Original Article

J Prev Med Public Health 2015;48:239-248 • http://dx.doi.org/10.3961/jpmph.14.049

pISSN 1975-8375 eISSN 2233-4521

Journal of Preventive Medicine & Public Health

Assessing Reliability of Medical Record Reviews for the Detection of Hospital Adverse Events

Minsu Ock[1], Sang-il Lee[1], Min-Woo Jo[1], Jin Yong Lee[2], Seon-Ha Kim[3]

[1]Department of Preventive Medicine, University of Ulsan College of Medicine, Seoul; [2]Public Health Medical Service, Seoul National University Boramae Medical Center, Seoul; [3]Department of Nursing, Dankook University, Cheonan, Korea

Objectives: The purpose of this study was to assess the inter-rater reliability and intra-rater reliability of medical record review for the detection of hospital adverse events.

Methods: We conducted two stages retrospective medical records review of a random sample of 96 patients from one acute-care general hospital. The first stage was an explicit patient record review by two nurses to detect the presence of 41 screening criteria (SC). The second stage was an implicit structured review by two physicians to identify the occurrence of adverse events from the positive cases on the SC. The inter-rater reliability of two nurses and that of two physicians were assessed. The intra-rater reliability was also evaluated by using test-retest method at approximately two weeks later.

Results: In 84.2% of the patient medical records, the nurses agreed as to the necessity for the second stage review (kappa, 0.68; 95% confidence interval [CI], 0.54 to 0.83). In 93.0% of the patient medical records screened by nurses, the physicians agreed about the absence or presence of adverse events (kappa, 0.71; 95% CI, 0.44 to 0.97). When assessing intra-rater reliability, the kappa indices of two nurses were 0.54 (95% CI, 0.31 to 0.77) and 0.67 (95% CI, 0.47 to 0.87), whereas those of two physicians were 0.87 (95% CI, 0.62 to 1.00) and 0.37 (95% CI, -0.16 to 0.89).

Conclusions: In this study, the medical record review for detecting adverse events showed intermediate to good level of inter-rater and intra-rater reliability. Well organized training program for reviewers and clearly defining SC are required to get more reliable results in the hospital adverse event study.

Key words: Adverse event, Patient safety, Intra-rater reliability, Inter-rater reliability, Medical record review

INTRODUCTION

Measuring the size of patient safety problem is the first step for enhancing patient safety [1]. Various methods and indicators are used to measure patient safety problems [2,3]. Incidence of adverse event (AE) is a representative indicator widely used for this measurement. The Harvard Medical Practice Study (HMPS), which provided a basis for raising patient safety as an important policy agenda in the US, defined AE as "an injury that was caused by medical management and that prolonged the hospitalization, produced a disability at the time of discharge,

Received: November 14, 2014 Accepted: July 22, 2015

Corresponding author: Sang-il Lee, MD, PhD
88 Olympic-ro 43-gil, Songpa-gu, Seoul 05505, Korea
Tel: +82-2-3010-4284, Fax: +82-2-477-2898
E-mail: sleemd@amc.seoul.kr

or both" [4]. As AEs are related to direct treatment outcomes, such as patient outcomes, length of stay, and medical expenditure, they are useful for measuring patient safety levels [5]. Furthermore, AEs include medical errors such as medication errors, comprehensively representing patient safety levels. Additionally, as AEs have been widely and continuously used as indicators of patient safety, the concept has been well established.

The HMPS examined patient medical records in New York State hospitals and evaluated AE occurrence, medical mistake or error occurrence, and patient disability level caused by AEs [4,6]. Many studies have since determined the incidence of AEs based on the HMPS methodology [7]. The biggest commonality is that evidence for AE occurrence was found by examining medical records. Specifically, previous studies applied a 2-stage examination method in which 2 nurses in the first stage and 2 physicians in the second stage examined medical records individually, taking the form of a sequential and independent review. Medical record review is commonly used to identify AE occurrence as medical records are easily accessible. However, it requires much time and necessitates commitment from medical professionals, resulting in higher cost than other methods [8]. Furthermore, unfaithful documentation can lead to underestimation of occurrence of AEs. However, the biggest shortcoming of medical record review for determining AEs is that the results may not be reliable if the consistency among different reviewers or even in a single reviewer is lacking [9]. Therefore, medical record review reliability should be examined and enhanced before conducting a study to measure AE incidence in hospitals [10].

In this study, we evaluated both intra-rater and inter-rater reliability for screening criteria (SC) detection and AE identification through medical record review.

METHODS

Study Design

A retrospective medical record review was conducted in a general hospital with approximately 500 beds. Review for AE identification was based on the HMPS methodology including a 2-stage examination (Figure 1) [4,6]. First, specific dates in 2007 were randomly selected using a random number table, and admissions of all patients discharged on those dates were selected as index admission. Psychiatric department admissions were excluded. Medical records from the entire duration of hospitalization and 1 year before and after were reviewed using a case review form developed in a previous study [11]. The first-stage review conducted by nurses screened medical records for 41 SC. These criteria encompassed events with high possibility of AE occurrence, such as antidote use, or those highly likely to lead to occurrence of additional AEs. The case

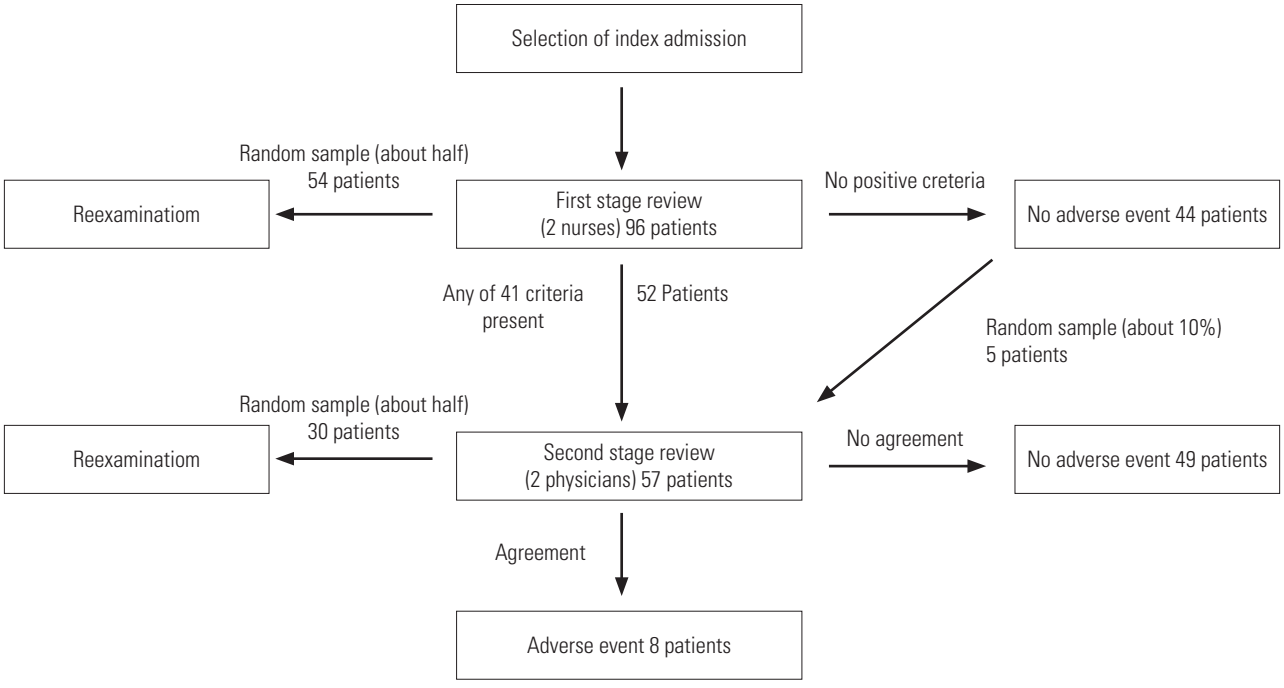

Figure 1. Process and results of medical record review for detecting hospital adverse events.

review form included 41 SC chosen through the modified Delphi method from a combination of previously used SC [11]. These criteria can be divided into 19 items from the HMPS-format studies and 28 items from the Global Trigger Tool (GTT) (Supplemental Table 1). As opposed to the HMPS-type SC, which rely more heavily on reviewer's clinical judgment, the GTT-type SC are more heavily based on objective clinical examination results [11]. For convenience of physicians while reviewing, nurse reviewers marked the corresponding medical record using a post-it when they found entries meeting the SC.

Two nurse reviewers independently reviewed medical records of the same patients. All cases determined by at least one nurse to meet the SC were included in the second-stage review for AE identification. Furthermore, to determine the occurrence of false negative results in the first-stage review, approximately 10% of cases that did not meet the SC during the first-stage review were randomly selected and included in the second-stage review. Two physician reviewers also independently reviewed medical records of the same patients to identify AE occurrence. Cases identified by both physician reviewers as having AEs were determined as AEs. In cases where opinions of the physician reviewers differed, a panel comprising professionals made the final call.

For evaluation of inter-rater reliability, approximately half of the patients from the primary investigation were selected for the secondary investigation using the previously discussed method. However, in the secondary investigation, the second-stage review was not conducted on cases that did not meet the SC during the first-stage review.

Reviewers and Reviewer Training

A total of four reviewers participated in this study: two physicians (P1 and P2) and two nurses (N1 and N2). First-stage reviewers were nurses with five or more years of clinical experience, and second-stage reviewers were specialists with ten or more years of clinical experience. Reviewers were trained for approximately two hours. Educational content consisted of understanding the concept of AEs, instructions on using the case review form, and a practice exercise using an actual medical record.

Statistical Analysis

Inter-rater and intra-rater reliability of reviews of nurses and physicians was examined (Table 1). First, inter-rater reliability of reviews of nurses was analyzed for two different purposes: to assess the presence of SC and the necessity of the second-stage review. In evaluating the SC, agreements and kappa values were calculated for each of the 41 items of the SC. This represents reliability of decision of two nurses on whether or not the same patient meets a specific screening criterion. As patients determined to meet at least one screening criterion are subjected to the second-stage review, inter-rater reliability of nurses was also shown using agreements and kappa value for second-stage review necessity (last rows on Tables 2 and 3). It was then confirmed whether the reliability of reviews of nurses varied depending on SC characteristics. Using the aforementioned analytical method, differences in inter-rater reliability of nurses regarding second-stage review necessity based on SC characteristics were examined. In other words, cases corresponding to more than one screening criterion used in the HMPS-type studies and cases corresponding to more than one GTT-type screening criterion were examined separately. Six overlapping criteria were included in both groups.

Next, cases that at least one nurse determined to have at least one screening criterion and five randomly selected cases from those determined to have no screening criterion by reviews of two nurses were selected to investigate inter-rater reliability of physicians. AE occurrence was confirmed when calculating physicians' inter-rater reliability. When one patient was determined to have 2 AEs, no examinations as to whether the contents were consistent with one another were done. Inter-rater reliability of physicians was calculated using agreement and kappa value for AE identification.

Using the results from the primary and secondary investigations, intra-rater reliability between reviews of nurses and physicians was determined. The time period for N1 and N2 was 13 to 14 days; 11 to 12 days for P1; and 14 days for P2. Using the same method as for inter-rater reliability, agreements and kappa values were determined. In all cases, agreements were calculated using the total number of cases in which decisions of

Table 1. Analytic framework for assessing reliability

	First stage (nurse review)	Second stage (physician review)
Inter-rater reliability	Agreement and kappa for the presence of each screening criteria Agreement and kappa for the determination of second-stage review necessity	Agreement and kappa for adverse event identification
Intra-rater reliability	Agreement and kappa for the presence of each screening criteria Agreement and kappa for the determination of second-stage review necessity	Agreement and kappa for adverse event identification

Table 2. Inter-rater reliability of nurse reviewers

Item No.	SC	SC type	Primary investigation (n=95)			Secondary investigation (n=54)		
			Agreement (%)	Kappa	95% CI	Agreement (%)	Kappa	95% CI
1	The index admission was an unplanned admission related to previous healthcare management	H	99	-		100	-	
2	Length of index was over 30 days	H	99	0.79	0.40, 1.00	100	-	
3	Unplanned readmission after discharge from index admission[1]	H	100	-		100	-	
4	Revisiting emergency room within 72 hours after discharge from index admission	G	96.8	0.71	0.40, 1.00	98.2	0.85	0.55, 1.00
5	Unplanned transfer to another acute care hospital	H	97.9	0.74	0.40, 1.00	98.2	0.66	0.03, 1.00
6	Temperature higher than 38.3°C at the point of discharge	H	99	-		100	-	
7	Unplanned transfer from general care to intensive care[1]	H, G	100	-		100	-	
8	Speciality consult	G	93.7	0.73	0.53, 0.94	88.9	0.35	-0.04, 0.74
9	Cardiac or respiratory arrest, rapid response team activation[1]	H	100	-		100	-	
10	Unexpected death[1]	H, G	100	-		100	-	
11	Healthcare-associated infection	H, G	97.9	0.49	-0.11, 1.00	98.2	-	
12	Hospital incurred patient injury[1]	H, G	100	-		98.2	-	
13	Over-sedation/hypotension	G	100	1.00		100	-	
14	Restraint use[1]	G	100	-		100	-	
15	Acute dialysis[1]	G	100	-		100	-	
16	In-unit procedure	G	96.8	0.87	0.73, 1.00	92.6	0.68	0.39, 0.97
17	Treatment for organ damage after an invasive procedure[1]	H	100	-		100	-	
18	Acute myocardial infarction, cerebrovascular accident, or pulmonary embolus during or after an invasive procedure[1]	H, G	100	-		100	-	
19	Transfusion or use of blood products	G	99	0.79	0.40, 1.00	98.2	0.66	0.03, 1.00
20	Avil (pheniramine) or Benadryl (diphenhydramine) use by intramuscular or intravenous rute	G	95.8	0.78	0.57, 0.99	98.2	0.9	0.70, 1.00
21	Abrupt medication stop	G	94.7	0.24	-0.15, 0.62	94.4	0.55	0.10, 0.99
22	Antidotes use	G	94.7	0.26	-0.19, 0.71	98.2	0.66	0.03, 1.00
23	Adverse drug reaction	H	98.9	0.66	0.04, 0.10	98.2	-	
24	Decrease in hemoglobin or hematocrit of 25% or greater	G	97.9	0.66	0.22, 0.10	94.4	-0.02	-0.06, 0.01
25	Hypoglycemic symptom	G	99	-		98.2	-	
26	Bleeding tendency	G	98.9	0.66	0.04, 1.00	98.2	-	
27	Rising BUN or serum creatine greater than 2 times baseline	G	99	-		100	-	
28	*Clostridium difficile* positive stool[1]	G	100	-		100	-	
29	Post-op troponin level greater than upper normal limit[1]	G	100	-		100	-	
30	Mechanical ventilation greater than 24 hours post-op[1]	G	100	-		100	-	
31	Unplanned return to the operating theatre[1]	H	100	-		100	-	
32	Unplanned removal, injury or repair of organ during surgery	H	99	-		98.2	-	
33	Intra-op epinephrine, norepinephrine, naloxone, or romazicon[1]	G	100	-		100	-	
34	Unplanned change in procedure or surgery[1]	H	100	-		100	-	
35	Intubation, reintubation, BiPap in post anesthesia care unit[1]	G	100	-		100	-	
36	X-ray in post anesthesia care unit[1]	G	100	-		100	-	
37	Terbutaline use in obstetrics	G	99	-		100	-	
38	Oxytocic agents in obstetrics	G	100	1.00		100	1.00	
39	Neonatal complications such as 5-minute Apgar score<6, or complication of abortion, amniocentesis or labor and delivery[1]	H, G	100	-		100	-	
40	Documentation or correspondence indicating litigation, dissatisfaction[1]	H	100	-		100	-	
41	Any other undesirable outcomes not covered above	H	91.6	0.16	-0.17, 0.50	90.7	0.26	-0.15, 0.68
	Second-stage review necessity		84.2	0.68	0.54, 0.83	77.8	0.55	0.33, 0.77

SC, screening criteria; CI, confidence interval; H, Harvard Medical Practice Study; G, Global Trigger Tool; BUN, blood urea nitrogen; op, operation; BiPap, bilevel positive airway pressure.

[1]Both nurse reviewers could not detect the presence of SC from all records.

Table 3. Intra-rater reliability of nurse reviewers

Item No.	SC	SC type	Nurse reviewer N1 (n=54)			Nurse reviewer N2 (n=54)		
			Agreement (%)	Kappa	95% CI	Agreement (%)	Kappa	95% CI
1	The index admission was an unplanned admission related to previous healthcare management	H	98.2	-		100.0	-	
2	Length of index was over 30 days	H	100.0	-		100.0	-	
3	Unplanned readmission after discharge from index admission[1]	H	100.0	-		100.0	-	
4	Revisiting emergency room within 72 hours after discharge from index admission	G	98.2	0.85	0.55, 1.00	96.3	0.65	0.19, 1.00
5	Unplanned transfer to another acute care hospital	H	100.0	1.00		100.0	1.00	
6	Temperature higher than 38.3°C at the point of discharge	H	100.0	-		100.0	-	
7	Unplanned transfer from general care to intensive care[1]	H, G	100.0	-		100.0	-	
8	Speciality consult	G	88.9	0.19	-0.22, 0.61	92.6	0.67	0.37, 0.97
9	Cardiac or respiratory arrest, rapid response team activation[1]	H	100.0	-		100.0	-	
10	Unexpected death[1]	H, G	100.0	-		100.0	-	
11	Healthcare-associated infection	H, G	98.2	-		100.0	-	
12	Hospital incurred patient injury[1]	H, G	98.2	-		100.0	-	
13	Over-sedation/hypotension	G	100.0	-		100.0	-	
14	Restraint use[1]	G	100.0	-		100.0	-	
15	Acute dialysis[1]	G	100.0	-		100.0	-	
16	In-unit procedure	G	100.0	1.00		96.3	0.85	0.66, 1.00
17	Treatment for organ damage after an invasive procedure[1]	H	100.0	-		100.0	-	
18	Acute myocardial infarction, cerebrovascular accident, or pulmonary embolus during or after an invasive procedure[1]	H, G	100.0	-		100.0	-	
19	Transfusion or use of blood products	G	100.0	1.00		98.2	0.66	0.63, 1.00
20	Avil (pheniramine) or Benadryl (diphenhydramine) use by intramuscular or intravenous route	G	92.6	0.47	0.05, 0.89	98.2	0.90	0.70, 1.00
21	Abrupt medication stop	G	96.3	-		96.3	0.73	0.38, 1.00
22	Antidotes use	G	98.2	0.66	0.03, 1.00	96.3	0.49	-0.11, 1.00
23	Adverse drug reaction	H	98.2	-		100.0	1.00	
24	Decrease in hemoglobin or hematocrit of 25% or greater	G	96.3	0.48	-0.13, 1.00	100.0	1.00	
25	Hypoglycemic symptom	G	100.0	1.00		100.0	-	
26	Bleeding tendency	G	100.0	1.00		100.0	-	
27	Rising BUN or serum creatine greater than 2 times baseline	G	100.0	-		100.0	-	
28	*Clostridium difficile* positive stool[1]	G	100.0	-		100.0	-	
29	Post-op troponin level greater than upper normal limit[1]	G	100.0	-		100.0	-	
30	Mechanical ventilation greater than 24 hours post-op[1]	G	100.0	-		100.0	-	
31	Unplanned return to the operating theatre[1]	H	100.0	-		100.0	-	
32	Unplanned removal, injury or repair of organ during surgery	H	100.0	-		100.0	1.00	
33	Intra-op epinephrine, norepinephrine, naloxone, or romazicon[1]	G	100.0	-		100.0	-	
34	Unplanned change in procedure or surgery[1]	H	100.0	-		100.0	-	
35	Intubation, reintubation, BiPap in post anesthesia care unit[1]	G	100.0	-		100.0	-	
36	X-ray in post anesthesia care unit[1]	G	100.0	-		100.0	-	
37	Terbutaline use in obstetrics	G	98.2	-		100.0	-	
38	Oxytocic agents in obstetrics	G	100.0	1.00		100.0	1.00	
39	Neonatal complications such as 5-minute Apgar score<6, or complication of abortion, amniocentesis or labor and delivery[1]	H, G	100.0	-		100.0	-	
40	Documentation or correspondence indicating litigation, dissatisfaction[1]	H	100.0	-		100.0	-	
41	Any other undesirable outcomes not covered above	H	96.3	-0.02	-0.05, 0.01	94.5	0.70	0.37, 1.00
	Second-stage review necessity		77.8	0.54	0.31, 0.77	83.3	0.67	0.47, 0.87

SC, screening criteria; CI, confidence interval; H, Harvard Medical Practice Study; G, Global Trigger Tool; BUN, blood urea nitrogen; op, operation; BiPap, bilevel positive airway pressure.

[1]Both nurse reviewers could not detect the presence of SC from all records.

both reviewers corresponded as a numerator and the total number of patients as a denominator.

SPSS version 21.0 (IBM Corp., Armonk, NY, USA) was used for all statistical analyses. The study was approved by the institutional review board of the National Evidence-based Healthcare Collaborating Agency (NECAIRB12-016-1).

RESULTS

Medical record review results are shown in Figure 1. The review was conducted on ninety-six patients discharged on three dates in 2007. Fifty-two patients were determined to have events satisfying at least one screening criterion by at least one nurse, and forty-four were determined by both nurses to meet none. Including the additional five individuals randomly selected from cases determined to have no screening criterion by reviews of the two nurses, physicians reviewed a total of fifty-seven patients. Eight patients were determined to have experienced AEs by both physician reviewers. Of the total 96 patients, 54 were re-investigated by nurses and 30 by physicians.

First, in looking at inter-rater reliability of nurses, agreements for the presence of each screening criterion ranged from 91.6% to 100% and the kappa values from 0.16 to 1.00 (Table 2). Due to the reviewers' lack of observed frequency, kappa values were not proposed for some SC. Nineteen of the SC were not found in any of the ninety-six patients during the first-stage review. SC yielding kappa values of 0.4 or lower in the primary investigation were No. 41 (kappa, 0.16), No. 21 (kappa, 0.24), and No. 22 (kappa, 0.26). The kappa value for determination of second-stage review necessity in the primary investigation was 0.68 (95% confidence interval [CI], 0.54 to 0.83). In the secondary investigation, SC with kappa values of 0.4 or lower were No. 24 (kappa, -0.02), No. 41 (kappa, 0.26), No. 8 (kappa, 0.35). In the secondary investigation, the kappa value for determination of second-stage review necessity was 0.55 (95% CI, 0.33 to 0.77).

For intra-rater reliability, agreements for the presence of each screening criterion ranged from 88.9% to 100% for N1, and the kappa values were -0.02 to 1.00 (Table 3). SC with kappa values lower than 0.4 were No. 41 (kappa, -0.02), No. 8 (kappa, 0.19). The kappa value for second-stage review necessity was 0.54 (95% CI, 0.31 to 0.77). In the case of N2, agreements for the presence of each screening criterion ranged from 90.7% to 100%, and the kappa values ranged from 0.49 to 1.00. No SC had kappa values of 0.4 or less. The kappa value for second-stage review necessity was 0.67 (95% CI, 0.47 to 0.87).

Looking at differences in reliability of reviews of nurses based on SC characteristics, the kappa values for reviews of nurses of SC used in the HMPS-type studies were 0.56 (95% CI, 0.32 to 0.80) for the primary investigation and 0.23 (95% CI, -0.11 to 0.57) for the secondary investigation (Table 4). On the other hand, the kappa values for reviews of nurses of the GTT-type SC were 0.72 (95% CI, 0.58 to 0.86) for the primary investigation

Table 4. Reliability of nurse reviewers by characteristics of screening criteria (SC)

SC type	Types of review	Agreement (%)	Kappa	95% CI
HMPS	Primary investigation	89.5	0.56	0.32, 0.80
	Secondary investigation	83.3	0.23	-0.11, 0.57
GTT	Primary investigation	86.3	0.72	0.58, 0.86
	Secondary investigation	81.5	0.62	0.40, 0.83
Total	Primary investigation	84.2	0.68	0.54, 0.83
	Secondary investigation	77.8	0.55	0.33, 0.77

CI, confidence interval; HMPS, Harvard Medical Practice Study; GTT, Global Trigger Tool.

Table 5. Reliability of physician reviewers for adverse event identification

	Types of review	Agreement (%)	Kappa	95% CI
Inter-rater reliability	Primary investigation (n=57)	93.0	0.71	0.44, 0.97
	Secondary investigation (n=30)	86.7	0.29	-0.16, 0.75
Intra-rater reliability	Physician reviewer P1 (n=30)	96.7	0.87	0.62, 1.00
	Physician reviewer P2 (n=30)	90.0	0.37	-0.16, 0.89
Total	Primary investigation	84.2	0.68	0.54, 0.83
	Secondary investigation	77.8	0.55	0.33, 0.77

CI, confidence interval.

and 0.62 (95% CI, 0.40 to 0.83) for the secondary investigation (Table 4).

The kappa values for inter-rater reliability of physicians for detecting AE occurrence were 0.71 (95% CI, 0.44 to 0.97) for the primary investigation and 0.29 (95% CI, -0.16 to 0.75) for the secondary investigation (Table 5). In addition, intra-rater reliability of P1 for detecting AE occurrence for the secondary investigation of 30 individuals was represented by the kappa value of 0.87 (95% CI, 0.62 to 1.00) and that of P2 by the kappa value of 0.37 (95% CI, -0.16 to 0.89) (Table 5).

DISCUSSION

In this study, medical records reviews were performed to identify inter-rater and intra-rater reliability on detection of SC and AE occurrence by reviewers. In the case of inter-rater reliability of nurses, agreements for decision of second-stage review necessity were 84.2% in the primary investigation and 77.8% in the secondary investigation, and the kappa values were 0.68 (95% CI, 0.54 to 0.83) for the primary investigation and 0.55 (95% CI, 0.33 to 0.77) for the secondary investigation. Previous studies looking at inter-rater reliability of nurses using the same method showed an agreement of 84% and kappa value of 0.67 [12], an agreement of 82% and kappa value of 0.62 (95% CI, 0.54 to 0.69) [13], and a kappa value of 0.73 [14], all similar to inter-rater reliability of nurses determined in this study.

For inter-rater reliability of physicians, agreements on AE occurrence identification were 93.0% for the primary investigation and 86.7% for the secondary investigation, and the kappa values were 0.71 (95% CI, 0.44 to 0.97) for the primary investigation and 0.29 (95% CI, -0.16 to 0.75) for the secondary investigation. Previous studies analyzing inter-rate reliability of physicians using the same methods showed an agreement of 89% and a kappa value of 0.61 [4]; an agreement of 79% and a kappa value of 0.4 (95% CI, 0.3 to 0.5) [15]; an agreement of 76% and a kappa value of 0.25 (95% CI, 0.05 to 0.45) [13]; and a kappa value of 0.74 [14]. Comparing former results to this study, inter-rater reliability of physicians is shown to be relatively higher in the primary investigation and lower in the secondary investigation. Furthermore, kappa values for the secondary investigation for both nurse and physician raters were lower compared to those for the primary investigation. As CIs overlap, one must be cautious in interpreting such findings as significant, and a repeat of medical record reviews will be needed for future studies.

For intra-rater reliability, the kappa values of all reviewers except P2 were 0.4 or higher, showing intermediate to good agreement for N1 and N2, and excellent agreement for P1 [16]. This may be attributed to differences in reviewers' medical record review experience. The 3 reviewers who showed intermediate to excellent agreement had more experience than the reviewer with a lower agreement level. As such, intra-rater reliability may increase as experience and training on medical record review for AE occurrence identification increase. In general, reliability of a measurement is affected by instrument variability, subject variability, and observer variability [17]. Reliability results in this study were analyzed considering such factors.

First, evaluation item characteristics were considered as instrument variability could have affected reliability. Intra-rater reliability of nurses was measured by determining the absence or presence of SC, which is less likely to involve subjective judgment of reviewers. Intra-rater reliability of physicians was determined by the absence or presence of AEs, which involves clinical decisions more so than determining the absence of presence of SC. Therefore, differences in identification of AEs were shown based on the clinical experience and knowledge of the physicians reviewers [18]. Considering such differences, setting specific examples of various AEs will be needed to ensure more consistent clinical decisions in physician reviewer training.

Second, as a subject variability, independence of medical record review affected reliability. When cases that met the SC were found, the nurse reviewers marked the corresponding medical records using a post-it for the convenience of physician reviewers. This could have given away clues while the nurses successively reviewed the medical records to find cases meeting the SC. Creating 2 copies of the medical records or utilizing electronic medical records with an anonymous system will solve this problem.

Third, as observer variability, experience and training of reviewers could have affected the reliability. Much like intra-rater reliability, differences in experience in reviewing medical records led to a reducing factor of inter-rater reliability. As the concept of AEs or SC are still new to health care professionals, relying only on reviewers with sufficient experience and training on patient safety-related medical record review will increase inter-rater reliability.

Overall, it can be indirectly inferred that studies determining AE incidence through medical record review can be conduct-

ed in South Korea (hereafter Korea). In addition, AE incidence in hospitals was found to be 8.3% (8/96), similar to the level in other studies [7,14]. This finding suggests that hospitals in Korea must pay more attention to patient safety and conduct additional studies to determine the size of patient safety problems within hospitals using sample patients or hospitals that can represent the whole nation.

Moreover, considering criticism on medical record review reliability [9], future studies must apply measures to increase reliability in determining AE occurrence. Such measures include standardizing measurement methods, training reviewers and raising their qualification levels, refining measurement tools, automating measurement tools, and repeating the measurements [17]. Considering these, measures to increase reliability in this study were determined as follows.

First, more thorough reviewer training will be needed to ensure their training and qualification levels. This study was thought to lack sufficient reviewer training, compared to other study, which conducted an education program for 3 days [19]. Reviewers with less experience and training in medical record review for detecting AEs are thought to be more greatly affected by it. In future studies, a much more thorough and standardized reviewer training should be conducted, such as reiterating the concept of AEs and repeatedly conducting medical record review exercises.

Second, to refine the measuring tool, clearer identification of each screening criterion will be needed. For example, for SC that may involve subjective evaluation, such as "No. 21. Abrupt medication stop", clarifying operational definitions will help in increasing inter-rater reliability [20]. In this study, the GTT-type SC showed a slightly higher trend of reliability of reviews of nurses compared to the HMPS-type SC relying more heavily on the clinical experience of reviewers. The case review form should be improved by including more GTT-type SC while ensuring no reduction in its sensitivity. Furthermore, adding a question judging AE occurrences in reviews of nurses to aid decision-making of physicians should be considered to increase inter-rater reliability of physicians. Introducing scoring system of short-answer questions to judge AE occurrences may be considered, like the questions on causality and preventability of AE [21].

This study has limitations. First, there is a limitation in evaluating inter-rater reliability of nurses on each screening criterion due to the sample size. Nineteen of the SC were not found even once in the process of reviewing medical records of 96 individuals. It was difficult to determine whether this was due to SC characteristics or the small sample size. However, SC not commonly seen in other studies, such as intra-operative or post-operative death, may not have been witnesses even with a bigger sample size due to their characteristics [20]. Furthermore, it was difficult to evaluate reliability and validity of decisions on specific details regarding AEs due to the sample size.

Second, as the medical record review was conducted in a single hospital, hospital characteristics may have had an effect. For example, this study largely lacked SC related to surgeries; therefore, inter-rater reliability for such SC was difficult to evaluate. Future studies should further evaluate inter-rater reliability of nurses for specific SC. Moreover, the quality of medical records in the hospital may have affected the medical record review, which may ultimately impact the inter-rater reliability.

In conclusion, although reliability was not good in certain cases, intermediate or higher levels were shown for inter-rater and intra-rater reliability in AE identification through medical record review. In the future, when conducting a large-scale medical record review for AE identification, measures such as reviewer training enhancement and further clarification of the definition of SC should be considered to increase inter-rater reliability in detecting AEs.

ACKNOWLEDGEMENTS

This study was supported by a grant from the National Evidence-based Healthcare Collaborating Agency (no. NECA-M-12-002).

CONFLICT OF INTEREST

The authors have no conflicts of interest with the material presented in this paper.

REFERENCES

1. World Health Organization, The research cycle: measuring harm [cited 2014 Sep 23]. Available from: http://www.who.int/patientsafety/research/strengthening_capacity/measuring_harm/en/.
2. Thomas EJ, Petersen LA. Measuring errors and adverse events in health care. J Gen Intern Med 2003;18(1):61-67.
3. World Health Organization. Assessing and tackling patient harm: a methodological guide for data-poor hospitals. Gene-

va: World Health Organization; 2010, p. 35-36.
4. Brennan TA, Leape LL, Laird NM, Hebert L, Localio AR, Lawthers AG, et al. Incidence of adverse events and negligence in hospitalized patients. Results of the Harvard Medical Practice Study I. N Engl J Med 1991;324(6):370-376.
5. Rivard PE, Luther SL, Christiansen CL, Shibei Zhao, Loveland S, Elixhauser A, et al. Using patient safety indicators to estimate the impact of potential adverse events on outcomes. Med Care Res Rev 2008;65(1):67-87.
6. Leape LL, Brennan TA, Laird N, Lawthers AG, Localio AR, Barnes BA, et al. The nature of adverse events in hospitalized patients. Results of the Harvard Medical Practice Study II. N Engl J Med 1991;324(6):377-384.
7. de Vries EN, Ramrattan MA, Smorenburg SM, Gouma DJ, Boermeester MA. The incidence and nature of in-hospital adverse events: a systematic review. Qual Saf Health Care 2008;17(3):216-223.
8. Vincent C. Patient safety. 2nd ed. Chichester: Wiley-Blackwell; 2010, p. 49-74.
9. 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.
10. 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.
11. Lee SI, Kim Y, Lee JH, Lee JY, Jo MW, Ock M, et al. Korean protocol development and assessment to secure the safety of patients. Seoul: National Evidence-based Healthcare Collaborating Agency; 2012, p. 144-201 (Korean).
12. Wilson RM, Runciman WB, Gibberd RW, Harrison BT, Newby L, Hamilton JD. The quality in Australian health care study. Med J Aust 1995;163(9):458-471.
13. Zegers M, de Bruijne MC, Wagner C, Hoonhout LH, Waaijman R, Smits M, et al. Adverse events and potentially preventable deaths in Dutch hospitals: results of a retrospective patient record review study. Qual Saf Health Care 2009;18(4):297-302.
14. Hwang JI, Chin HJ, Chang YS. Characteristics associated with the occurrence of adverse events: a retrospective medical record review using the Global Trigger Tool in a fully digitalized tertiary teaching hospital in Korea. J Eval Clin Pract 2014;20(1):27-35.
15. Aranaz-Andrés JM, Aibar-Remón C, Vitaller-Murillo J, Ruiz-López P, Limón-Ramírez R, Terol-García E, et al. Incidence of adverse events related to health care in Spain: results of the Spanish National Study of Adverse Events. J Epidemiol Community Health 2008;62(12):1022-1029.
16. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33(1):159-174.
17. Hulley SB, Cummings SR, Browner WS, Grady D, Newman TB. Designing clinical research. 4th ed. Philadelphia: Lippincott Williams & Wilkins; 2013, p. 32-42.
18. Mendes W, Martins M, Rozenfeld S, Travassos C. The assessment of adverse events in hospitals in Brazil. Int J Qual Health Care 2009;21(4):279-284.
19. Soop M, Fryksmark U, Köster M, Haglund B. The incidence of adverse events in Swedish hospitals: a retrospective medical record review study. Int J Qual Health Care 2009;21(4):285-291.
20. Naessens JM, O'Byrne TJ, Johnson MG, Vansuch MB, McGlone CM, Huddleston JM. Measuring hospital adverse events: assessing inter-rater reliability and trigger performance of the Global Trigger Tool. Int J Qual Health Care 2010;22(4):266-274.
21. Davis P, Lay-Yee R, Briant R, Ali W, Scott A, Schug S. Adverse events in New Zealand public hospitals II: preventability and clinical context. N Z Med J 2003;116(1183):U624.

Supplemental Table 1. Types of screening criteria (SC)

Item No.	SC	SC type
1	The index admission was an unplanned admission related to previous healthcare management	HMPS
2	Length of index was over 30 days	HMPS
3	Unplanned readmission after discharge from index admission	HMPS
4	Revisiting emergency room within 72 hours after discharge from index admission	GTT
5	Unplanned transfer to another acute care hospital	HMPS
6	Temperature higher than 38.3˚C at the point of discharge	HMPS
7	Unplanned transfer from general care to intensive care	HMPS & GTT
8	Speciality consult	GTT
9	Cardiac or respiratory arrest, rapid response team activation	HMPS
10	Unexpected death	HMPS & GTT
11	Healthcare-associated infection	HMPS & GTT
12	Hospital incurred patient injury	HMPS & GTT
13	Over-sedation/hypotension	GTT
14	Restraint use	GTT
15	Acute dialysis	GTT
16	In-unit procedure	GTT
17	Treatment for organ damage after an invasive procedure	HMPS
18	Acute myocardial infarction, cerebrovascular accident, or pulmonary embolus during or after an invasive procedure	HMPS & GTT
19	Transfusion or use of blood products	GTT
20	Avil (pheniramine) or Benadryl (diphenhydramine) use by intramuscular or intravenous route	GTT
21	Abrupt medication stop	GTT
22	Antidotes use	GTT
23	Adverse drug reaction	HMPS
24	Decrease in hemoglobin or hematocrit of 25% or greater	GTT
25	Hypoglycemic symptom	GTT
26	Bleeding tendency	GTT
27	Rising BUN or serum creatine greater than 2 times baseline	GTT
28	*Clostridium difficile* positive stool	GTT
29	Post-op troponin level greater than upper normal limit	GTT
30	Mechanical ventilation greater than 24 hours post-op	GTT
31	Unplanned return to the operating theatre	HMPS
32	Unplanned removal, injury or repair of organ during surgery	HMPS
33	Intra-op epinephrine, norepinephrine, naloxone, or romazicon	GTT
34	Unplanned change in procedure or surgery	HMPS
35	Intubation, reintubation, BiPap in post anesthesia care unit	GTT
36	X-ray in post anesthesia care unit	GTT
37	Terbutaline use in obstetrics	GTT
38	Oxytocic agents in obstetrics	GTT
39	Neonatal complications such as 5-minute Apgar score<6, or complication of abortion, amniocentesis or labor and delivery	HMPS & GTT
40	Documentation or correspondence indicating litigation, dissatisfaction	HMPS
41	Any other undesirable outcomes not covered above	HMPS

HMPS, Harvard Medical Practice Study; GTT, Global Trigger Tool; BUN, blood urea nitrogen; op, operation; BiPap, bilevel positive airway pressure.