American Journal of Emergency Medicine 34 (2016) 1043–1048

Contents lists available at ScienceDirect

American Journal of Emergency Medicine

journal homepage: www.elsevier.com/locate/ajem

Original Contribution

Can medical record reviewers reliably identify errors and adverse events in the ED?☆

Richard S. Klasco, MD *,[1], Richard E. Wolfe, MD [1], Terrance Lee, MD, Philip Anderson, MD, Lee S. Jacobson, MD, PhD, Joshua Solano, MD, Jonathan Edlow, MD, Shamai A. Grossman, MD

Beth Israel Deaconess Medical Center, Department of Emergency Medicine

ARTICLE INFO

Article history:
Received 28 December 2015
Received in revised form 29 February 2016
Accepted 1 March 2016

ABSTRACT

Background: Chart review has been the mainstay of medical quality assurance practices since its introduction more than a century ago. The validity of chart review, however, has been vitiated by a lack of methodological rigor.

Objectives: By measuring the degree of interrater agreement among a 13-member review board of emergency physicians, we sought to validate the reliability of a chart review–based quality assurance process using computerized screening based on explicit case parameters.

Methods: All patients presenting to an urban, tertiary care academic medical center emergency department (annual volume of 57,000 patients) between November 2012 and November 2013 were screened electronically. Cases were programmatically flagged for review according to explicit criteria: return within 72 hours, procedural evaluation, floor-to-ICU transfer within 24 hours of admission, death within 24 hours of admission, physician complaints, and patient complaints. Each case was reviewed independently by a 13-member emergency department quality assurance committee all of whom were board certified in emergency medicine and trained in the use of the tool. None of the reviewers were involved in the care of the specific patients reviewed by them. Reviewers used a previously validated 8-point Likert scale to rate the (1) coordination of patient care, (2) presence and severity of adverse events, (3) degree of medical error, and (4) quality of medical judgment. Agreement among reviewers was assessed with the intraclass correlation coefficient (ICC) for each parameter.

Results: Agreement and the degree of significance for each parameter were as follows: coordination of patient care (ICC = 0.67; $P < .001$), presence and severity of adverse events (ICC = 0.52; $P = .001$), degree of medical error (ICC = 0.72; $P < .001$), and quality of medical judgment (ICC = 0.67; $P < .001$).

Conclusion: Agreement in the chart review process can be achieved among physician-reviewers. The degree of agreement attainable is comparable to or superior to that of similar studies reported to date. These results highlight the potential for the use of computerized screening, explicit criteria, and training of expert reviewers to improve the reliability and validity of chart review–based quality assurance.

1. Background

Chart review has been the mainstay of medical quality assurance (QA) activities since its introduction by Codman [1] more than a century ago at the Massachusetts General Hospital. The validity of chart review, however, has been vitiated by a lack of methodological rigor [2].

Commonly cited weaknesses include unclear inclusion criteria, unsystematic case identification, implicit methods, subjective standards, inadequate reviewer training, lack of internal consistency, and conflation of correlation and causation (ie, the imprecise use of scientific terms such as *dependence*, *association*, and *correlation*) [3–6]. Reviews have also suffered from confusion between assessments of process (ie, physician performance) and outcome (ie, results of care) [7]. Remediation of these weaknesses began in the 1960s with the work of Donabedian [8], who differentiated between assessments of process and assessments of outcome.

In previous studies, having higher numbers of reviewers and more experienced reviewers leads to greater agreement. Brennan et al [18] demonstrated good agreement (kappa = 0.57) between senior physicians and physician-reviewers trained in the use of an explicit adverse event form for the presence of adverse events. This comports with our process and agrees well with our finding of an intraclass correlation

☆ Prior presentations: National Meeting of the Society for Academic Emergency Medicine in Dallas, TX, June 2014. New England Regional Meeting for Society for Academic Emergency Medicine, April 2014, New Haven, CT.

* Corresponding author at: Department of Emergency Medicine, Beth Israel Deaconess Medical Center, West, Clinical Center 2, One Deaconess Rd, Boston, MA 02215. Tel.: +1 303 900 3055.

E-mail address: rklasco@bidmc.harvard.edu (R.S. Klasco).

[1] These authors contributed equally to this study as first authors.

http://dx.doi.org/10.1016/j.ajem.2016.03.001
0735-6757/

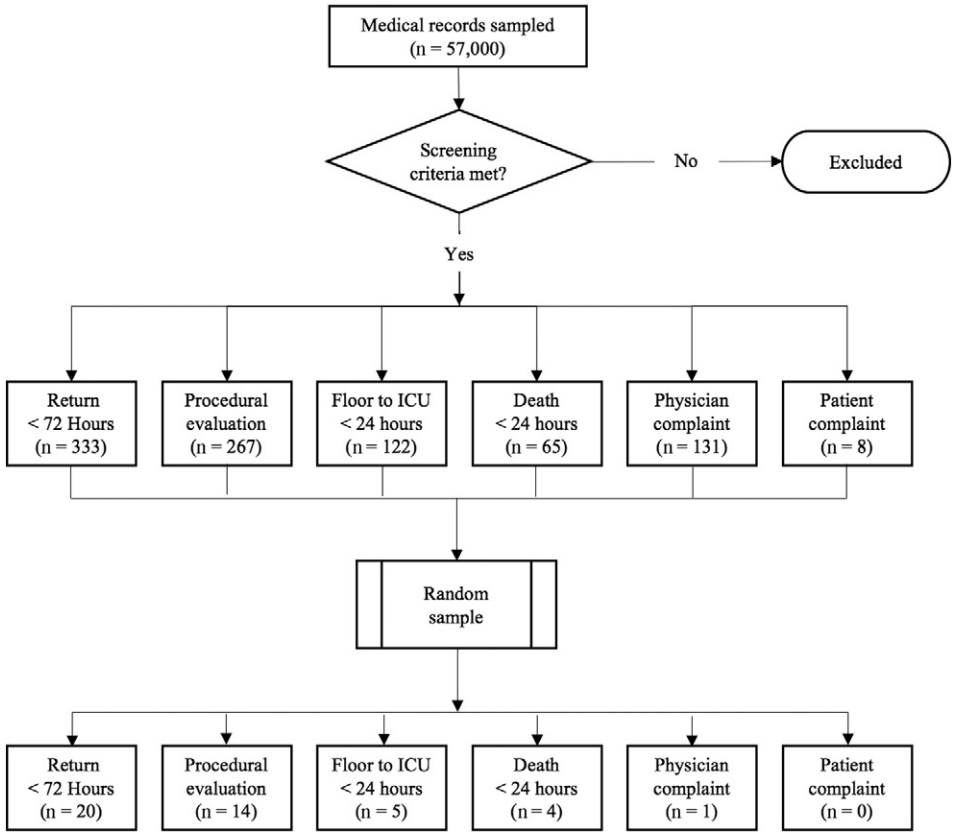

Fig. 1. Case identification flowchart.

coefficient (ICC) = 0.52 for "presence and severity of adverse events." Although kappa and ICC are not synonymous, Fleiss and Cohen [19] consider the weighted kappa to be equivalent to the ICC.

In a later study of similarly trained reviewers, Brennan et al [20] found good agreement (kappa = 0.57) for causation of an adverse event and (kappa = 0.62) for negligence. Our findings of ICC = 0.72 for "degree of medical error" in conjunction with our finding of ICC = 0.52 for "presence and severity of adverse events" seem similar in intent and magnitude.

Hayward et al [21] noted poorer levels of agreement among physician-reviewers for focused quality problems and resource utilization (kappa ≤ 0.2) than were observed in our study. Hayward et al measured agreement among pairs of reviewers and hypothesized that higher levels of agreement might have been achieved with larger numbers of reviewers. Thomas et al [22] found moderate to poor interrater reliability among 3 physicians for adverse events. Therefore, our use of 13 reviewers per case may partially explain our superior results. The implicit (ie, subjective) criteria of Hayward et al, in contrast to our explicit methods, may have also contributed to the difference in results. In further support of this notion, Hofer et al [23] noted an intermediate level of agreement (ICC = 0.16-0.46) with the use of structured implicit criteria, perhaps suggesting a dose-response phenomenon for methodological rigor.

Localio et al [24] observed frequent disagreement regarding the occurrence of adverse events. Such a finding might be expected, however, as that study used pairs of reviewers and implicit criteria. Localio et al also hypothesized that greater reviewer experience might have produced greater levels of agreement. Our use of experienced, trained reviewers may then have contributed to our superior results. The work of Allison et al [4] supports this contention. In their study, several rounds of training and refinement improved interrater reliability from 80% to 96%.

We sought to address these issues through the use of explicit case definitions and programmatic case identification within a robust, protocol-driven QA process in which all cases were reviewed by a 13-member board of board-certified emergency physicians.

2. Methods

2.1. Study design, goals, and oversight

The study sample was a prospective cohort comprised of all patients presenting to a tertiary care academic emergency department (ED) between November 2012 and November 2013. The ED has an annual census of 57,000 patients. Institutional review board jurisdiction was waived by the study hospital institutional review board.

To assess the degree of agreement among reviewers engaged in chart review, we used 6 predefined high-risk conditions that are commonly used in QA processes: return to the ED within 72 hours, procedures performed in the ED (eg, intubation, tube thoracostomy), transfer from floor-to-ICU within 24 hours of admission, death within 24 hours of admission, complaints from physicians outside of the ED, and complaints from patients.

The ED QA committee provided oversight. The QA committee is integrated into the medical center's overall QA operations through formal processes and procedures as described previously [9].

2.2. Selection of participants

An electronic medical record was created for each patient. The database of all electronic medical records was searched for cases meeting 1 or more of the above 6 predefined high-risk criteria using an electronic QA dashboard that interfaced with a commercially available health information system [10]. Cases meeting criteria were flagged for