HHS Public Access

Author manuscript

Med Care. Author manuscript; available in PMC 2018 April 01.

Published in final edited form as:

Med Care. 2017 April ; 55(4): 436–441. doi:10.1097/MLR.0000000000000679.

Measuring harm in healthcare: optimizing adverse event review

Kathleen E. Walsh, MD, MSc[1,2], **Polina Harik, PhD**[3], **Kathleen M. Mazor, EdD**[4], **Deborah Perfetto, PharmD**[5], **Milena Anatchkova, PhD**[6], **Colleen Biggins, BA**[4], **Joann Wagner, MSW**[4], **Pamela J. Schoettker, MS**[1], **Cassandra Firneno, BA**[4], **Robert Klugman, MD**[7], and **Jennifer Tjia, MD, MSCE**[6]

[1]James M. Anderson Center, Cincinnati Children's Hospital, Cincinnati, Ohio [2]Department of Pediatrics, Cincinnati Children's Hospital, Cincinnati, Ohio [3]National Board of Medical Examiners, Philadelphia, Pennsylvania [4]Meyers Primary Care Institute, Worcester, Massachusetts [5]Agency for Healthcare Quality and Research, Rockville, Maryland [6]Department of Quantitative Health Sciences, University of Massachusetts Medical School, Worcester, Massachusetts [7]Kindred Healthcare, Wellesley, Massachusetts

Abstract

Objective—The objective of this study was to identify modifiable factors that improve the reliability of ratings of severity of healthcare-associated harm in clinical practice improvement and research.

Methods—A diverse group of clinicians rated eight types of adverse events: blood product, device or medical/surgical supply, fall, healthcare-associated infection, medication, perinatal, pressure ulcer, surgery. We used a Generalizability theory framework to estimate the impact of number of raters, rater experience, and rater provider type on reliability.

Results—Pharmacists were slightly more precise and consistent in their ratings than either physicians or nurses. For example, to achieve high reliability of 0.83, 3 physicians could be replaced by 2 pharmacists without loss in precision of measurement. If only 1 rater was available for rating, approximately 5% of the reviews for severe harm would have been incorrectly categorized. Reliability was greatly improved with 2 reviewers.

Conclusion—We identified factors that influence the reliability of clinician reviews of healthcare-associated harm. Our novel use of generalizability analyses improved our understanding of how differences affect reliability. This approach was useful in optimizing resource utilization when selecting raters to assess harm and may have similar applications in other settings in healthcare.

Corresponding Author: Kathleen Walsh, MD, MSc 3333 Burnet Avenue, Cincinnati, Ohio, 45229. Kathleen.walsh@cchmc.org. phone: 513-803-2187.

Conflicts of Interest: The authors have no conflicts of interest to declare.

BACKGROUND

Increasingly, healthcare systems and researchers have focused on the elimination of patient harm due to healthcare. An essential part of this effort involves identifying and measuring the incidence and severity of healthcare-associated harm in a valid and reliable way.[1] In hospitals, this process involves identifying the occurrence of severe adverse events, determining their causes, and making system-wide changes.[1] This is a time-, labor-, and cost-intensive process because it relies on the labor of highly trained clinicians. Distinguishing between more and less serious adverse events is critical because less serious events, while included in hospital error rates, are typically not considered after severity is assigned. (Appendix 1) For example, there are different implications for internal review and risk mitigation if a medication infusion error resulting in shortness of breath is determined to be “life-threatening harm” or “not severe”. (Appendix 2) The Institute of Medicine has stated, “if root-cause analyses are not focused on a critical subset, then 1. useless analyses will be carried out because there is no time to do them properly, and 2. effort will be devoted to performing root-cause analyses at the expense of testing and implementing real system changes that can reduce injury rates.”[1]

Unfortunately, evidence to date suggests that clinician ratings of severity for adverse events are highly variable, with Cohen’s Kappa coefficients ranging from 0.4 to 0.76.[2–11] In spite of the importance of adverse event ratings, there has been little information on how to optimize the reliability of ratings. For example, it is unclear whether more raters or more experienced raters improves reliability. Few prior studies address this issue. One was limited in focus to medication errors.[12] Another did not account for multiple sources of reliability error such as type of case (i.e. surgical or infection) or rater experience because this is not possible with the traditional inter-rater reliability (IRR) approach.[13] An alternative approach is Generalizability theory (G-theory), a psychometric framework that provides a means of identifying multiple varying sources of error in a rating and that can be used to predict improvements in reliability under different measurement conditions (e.g., if more raters are used).[14] One of the major advantages of the generalizability framework is that results can be used to predict the reliability of ratings that would be obtained under alternative measurement conditions (e.g., to estimate the number and type of raters needed to achieve a given level of rating precision or reliability).

Because hospital safety officers and researchers make poorly informed decisions about the appropriate number, experience, and provider type of clinician raters best suited to review healthcare-associated harm events, we conducted this study to identify modifiable factors to improve the reliability of harm severity ratings in clinical practice improvement and research. To achieve this objective, we used G-theory to estimate the impact of number of raters, rater experience, and rater provider type on reliability. We seek to inform hospitals and researchers making decisions around resource utilization to optimize rater selection for adverse event reviews. This study also makes a methodological contribution to the literature by illustrating the value of G-theory to health services research.

METHODS

The protocol and interview guide were reviewed and approved by the University of Massachusetts Medical School's Institutional Review Board.

Overview

For this prospective study, a diverse group of clinician raters rated eight types of adverse events. The types of adverse events evaluated were selected to reflect the Agency for Healthcare Research and Quality (AHRQ) common formats: blood product, device or medical/surgical supply, fall, healthcare-associated infection, medication, perinatal, pressure ulcer, surgery.[15] In practice, hospitals collect such events and assign 1 or 2 raters to assess the level of using one of several severity rating scales (e.g., AHRQ Harm Scale[16] or the National Coordinating Council for Medication Error Reporting and Prevention Index[17]). For the current study, we use Generalizability theory to analyze the reliability of severity ratings conducted by multiple raters. Generalizability theory provides a framework for extending the results obtained under specific conditions (e.g., with a set number of raters) to other conditions (e.g. with a different number of raters).[14] Cohen's kappa was also calculated to allow comparisons with existing literature.

Study Setting and Raters

We conducted this study at a large academic medical center in the northeastern United States. From this medical center, we recruited nine clinician raters to represent clinical specialties and levels of experience typically available in the US healthcare system for harm assessment. These included 3 nurses, 3 pharmacists, and 3 physicians, each with either a low, moderate, or high level of experience. We defined low experience as no prior experience with adverse event rating, moderate as some prior experience, and high as prior experience with more than one project and/or ongoing responsibility in patient safety management.

Selection of Adverse Event Cases

We selected sample adverse events from the same institution's adverse event reporting system, from the AHRQ web Morbidity & Mortality reports, and from published case reports. We adapted cases so that each type had a representative variation in each level of harm (e.g., the same type of case would have a variant with 'no harm' and 'severe harm'). There were 50 adverse events in each type of common format adverse event, for a total of 400 cases.

Rater Assessment of Harm Severity

Prior to conducting ratings, we trained clinician raters using a presentation containing information about medical errors, sample adverse event reports, and guidelines for assigning harm severity, consistent with the type of training performed in research and in healthcare.[2,4–6,8–12] Raters received written instructions for using the revised AHRQ Harm Scale and exemplar cases for each severity level of harm within each adverse event type. The revised AHRQ Harm Scale has 6 options to assign severity of harm (death, severe, moderate, mild, no harm, unknown).[16] Prior to beginning actual study ratings, all raters completed

practice rating sessions with sample adverse event reports during which they compared and discussed their ratings. Raters performed their actual adverse event reviews independently online, entering their ratings into a secure REDCap database.[18]

Study Design

We conducted two generalizability studies with identical experimental designs. Study 1 examined differences in reliability between rater specialties (physicians, nurses and pharmacists), (Table 1 shows study design visually) and study 2 compared reliability for raters with varying level of experience (high, medium and low).

Generalizability analysis

G-theory is an extensive conceptual framework for disentangling multiple sources of measurement error through an analysis of variance. As applied here, generalizability analysis is used to examine how well the results of this study generalize to other data. Specifically, we sought to examine how the reliability of the harm rating would change if: 1. the number of raters increased or decreased, 2. only one type of rater was used, and 3. raters had a specific level of experience. G-theory provides a framework that allows us to answer these types of questions. For an in-depth discussion of the G-theory framework, see Brennan.[14]

Within the G-theory framework, each event has an inherent or 'true' severity of harm that is unknown. What is known are the judgments made by individual raters. Each individual rating is conceptualized as a sum of *true severity* and *measurement error*. Measurement error can have various sources of error; in the present study, sources of error could be systematic, such as differences in rater stringency, and the remaining random error, or residual. Residual error is due to factors that are usually not under our control, such as rater fatigue, environmental disruptions, etc. Systematic measurement errors, such as differences in rater stringency, can be controlled in different ways, such as more training, measurement adjustment, or more raters.

True variance, $\sigma_t^2 = \sigma_{ev}^2$, refers to variability due to the actual 'true' differences in the severity of harm among the events, and error variance, σ_e^2, which is the portion of variance in the ratings that might be due to differences in more stringent or lenient raters or unknown random factors. 'True' variance and 'error' variance in G-theory are analogous to 'signal' and 'noise,' the statistical terms familiar in health services research. 'Error' and 'noise' refer to undesirable variability, while 'true' and 'signal' refer to the actual differences in harm.

The underlying statistical methodology of the G-framework is analyses of variance. In Study 1, all events were rated by three raters within each provider type (Table 1). In Study 2, the same events were analyzed by all raters, but raters were grouped by levels of experience. These designs can be symbolically represented as EV × R, indicating that events (EV) are crossed with raters (R). Variability in ratings for the EV × R design can be partitioned, or decomposed, into the following variance components:

$$\sigma^2{}_{\mathrm{Xevr}}=\sigma^2{}_{\mathrm{ev}}+\sigma^2{}_{\mathrm{r}}+\sigma^2{}_{\mathrm{evr}} \quad (1)$$

where $\sigma^2{}_{\mathrm{Xevr}}$ is the total observed variance of ratings X for event ev given by rater r, $\sigma^2{}_{\mathrm{ev}}$ is the true variability in harm among the events, $\sigma^2{}_{\mathrm{r}}$ is variance due to differences in rater stringencies, and $\sigma^2{}_{\mathrm{evr}}$ is the remaining residual variance due to unexplained interaction effects between events and raters. In G-theory, all variance components except for the true differences in event harm $\sigma^2{}_{\mathrm{ev}}$ are undesirable and are considered error variance, $\sigma^2{}_{\mathrm{e}}$. Minimizing the part of the error variance that is due to controllable and systematic factors, such as differences in rater stringencies, leads to an increase in reliability of an instrument, as will become apparent in the next few paragraphs.

The error variance $\sigma^2{}_{\mathrm{e}}$ in G-theory framework can be expressed as a sum of rater variance and a residual:

$$\sigma^2{}_{\mathrm{e}}=\sigma^2{}_{\mathrm{r}}+\sigma^2{}_{\mathrm{evr}} \quad (2)$$

and is estimated for individual ratings (as opposed to ratings averaged across all raters for an event). This stage of analysis is referred to as Generalizability analysis (G study). Once variance components are estimated, additional analyses are conducted to estimate the impact of changing the measurement conditions (e.g. fewer or more raters). Using the G-theory framework, this stage of analysis is referred to as a decision study (D study).

To examine how different number of raters would affect reliability, rater variance $\sigma^2{}_{\mathrm{r}}$ and residual error $\sigma^2{}_{\mathrm{evr}}$ are computed by dividing the individual variance components by the desired number of raters n_r:

$$\begin{aligned}\sigma^2_{r(forn_r)}&=\sigma^2_r/n_r\\ \sigma^2_{evr(forn_r)}&=\sigma^2_{evr}/n_r\end{aligned} \quad (3)$$

These derived variance components are then used in a D study to estimate various reliability indexes for n_r number of raters.

One of the reliability-type coefficients in G-theory that is suitable to our study is called ‘index of dependability,’ which is denoted as Phi – Φ. The dependability coefficient Phi is often interpreted as the chance-corrected proportion of maximum possible agreement, given the data.[18] Computationally, Φ is a ratio of the ‘true’ variance to itself plus error variance or, in more common terms, *signal/(signal+noise).* For our study, the index of dependability is the ratio of the ‘true’ variance in harm among the events over the total variance (‘true’ plus the error variance):

$$\Phi = \frac{\sigma^2_{ev}}{\sigma^2_{ev} + \sigma^2_{e}} \quad (4)$$

where error variance σ^2_e is defined in equation 1. Combining equations 1 and 3 we arrive at the formula for dependability coefficient Phi for any number of raters, $n_{r:}$

$$\Phi = \frac{\sigma^2_{ev}}{\sigma^2_{ev} + \frac{\sigma^2_{r}}{n_r} + \frac{\sigma^2_{evr}}{n_r}} \quad (5)$$

We used mGENOVA for all analyses.[19]

RESULTS

The Effect of Rater Provider type

Variance components from the generalizability analysis for physicians, nurses and pharmacists are shown in Table 2. Total variance was broken into variance components attributable to the different sources of variability – differences in 'true' severity of harm for events, differences in rater stringency, and differences due to the interaction between raters and events (residual error variance). For example, for physicians, most of the variance was due to the 'true' differences in event severity (0.502); there was less of an effect in the interaction between event and rater (0.223) and least effect of the raters themselves (0.087). This same pattern held true for nurses and pharmacists.

Event variance represents the 'true' (if it was possible for us to know the absolute truth) differences in event severity of harm. The variability in event variance estimates across specialties was similar. Interaction effects between raters and events were also very similar across all groups.

All of the rater-effect variance components were very low, only around 10%, of the total variance. For example, for physicians it was 11% (0.087/(0.502+0.087+0.223)). The combined error variance (rater and rater by event) was somewhat low (38% for physicians) compared to the total variance. As evident from equations 4 and 5, lower error variance results in higher reliability. Based on the estimated variance components for three specialties with three raters in each, the physicians, nurses, and pharmacists had Phi coefficients of 0.76, 0.80, and 0.83, respectively (assuming two raters within each provider type, as used in this study), indicating a high level of agreement in ratings. Pharmacists were the most accurate and consistent of all specialties. If only one rater was available for rating, the Phi coefficients would have been 0.63, 0.66, and 0.70, for physicians, nurses, and pharmacists respectively. (Figure 1)

As Figure 1 illustrates, pharmacists were slightly more precise and consistent in their ratings than either physicians or nurses. Although the differences in reliability seem very small, they are practically important for selecting raters. For instance, to achieve a reliability of about

0.91, 6 physicians could be replaced by 5 nurses, or 4 pharmacists without loss in precision of measurement. Kappas (calculated for each pair of raters and averaged) were 0.37, 0.45, and 0.52 for physicians, nurses and pharmacists respectively.

The Effect of Rater Experience

Table 3 shows the variance component estimates from the generalizability analysis for low, medium, and high experienced raters. Within each rating effect (event, rater, and event by rater), the values represent the estimates of variance components within each level of experience.

Based on the estimated variance components for the three specialties with three raters in each, the low, medium, and high experience raters Phi coefficients were 0.85, 0.85, and 0.88, respectively. If only one rater was available for rating, the Phi coefficients would have been 0.65, 0.66, and 0.72, for low, medium and high levels of experience respectively. These results suggest that high experience raters are more precise and consistent in their ratings than either medium or low experience raters. Although the differences in reliability seem very small, they are practically important in selecting raters for the task, as 2 high experienced raters would need to be replaced by 3 medium or low experience raters to maintain the same measurement precision. However, 2 medium or low experienced raters would be more precise than 1 highly experienced rater.

Kappas (calculated for each pair of raters and averaged) were 0.45, 0.42 and 0.45 for low, medium and high experience raters respectively.

Using Conditional Standard Errors of Measurement to Assess Rating Precision by Degree of Harm

A practical application of the G-theory framework is the use of conditional standard errors of measurement (CSEM). Standard errors of measurement indicate the accuracy (or lack of) with which a mean score of a distribution is estimated. *Conditional* SEMs indicate the precision with which *each point* of the scale is estimated. Figure 2 displays CSEMs for each point on the Harm Scale. Each line in Figure 2 summarizes conditional standard errors for an average rating based on a different number of raters, from 1 to 9. As this figure illustrates, when the observed severity is ‘death (1)’, there is no disagreement among the raters and the CSEM is practically zero, regardless of the number of raters. When the observed severity is ‘severe harm (2)’, the conditional SEM based on 9 ratings is about 0.02 and it is 0.03 for ‘mild harm (4)’. However, as the number of raters goes down, the CSEMs increase, indicating lower accuracy. For example, for ‘severe harm’ the error increases nearly ten-fold from 0.02 to 0.18 when the number of raters is decreased from nine to one. Similarly, with one, two, or three raters, error is higher for ‘moderate harm (3)’, ‘mild harm (4)’, and ‘no harm (5)’.

Conditional SEMs can be interpreted in terms of confidence intervals. For example, for a rating of ‘severe harm (2)’ with 9 raters, the 95% confidence interval is between two standard errors below to two standard errors above the mean (2 +/− 0.02*2). It follows that given an unlimited number of events and multiple samples of 9 raters, 95% of the time the average rating will fall between 1.96 and 2.04. This means that given the conditional SEM of

0.02, there is practically no chance that any resulting average rating (across 9 raters) would not be averaged to a 2 because 1.499 and 2.5 (averaged ratings that would be rounded to 1 or 3) are more than 20 SEMs away from 2. However, when the number of raters is reduced to one, for a rating of '2', 98% of the time the rating will fall between 1.5 and 2.49 (2 +/ − 0.18*2.7) allowing a 2% chance that a true 'severe harm (2)' event will not be rated as a '2'. Examining the CSEMs for 2–8 raters, we can show that in order to reduce misclassification of severe event harm to zero, there is no need to increase the number of raters to 9. Increasing the number of raters to 2, which is more feasible in the health care setting, will achieve the same result of reducing misclassification of severe event harm to zero. The highest CSEM on the harm scale for a rating averaged between 2 raters is 0.15 at the mild harm category, which places average ratings of 3.5 and 4.499 more than 3 CSEMs away from '4', indicating that there is a lower than 1% chance of a rating averaged between 2 raters to be anything other than '4' when the true event harm is 'mild'. Two raters are even more precise than this for the other remaining categories: death, severe harm, moderate harm, and no harm.

DISCUSSION

When a patient is harmed by healthcare, an accurate assessment of severity of harm is critical: allocation of resources may vary from none to a full review of root causes and action steps to prevent future harm depending on the harm rating. We demonstrate that increasing the number of raters results in substantial increases in reliability across the specialties and levels of experience examined. Of the three specialties included in this study, pharmacists were most consistent in their ratings. Not surprisingly, more experienced raters provided more consistent ratings.

When moving from one rater to two, the reliability of ratings improved for all types of raters. This increase was most dramatic for less experienced raters (low and medium experience), where the reliability (phi) if only one rater is used was only 0.65; this increases to 0.8 if two raters are used. This differs from another study that reported that reliability did not improve with two versus one reviewers per record, but that study used a traditional Kappa-based approach.[13]

We found that pharmacists were slightly more precise in their ratings, with higher Phi coefficients. Few other studies compare adverse event severity ratings by healthcare professionals. Williams and Ashcroft reported that pharmacy technicians and nurses were more likely to assign higher severity ratings for medication errors than pharmacists or physicians, but did not examine reliability across professions.[20] Another study found that pharmacists had higher agreement for harm severity ratings than physicians when evaluating preventable medication errors ($\kappa_{\text{pharmacists}} = 0.49$ vs $\kappa_{\text{physicians}} = 0.36$) and overall medication errors ($\kappa_{\text{pharmacists}} = 0.34$ vs $\kappa_{\text{physicians}} = 0.25$).[21] Our study extends these findings to document reliability comparisons that also include nurses and non-medication related harms.

Our study indicates that at least two clinicians should be used to rate each harm event to ensure adequate precision of the ratings and that the choice of clinician profession

performing the ratings matters. Cultures of medical, nursing and pharmacy trainees differ in how they believe clinical work should be organized and whether clinical work is the responsibility of individuals or should be systemized; this may explain these differences.[22–24] In the future, information technology may have a role in ensuring the reliability of adverse event ratings.

G-theory provides a framework for examining score precision at specific points of the rating scale because it allows computation of conditional standard errors of measurement. Our results suggest that for death CSEMs are very small, indicating a very low likelihood of misclassification. However, the CSEMs differ across the levels of harm, and reveal that the likelihood of misclassification is higher for all other levels of harm. Increasing the number of raters reduces the CSEM in these instances. For example, using two reviewers instead of one decreases the chance of incorrectly rating severe harm as less or more harm from 2% to essentially zero.

To date, G-theory has been underutilized in health services research. We are aware of only one prior study using G-theory to assess the impact of the number and provider type of raters on the reliability of harm severity review, but that study focused only on medication errors.[12] In the present study, G-theory informed our understanding of different factors that contribute to unreliability of severity ratings. We were able to estimate the effects of both number and experience level of raters and to determine the best design for future adverse event ratings. G-theory may also be useful in the study of other situations where consistency and accuracy across ratings are important but poor. For example, G-theory may be valuable in patient satisfaction scoring, to indicate the number of patients and items needed to achieve reliable ratings. G-theory could also be used to evaluate whether an actual adverse event occurred in patient safety work. Neither G-theory nor the more commonly used kappa measures of interrater reliability provide an absolute estimate of “good” or “bad” reliability. Reliability is, by nature, a relative term, with the degree of reliability needed varying based on the situation in which judgments are made. Health system leaders and researchers must balance level of reliability needed with cost of personnel needed to perform reviews. We used the CSEM measure to help health system leaders and researchers understand how the chance of misclassifying event harm can be reduced to essentially zero by using two reviewers instead of one.

While this study performed a rigorous comparison of reviews of 400 adverse events by 9 raters, interpretation of the results is subject to some limitations. We used real cases, supplemented with cases from the literature when numbers in certain categories (e.g., very severe events) were small. We did not perform a two-step abstraction and adjudication process, as is often performed using trigger tools or in research studies.[25] The impact of the abstraction process on reliability of case review was not assessed. While there was interaction between adverse event type and reliability, we chose to report pooled data rather than reporting all data in eight strata, one for each adverse event type, to simplify interpretation of results. We used the AHRQ Harm Scale, which has a reliability ranging from K=0.47–0.58 in a previous publication, consistent with our findings.[16] We did not assess the impact of rater characteristics using other types of harm scales, such as the

National Coordinating Council Medication Error Reporting and Prevention scale.[17] However, the reliability of the AHRQ Harm Scale is comparable to other harm scales.[12,26]

In summary, this study provides important clinical and methodological contributions to the field of health services research. First, our study indicates that at least two experienced clinician raters should review each adverse event reported in order to optimize the accuracy of adverse event review. Second, pharmacists seem to be more consistent in their ratings, thus requiring fewer raters than nurses or physicians to achieve the same reliability of ratings. Keeping this in mind, health system leaders can ensure that resources are directed toward prevention of the most severe events. Finally, this study illustrates how generalizability theory can contributes to a better understanding of the sources and magnitude of error in adverse event measurement and, in this way, makes a unique and complementary contribution to the standard approaches to reliability assessment.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Funding: Agency for Healthcare Quality and Research Contract No. HHSA290201000022I

References

1. Aspden, P., Corrigan, JM., Wolcot, J., Erickson, S. Patient Safety: Achieving a New Standard for Care. The National Academies Press; Washington, DC: 2004.
2. Bates D, Cullen D, Laird N, et al. Incidence of adverse drug events and potential adverse drug events. Implications for prevention. ADE Prevention Study Group. J Am Med Assoc. 1995; 274:29–34. DOI: 10.1016/S1075-4210(05)80011-2
3. Field TS, Tjia J, Mazor KM, et al. Randomized trial of a warfarin communication protocol for nursing homes: An SBAR-based approach. Am J Med. 2011; 124(2):179.e1–e179.e7. DOI: 10.1016/j.amjmed.2010.09.017 [PubMed: 21295198]
4. Gurwitz JH, Field TS, Avorn J, et al. Incidence and preventability of adverse drug events in nursing homes. Am J Med. 2000; 109(2):87–94. DOI: 10.1016/S0002-9343(00)00451-4 [PubMed: 10967148]
5. Gurwitz JH, Field TS, Harrold LR, et al. Incidence and preventability of adverse drug events among older persons in the ambulatory setting. JAMA. 2003; 289(9):1107–1116. DOI: 10.1001/jama.289.9.1107 [PubMed: 12622580]
6. Gurwitz JH, Field TS, Judge J, et al. The incidence of adverse drug events in two large academic long-term care facilities. Am J Med. 2005; 118(3):251–258. DOI: 10.1016/j.amjmed.2004.09.018 [PubMed: 15745723]
7. Gurwitz JH, Field TS, Radford MJ, et al. The safety of warfarin therapy in the nursing home setting. Am J Med. 2007; 120(6):539–544. DOI: 10.1016/j.amjmed.2006.07.045 [PubMed: 17524757]
8. Gurwitz JH, Field TS, Rochon P, et al. Effect of computerized provider order entry with clinical decision support on adverse drug events in the long-term care setting. J Am Geriatr Soc. 2008; 56(12):2225–2233. DOI: 10.1111/j.1532-5415.2008.02004.x [PubMed: 19093922]
9. Landrigan CP, Parry GJ, Bones CB, Hackbarth AD, Goldmann DA, Sharek PJ. Temporal trends in rates of patient harm resulting from medical care. N Engl J Med. 2010; 363(22):2124–2134. DOI: 10.1056/NEJMsa1004404 [PubMed: 21105794]

10. Walsh KE, Adams WG, Bauchner H, et al. Medication errors related to computerized order entry for children. Pediatrics. 2006; 118(5):1872–1879. DOI: 10.1542/peds.2006-0810 [PubMed: 17079557]
11. Walsh KE, Landrigan CP, Adams WG, et al. Effect of computer order entry on prevention of serious medication errors in hospitalized children. Pediatrics. 2008; 121(3):e421–e427. DOI: 10.1542/peds.2007-0220 [PubMed: 18310162]
12. Dean BS, Barber ND. A validated, reliable method of scoring the severity of medication errors. Am J Heal Pharm. 1999; 56:57–62.
13. Zegers M, de Bruijne MC, Wagner C, Groenewegen PP, van der Wal G, de Vet HCW. 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. DOI: 10.1016/j.jclinepi.2009.03.004 [PubMed: 19473812]
14. Brennan, R. Generalizabfaiity Theory. New York: Springer-Verlag; 2001.
15. Agency for Healthcare Quality and Research. Patient Safety Organization Common Formats. 2012. http://www.ahrq.gov/policymakers/measurement/common-formats/index.html. Accessed on May 4, 2016
16. Williams T, Szekendi M, Pavkovic S, Clevenger W, Cerese J. The Reliability of AHRQ Common Format Harm Scales in Rating Patient Safety Events. J Patient Saf. 2013; 00(1):1–8. DOI: 10.1097/PTS.0b013e3182948ef9
17. National Coordinating Council for Medication Error Reporting and Prevention. Types of Medication Errors. http://www.nccmerp.org/medErrorCatIndex.html. Accessed February 27, 2014
18. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform. 2009; 42(2):377–381. DOI: 10.1016/j.jbi.2008.08.010 [PubMed: 18929686]
19. Brennan, R. Manual for mGENOVA. Iowa City: Iowa Testing Programs; 2001.
20. Williams SD, Ashcroft DM. Medication errors: how reliable are the severity ratings reported to the national reporting and learning system? Int J Qual Health Care. 2009; 21(5):316–320. DOI: 10.1093/intqhc/mzp034 [PubMed: 19679598]
21. van Doormaal JE, Mol PG, van den Bemt PM, et al. Reliability of the assessment of preventable adverse drug events in daily clinical practice. Pharmacoepidemiol Drug Saf. 2008; 17(7):645–654. DOI: 10.1002/pds.1586 [PubMed: 18338767]
22. Hall P. Interprofessional teamwork: professional cultures as barriers. J Interprof Care. 2005; 19(1):188–196. DOI: 10.1080/13561820500081745 [PubMed: 16096155]
23. Horsburgh M, Perkins R, Coyle B, Degeling P. The professional subcultures of students entering medicine, nursing and pharmacy programmes. J Interprof Care. 2006; 20(4):425–431. DOI: 10.1080/13561820600805233 [PubMed: 16905490]
24. Pecukonis E, Doyle O, Bliss DL. Reducing barriers to interprofessional training: Promoting interprofessional cultural competence. J Interprof Care. 2008; 22(4):417–428. DOI: 10.1080/13561820802190442 [PubMed: 18800282]
25. Griffin, F., Resar, R. IHI Global Trigger Tool for measuring adverse events; IHI Innov Ser white Pap. 2007. p. 1-44.http://www.ihi.org/resources/Pages/IHIWhitePapers/IHIGlobalTriggerToolWhitePaper.aspx. Accessed on May 4, 2016
26. Garfield S, Reynolds M, Dermont L, Franklin BD. Measuring the severity of prescribing errors: A systematic review. Drug Saf. 2013; 36:1151–1157. DOI: 10.1007/s40264-013-0092-0 [PubMed: 23955385]

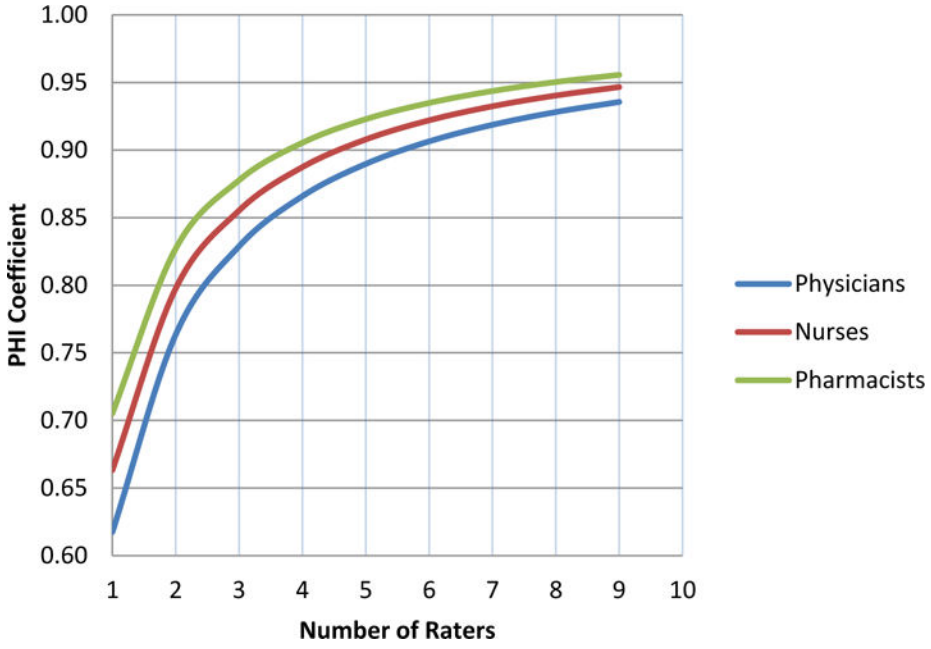

Figure 1.
Reliability Coefficient Phi (Generalizability theory framework) for Physicians, Nurses and Pharmacists

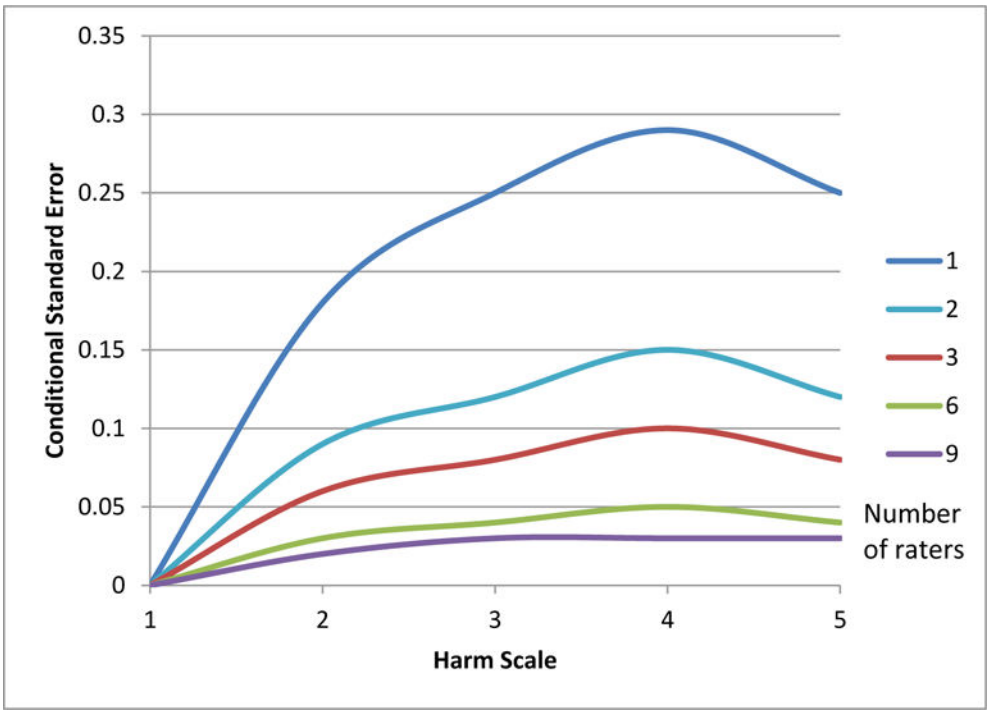

Figure 2. Conditional Standard Errors of Measurement for Groups of 1, 2, 3, 6, and 9 raters
Harm scale ratings range from 1 to 5, with lower numbers indicating greater severity of harm.

Table 1

Study 1: Events crossed with raters, for each specialty.

Events	Physicians			Nurses			Pharmacists		
	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Rater 6	Rater 7	Rater 8	Rater 9
1	X	X	X	x	X	X	x	x	x
2	X	X	x	x	X	X	x	x	x
.	.	.	.	.	.	.	.	.	.
400	X	X	x	x	X	X	x	x	x

Table 2

Variance Components and Reliability coefficient Phi, by Specialty

Effect	Physicians	Nurses	Pharmacists
Event	0.502	0.465	0.579
Rater	0.087	0.022	0.003
Event × Rater	0.223	0.214	0.238
Phi	0.83	0.86	0.88

Table 3

Variance Components and Reliability Coefficient Phi, by Level of Expertise

Effect	Low	Medium	High
Event	0.468	0.513	0.589
Rater	0.002	0.042	0.050
Event × Rater	0.239	0.230	0.183
Phi	0.85	0.85	0.88