HHS Public Access

Author manuscript

Prehosp Emerg Care. Author manuscript; available in PMC 2015 October 01.

Published in final edited form as:

Prehosp Emerg Care. 2014 ; 18(4): 495–504. doi:10.3109/10903127.2014.916022.

A Comparative Assessment of Adverse Event Classification in the Out of Hospital Setting

P. Daniel Patterson, PhD, MPH, MS, NREMT-P, **Judith R. Lave, PhD**, **Matthew D. Weaver, MPH, EMT-P**, **Francis X. Guyette, MD**, **Robert M. Arnold, MD**, **Christian Martin-Gill, MD**, **Jon C. Rittenberger, MD**, **David Krackhardt, PhD**, **Vincent N. Mosesso, MD**, **Ronald N. Roth, MD**, **Richard J. Wadas, MD**, and **Donald M. Yealy, MD**

Department of Emergency Medicine, University of Pittsburgh School of Medicine, Pittsburgh, PA (PDP, DMY, CMG, MDW, RJW, RNR, VNM, FXG, JCR). Division of General Internal Medicine, University of Pittsburgh, Pittsburgh, PA (RMA). Graduate School of Public Health, Department of Health Policy and Management, University of Pittsburgh, Pittsburgh, PA (JRL). Carnegie Mellon University, Heinz School of Policy, Management, Information Technology (DK).

Abstract

Objectives—We sought to test reliability of two approaches to classify adverse events (AEs) associated with Helicopter EMS (HEMS) transport.

Methods—The first approach for AE classification involved flight nurses and paramedics (RN/Medics) and mid-career emergency physicians (MC-EMPs) independently reviewing 50 randomly selected HEMS medical records. The second approach involved RN/Medics and MC-EMPs meeting as a group to openly discuss 20 additional medical records and reach consensus-based AE decision. We compared all AE decisions to a reference criterion based on the decision of three senior emergency physicians (Sr-EMPs). We designed a study to detect an improvement in agreement (reliability) from fair (kappa=0.2) to moderate (kappa=0.5). We calculated sensitivity, specificity, percent agreement, and positive and negative predictive values (PPV / NPV).

Results—For the independent reviews, the Sr-EMP group identified 26 AEs while individual clinician reviewers identified between 19 and 50 AEs. Agreement on the presence/absence of an AE between Sr-EMPs and three MC-EMPs ranged from K=0.20 to K=0.25. Agreement between Sr-EMPs and three RN/Medics ranged from K=0.11 to K=0.19. For the consensus/open-discussion approach, the Sr-EMPs identified 13 AEs, the MC-EMP group identified 18 AEs, and

Address correspondence and requests for reprints to: P. Daniel Patterson, Department of Emergency Medicine, University of Pittsburgh School of Medicine, 3600 Forbes Avenue, Iroquois Bldg., Suite 400A, Pittsburgh, PA 15261, (T) 412-647-3183, (F) 412-647-6999, pattersond@upmc.edu..

AUTHOR CONTRIBUTION STATEMENT: Drs. Daniel Patterson (PDP), Judith Lave (JL), Robert Arnold (RA), David Krackhardt (DK), and Donald M. Yealy (DMY) conceived the study idea and hypotheses. Drs. Patterson (PDP), Francis Guyette (FG), Christian Martin-Gill (CMG), Jon Rittenberger (JR), Vincent Mosesso (VM), Richard Wadas (RW), Ronald Roth (RR), and Mr. Matthew Weaver (MW) developed the study protocol and methodology. PDP and MW obtained institutional review board approval, carried out the study protocol, and collected study data. PDP and MW analyzed study data and fashioned study tables and figures. All authors/investigators examined study findings, interpreted study findings, and synthesized study findings for reporting in a peer-reviewed manuscript. All authors participated in development and editing of the draft manuscript, providing input and editing for all components, including the introduction, methods, results, discussion and conclusions.

COMPETING INTERESTS

The authors report no competing interests or conflicts of interests.

RN/Medic group identified 36 AEs. Agreement between Sr-EMPs and MC-EMP group was (K=0.30 95%CI −0.12, 0.72), whereas agreement between Sr-EMPs and RN/Medic group was (K=0.40 95%CI 0.01, 0.79). Agreement between all three groups was fair (K=0.33, 95%CI 0.06, 0.66). Percent agreement (58-68%) and NPV (63-76%) was moderately dissimilar between clinicians, while sensitivity (25-80%), specificity (43-97%), and PPV (48-83%) varied.

Conclusions—We identified a higher level of agreement/reliability in AE decisions utilizing a consensus-based approach for review rather than independent reviews.

Keywords

Safety; Adverse Events; Reliability; Measurement

INTRODUCTION

One-third of all hospital admissions involve an Adverse Event (AE),[1] defined as an unintended injury caused by medical management.[2] Little data exist on AEs outside the hospital setting. Helicopter Emergency Medical Services (HEMS) are an important setting for the study of AEs. Care in HEMS is complicated because its patients, who are critically ill or injured with time-sensitive and complex conditions, often present with limited information following transfer(s) of care in an uncontrollable setting. The United States National EMS Culture of Safety Strategy project (www.emscultureofsafety.org) established a need for safety awareness and improvement in EMS. Despite this effort, there has been limited research on identifying and categorizing AEs in prehospital ground or helicopter care delivery.

Approaches to detecting and categorizing AEs include review of medical records, direct observation, identification of sentinel events and anonymous reporting.[3] Of these, direct observation has many advantages over other approaches and may be considered to have the highest level of predictive validity.[3] It eliminates potential recall bias of clinicians associated with self-report and reduces the need to perform time-consuming retrospective record review. However, the use of direct observation is limited because it is costly.[3] The most common approach is medical record review (supported by a trigger tool) conducted by nurses or physicians to determine if the delivery of care led to measurable harm.[2,4] Although this approach is time consuming, it is the accepted quality assurance practice in most healthcare settings.[3]

A challenge with medical record review for AE identification in the prehospital setting is the lack of a standardized technique considered reliable, valid, and applicable. Current quality/safety assurance in use by diverse EMS systems is physician-led review of medical records to identify excellence in care, medical error, and AEs,[5-7] with no standardization across EMS systems. We have developed a content valid framework for AE identification using medical record review that incorporates an EMS-specific AE definition, use of a trigger tool, and a procedure for rating outcome categories of AEs.[8] In this paper, we evaluate this tool by assessing reliability (agreement) with a criterion standard. We hypothesized that an approach to medical record review using consensus/open-discussion would result in higher reliability than reviews utilizing traditional independent clinician review.

METHODS

Overview of Study and Design

The PittAETool provides a framework for performing medical record reviews.[8] It includes a trigger tool and multi-step process for reviewing a medical record that is similar in overall structure to other processes used for medical record review, including those used in the Harvard Medical Practice studies.[9] Use of these frameworks and a common structure promotes consistency, use of a common taxonomy, and improves the possibility of comparing results between different organizations/settings. Inconsistencies in clinician agreement regarding the presence/absence or severity of AEs after using a specific framework raise questions about the most appropriate or optimal application of a framework for reliability and validity purposes.

In our study, we used two approaches to assessing the reliability of the PittAETool. In the first approach, we asked emergency clinicians to review medical records independently and render an AE decision based only on their independent review (Figure 1). In the second approach, we separated the clinicians into two groups. One group of emergency clinicians included three prehospital flight nurses / flight paramedics (RN/Medics). The three RN/Medics were certified flight paramedics or prehospital flight nurses with responsibilities as QA officers in a large HEMS organization. The second group included three mid-career emergency medicine physicians (MC-EMPs). The three MCEMPs were licensed emergency medicine physicians with an average of 15 years of EMS and emergency medicine experience. Each group met together to discuss a medical record and produce a consensus-based decision regarding an AE. We identified a third group of senior emergency medicine physicians (Sr-EMPs) that we used to set our criterion standard for AE decisions. All Sr-EMPs had prior experience leading quality assurance committees and a minimum of 15 years experience with EMS organizations as a medical director or providing medical care and oversight. Analogous to previous research,[9] we considered the decisions made by senior emergency medicine physicians as the reference criterion (aka: gold standard) for this study and for the purpose of comparing the reliability of two different approaches to medical record review. Adjudication by senior physicians is a criterion (gold) standard in medical decision-making and AE identification.[9-13] Senior physician adjudication was the criterion standard for the Harvard Medical Practice studies.[9]

We powered the study to detect an improvement in agreement (reliability) from an estimated kappa=0.2 to 0.4 in the independent/blinded reviews to an estimated kappa=0.7 in the approach using open discussion and consensus-based AE decisions. We based our power calculations on data collected from a prior study of AEs in the EMS setting.[14] For all power calculations, we used the KappSize package in R and a fixed alpha level of 0.05 and power of 0.80. Assuming three unique reviewers and a 50% likelihood of an AE in our sample of medical records, we required a minimum of 38 medical records to detect an improvement in agreement from kappa of 0.2 to 0.5 for the first approach (independent medical record reviews). To test our main hypothesis that agreement (reliability) would be higher utilizing a consensus-based (open-discussion) approach, a minimum of 14 cases would be required to detect an improvement from a kappa of 0.2 to 0.7. To ensure adequate power, we randomly

sampled 50 medical records for the approach using independent reviews, and 20 medical records for the open-discussion, consensus-based review approach. The University of Pittsburgh Institutional Review Board approved this study.

Study Setting and Data Source

We collected medical records from STAT MedEvac, a large HEMS agency with 17 base sites spread across Eastern Ohio, Pennsylvania, Northern Maryland, and the District of Columbia. The annual patient volume for STAT MedEvac is approximately 10,000 patients and includes emergent scene to hospital and interfacility air medical transports. STAT MedEvac follows a multi-step process for medical record review that includes a review of all medical records.

We obtained two random samples of STAT MedEvac medical records from 2008. We designed the first sample of 50 records to include an equivalent number of medical records from each month of 2008. Forty-percent of the sampled medical records had a preexisting Quality Assurance (QA) marker (A “QA flag”). A QA flag is a memo associated with a particular medical record by a medical director physician or his/her designee who is in charge of quality assurance and medical record review. A QA flag is most often used to direct the attention of the flight paramedic or nurse to a deviation in protocol, a documentation error, other errors or irregularities identified in the medical record during routine review. A QA flag may also be used to acknowledge exceptional performance on a challenging patient encounter. We designed the sampling of the second sample of 20 medical records to include 10 with a QA flag. We purposely included both medical records with and without QA flags to increase the likelihood that some of our charts would include an AE. We selected records from 2008 in order to reduce the likelihood that any of our clinician reviewers would have had prior exposure to or memory of the selected medical records.

Study Protocol

We trained all nine clinicians to use our recently published framework for reviewing HEMS medical records.[15] The PittAETool is a valid framework that includes a four-step process for medical record review and AE detection and classification (Figure 2). The advantages of using this framework for medical record review and AE detection include: 1) it incorporates a standard lexicon of safety terminology,[16] 2) it follows a standardized protocol based on previous research and preferred practice in patient safety,[16] and 3) it is content valid.[15] All clinicians in this study were involved in the development of the PittAETool and were familiar with it and the intended use. The first author (PDP) led a practice session with three actual HEMS medical records to ensure that all clinicians understood how to apply the PittAETool. Clinicians practiced documenting their AE decisions using a standard data-recording sheet that can be located in a prior publication.[15] During the practice session, we addressed questions regarding use of the PittAETool and emphasized that a “triggered” medical record should not – by itself – imply the presence of an AE, but rather prompt a detailed review of the entire medical record. We also emphasized the sensitive characteristics of our adopted AE definition in comparison to previously published definitions. That definition reads as: An adverse event in EMS is a harmful or potentially

harmful event occurring during the continuum of EMS care that is potentially preventable and thus independent of the progression of the patient's condition."[14] In total, clinician reviewers accumulated an estimated 20+ hours of exposure to and training in use of the PittAETool; well beyond the 2-3 hours reported in other studies.[9,17]

As noted earlier, we asked the clinician reviewers to apply the PittAETool in two different ways. First, we asked the three MC-EMPs and three RN/Medics to follow the more traditional approach utilized in air-medical and ground-based EMS record reviews and to independently review 50 HEMS medical records and provide their own decisions about AEs. We referred to this approach as the "independent approach." Second, we asked the three MC-EMPs to meet as a group and review 20 medical records and come to a consensus-based decision regarding AEs using the PittAETool. We asked the RN/Medic group to do the same thing. We refer to this approach as the "consensus-based approach." Finally, we asked our three Sr-EMPs to meet as a group and apply the PittAETool to all 70 medical records reviewed by both the MC-EMPs and RN/Medics. We considered the AE decisions of the Sr-EMPs as the reference criterion (aka: gold standard) to which the independent and consensus-based approaches would be compared.

Outcome Measures

Our primary measure of interest was agreement (reliability) of AEs decisions. Specifically, we were interested in determining if the agreement observed between the Sr-EMPs (our reference criterion) and independent reviews was better or worse than the agreement between the Sr-EMPs and the consensus-based reviews performed by MC-EMPs and RN/Medics.

Data Analysis

We used Cohen's Kappa and calculated percent agreement to quantify agreement between AE decisions made by clinician reviewers and the decisions made by our reference criterion Sr-EMPs. We used Fleiss' multi-rater kappa and corresponding 95% confidence to calculate agreement between multiple (>2) clinicians.[18] The percentage agreement, sensitivity, specificity, positive and negative predictive values of AE decisions made by MC-EMPs and RN/Medics were calculated by treating AE decisions as binary (yes/no AE present in a medical record). We used counts and frequencies to describe the number of unique events identified by reviewers, frequency of trigger selection, frequency of proximal cause selection, and frequency of each type of AE. Stata version 10 SE (StataCorp LLP, College Station, TX) and SAS version 9.2 (SAS Institute Inc., Cary, NC) were used to analyze study data.

RESULTS

In the first set of 50 medical records, the reference criterion Sr-EMPs identified one AE with evidence of harm and 25 AEs with potential for harm (Table 1). In the second set of 20 medical records, the Sr-EMPs identified one AE with evidence of harm and 12 AEs with the potential for harm (Table 1). The first AE with harm was described as "a worsening trend in hemodynamics or mental status after administration of labetalol and fentanyl" (Medical

Record #11; Table 2). The trigger for this event was [T5] – "A worsening trend (deterioration) in patient hemodynamic or mental status indicators." The Sr-EMPs determined that the proximal cause for this event involved clinicians at the referring or receiving facility (labeled the Non-HEMS crew in the PittAETool framework). The second AE with harm identified by Sr-EMPs was discovered in Medical Record #109 and described as "Cardizem bradycardic, hypotensive patient by Non-HEMS." The trigger for this event was [T10] – "Suggestive evidence of deviation from standard of care by performing an intervention or administering a medication that appears to be outside of protocol, or failure to perform an intervention or provide a medication that is within the standard of care." The Non-HEMS crew was again designated as the proximal cause for this event.

The Independent Review Approach

There was wide variation in the total number of medical records that each of the three RN/Medic reviewers and MC-EMP reviewers determined to have an AE. There was also wide variation between reviewers in the total number of AEs identified (Table 1).

As noted above, in the first 50 medical records, there was one AE with harm identified by the reference criterion Sr-EMP group (Medical Record #11; Table 2). This specific AE was identified by all MC-EMPs, with two of the three MC-EMPs categorizing the AE as having evidence of harm, and the third MC-EMP categorizing it as having potential for harm. One RN/Medic identified an event in the same medical record (#11), yet did not classify the event as an AE (Table 2).

In the first 50 medical records, two additional AEs with harm were identified in Medical Record #23. Two of the RN/Medics and two of the MC-EMPs identified the first of two administrations of vecuronium (a neuromuscular blocking agent) and classified it as an AE. Both MC-EMPs classified this AE as "potential for harm." One of the RN/Medics classified this AE as harm identified. Two RN/Medics and one MC-EMP identified the second AE in this medical record (#23) – a second dose of vecuronium to the same patient without adequate sedation. One RN/Medic (RN/Medic3) classified this second AE as harm identified, whereas RN/Medic2 and MC-EMP1 classified this event as potential for harm (Table 2). Notably, the Sr-EMP group did not identify these two additional AEs, both of which involved the administration of a vecuronium without adequate sedation (Table 2).

We also observed differences in the total number of AEs with potential for harm identified by the three RN/Medics and MC-EMPs. This number ranged from 19 to 36 for the three RN/Medics, and from 5 to 50 for among the three MC-EMPs.

Agreement on the presence/absence of any AE (this includes those with evidence or potential for harm) ranged from (K=0.11; 95%CI −0.16-0.37) to (K=0.19; 95%CI −0.08-0.46) between the reference criterion Sr-EMPs and three RN/Medics (Table 1). Agreement between the Sr-EMPs and all independent decisions by the three RN-Medics was 'slight' (Fleiss' K=0.16; 95%CI 0.03-0.32; Figure 3). The mean percent agreement between the three RN/Medics and the reference criterion Sr-EMPs was 61.3% (min=58%, max=66%; Table 1). The mean sensitivity and specificity between the three RN/Medics and the reference criterion Sr-EMPs was 46.7% (sensitivity min=30%, max=60%) and 67.7%

(specificity min=60%, max=80%), respectively (Table 1). The positive and negative predictive values for RN/Medic1, RN/Medic2, and RN/Medic3 were comparable (Table 1).

Agreement on the presence/absence of any AE between the reference criterion Sr-EMPs and the three individual MC-EMP clinicians ranged from (K=0.20; 95%CI −0.07-0.48) to (K=0.25; 0.02-0.47; Table 1). The multi-rater agreement between the Sr-EMPs and all three MC-EMPs was 'slight' (Fleiss' K=0.15; 95%CI 0.01-0.31; Figure 3). The mean percent agreement of MC-EMPs with the reference criterion Sr-EMPs was 62.7% (min=58%, max=68%; Table 1). The mean sensitivity and specificity of MC-EMPs with the reference criterion Sr-EMPs was 51.7% (sensitivity min=25%, max=80%) and 70% (specificity min=43%, max=97%), respectively (Table 1). Positive and negative predictive values varied widely between the three MC-EMPs (Table 1).

The three most common triggers selected by the three MC-EMPs and three RN/Medics included: 1) [T10] – deviation from protocol, or failure to perform an intervention, 2) [T1] – missing, incomplete, or unclear documentation; and 3) [T9] – use of medications or fluids: (e.g., blood products, vasopressors or inotrope). The three most common proximal causes for the described events with identifiable harm listed in Table 2 include: 1) HEMS provider (the flight paramedic or flight nurse); 2) Undetermined, and 3) Non-HEMS provider (a nurse or other clinician at the referring or receiving facility).

The Consensus-Based Review Approach

The reference criterion Sr-EMP group identified one AE with evidence of harm in the second set of 20 medical records (Medical Record #109; Table 2). The RN/Medic group identified this same AE in Medical Record #109 and also classified it as having evidence of harm. The MC-EMP group did not detect this AE and did not trigger this medical record for detailed review (Table 2). In this second set of 20 medical records, the RN/Medic group and MC-EMP group identified two additional AEs not identified by the reference criterion Sr-EMPs. The first of two additional AEs not detected by the Sr-EMP group was identified in Medical Record #102 and described as "*Failure to adequately manage hypotension"* by the MC-EMP group and as *"Norepinephrine dose below therapeutic range"* by the RN/Medic group (Table 2). The MC-EMP group classified this AE has having evidence of harm, whereas the RN/Medic group classified it as having the potential for harm. The second AE not detected by the Sr-EMP group was identified in Medical Record #107 and described by both the RN/Medic group and MC-EMP group as a "Worsening trend-blood pressure, Vitals deterioration-hypotension." The RN/Medic group and MC-EMP group classified the severity of this AE differently. The RN/Medic group classified this event as having the potential for harm, whereas the MC-EMP group classified it as harm identified (Table 2). Notably, the Sr-EMP group triggered this same medical record (#107) and described the event similarly to the RN/Medics and MC-EMPs, yet the Sr-EMPs did not find this to be an AE (Table 2).

For the second set of 20 medical records, agreement between reference criterion Sr-EMP group and the RN/Medic group on identification of any AE was fair to moderate (K=0.40, 95% CI 0.01-0.79; Table 1). Agreement between the reference criterion Sr-EMP group and the MC-EMP group was 'fair' (K=0.30, 95% CI −0.12-0.72; Table 1). The multi-rater

agreement between all three groups (RN/Medics, MC-EMPs, and Sr-EMPs) on the presence/absence of any AE at the medical record level was 'fair' (K=0.33; 0.06-0.66; Figure 3). The percent agreement between the reference criterion Sr-EMPs and the two review groups (RN/Medics and MC-EMPs) was comparable; 70% and 65%, respectively (Table 1). The sensitivity between the Sr-EMPs and the two review groups (RN/Medics and MC-EMPs) was comparable; 80% and 70%, respectively (Table 1). Specificity between the Sr-EMPs and the two review groups (RN/Medics and MC-EMPs) was also similar at 60% each (Table 1). Positive and negative predictive values were comparable for the RN/Medic group and MC-EMP groups (Table 1).

The three most common triggers were: 1) [T8] – A failed intervention or procedure; 2) [T1] – missing, incomplete, or unclear documentation; and 3) [T7] – use of advanced interventions (e.g., cardioversion, defibrillation, transcutaneous pacing). The most common causes for the triggered events included HEMS provider (the flight paramedic or flight nurse) and Non-HEMS provider (a nurse or other clinician at the referring or receiving facility).

DISCUSSION

Our data suggest that many HEMS medical records contain events that would be triggered for follow-up by a physician or other clinician reviewers; very few of these "triggers" appear to indicate an AE with evidence of harm to the patient. There is wide variation in assessment across reviewers, a finding observed in other settings.[14,15,19-24] We observed that a consensus-based approach to medical record review utilizing open-discussion between reviewers is more reliable than an approach that relies solely on a clinician identifying AEs independently.

The degree of reliability (agreement) that we found for each approach was comparable to those of prior investigations, most of which focused on inpatient medical record review. For example, Schildmeijer and colleagues reported "slight agreement" on use of triggers and "fair agreement" on total number of AEs between five experienced teams comprised of two nurses and one physician.[15] In the original study of the Institute for Healthcare Improvement Global Trigger Tool (GTT), investigators found a mean agreement on the presence/absence of an AE between physician and non-physician reviewers of K=0.49 after extensive training of clinician raters.[17] In previous research of EMS medical records, four emergency physicians independently reviewed 250 ground-based EMS and achieved "fair" agreement (K=0.24; 95%CI 0.19, 0.29) on the presence/absence of an AE.[14] Schildmeijer and colleagues also reported that out of 42 total AEs identified, only eight of the same AEs were detected by three of five clinician teams.[15] A previous report in which 30 physicians reviewed 319 medical records concluded that complete agreement between independent reviewers is an unrealizable expectation.[19] Forster and colleagues investigated medical record misclassification and concluded that relying on one clinician to review and identify AEs is unreliable.[25] We too detected rather poor agreement between the reference criterion Sr-EMP group and other clinician reviewers regarding the same AE for both the independent approach and group-based approach to medical record review (Table 2).

Our study raises a number of important questions regarding patient safety outcome measurement in the prehospital setting. The first has to do with the actual lack of reliability in AE measurement. How good can we get? What is our benchmark for reliability? The findings across all studies (in both the prehospital and hospital setting) are not very encouraging. The PittAETool was developed by emergency medicine clinicians, is content valid, and yet reliability in its application remains low. We cannot speculate on further uptake and effectiveness absent more data.

The second is a process question. Should AE identification in EMS clinical care be based on medical record review performed by groups of emergency medicine physicians? While our findings suggest "yes," the factors that affect medical oversight and medical record review are complex and diverse across the estimated 19,000 EMS organizations nationwide.[26] Most EMS agencies receive medical oversight from a single physician. Two national emergency medicine societies have published position statements identifying record review as a core/routine function of the physician overseeing the agency.[6,7] A survey of 1,425 local EMS administrative directors found that less than half of designated medical directors routinely reviewed medical records.[27] This finding may be a resource issue, where designated medical directors are not afforded the compensation and resources needed to perform routine audits. The capability to perform routine medical record review may be more limited in rural communities and settings with constraints on EMS resources. One tactic that may be supported by our findings is the sharing of medical oversight between EMS systems, organized in a regionalized arrangement of medical oversight and AE detection. A potentially feasible first step option may be the embedding of the trigger tool component of the PittAETool into electronic medical record systems as a method of filtering only those medical records that may require detailed reviews. It may be possible to refine the trigger tool component to identify AEs with harm and to limit the number of "potential" AEs to those that are clearly deviations from protocol, but have an acceptably low likelihood of a negative outcome (i.e., missing or incomplete documentation).

Research addressing patient and provider safety in EMS is nascent, but improving patient safety in healthcare is a national priority and endorsed by the "National EMS Culture of Safety Strategy" (www.emscultureofsafety.org). Having reliable and valid techniques to measure safety in the EMS setting is a key first step to improvement. The EMS setting poses unique challenges not experienced in the in-hospital setting, where most currently used safety outcome measurement is based. In contrast to larger teams seen in hospital care, EMS care teams are almost exclusively two clinicians (a dyad). Patients are acutely ill or injured, information is often limited, resources to care for complex conditions are scarce, and odds of error or AE are high.[28] Despite the inherent risk exposures in EMS, there is limited research on safety with a poorly defined total burden. Our study exposes a number of challenges associated with AE measurement and represents what we consider formative work in patient safety outcome measurement for the prehospital setting.

LIMITATIONS

We are limited by our convenience sample of clinicians affiliated with one large HEMS organization. Clinicians in other settings may have approached medical record reviews

differently impacting agreement observed in this study. We reduced potential bias associated with selection of our main data source (medical records) by incorporating random sampling, though we cannot entirely eliminate the threat of selection bias. Similarly, we reduced the potential impact of recall bias by the reviewers through limiting the inclusion of medical records to 2008 rather than a time period proximal to present day. We are also limited by human error and reviewer fatigue.

We did not formally collect the exact time a clinician devoted to each review of a medical record, but informally, many clinician reviewers self-reported to the lead investigator an average of 20 minutes per medical record; a time consistent with prior research.[15] While 20 minutes appears to be a norm, automation of the trigger tool component may helpful in preventing the potential for reviewer fatigue in future applications of the PittAETool. In addition, we also did not collect data on disagreements between physicians in the senior group. This group was instructed to produce a consensus-based decision for each medical record and each AE identified. In no instance did this group raise concerns of decisions that could not be resolved.

We used prior research as a model for this study,[9] adapted based on limitations in time and resources. For example, unlike prior studies like that of Brennan and colleagues,[9] we allowed for all three groups of reviewers (RN/Medics, MC-EMPs, and Sr-EMPs) to review all medical records simultaneously. Other studies followed a process whereby medical record administrators of nurses “screened” a medical record, sending on only those records selected at screening considered to possibly have an AE or error present.[9,15] Our decision to have all three groups review all records may have contributed, in some way, to the finding where several AEs with harm were identified by RN/Medics or MC-EMPs and missed by Sr-EMPs. In a stepwise approach like that used in many previous studies,[9,15] the physician may have been more attentive and detected AEs based solely on the impression that some error or AE may be present because at screening, the record was tagged for further review.

Many may not embrace our choice for a reference criterion. The Harvard Medical Practice studies are the cornerstone of AE research, and these investigators considered decisions by the “senior physicians” using open discussion and consensus to be the best approximation of a “gold standard”.[9]

Some may not agree methodologically with our choice to make direct comparisons between two diverse approaches to medical record review when the criterion standard is analogous to one of the approaches tested. Some may feel agreement between groups using consensus-based approaches will be higher than other approaches because the criterion standard also uses consensus. We believe our approach is reasonable and the structure of our comparisons did not predetermine that the group / consensus-based approach would produce higher reliability. We reference research by Schildmeijer and colleagues that shows agreement between groups regarding AEs varies widely despite use of a grouped-based, consensus decision approach.[15] Schildmeijer and colleagues studied agreement in AE decisions between five teams of two nurses and a physician from five different hospitals reviewing 50 medical records.[15] Clinicians used a trigger tool methodology analogous to the tool employed in the current study, and elements of their review approach required open-

discussion and consensus-based decisions. Authors noted that all clinician reviewers were familiar with review methodology and had at least three years of experience using the trigger tool approach. Despite a high level of familiarity with a standardized approach and use of open-discussion, findings revealed wide variation in agreement between the teams (kappa=0.26 to 0.77) regarding the number of AEs detected.[15]

Another potential limitation is reviewer training. Numerous studies show poor agreement between clinicians attempting to identify and adjudicate AEs.[9,17,29-31] Prior research suggests that training can lead to improved agreement.[17] Studies that led to commonly used tools (e.g., the Global Trigger Tool) cite 2-3 hours of training and familiarity with an AE identification process as adequate.[9,17] Other studies show agreement between reviewers may be poor despite years of familiarity with a specific method.[15] We addressed the issue of training by having nine of the 10 clinicians involved in the development of the PittAETool take part in this study. We feel this strengthened our study because all clinicians were intimately familiar with the tool and intended use. We estimate a 20+ hour total period of training with tool development and practice sessions with actual medical records. Additional research with additional training may result in improved agreement.

Our data highlight the need to further the development of clear and measureable definitions of an adverse event and creation of a valid reference criterion. While the AE definition adopted for this study has face validity and was developed by emergency physicians,[14] inclusion of “potential” harm introduces potential ambiguity and threatens AE identification and reliability. A more clear and measurable “case definition” of an AE is needed, as is data that supports its application and utility for AE detection, no matter the approach or source of data (e.g., medical record review, direct observation, or other).

CONCLUSIONS

We identified a higher level of agreement in AE decisions utilizing a consensus/grouped-based format for review as compared to an independent review format. Despite this observation, reliable and efficient AE detection remain a challenge.

ACKNOWLEDGEMENTS, FUNDING

Work on this study was supported by Dr. Patterson’s KL2 training grant from the National Center Research Resources and the National Institutes of Health (NIH/NCATS Grant no: KL2 TR000146 (Dr. Reis PI). The conclusions, views, opinions, and content in this paper should not be interpreted as reflecting the opinions of the NIH.

REFERENCES

1. Classen DC, Resar R, Griffin F, et al. ‘Global trigger tool’ shows that adverse events in hospitals may be ten times greater than previously measured. Health Aff (Millwood). 2011; 30(4):581–589. [PubMed: 21471476]
2. Brennan TA, Leape LL, Laird NM, et al. Incidence of adverse events and negligence in hospitalized patients. Results of the Harvard Medical Practice Study I. N Engl J Med. Feb 7; 1991 324(6):370–376. [PubMed: 1987460]
3. Michel, P. Strengths and weaknesses of available methods for assessing the nature and scale of harm caused by the health system: literature review. World Health Organization; Geneva, Switzerland: 2003.

4. Baker GR, Norton PG, Flintoft V, et al. The Canadian Adverse Events Study: the incidence of adverse events among hospital patients in Canada. CMAJ. May 25; 2004 170(11):1678–1686. [PubMed: 15159366]
5. Swor, RA.; Brennan, JA.; Krohmer, JR. Principles of EMS Systems. Jones and Bartlett Publishers; Sudbury, MA: 2006. Medical Oversight and Accountability; p. 64-73.
6. Alonso-Serra H, Blanton D, O'Connor RE. Physician medical direction in EMS. National Association of EMS Physicians. Prehosp Emerg Care. 1998; 2(2):153–157. [PubMed: 9709337]
7. Polsky S, Krohmer J, Maningas P, McDowell R, Benson N, Pons P. Guidelines for medical direction of prehospital EMS. American College of Emergency Physicians. Ann Emerg Med. 1993; 22(4):742–744. [PubMed: 8457108]
8. Patterson PD, Lave JR, Martin-Gill C, et al. Measuring adverse events in helicopter emergency medical services: Establishing content validity. Prehosp Emerg Care. 2013
9. Brennan TA, Localio RJ, Laird NL. Reliability and validity of judgments concerning adverse events suffered by hospitalized patients. Med Care. 1989; 27(12):1148–1158. [PubMed: 2593729]
10. Gratton MC, Ellison SR, Hunt J, Ma OJ. Prospective determination of medical necessity for ambulance transport by paramedics. Prehosp Emerg Care. 2003; 7(4):466–469. [PubMed: 14582100]
11. Pointer JE, Levitt MA, Young JC, Promes SB, Messana BJ, Ader ME. Can paramedics using guidelines accurately triage patients? Ann Emerg Med. 2001; 38(3):268–277. [PubMed: 11524646]
12. Sasser SM, Brokaw M, Blackwell TH. [CONFERENCE ABSTRACT] Paramedic vs emergency physician decisions regarding the need for emergency department evaluation. Acad Emerg Med. 1998; 5(5):391.
13. Patterson PD. Use of ED diagnosis to determine medical necessity of EMS transports. Prehosp Emerg Care. 2006; 10(4):6.
14. Patterson PD, Weaver MD, Abebe K, et al. Identification of adverse events in ground transport emergency medical services. Am J Med Qual. 2012; 27(2):139–146. [PubMed: 21816967]
15. Schildmeijer K, Nilsson L, Arestedt K, Perk J. Assessment of adverse events in medical care: lack of consistency between experienced teams using the global trigger tool. BMJ Qual Saf. 2012; 21(4):307–314.
16. Chang A, Schyve PM, Croteau RJ, O'Leary DS, Loeb JM. The JCAHO patient safety event taxonomy: a standardized terminology and classification schema for near misses and adverse events. Int J Qual Health Care. 2005; 17(2):95–105. [PubMed: 15723817]
17. Classen DC, Lloyd RC, Provost L, Griffin FA, Resar R. Development and evaluation of the Institute for Healthcare Improvement Global Trigger Tool. J Patient Saf. 2008; 4(3):169–177.
18. Fleiss, JL. Statistical Methods for Rates and Proportions. Wiley; New York, NY: 1981.
19. Forster AJ, Taljaard M, Bennett C, van Walraven C. Reliability of the peer-review process for adverse event rating. PLoS One. 2012; 7(7):e41239. [PubMed: 22844445]
20. Pines J, Uscher Pines L, Hall A, Hunter J, Srinivasan R, Ghaemmaghami C. The interrater variation of ED abdominal findings in patients with acute abdominal pain. American Journal of Emergency Medicine. 2005; 23(4):483–487. [PubMed: 16032616]
21. Erling BF, Perron AD, Brady WJ. Disagreement in the interpretation of electrocardiographic ST segment elevation: a source of error for emergency physicians? American Journal of Emergency Medicine. 2004; 22(2):65–70. [PubMed: 15011215]
22. Foldes S, Fischer L, Kaminsky K. What Is an Emergency? The Judgments of Two Physicians. Ann Emerg Med. 1994; 23(4):833–840. [PubMed: 8161055]
23. Localio AR, Weaver SL, Landis JR, et al. Identifying adverse events caused by medical care: degree of physician agreement in a retrospective chart review. Ann Intern Med. Sep 15; 1996 125(6):457–464. [PubMed: 8779457]
24. Zegers M, de Bruijne MC, Wagner C, Groenewegen PP, van der Wal G, de Vet HC. The inter-rater agreement of retrospective assessments of adverse events does not improve with two reviewers per patient record. J Clin Epidemiol. 2010; 63(1):94–102. [PubMed: 19473812]

25. Forster AJ, O'Rourke K, Shojania KG, van Walraven C. Combining ratings from multiple physician reviewers helped to overcome the uncertainty associated with adverse even classification. J Clin Epidemiol. 2007; 60(9):892–901. [PubMed: 17689805]

26. Mears, G.; Armstrong, B.; Fernandez, AR., et al. National EMS Assessment. 2011. http://www.ems.gov/pdf/2011/National_EMS_Assessment_Final_Draft_12202011.pdf

27. Slifkin RT, Freeman VA, Patterson PD. Designated medical directors for emergency medical services: recruitment and roles. J Rural Health. 2009; 25(4):392–398. [PubMed: 19780921]

28. Brice JH, Studnek JR, Bigham BL, et al. EMS provider and patient safety during response and transport: Proceedings of an ambulance safety conference. Prehosp Emerg Care. 2012; 16(1):3–19. [PubMed: 22023217]

29. Thomas EJ, Lipsitz SR, Studdert DM, Brennan TA. The reliability of medical record review for estimating adverse event rates. Ann Intern Med. Jun 4; 2002 136(11):812–816. [PubMed: 12044129]

30. Thomas EJ, Brennan TA. Incidence and types of preventable adverse events in elderly patients: population based review of medical records. BMJ. Mar 18; 2000 320(7237):741–744. [PubMed: 10720355]

31. Vincent C, Neale G, Woloshynowych M. Adverse events in British hospitals: preliminary retrospective record review. BMJ. Mar 3; 2001 322(7285):517–519. [PubMed: 11230064]

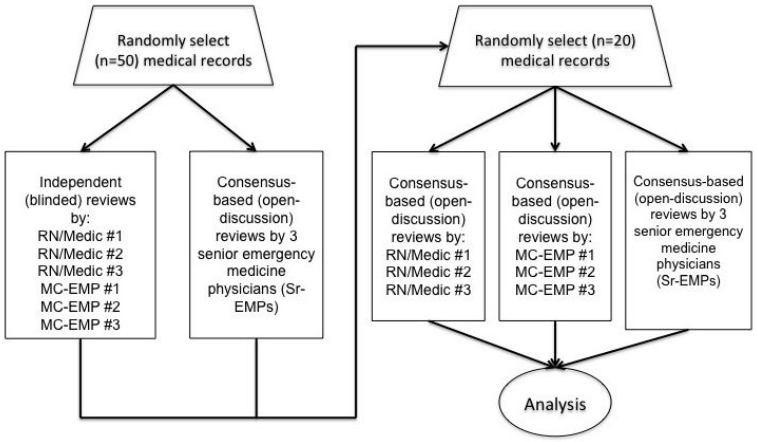

Figure 1.
Illustration of medical record review process

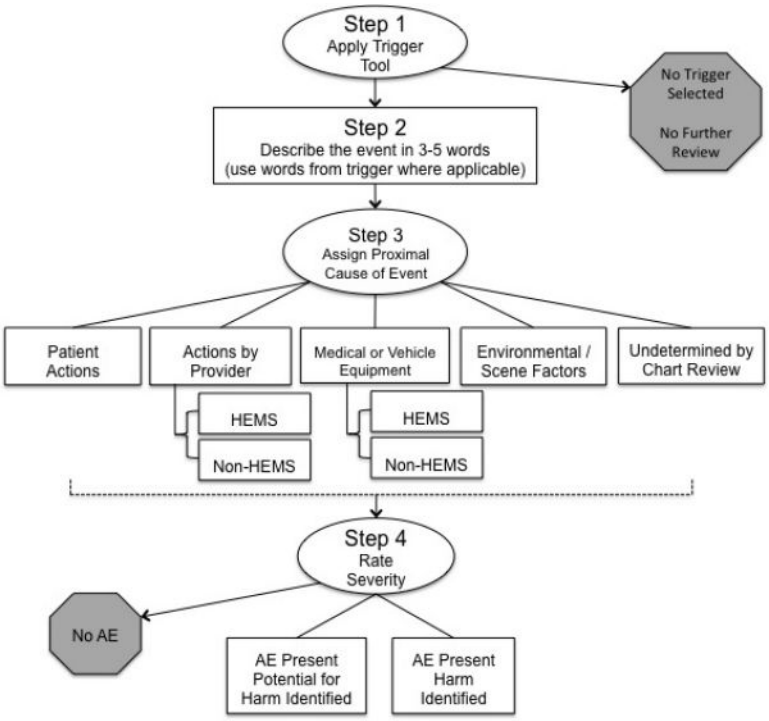

Figure 2.
The PittAETool Framework for detecting and classifying Adverse Events

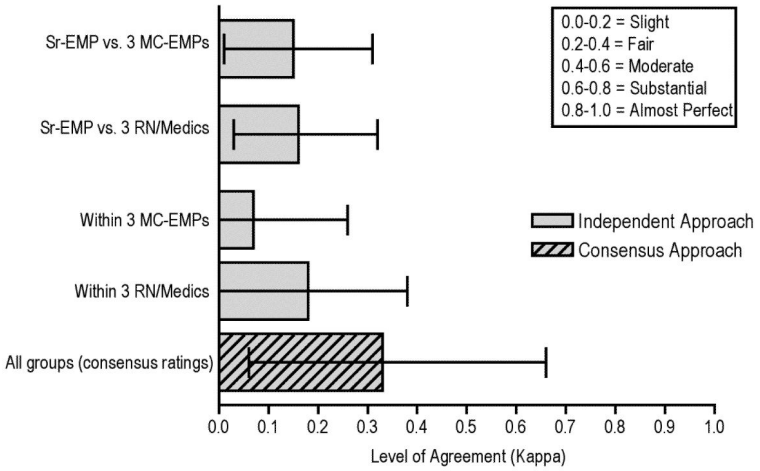

Figure 3.
Agreement on presence/absence of any Adverse Event in medical record

Table 1

Number of AEs identified by clinician reviewers using two approaches to AE classification

Approach Used for Medical Record Review	RN/Medic1	RN/Medic2	RN/Medic3	MC-EMP1	MC-EMP2	MC-EMP3	Reference criterion Sr-EMPs
Approach 1: Independent Reviews (n=50 charts)							
Medical Records Triggered	21 (42%)	34 (68%)	33 (66%)	28 (56%)	38 (76%)	31 (62%)	32 (64%)
Total Number of Triggered Events	33	65	58	63	69	47	60
Medical Records with AE	12 (24%)	21 (42%)	24 (48%)	18 (36%)	33 (66%)	5 (10%)	20 (40%)
Total Number of AEs	19	30	38	31	50	5	26
-AEs Potential for Harm	19	30	36	30	50	5	25
-AEs Harm Identified	0	0	2	1	0	0	1
Kappa (95% CI)	0.11 (−0.16-0.37)	0.13 (−0.14-0.41)	0.19 (−0.08-0.46)	0.20 (−0.07-0.48)	0.21 (−0.02-0.44)	0.25 (0.02-0.47)	Ref
Percent Agreement	66%	58%	60%	62%	58%	68%	Ref
Sensitivity | Specificity	30% | 80%	50% | 63%	60% | 60%	50% | 70%	80% | 43%	25% | 97%	Ref
PPV | NPV	50% | 63%	48% | 66%	50% | 69%	53% | 68%	48% | 76%	83% | 66%	Ref
Approach 2: Grouped Reviews (n=20 charts)							
Medical Records Triggered		17 (85%)			19 (95%)		17 (85%)
Total Number of Triggered Events		50			42		33
Medical Records with AE		12 (60%)			11 (55%)		10 (50%)
Total Number of AEs		36			18		13
-AEs Potential for Harm		35			16		12
-AEs Harm Identified		1			2		1
Kappa (95% CI)		0.40 (0.01-0.79)			0.30 (−0.12-0.72)		Ref
Percent Agreement		70%			65%		Ref
Sensitivity | Specificity		80% | 60%			70% | 60%		Ref
PPV | NPV		67% | 75%			64% | 66%		Ref

Table 2

Agreement on adverse events designated by any clinician or group to have resulted in identifiable harm

Medical Record #	Clinician Reviewer	Triggers Selected	Event Description	Proximal Cause	AE Y/N	Severity
11	RN/Medic1	---	---	---	---	---
11	RN/Medic2	T5	Worsening level of consciousness	HEMS provider	No	---
11	RN/Medic3	---	---	---	---	---
11	MC-EMP1	T5	Decrease in mental status after fentanyl dose (>1 mcg/kg)	Non-HEMS	Yes	Harm identified
11	MC-EMP2	T10	Labetalol dose immediately precipitated bradycardia	HEMS provider	Yes	Potential for harm
11	MC-EMP3	T5	ED gave fentanyl and labetalol, patient decompensated	Non-HEMS	Yes	Potential for harm
11	Sr-EMPs	T5	A worsening trend in hemodynamics or mental status after administration of labetalol and fentanyl by ED staff	Non-HEMS	Yes	Harm identified
			Medical Record #23 – Event #1			
23	RN/Medic1	---	---	---	---	---
23	RN/Medic2	T10	Deviation from standard – no indication for vecuronium at 1932	HEMS provider	Yes	Potential for harm
23	RN/Medic3	T10	Administered paralytic without sedation	HEMS provider	Yes	Harm identified
23	MC-EMP1	T10	Vecuronium administered with inadequate sedation	Non-HEMS	Yes	Potential for harm
23	MC-EMP2	T11	Vecuronium administered for sedation, readministered without consult	HEMS provider	Yes	Potential for harm
23	MC-EMP3	---	---	---	---	---
23	Sr-EMPs	---	---	---	---	---
			Medical Record #23 – Event #2			
23	RN/Medic1	---	---	---	---	---
23	RN/Medic2	T10	Deviation from standard – Vecuronium at 2030 for facial twitching	HEMS provider	Yes	Potential for harm
23	RN/Medic3	T10	Administered paralytic without sedation	HEMS provider	Yes	Harm identified
23	MC-EMP1	T10	Vecuronium administered with inadequate sedation	HEMS provider	Yes	Potential for harm
23	MC-EMP2	---	---	---	---	---
23	MC-EMP3	---	---	---	---	---
23	Sr-EMPs	---	---	---	---	---
102	RN/Medics	T10	Levophed dose below therapeutic range	Non-HEMS	Yes	Potential for harm
102	MC-EMPs	T8	Failure to adequately manage hypotension	HEMS Provider	Yes	Harm identified
102	Sr-EMPs	---	---	---	---	---

Medical Record #	Clinician Reviewer	Triggers Selected	Event Description	Proximal Cause	AE Y/N	Severity
107	RN/Medics	T5	Worsening trend – blood pressure	HEMS provider	Yes	Potential for harm
107	MC-EMPs	T5	Vitals deterioration – hypotension	Undetermined	Yes	Harm identified
107	Sr-EMPs	T5	Declining vital signs	HEMS provider	No	---
109	RN/Medics	T10	Cardizem infusion continued on bradycardic/hypotensive pt.	Non-HEMS	Yes	Harm identified
109	MC-EMPs	---	---	---	---	---
109	Sr-EMPs	T10	Cardizem bradycardic, hypotensive patient by non-hems	Non-HEMS	Yes	Harm identified