Inter-Observer Agreement Among Medical Professionals in Critical Care of Neonates and Children

Tourgeman-Bashkin[1] Osnat*, Shinar[1] David, Parmet[1] Yisrael, and Zmora[2] Ehud

[1]Department of Industrial Engineering & Management, Faculty of Engineering, Ben Gurion University of the Negev, PO Box 653, Beer Sheva, Israel
osnatb@tasmc.health.gov.il
shinar@bgu.ac.il
iparmet@bgu.ac.il
[2]Department of Neonatology, Soroka Medical Center, Faculty of Health Sciences, PO Box 653, Beer Sheva, Israel
ezmora@bgu.ac.il

ABSTRACT

Inter-observer agreement is essential to medical staff members and has a major effect on communication. The goal of the study was to examine the way medical professionals evaluate the potential severity of Almost Adverse Events (AAEs) that were observed in two intensive care units (ICUs). One hundred and fourteen AAEs were observed and recorded in both units by engineering students. Each AAE was rated independently by five senior medical staff members from each ICU, chosen by the unit manager, on a three- point severity level scale. Statistical analysis (*K* statistic and Cohen's Kappa) yielded relatively low levels of agreement among raters in both ICUs (< 0.3), but significantly greater agreement was found among nurses than among physicians in both ICUs. Low levels of agreement are attributed to the nature of work and characteristics of each ICU. Recommendations for improving agreements including forming shared mental models are specified.

Keywords: inter-observer agreement, kappa coefficients, almost adverse events, communication, mental model

1. INTRODUCTION

The Institute of Medicine's 2000 report on patient safety identified medical errors as a significant contributor to patient morbidity and mortality in adults, with an estimated at least 44,000 deaths per year caused by medical adverse events as a result of erroneous

*Corresponding author, the unit for clinical performance research, Tel- Aviv Sourasky Medical Center. Tel: +972-8-9466643, osnatb@tasmc.health.gov.il

medical interventions [1]. One way to minimize the prevalence of medical errors is to learn more from near misses. Near misses, also termed almost adverse events (AAEs), are defined as errors that did not have severe clinical sequelae for the patient [2]. AAEs can be used as a useful knowledge base that enables practitioners to study and prevent medical human errors before they cause severe outcomes.

Severity assessment of medical cases is one of the important tasks that medical practitioners are involved in as it determines the course of patient care. Nevertheless, there is a question of how reliable medical staff members are in assessing the potential severity of medical events. Wrong assessment of severity can lead to inappropriate medical treatment. Measuring Inter-observer agreement can be used as a method to examine this issue.

In general, studies that examined inter-observer agreements tend to focus on discrepancies in observations of signs and symptoms and in medical diagnosis [3–5]. The few studies that addressed the issue of severity evaluations of medical adverse events did so by comparing two or more severity models or objective scoring systems [6–8].

Possible sources of inter-observer discrepancies include differences in the training, culture and roles that different observers may have. When these differences are significant, they can have significant clinical implications such as communication problems that can lead to adverse events.

Badihi [9] investigated the formal structure of thinking and its implications for communication and performance errors by examining the thinking style and knowledge structure of physicians and nurses in an ICU, and found differences in thinking style between nurses and physicians; physicians tended to filter information according to their initial perception of the situation, whereas nurses updated their understanding of the clinical situation more often. It is suggested that communication failures between physicians and nurses may stem from differences in the way of processing and collecting data, and from the fact that the information exchanged in physicians-nurses interactions is not compatible with the information needed [9].

Joshua et al. [10] reviewed the literature on the accuracy and reliability of physical examinations and common clinical signs in areas such as cardiovascular, respiratory, gastroenterological, and neurological examinations. They found wide variations in inter-observer agreements for different physical signs including sensitivity, specificity and reproducibility. Variations in inter-observer agreement reflects factors such as difficulty in perceiving signs, differences in training and experience, task complexity and increasing reliance on diagnostic tests. This variation is added to the variation within and among patients and circumstances to create a complex setting for medical decision making and a possibility for error to occur [10]. Another research showed that physicians and nurses significantly disagree on who is responsible for patient safety, what constitutes a medical error, how these errors should be reported, and what actions are needed to prevent them in the future [11].

The current study was part of a long term project that assessed the nature of AAEs in ICUs [12]. The study relied on multi-disciplinary cooperation between medical professionals and human factors engineers to examine the way medical professionals

evaluate the potential severity of AAEs in two ICUs. In this study, we referred to low inter-observer agreement as an aspect of communication failures, and hypothesized that nurses and physicians in ICUs disagree on the severity assessment of AAEs.

2. METHODS

2.1. Study Environment

The research was conducted in the Neonatal ICU (NICU) and the Pediatric ICU (PICU) in a large medical center. The NICU serves premature neonates and neonates with special health conditions that require intensive care. Thus, unlike the PICU, the NICU also includes infants that do not require intensive care. The unit admits approximately 600 infants per year, and has a yearly occupancy rate of 100%. Thirty registered nurses, all trained in neonatal intensive care, in addition to eight senior physicians, work in the NICU. Three physicians and 7 nurses work in the morning shift, and 1 physician and 5 nurses in the noon and night shifts in the NICU. There are 18 incubators and 6 beds in the NICU.

The PICU has 8 beds for infants and children up to 18 years old, mostly with severe injuries or conditions that require continuous medical supervision. There are approximately 450 admissions per year with yearly occupancy rate over 100%. The unit has 20 registered nurses; most of them had specialized training in intensive care. The unit also has 4 physicians: 2 senior physicians trained in pediatric intensive care, a resident in Pediatrics, and a fellow in critical care Pediatrics. Three physicians and 4 nurses work in the morning shift, and 3 nurses and 1 physician in the noon and night shifts in the PICU.

2.2. Study Design and Procedures

In a period of 4 months, three senior Industrial Engineering & Management students performed a total of 500 hours of observations, in 62 8-hours shifts in both ICUs. Prior to data collection, the observers met several times with the medical staff of the ICUs to develop rapport and consistency in the observation and coding process. Meetings with staff members from both ICUs helped observers understand and recognize when an AAE occurred. To ensure consistency among the three observers, the initial observations were conducted in pairs. The senior engineering students were selected to conduct observation as bias-free as possible. The observers received permission to perform the observations from the local research ethics committee, the hospital management, and the medical staff members. The observers did not interfere with the daily work and did not speak with the staff members while observing their work; they documented every activity they saw in the units. The observations only focused on the medical staff members while they were performing their tasks, and not on the patients in the units; therefore permission from parents was unnecessary.

For the purpose of the study, an AAE was defined as an error that did not result in clinically significant adverse outcome [2]. The severity of each AAE observed in the NICU and the PICU was rated independently and anonymously by five medical staff members at each ICU, including three senior physicians and two senior nurses chosen by the ICU managers. The severity was rated on a three-point severity scale, where 1

represents an event that has no severe clinical implications or an event that could only cause a minor inconvenience to the patient, 2 represents an event that could potentially prolong the hospitalization or cause a severe illness, and 3 represents an event that could potentially result in the loss of a body organ, or a significant deterioration in patient's status that can lead to death. The severity scale was based on expert opinion and developed by medical professionals and human factors researchers for the specific purpose of the research and was not statistically validated against actual outcomes.

Each medical staff member received the AAEs descriptions prepared by the observers, and rated the severity based on personal opinion and experience. As soon as the task was completed, the files were returned to the researchers.

2.3. Data Analysis

Microsoft Excel and SPSS-11 were used to perform the statistical analysis and to assess numeric trends. Intraclass Correlation (ICC) was used to measure the level of agreement among physicians and nurses. There are two approaches to ICC: consistency and absolute agreement. The difference between consistency and absolute agreement measures how the systematic variability due to raters or measures is treated. If that variability is considered irrelevant, it is not included in the denominator of the estimated ICCs, and measures of consistency are produced. If systematic differences among levels of ratings are considered relevant, rater variability contributes to the denominators of the ICC estimates, and measures of absolute agreement are produced. In the current study, we used the consistency approach due to the fact that it is more suitable to Kappa statistic in our later analysis. *K* statistic was employed to measure the level of agreement among the physicians themselves and among the nurses themselves (quadratic weighting). The K statistic is based on a formula developed by Fleiss [13], which provides a numerical measure of agreement among multiple raters. Cohen's Kappa coefficient was used to test levels of agreement between the two nurses in each unit, Cohen's Kappa is more suitable than Fleiss[13] *K* statistic to examine inter-observer agreement between two raters. The Kappa statistic measures the observed amount of agreement adjusted for the amount of agreement expected by chance alone. A value of −1.00 indicates complete disagreement, a value of 0 indicates that the agreement is no better than chance, and a value of +1.00 indicates a perfect agreement. In addition, Chi square analysis was performed in order to examine the differences between the two units in the staff members' ratings**.**

3. RESULTS

A total of 114 AAEs were observed and recorded: 49 events in the NICU and 65 in the PICU. It is important to note that descriptions of the AAEs were recorded without personal identifying information. Because of the different nature of the two ICU's, their results are presented separately.

3.1. Neonatal ICU

The inter-rater correlations between the physicians and the nurses were quite low in terms of inter-observer reliability. The intraclass correlation between the two nurses was

0.49 (with 95% confidence interval from 0.25 to 0.68), the intraclass correlation among the three physicians was 0.36 (with 95% confidence interval from 0.18 to 0.53), and the intraclass correlation among all five caretakers was 0.37 (with 95% confidence interval from 0.24 to 0.52). The analysis of agreement yielded a Fleiss Kappa coefficient = 0.18 for the three physicians and Cohen Kappa coefficient = 0.27 for the two nurses. Both values were quite low indicating slight or poor agreement among the raters.

A closer examination of the ratings showed that in 82% of the events, there was a disagreement among the five raters. In 31% of the AAEs, there was a disagreement in whether the event should be rated as severity scale 1 or 2. In 16% of the events, there was a disagreement in whether the event should be rated as severity scale 2 or 3. In 35% of the events, there was a disagreement in whether the event should be rated as severity scale 1 or 3.

Table 1 summarizes the levels of agreement on the severity rating of the AAEs among the five raters. The table shows that there is a higher level of agreement between nurses than among physicians. Nurses disagreed by two points on the 3-point severity scale in only 6% of the AAEs, whereas physicians disagreed by two points on the severity scale in 22% of the AAEs.

Figure 1 presents the distribution of severity ratings of AAEs by the NICU physicians and nurses. Note that for the data in Figure 1, the total sample consisted of 98 data provided by the nurses (49 events × 2 nurses) and 147 data provided by the physicians (49 events × 3 physicians).

3.2. Pediatric ICU

The inter-rater correlations among the physicians and nurses were quite low in terms of measures of inter-observer reliability. The intraclass correlation between the two nurses was 0.62 (with 95% confidence interval from 0.44 to 0.75), the intraclass correlation among the three physicians was 0.16 (with 95% confidence interval from 0.01 to 0.33), and the intraclass correlation among all five raters was 0.37 (with 95% confidence interval from 0.25 to 0.49). The analysis of agreement yielded a Fleiss Kappa coefficient = 0.14 for the three physicians and Cohen Kappa coefficient = 0.29 for the two nurses. Both values were quite low indicating slight or poor agreement among the raters.

Table 1. Levels of agreement on the AAEs' severity among NICU raters

	Nurses n = 2	**Physicians n = 3**	**Nurses & Physicians n = 5**
Total Agreement	53%	31%	18%
Disagreement by one point*	41%	47%	47%
Disagreement by two points*	6%	22%	35%

*On a 3-point severity scale.

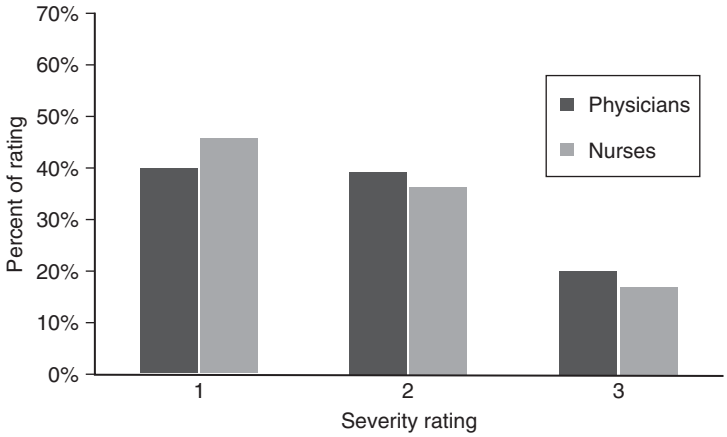

Figure 1. Distributions of severity rating of AAEs by NICU physicians and nurses.

A closer examination of the ratings showed a disagreement among the five raters in 78% of the events. In 40% of the AAEs, there was a disagreement in whether the event should be rated as severity scale 1 or 2. In 10% of the ratings, there was a disagreement in whether the event should be rated as severity scale 2 or 3. In 28% of the ratings, there was a disagreement in whether the event should be rated as severity scale 1 or 3.

Table 2 summarizes the levels of agreement among the five raters on the severity rating of the AAEs. It shows no (0%) disagreements by two points of severity scale between nurses, and 9% of disagreements by two points of severity scale among physicians.

Figure 2 presents the distributions of severity rating of the AAEs by the PICU physicians and nurses. Note that for the data in Figure 2, the total sample consisted of 130 data from the nurses (65 events × 2 nurses) and 195 data from the physicians (65 events × 3 physicians).

At the PICU, 6% of the events were rated as category 3 by physicians, while at NICU, 20% of the events were rated as category 3 by physicians ($\chi^2 = 5.27$, $p < 0.05$).

Table 2. Levels of agreement on the AAEs' severity among PICU raters

	Nurses n = 2	Physicians n = 3	Nurses & Physicians n = 5
Total Agreement	49%	36%	22%
Disagreement by one point*	51%	55%	50%
Disagreement by two points*	0%	9%	28%

*On a 3-point severity scale.

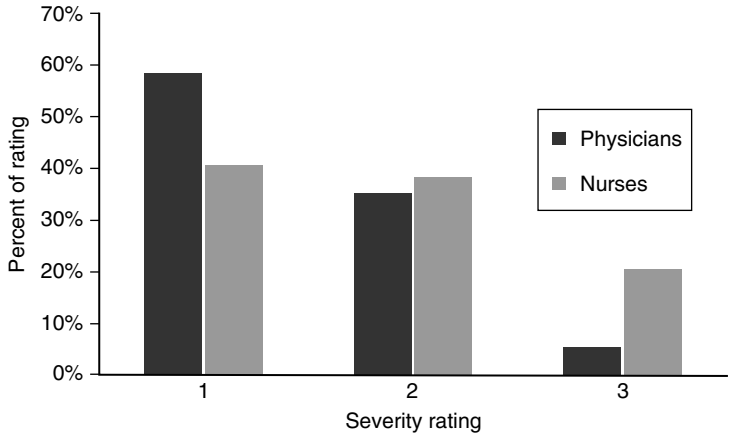

Figure 2. Distributions of severity rating of AAEs by PICU physicians and nurses.

At the PICU, 21% of the events were rated as category 3 by nurses, while at NICU, 17% of the events were rated as category 3 by nurses (there were no statistically significant differences between the two units in the nurses' ratings of category 3).

4. DISCUSSION

The findings of the current study are important as they reflect a problematic situation in which staff members working in the same medical environment perceive and assess the same clinical events quite differently. Nevertheless, there are a few possible limitations to the study. Medical staff members who rated the AAEs were part of the ICU teams and it is possible that they were involved in an AAE that they rated. Directly taking care of an AAE may affect the rater's severity rating.

The present results show relatively poor agreement in AAE rating among caretakers in both ICUs, with better agreement between nurses than among physicians in both ICUs. For example, in one AAE in the NICU, all three physicians rated the event as one without severe clinical implications, whereas the two nurses rated it as a most severe event. In an AAE that occurred in the PICU, one of the physicians rated it as having no severe clinical meaning, whereas one of the nurses rated it as a most severe event. The other two physicians and the second nurse rated the event as one that could potentially prolong the hospitalization or cause a severe illness. Such levels of disagreement reflect differences in attitudes and views of the staff members towards the patient and the medical work procedures, and show that each staff member has a different mental schema or map regarding the situation or the work environment.

To understand the reasons for the differences between nurses and physicians and between units, we must consider the characteristics of each unit investigated, and the significant differences between those two medical work environments. Unlike the

PICU, the NICU is not officially defined as an ICU because it also admits infants that do not require intensive care. That means much of the NICU daily work includes medical care of patients that are not necessarily in critical or severe condition. This, in turn, can influence the general attitude towards the perceived severity of different situations in the NICU. Nevertheless, Physicians in the PICU rated the AAEs as severity scale 3 much less frequently (only 6% of the cases) than the nurses in the PICU and the staff (nurses and physicians) in the NICU. This finding may reflect the flip side of the coin when discussing the effects of working in a medical environment that is characterized by treating very critical and severe patients. Working in such an environment could affect judgment in the long term, resulting in underestimation of the more severe AAEs. However, it is necessary to emphasize that this is a hypothesis that requires a separate examination, as our findings showed that nurses in the PICU tended to rate AAEs as more severe in their judgments.

The origins of the discrepancies found in our study may be rooted in the fundamental differences between nurses and physicians, including status/authority, gender, training, experience, patient care responsibilities, and other personal characteristics [14]. Croser [15] suggests that nurses tend to focus on the social histories of patients, to be more personally involved with patients, and to focus more on caring than curing. These differences in culture and the way nurses and physicians communicate can influence judgment in general, and severity assessment of clinical events in particular. Unfortunately, we do not have the data available to link the personal characteristics of the participating raters to their severity assessments of AAEs. Inter-observer discrepancies can have severe clinical implication because they reflect differences among care givers that can result in communication barriers.

Several countermeasures can be implemented to reduce disagreements among medical professionals, including teaching them how to share information and training in clinical working skills that can reduce individual differences and disagreements. When engaged in specific medical tasks, experienced, appropriately-trained physicians tend to agree among themselves more often than less experienced and less trained physicians [16, 17]. Physicians and nurses need to acknowledge the different ways to gather as much relevant information and signs as they can in the complicated process of diagnosing and assessing patients' state; this should lead to a shared mental model. Mental models are cognitive constructs allowing people to understand and predict the world around them, to recognize relationships among components of the environment, and to construct expectations regarding the near future [18]. Shared mental models are tools that provide team members with a common understanding of who is responsible for which tasks and what the information requirements for each task are. Shared mental models allow team members to anticipate one another's needs and to work in synchrony [19]. Minionis et al. [20] reported that when teams were completing subtasks that required teamwork, having a shared mental model resulted in improved performance. Forming shared mental models of the team, the task, and the informational requirements serves as an important mechanism for achieving efficient communications and improving team performance [19].

5. CONCLUSIONS

The results of the current research reflect the complexity that exists in healthcare environments. The results show large discrepancies in the way professionals perceive different situations and assess their potential severity. Significant disagreements regarding the severity of AAEs suggest differences in how staff members perceive and eventually treat the patient. One possible effective measure to improve inter-observer reliability or agreement is to establish a detailed written consensus of definitions and examples of different levels of severity of each AAE. Forming a special committee to address the issue should be a high priority for healthcare systems because disagreement among caregivers reflects communication problems that may eventually cause adverse human errors.

ACKNOWLEDGMENTS

This research was partially funded by the Israel National Institute for Health Policy and Health Service Research. We thank the PICU and the NICU medical staff who participated in this study.

REFERENCES

[1] Institute of Medicine, Committee on Quality of Health Care in America, In: Kohn L.T., Corrigan J.M., Donaldson M.S., eds. To Err is Human Building a Safer Health System, *Washington, DC: National Academy Press*, 2000.

[2] Bogner M.S., Human error in medicine, *Hillsdale, NJ: Lawrence Erlbaum Associates*, 1994.

[3] Theodossi A., Knill-Jones R.P., Skene A. et al., Inter-observer variation of symptoms and signs in jaundice, *Liver*, 1981, 1, 21–32.

[4] Lindsay K.W., Teasdale G.M., Knill-Jones R.P., Observer variability in assessing the clinical features of subarachnoid hemorrhage, *Journal of Neurosurgery*, 1983, 58, 57–62.

[5] Tomasello F., Mariani F., Fieschi C. et al., Assessment of inter-observer differences in the Italian multicenter study on reversible cerebral ischemia, *Stroke*, 1982, 13, 32–35.

[6] Cowen J.S., Kelly M.A., Errors and bias in using predictive scoring systems, *Critical Care Clinics*, 1994, 10, 53–78.

[7] Teres D., Comment on "The case for using objective scoring systems to predict intensive care outcome.", *Critical Care Clinics*, 1994, 10, 91–92.

[8] Teres D., Lemeshow S., Why severity models should be used with caution, Critical *Care Clinics*, 1994, 10, 93–110.

[9] Badihi Y., The Formal Structure of Thinking and Its Implications to Communication and Performance Errors, *Research Thesis for the Degree of Doctor of Science. Technion, Israel*. 1993.

[10] Joshua A.M., Celermajer D.S., Stockler M.R., Beauty is in the eye of the examiner: reaching agreement about physical signs and their value, *Internal Medicine Journal*, 2005, 35, 178–787.

[11] Cook A.F., Disagreement on Medical Error Reporting May Place Patients at Risk, *New American Journal of Nursing Study. Journal to Convene Special Symposia to Review and Address Challenges*, Advancing Nursing Practice Excellence: State of the Science, 2004.

[12] Tourgeman-Bashkin O., Shinar D., Zmora E., causes of near misses in critical care of neonates and children, *Acta Pediatrica*, 2008, 97, 299–303.

[13] Fleiss J.L., Measuring nominal scale agreement among many raters, *Psychological Bulletin*, 1971, 76, 378–382.

[14] Thomas E.J., Sexton J.B., Helmreich R.L., Discrepant attitudes about teamwork among critical care nurses and physicians, *Critical Care Medicine*, 2003, 31(3), 956–959.

[15] Croser W.D., The contemporary nurse-physician relationship: Insights from scholars outside the two professions, *Nurse Outlook*, 2000, 48, 263–268.

[16] Raftery E.B., Holland W.W., Examination of the Heart: an investigation into variation, *American Journal of Epidemiology*, 1967, 85, 438–44.

[17] Tomasello F., Mariani F., Fieschi C., Argentino C., Bono G., De Zanche L. et al., Assessment of interobserver differences in the Italian multicenter study on reversible cerebral ischemia, *Stroke*, 1982, 13, 32–5.

[18] Rouse W.B., Morris N.M., On looking into the black box: prospects and limits in the search for mental models, *Psychological bulletin*, 1986, 100, 349–363.

[19] Stout R.J., Cannon-Bowers J.A., Salas E., Milanovich D.M., Planning, Shared Mental Models, and Coordinated Performance: An Empirical Link Is Established, *Human Factors*, 1999, 41, 61–71.

[20] Minionis, D.P., Zaccaro, S.J., Perez, R., Shared mental models, team coordination, and team performance, *Paper presented at the 10th Annual Conference of the Society for Industrial/Organizational Psychology, Orlando, FL*. 1995.