adoption of standard terminologies and a clinical document architecture may improve performance and generalizability.

ACKNOWLEDGMENT

This study was funded by Agency for Healthcare Research and Quality grant R18 HS11806 “Mining complex clinical data for patient safety research” and National Library of Medicine grant R01 LM06910 “Discovering and applying knowledge in clinical databases.”

REFERENCES

1. Kohn KT, Corrigan JM, Donaldson MS (editors) Committee on Quality of Health Care in America. To Err is Human: Building a Safer Health System. Institute of Medicine. National Academy Press, 1999
2. Brennen TA, Leape LL, Laird NM, Herbert L, Localio AR, Lawthers, AG et al. Incidence of adverse events and negligence in hospitalized patients: Results of the Harvard Medical Practice Study I. NEJM 1991;324:370-6.
3. Thomas EJ, Studdert DM, Burstin HR, Orav EJ, Zeena T, Williams EJ et al. Incidence and types of adverse events and negligent care in Utah and Colorado. Medical Care 2000;38:261-71.
4. Leape LL, Brennan TA, Laird N, Lawthers AG, Localio AR, Barnes BA, et al. The nature of adverse events in hospitalized patients. Results of the Harvard Medical Practice Study II. N Engl J Med 1991;324:377-84.
5. Wilson RMcL, Runciman WB, Gibberd RW, Harrison BT, Newby L, Hamilton JD. The Quality in Australian Health Care Study. Med. Journal of Australia 1995;163: 458-71.
6. Zapt D, ReasonJT. Introduction to error handling. Appl Psychol 1994;43:427-32.

7. Novello AC. NYPORTS: The New York Patient Occurrence and Tracking System Annual Report 1999. New York State Health Department, 2001.

8. Kaplan HS, Battles JB, Van der Schaaf TW, Shea CE, Mercer SQ. Identification and classification of the causes of events in transfusion medicine. Transfusion 1998;38:1071-1081.

9. Hiatt HH, Barnes BA, Brennan TA, et.al. "A study of Medical Injury and Medical Malpractice." N Engl J Med, 1989:321:480-4.

10. O'Neil AC, Petersen LA, Cook EF, Bates DW, Lee TH, Brennan TA. "Physician Reporting Compared to Medical Record Review To Identify Adverse Medical Events." Ann Intern Med 1993;119:370-376.

11. Berry LL, Berry RS, Sherrin TP, Fudge KA. "Sensitivity and Specificity of three methods of detecting adverse drug reactions." Am J of Hosp Pharm, 1988:45:1534-9.

12. Evans RS, Pestotik SL, Classen DC, et.al. "Preventing adverse drug events in hospitalized patients." Ann Pharmachother, 1994;28:523-27.

13. Jha AK, Kuperman GJ, Teich JM, Leape L, Shea B, Rittenberg E, et al. Identifying adverse drug events: Development of a computer-based monitor and comparison with chart review and stimulated voluntary report. J Am Med Inform Assoc 1998;5:305-14.

14. Evans RS, Pestotnik SL, Classen DC, Burke JP. Evaluation of a computer-assisted antibiotic-dose monitor. Ann Pharmacother 1999;33:1026-31.

15. Evans RS, Larsen RA, Burke JP, Gardner RM, Meier FA, Jacobson JA, et al. Computer surveillance of hospital-acquired infections and antibiotic use. JAMA 1986;256:1007-11.

16. Bates DW, O'Neil AC, Boyle D, Teich J, Chertow GM, Komaroff AL, Brennana TA. "Potential Identifiabilility and Preventability of Adverse Events Using Information Systems." JAMIA, 1994; 1:404-11.
17. Bates DW, Leape LL, Cullen DJ, et.al. "Effect of Computerized Physician Order Entry and a Team Intervention on Prevention of Serious Medical Errors." JAMA 1998;280:1311-16.
18. Bates DW, Teich JM Lee J, Seger D, et.al. "The impact of Computerized Physician Order Entry on Medication Error Prevetion." JAMIA 1999:6:313-21.
19. Kossovsky MP, Sarasin FP, Bolla F, Gaspoz JM, Borst F. Distinction between planned and unplanned readmissions following discharge from a department of internal medicine. Methods Inf Med 1999;38:140-3.
20. Roos LL, Jr., Cageorge SM, Austen E, Lohr KN. Using computers to identify complications after surgery. Am J Public Health 1985;75:1288-95.
21. Roos LL, Stranc L, James RC, Li J. Complications, comorbidities, and mortality: improving classification and prediction. Health Serv Res 1997;32:229-38.
22. Iezzoni LI, Foley SM, Heeren T, Daley J, Duncan CC, Fisher ES, et al. A method for screening the quality of hospital care using administrative data: Preliminary validation results. QRB Qual Rev Bull 1992;18:361-71.
23. Iezzoni LI. Assessing quality using administrative data. Ann Intern Med 1997;127:666-74.
24. Honigman B, Light P, Pulling RM, Bates DW. A computerized method for identifying incidents associated with adverse drug events in outpatients. Int J Med Inf 2001;61:21-32.

25. Committee on Quality of Health Care in America, Institute of Medicine. Crossing the Quality Chasm: A New Health System for the 21st Century. Washington, D.C.: National Academy Press, 2001.

26. Knirsch CA, Jain NL, Pablos-Mendez A, Friedman C, Hripcsak G. Respiratory isolation of tuberculosis patients using clinical guidelines and an automated clinical decision support system. Infection Control and Hospital Epidemiology 1998;19:94-100.

27. Lau LM, Warner HR. "Performance of a Diagnostic System (Iliad) as a Tool for Quality Assurance." Computers and Biomedical Research, 1992;25:314-23.

28. Bates DW, Evans RS, Murff H, Stetson PD, Pizziferri L, Hripcsak G.Detecting Adverse Events Using Information Technology. J Am Med Inform Assoc 2003;10:11528.

29. Jollis JG, Ancukiewicz M, DeLong ER, Pryor DB, Muhlbaier LH, Mark DB. Discordance of databases designed for claims payment versus clinical information systems. Implications for outcomes research. Ann Intern Med. 1993;119:844-50.

30. Friedman C, Hripcsak G. Natural language processing and its future in medicine. Acad Med 1999;74:890–5.

31. Hersh WR. Information Retrieval: A Health Care Perspective. New York: Springer-Verlag, 1996.

32. Johnson SB, Chatziantoniou D. Extended SQL for manipulating clinical warehouse data. Proc AMIA Symp. 1999;819–23.

33. Evans RS, Larsen RA, Burke JP, Gardner RM, Meier FA, Jacobson JA, et al. Computer surveillance of hospital-acquired infections and antibiotic use. JAMA 1986;256:1007-11.

34. Classen DC, Pestontnik SL, Evans RS, Burke JP. "Computerized Surveillance of Adverse Drug Events in Hospitalized Patients." JAMA, 1991;266:2847-51.

35. Bates DW, O'Neil AC, Petersen LA, Lee TH, Brennan TA. Evaluation of screening criteria for adverse events in medical patients. Med Care 1995;33:452-62.

36. Johnson SB, Hripcsak G, Chen J, Clayton P. Accessing the Columbia Clinical Repository. Proc Annu Symp Comput Appl Med Care 1994:281-5.

37. Hripcsak G, Clayton PD, Jenders RA, Cimino JJ, Johnson SB. Design of a clinical event monitor. Comput Biomed Res 1996;29:194-221.

38. Hripcsak G, Ludemann P, Pryor TA, Wigertz OB, Clayton PD. Rationale for the Arden Syntax. Comput Biomed Res 1994;27:291-324.

39. Hripcsak G. Writing Arden Syntax medical logic modules. Comput Biol Med 1994;24:331-63.

40. Jollis JG, Ancukiewicz M, DeLong ER, Pryor DB, Muhlbaier LH, Mark DB. Discordance of databases designed for claims payment versus clinical information systems. Implications for outcomes research. Ann Intern Med 1993;119:844-50.

41. Spyns P. Natural language processing in medicine: an overview. Methods Inf Med 1996;35:285-301.

42. Baud RH, Rassinoux AM, Scherrer JR. Natural language processing and semantical representation of medical texts. Methods Inf Med 1992;31:117-125.

43. Haug PJ, Ranum DL, Frederick PR. Computerized extraction of coded findings from free-text radiologic reports: work in progress. Radiology 1990; 174:543-548.

44. Friedman C, Hripcsak G, DuMouchel W, Johnson SB, Clayton PD. Natural language processing in an operational clinical information system. Natural Language Engineering 1995;1:83-108.

45. Hripcsak G, Friedman C, Alderson PO, DuMouchel W, Johnson SB, Clayton PD. Unlocking clinical data from narrative reports: a study of natural language processing. Ann Intern Med. 1995;122:681-8.

46. Sager N, Lyman M, Bucknall C, Nhan N, Tick LJ. Natural language processing and the representation of clinical data. J Am Med Inform Assoc. 1994;1:142-60.

47. Zweigenbaum P, Bouaud J, Bachimont B, Charlet J, Boisvieux JF. Evaluating a normalized conceptual representation produced from natural language patient discharge summaries. Proc AMIA Annu Fall Symp. 1997:590-4.

48. Goldman JA, Chu WW, Parker DS, Goldman RM. Term domain distribution analysis: a data mining tool for text databases. Methods Inf Med 1999 Jun; 38(2):96-101.

49. Rind DM, Yeh J, Safran C. Using an electronic medical record to perform clinical research on mitral valve prolapse and panic/anxiety disorder. Proc Annu Symp Comput Appl Med Care 1995:961.

50. Giuse DA, Mickish A. Increasing the availability of the computerized patient record. Proc AMIA Annu Fall Symp. 1996;633-7.

51. Hripcsak G, Kuperman GJ, Friedman C. Extracting findings from narrative reports: software transferability and sources of physician disagreement. Methods Inf Med 1998;37:1-7.

52. Fiszman M, Chapman WW, Aronsky D, Evans RS, Haug PJ. Automatic detection of acute bacterial pneumonia from chest X-ray reports. J Am Med Inform Assoc 2000;7:593-604.

53. Gabrieli ER. Computer-assisted assessment of patient care in the hospital. J Med Syst. 1988;12:135-46.

54. Friedman C, Knirsch C, Shagina L, Hripcsak G. Automating a severity score guideline for community-acquired pneumonia employing medical language processing of discharge summaries. Proc AMIA Symp 1999:256-60.

55. Lenert LA, Tovar M. Automated linkage of free-text descriptions of patients with a practice guideline. Proc Annu Symp Comput Appl Med Care 1993:274-8.

56. Hersh WR, Leen TK, Rehfuss PS, Malveau S. Automatic prediction of trauma registry procedure codes from emergency room dictations. Medinfo 1998;9 Pt 1:665-9.

57. Delamarre D, Burgun A, Seka LP, Le Beux P. Automated coding of patient discharge summaries using conceptual graphs. Meth Inform Med 1995;34:345-51.

58. Spyns P, Nhan NT, Baert E, Sager N, De Moor G. Medical language processing applied to extract clinical information from Dutch medical documents. Medinfo. 1998;9 Pt 1:685-9.

59. Gabrieli ER. Automated processing of narrative medical text-a new tool for clinical drug studies. J Med Syst 1989;13:95-102.

60. Lyman M, Sager N, Tick L, Nhan N, Borst F, Scherrer JR. The application of natural-language processing to healthcare quality assessment. Med Decis Making 1991;11(suppl):S65-8.

61. Elkins JS, Friedman C, Boden-Albala B, Sacco RL, Hripcsak G. Coding neuroradiology reports for the Northern Manhattan Stroke Study: a comparison of natural language processing and manual review. Comput Biomed Res 2000;33:1-10.

62. Hripcsak G, Knirsch C, Jain NL, Stazesky RC, Pablos-Mendez A, Fulmer T. A health information network for managing inner-city tuberculosis: bridging clinical care, public health, and home care. Comput Biomed Res 1999;32:67–76.

63. Friedman C, Alderson PO, Austin JH, Cimino JJ, Johnson SB. A general natural-language text processor for clinical radiology. J Am Med Inform Assoc. 1994;1:161-74.

64. Stetson PD, MdKnight LK, Bakken S, Curran C, Kubose TT, Cimino JJ. Development of an ontology to model medical errors, information needs, and the clinical communication space. Proc AMIA Symp 2001;672–6

65. Adriaans P, Zantinge D. Data Mining. Harlow, England: Addison Wesley Longman; 1996.

66. Wilcox A, Hripcsak G. Medical text representations for inductive learning. Proc AMIA Symp 2000; 923-7.

67. Wilcox A. Automated Classification of Medical Text Reports [dissertation]. Columbia University; 2000.

68. Wilcox A, Hripcsak G. Classification algorithms applied to narrative reports. Proc AMIA Symp 1999; (1-2):455-9.

69. Wilcox A, Hripcsak G. Knowledge discovery and data mining to assist natural language understanding. Proc AMIA Annu Fall Symp 1998; 835-9.

70. Wilcox A, Hripcsak G, Knirsch C. Knowledge discovery using the electronic medical record (poster). Proc AMIA Symp 2002: 1198.

71. Chapman WW, Fizman M, Chapman BE, Haug PJ. A comparison of classification algorithms to automatically identify chest X-ray reports that support pneumonia. Journal of Biomedical Informatics 2001;34:4-14.

72. Van der Schaaf, TW. New Miss Reporting in the Chemical Process Industry. Eindhoven, The Netherlands: Eindhoven University of Technology; 1992. Thesis.

73. Battles JB, Kaplan HS, Van der Schaaf TW, Shea CE. The attributes of medical event-reporting systems: experience with a prototype medical event-reportign system for transfusion medicine. Arch Pathol Lab Med 1998;122:231-238.

74. Reason J. Generic error-modeling system (GEMS): a cognitive framework for locating common human error forms. In: Rasmussen J, Duncan K, Leplat J, eds. New Technology and Human error. London, England: John Wiley & Sons Ltd; 1987;63-83.

75. Rasmussen J. The definition of human error and a taxonomy for technicalsystems design. In: Rasmussen J, Duncan K, Leplat J, eds. New Technology and Human error. London, England: John Wiley & Sons Ltd; 1987;23-30.

76. Kintsch W. The role of knowledge in discourse comprehension: A construction-integration model. Psychological Review. 1988;95(2):163-182.

77. Kintsch W. Comprehension: A Paradigm for Cognition Cambridge: Cambridge University Press; 1998.

78. Bransford JD, Johnson MK. Contextual prerequisites for understanding: some investigation of comprehension and recall,. Journal of Verbal Learning and Verbal Behavior. 1971;11:717-726.

79. van Dijk TA, Kintsch W. Strategies of discourse comprehension New York, NY: Academic Press; 1983.

80. Kintsch W, Vipond D. Reading comprehension and readability in educational practice and psychological theory. In: Nilsson LG, ed. Perspectives on Memory Research. Hillsdale, NJ: Erlbaum; 1979.

81. Miller JR, Kintsch W. Readability and recall of short prose passages: A theoretical analysis. Journal of Experimental Psychology: Human Learning and Memory. 1980;6:335-354.

82. Frederiksen CH. Representing logical and semantic structure of knowledge acquired from discourse. Cognitive Psychology. 1975;7:371-458.

83. Sowa JF. Conceptual structures: Information processes in mind and machine Reading, MA: Addison-Wesley; 1983.

84. Sowa JF. Principles of semantic network explanations in the representation of knowledge San Mateo, CA: Morgan-Kaufmann; 1991.

85. Leprohon, J. & Patel, V.L. (1995) Decision making strategies for telephone triage in emergency medical services. Medical Decision Making, 15(3), 240-253.

86. Patel VL, Evans DA, Kaufman DR. Cognitive framework for doctor-patient interaction. In: Evans DA, Patel VL, eds. Cognitive science in medicine: Biomedical modeling. Cambridge, MA: MIT Press; 1989:253-308.

87.

http://www.hl7.org/Library/Committees/structure/CDA.ReleaseTwo.ClevelandDraft.April23.2003.zip

88. Federal Register. Vol. 88, No. 68; May 7, 2003.