Postprint Version	1.0
Journal website	http://dx.doi.org/10.1016/j.jclinepi.2009.03.004
Pubmed link	http://www.ncbi.nlm.nih.gov/pubmed/19473812
DOI	10.1016/j.jclinepi.2009.03.004

This is a NIVEL certified Post Print, more info at http://www.nivel.eu

The inter-rater agreement of retrospective assessments of adverse events does not improve with two reviewers per patient record

MARIEKE ZEGERS[a],*, MARTINE C. DE BRUIJNE[b], CORDULA WAGNER[a,b], PETER P. GROENEWEGEN[a,c,d], GERRIT VAN DER WAL[b,e], HENRICA C.W. DE VET[b,f]

[a]NIVEL, Netherlands Institute for Health Services Research, P.O. box 1568, 3500 BN Utrecht, The Netherlands
[b]Department of Public and Occupational Health, EMGO Institute, VU University Medical Center, Amsterdam, The Netherlands
[c]Department of Sociology, Utrecht University, The Netherlands
[d]Department of Human Geography, Utrecht University, The Netherlands
[e]Netherlands Health Care Inspectorate
[f]Department of Epidemiology and Biostatistics, EMGO Institute, VU University Medical Center, Amsterdam, The Netherlands

ABSTRACT

Objective: To evaluate the inter-rater agreement of the record review process of the Dutch Adverse Event study, which we aimed to improve by the involvement of two independent physician reviewers per record instead of one including a consensus procedure in case of disagreement.

Methods: The inter-rater agreement within pairs of physicians (independent review between physician A+B) and between pairs of physicians (independent review between physician A+B and C+D) was measured to evaluate the record review process with two physicians including a consensus procedure, with 4,272 and 119 records, respectively.

Results: The inter-rater agreement within pairs of physicians was substantial for the determination of adverse events (AEs) with a kappa of 0.64 (95% confidence interval [CI]: 0.61, 0.68). The inter-rater agreement between pairs of physicians was fair for the determination of AEs with a kappa of 0.25 (95% CI: 0.05, 0.45).

Conclusion: A record review process with two physicians per record including a consensus procedure to assess AEs is not more reliable than a record review process with one physician. Retrospective estimates of incidence of AEs from record review studies should be interpreted with caution. Improvement of the method is necessary for monitoring incidence of AEs over time at a national level.

1. INTRODUCTION

What's new ?

• The inter-rater agreement of record review to assess adverse events in hospital admissions does not improve by involvement of two independent physician reviewers per patient record including a consensus procedure in case of disagreement, instead of a single physician reviewer. Involving two physicians in the assessment of adverse events including a consensus procedure gives a false feeling of higher reliability.

• A record review process with two physician reviewers per record and a consensus procedure led, however, to more reported adverse events than a record review process with only one physician reviewer per record. Physician reviewers were more reluctant in their judgement without support of a collegial review.

If the aim is to investigate quality improvement, two reviewers may be preferred over one, to maximize the amount of information for quality improvement. But for routine assessment of adverse events, one physician reviewer per record may be considered and makes the assessment of adverse events more efficient and cheaper.

Patient record review of hospital admissions is by far the most widely applied and thoroughly studied method for measurement of patient safety [1], [2], [3], [4], [5], [6], [7], [8], [9] and [10]. It is a standard method by which adverse events (AEs) of clinical care and their degree of preventability are measured and it forms the basis for patient safety policy in several countries [11]. This method was proven valid to identify AEs and estimate their incidence in hospitals nationwide [2]. However, previous AE studies showed poor to moderate inter-rater agreement for the determination of AEs and their preventability [1], [2], [3], [5], [7], [8], [9] and [10]. Therefore, standing on the shoulders of our predecessors and keeping the method and instruments maximally comparable, we have tried to improve the inter-rater agreement of the measurement of AEs and their preventability within the Dutch Adverse Event study.

Inter-rater agreement refers to the consistency of ratings or to the ability of various raters to reach the same conclusion about a specific case [2] and [12]. Strategies to enhance inter-rater agreement are standardization of the measurement and consensus procedure between the reviewers [12] and [13]. To improve the inter-rater agreement for the assessment of AEs in the Dutch Adverse Event study all records were independently reviewed by two physicians instead of one and in case of disagreement, the two physicians discussed and reconsidered their review to obtain consensus. We hypothesized that the involvement of two physicians per patient record including a consensus procedure would give a more reliable assessment of AEs and their preventability. Within the Dutch Adverse Event study a reliability study was conducted to evaluate the inter-rater agreement of the patient record review. The objective was twofold. First, to examine the inter-rater agreement of the original review by two independent physician reviewers before the consensus procedure. This is called the inter-rater agreement within pairs of physicians (physician A vs. B). Second, to examine the inter-rater agreement of the complete record review process, including the consensus procedure, with a second pair of physicians. This is called the inter-rater agreement between pairs of physicians (physician A+B vs. C+D). The Harvard Medical Practice Study in the United States and the Australian study on the occurrence of AEs also involved two physician reviewers and the Australian study also used a consensus procedure in case of disagreement between the two physicians [3] and [10]. However, these studies only evaluated the inter-rater agreement of the original review within pairs of physicians (physician A and B) and not of the ultimate decisions made by pair of physicians. To gain insight in the reliability of the record review procedure with two physicians per patient record including a consensus procedure in case of disagreement, the inter-rater agreement between pairs of physicians is more relevant and has not yet been studied thoroughly.

2. METHODS

2.1. Study design and setting

A retrospective patient record review study was conducted to determine the incidence and preventability of AEs among hospitalized patients in the Netherlands [14]. The method of this study was based on a protocol and instruments originally developed by the Harvard Medical Practice Study. They studied the incidence of AEs in New York state hospitals in 1984, based on analysis of information in patient records [3] and [15]. This method, with modifications, was used in subsequent studies in Australia, the United Kingdom, New Zealand, the United States, Denmark, France, and Canada [1], [5], [8], [10], [15], [16], [17], [18] and [19]. This reliability study was conducted as part of the Dutch Adverse Event study. We have described the design and method of our main study in more detail elsewhere [14]. A random sample of 7,926 patient records, originating from 21 randomly selected Dutch hospitals, was reviewed using a three-stage retrospective patient record review process by trained nurses and physicians between August 2005 and October 2006 (Fig. 1). The sample of records (N = 7,926) was stratified for discharged and deceased hospital patients—3,943 records of discharged patients and 3,983 records of deceased patients in 2004. A large subsample of deceased hospital patients was included to determine the occurrence of potentially preventable deaths in hospitals more precisely than in previous studies [20].

[FIGURE 1]

2.2. Original record review process

In the first stage of the three-stage review process, a nurse screened the patient records by using 18 screening criteria indicating potential AEs. One or more criteria were fulfilled in 4,317 patient records, of which 65% were from deceased patients and 35% were from discharged patients. These were forwarded to the second stage of the review process for a medical review by two physicians independently to determine whether an AE had occurred and whether the AE had been preventable. An AE was defined as an unintended injury among hospitalized patients that results in disability, death, or prolonged hospital stay, and was caused by health care management. Preventability of an AE is defined as health care that fell below the current level of expected performance for practitioners or systems (Appendix A) [1], [5], [10] and [15].

The patient records were independently reviewed by two physicians of the same specialty (general internists, general surgeons, neurologists, or pediatricians). The physician reviewers had to focus on their own expertise. Records with, for example, screening criteria related to surgical events were reviewed by two surgeons. Moreover, records with screening criteria related to multiple specialties were reviewed by two different specialists.

In case of disagreement about the presence of an AE and/or degree of preventability between the two independent reviews of the physicians, they started a consensus procedure (stage 3). In this consensus procedure the physicians considered and discussed both reviews and reconsidered their reviews to obtain consensus. When they failed to reach agreement, a third trained reviewer gave a final judgment based on information of the first two reviews [14]. In 663 hospital admissions one or more AEs were identified (Fig. 1) [20].

2.3. Reliability study

2.3.1. Measurement inter-rater agreement within pairs

The inter-rater agreement of the original review between physician A and B (before consensus procedure), defined as the inter-rater agreement within pairs of physicians, was determined for 4,272 records; 45 records were excluded, because the AEs had occurred outside the participating hospitals (thus in another hospital) or were not related to the sampled admission (Fig. 1).

2.3.2. Measurement inter-rater agreement between pairs

To assess the inter-rater agreement of the complete review process by pairs of physicians, including the consensus procedure and if applicable a third review, a stratified random sample of 119 records was selected for an independent review by a second pair of physicians (physician C and D) (Fig. 1). Physician C and D independently reviewed the selected records and in case of disagreement, they started a consensus procedure and if applicable, a third reviewer gave a final judgment. Only records reviewed by two internists or two surgeons were selected.

Because the inter-rater agreement may vary for records that did and did not need a consensus procedure to decide about the presence of AEs, the sample was stratified for records with or without a consensus procedure between physician A and B. As most of the records did not show an AE, for efficiency reasons we aimed to include about an equal number of records with and without an AE in the sample (Table 1).

[TABLE 1]

2.4. Physician reviewers and training

In this study, 55 trained physicians reviewed in several different hospitals (average 5.2 hospitals per physician). The eligibility criteria for physicians to act as reviewer were as follows: more than 10 years post graduate general clinical experience, good reputation among colleagues, no longer than 5 years retired, experience or affinity with analysis of incidents, complaints and errors of clinical care, and availability for at least 1 day per week.

The physicians followed a 1-day training in small groups (maximum 12 participants), led by one researcher and one experienced physician. During the training, the study protocol, definitions, and electronic review forms were explained and examples of (preventable) AEs were discussed. The reviewers practiced with cases and the review forms and they were provided with a review manual in which the research protocol, instruments, and definitions were defined [14].

During the study, the physicians could consult an expert panel of medical specialists for questions about accepted clinical practice. After 1 month of reviewing, the reviewers had a half-day training session to discuss their problems concerning the review process and reviewers were updated with the latest insights about the review process. These training sessions were organized frequently during data collection. The discussed problems were collected and noted in a regularly updated Frequently Asked Questions (FAQ) document, which was spread via post and mail to all reviewers.

2.5. Statistical analysis

The judgment about the presence of an AE and degree of preventability were measured on a six-point scale (Appendix A). We used a score of at least four (>50% chance that medical management caused the AEs) to indicate the presence of AEs. For preventability we used a score two or higher (at least slight to modest evidence that the AE was preventable) [20].

The inter-rater agreement within pairs of physicians (physician A and B) and between pairs of physicians (physician A+B and C+D) was measured for the determination of AEs and for the determination of the degree of preventability of the AEs. Because no preventability score could be given for records without AEs, the inter-rater agreement for preventability was only estimated for records in which both (pairs of) physicians found an AE. The inter-rater agreement between pairs of physicians was adjusted for the stratified sampling procedure with respect to the oversampling of records with AEs and the oversampling of the presence of a consensus procedure.

The inter-rater agreement was expressed as a kappa (κ) statistic with 95% confidence interval (CI) and as the percentage of records for which there was agreement. A κ-value between 0.00 and 0.20 was classified as “slight”; between 0.21 and 0.40 as “fair”; between

0.41 and 0.60 as “moderate”; between 0.61 and 0.80 as “substantial”; and between 0.81 and 1.00 as “almost perfect” [21]. Data were analyzed using SPSS 14.0 for Windows.

3. RESULTS

3.1. Inter-rater agreement within pairs of physicians

The inter-rater agreement within pairs of physicians (physician A and B) was determined for 2,757 (65%) records of deceased patients and for 1,515 (35%) records of discharged patients. The inter-rater agreement for the determination of AEs was substantial ($\kappa = 0.64$, 95% CI: 0.61, 0.68). Also for the determination of their preventability the inter-rater agreement was substantial ($\kappa = 0.72$, 95% CI: 0.66, 0.79) (Table 2).

[TABLE 2]

Physician A and physician B separately determined 592 and 621 AEs before a consensus procedure (Table 2). After discussion and reconsideration of their reviews in a consensus procedure more AEs were determined (n = 663). Figure 1 shows that 373 (213 + 105 + 55) records were discussed in a consensus procedure between physician A and B about the determination of AEs and/or degree of preventability or were finally judged by third physician reviewer. This procedure resulted in 243 (213 + 30) AEs. Of all detected AEs, 37% (243/663) were determined after consensus procedure.

3.2. Inter-rater agreement between pairs of physicians

The inter-rater agreement between pairs of physicians (physician A+B and C+D) was determined for 77 (65%) records of deceased patients and for 42 (35%) records of discharged patients. The inter-rater agreement was fair for determination of presence of AEs ($\kappa = 0.25$, 95% CI: 0.05, 0.45) and for determination of preventability of AEs ($\kappa = 0.40$, 95% CI: 0.07, 0.73) (Table 3).

[TABLE 3.]

The second pair of physicians determined less AEs (n = 40) compared with the first pair of physicians (n = 62) (Table 3). For 46 records there was no agreement between the pairs of physicians. For 21 (46%) of these records a consensus procedure was needed between physician A and B.

The inter-rater agreement for AE determination and their preventability between pairs of physicians was lower than the inter-rater agreement within pairs of physicians. The kappa value for the AE determination was lower than the kappa value for the determination of the preventability of AEs.

3.3. Subgroup analysis

To get an indication of determinants for high or low agreement for the assessment of AEs we performed a post-hoc analysis of the inter-rater agreement separately for subgroups of records. The inter-rater agreement within pairs of physicians was higher for records of discharged patients compared with records of deceased patients. For records reviewed by two neurologists the inter-rater agreement within pairs of physicians was higher compared with records reviewed by two internists, surgeons, or pediatricians. The inter-rater agreement within pairs of physicians was higher for records that were reviewed by two physicians both of whom reviewed many records compared with records reviewed by physicians both of whom reviewed less records. However, the differences in kappa values were not statistically significant, except the kappa value within pairs of neurologists (Table 4).

[TABLE 4.]

We also analyzed the inter-rater agreement between pairs of physicians for subgroups of records. The inter-rater agreement between pairs of physicians was also higher for records of discharged patients ($\kappa = 0.55$, 95% CI: 0.16, 0.94) compared with records of deceased patients ($\kappa = 0.14$, 95% CI: −0.09, 0.36). For records reviewed by two internists the inter-rater agreement was higher ($\kappa = 0.27$, 95% CI: −0.01, 0.55) compared with records reviewed by two surgeons ($\kappa = 0.17$, 95% CI: −0.14, 0.47). The inter-rater agreement between pairs of physicians for records with a consensus procedure between physician A and B was lower ($\kappa = 0.07$, 95% CI: −0.43, 0.58) than the records without a consensus procedure ($\kappa = 0.23$, 95% CI: 0.01, 0.45). However, most of these kappa values had wide CI because of the small number of records in these subgroup analyses, meaning that these results were not statistically significant.

4. Discussion

We hypothesized that the involvement of two physicians per patient record including a consensus procedure in case of disagreement between their reviews would improve the reliability of the review process to assess AEs. However, the inter-rater agreement of the complete medical review process (inter-rater agreement between pairs of physicians), including the consensus procedure, was only fair, although the inter-rater agreement within pairs of physicians was substantial.

More consensus procedures during the study between the same physicians has probably led to more simultaneous reviews that increased the inter-rater agreement within pairs of physicians. The Dutch Adverse Event study measured the incidence of AEs at national level. For geographical reasons, physicians often reviewed in the same region and hospitals. After a number of independently reviewed records, physicians had consensus procedures to obtain consensus in case there was disagreement. If physicians reviewed in the same hospital, they may often have had consensus procedures with the same colleagues. The second pair of physicians (physician C+D) reviewed in another region and had seldom or never consensus procedures with physicians of the first pair (physician A+B). The within pair consensus procedure produced a pair-specific improvement in agreement but not in overall reliability between pairs of physicians who were part of different discussions. It even had negative consequences for the inter-rater agreement between pairs of physicians. Perhaps the consensus procedure did not improve the inter-rater agreement of assessment of AEs, because there was not enough mixture of reviewers within and across pairs. A possible explanation for the fact that a consensus procedure did not improve the assessment is that for a group judgment task (in this study the discussion between the two physicians during a consensus procedure), discussion is primarily used as a justification for the members' original positions rather than contributing any input to a group decision.[22] However, we did not find dominant physician reviewers who convinced other physician reviewers more often during consensus procedures. The physician reviewers evaluated the double review procedure and the consensus procedure to be meticulous and instructive. However, this comfort is deceptive and may lead to unwarranted confidence in the result.[22]

The results also showed that more AEs were found after the consensus procedure than with two independent reviews by the two physicians. One-third of all AEs (37%) were determined using a consensus procedure about the presence of AEs and/or their preventability. This implies that it was hard for the physician reviewers to judge about the presence of AEs and that physicians were more reluctant in their judgment without support of a collegial review.

Another finding was the higher kappa value for the assessment of preventability than for the assessment of AEs. An explanation is that the inter-rater agreement of preventability could only be estimated for records in which both (pairs of) physicians found an AE. Another explanation is that in the Dutch study the threshold for preventability was two and higher (Appendix A). Previous studies that maintained four and higher as a threshold for preventability scores showed lower kappa values for the assessment of preventability.

Poor inter-rater agreement could be caused by a lack of information or knowledge necessary to appropriately determine AEs [23]. Judgments require not only up-to-date clinical knowledge, but also consideration of standards of care and the recognition of distinction between AEs and unintended outcomes caused by the disease or patient condition [24]. The determination of AEs and their preventability in this study is based on a structured implicit method that relies on expert judgment. The structured part is that reviewers are guided in determining AEs and their preventability. The review is, however, implicit in that the reviewers are asked to judge on basis of relatively uncodified knowledge, held in their minds and perhaps tailored to the circumstances of a specific case [25]. The measurement procedure in implicit review requires the reviewers to form their criteria and apply them, and thus a source of variability is included in the measurement of reliability. Explicit methods for AEs assessment based on clearly defined criteria showed higher inter-rater agreement compared with implicit methods [25]. However, explicit methods are only applicable in the case of selected, homogeneous samples of cases. In the Dutch Adverse Event study, admissions were selected hospital-wide with a wide variation of diagnosis and treatments. Many hospital patients suffer from multiple and complex diseases and need complex treatment. A “gold standard” on good clinical practice is often lacking for each unique individual patient within his unique context. In addition, the AEs showed a wide range of origins and outcomes. So even if the standardized review process was perfectly applied by all reviewers, one would still expect a certain amount of disagreement about presence of AEs.

Poor inter-rater agreement could also be caused by lack of review experience. The inter-rater agreement was higher for records reviewed by physicians with more experience (assessed by the number of reviewed records) compared with records reviewed by physicians with less experience. On an average, a small group of neurologists (n = 5) reviewed more records per reviewer compared with the group internists (n = 25) or surgeons (n = 20). The inter-rater agreement of record reviewed by neurologists was higher compared with the inter-rater agreement of records reviewed by internists or surgeons. The judgment about the presence of AEs probably became more standardized within a smaller group of reviewers.

Low inter-rater agreement may imply an over- or underestimation of AEs in the national study on the occurrence of AEs in Dutch hospitals. Policy makers, hospital managers, and health care workers should be aware of this and interpret the results with caution. The moderate reliability of the review process is a well-known problem of record review studies to identify AEs and their preventability, in which kappa values ranged from 0.2 to 0.6 [3], [5], [7], [9] and [10]. The Harvard Medical Practice Study and the Australian study also involved two physicians for the medical review [3] and [10]. They found a kappa value of 0.61 and 0.55 for the AE identification, respectively. However, these studies only evaluated the inter-rater agreement within pairs of physicians. In our study, the inter-rater agreement within pairs of physicians was substantially higher compared with the inter-rater agreement between pairs of physicians. Thus, the inter-rater agreement presented in the previous studies limited to within pair agreement between two physicians could be an overestimation. However, the value of kappa depends on the prevalence of AEs [26] and [27]. Good care, meaning low AE rates, is likely to be associated with lower values of kappa [25]. This impairs comparison of kappa values between populations.

Our study had some limitations. We included more patient records of deceased patients compared with patient records of discharged patients in the reliability study. The proportion of deceased and discharged patients in the Dutch hospital population was 3% vs. 97% and in the sample for the determination of the inter-rater agreement between pairs of physicians, the proportion was 65% vs. 35%. The post-hoc analysis showed that the inter-rater agreement for records of discharged patients is much better compared with the records of deceased patients. Therefore, the kappa values of the inter-rater agreement between physicians may be underestimated in this study. However, because of the small number of records in the

subgroup analysis for discharged and deceased patients and the wide 95% CI, we refrained from adjustment for the oversampling of deceased patients.

Secondly, we selected 119 records to evaluate the inter-rater agreement between pairs of physicians including consensus procedure. The number of records was too small for appropriate analysis of the inter-rater agreement for subgroups of records.

Finally, we involved 55 physicians, which is more than in other studies. The more the heterogeneity in the raters and the conditions studied, the lower the reliability will be [6]. Table 4 showed that the inter-rater agreement for the internists (n = 25) was lower than the inter-rater agreement for the five neurologists who reviewed all records with neurologic events. Also the inter-rater agreement for physicians who reviewed more records was better than the inter-rater agreement for physicians who reviewed less records. However, for logistic reasons, many raters will be needed to estimate the incidence of AEs at a national level.

The involvement of a second physician and implementation of a consensus procedure did not improve the reliability of the patient record review method for the measurement of AEs. In future record review studies on the occurrence of AEs, one physician reviewer per record may be considered and makes the study much cheaper.

The suboptimal reliability of patient record review to identify AEs should be further improved to monitor patient safety in hospitals and hospital departments over time at national level. To improve the reliability a more explicit method based on specified and detailed checklists (using standards) for specific departments or patient groups may offer a solution. An approach that combines record review with prospective methods, in which clinical staff is interviewed about the origin of the AE, may also help to improve the inter-rater agreement to assess AEs. The team of physicians could be extended with more specialists from different disciplines, for example, cardiologists and neurosurgeons. However, the overall number of reviewers should be reduced to increase the experience per reviewer and to facilitate intensive training and standardization of the process. A wider spread of reviewers over hospitals across the country may help to avoid artificial enlargement of the inter-rater agreement within pairs of physicians. Also more training sessions during the study with all physician reviewers together should be organized devoted to comparison of reviews to standardize the review process and to enhance the overall inter-rater agreement.

5. Conclusion

Although judgment of presence of AEs is difficult, retrospective patient record studies currently offer the best method available to assess the incidence of AEs and their preventability, nature, and types [6]. The results of record review studies provide urgently needed insight in the current state of patient safety and possibilities for improvement of patient safety and are therefore generally highly appreciated.

Involvement of two physicians per patient record and consensus procedure in case of disagreement between physicians did not improve the inter-rater agreement of patient record review for the assessment of AEs. However, a record review process with two physician reviewers per record and a consensus procedure led to more reported AEs compared with a record review process with only one physician reviewer per record. If the aim is to investigate quality improvement, two reviewers may be preferred over one, to maximize the amount of information for quality improvement. But for routine assessment of AEs, one physician reviewer per record may be considered and makes the assessment of AEs more efficient and cheaper.

Retrospective estimates of incidence data of AEs should be interpreted with caution. Improvement of record review is necessary for monitoring incidence of AEs over time at a national level.

ACKNOWLEDGMENTS

The authors thank everyone who contributed to the study—the physicians who reviewed the patient records; the researchers for the coordination of the data collection; and the 21 participating hospitals and their staff who facilitated the patient records.

Funding: The Dutch Patient Safety Research Program has been initiated by the Dutch Society of Medical Specialists (in Dutch: Orde van Medisch Specialisten) and the Dutch Institute for Health care Improvement (CBO) with financial support from the Ministry of Health, Welfare, and Sport. The Program is carried out by EMGO Institute/VUmc and NIVEL.

REFERENCES

[1] G.R. Baker, P.G. Norton, V. Flintoft, R. Blais, A. Brown and J. Cox et al., The Canadian Adverse Events Study: the incidence of adverse events among hospital patients in Canada, CMAJ 170 (2004), pp. 1678–1686.

[2] T.A. Brennan, R.J. Localio and N.L. Laird, Reliability and validity of judgments concerning adverse events suffered by hospitalized patients, Med Care 27 (1989), pp. 1148–1158.

[3] T.A. Brennan, L.L. Leape, N.M. Laird, L. Hebert, A.R. Localio and A.G. Lawthers et al., Incidence of adverse events and negligence in hospitalized patients. Results of the Harvard Medical Practice Study I, N Engl J Med 324 (1991), pp. 370–376.

[4] P. Davis, R. Lay-Yee, S. Schug, R. Briant, A. Scott and S. Johnson et al., Adverse events regional feasibility study: methodological results, N Z Med J 114 (2001), pp. 200–202.

[5] P. Davis, R. Lay-Yee, R. Briant, W. Ali, A. Scott and S. Schug, Adverse events in New Zealand public hospitals I: occurrence and impact, N Z Med J 115 (2002), p. U271.

[6] R.J. Lilford, M.A. Mohammed, D. Braunholtz and T.P. Hofer, The measurement of active errors: methodological issues, Qual Saf Health Care 12 (Suppl. 2) (2003), pp. ii8–ii12.

[7] A.R. Localio, S.L. Weaver, J.R. Landis, A.G. Lawthers, T.A. Brennan and L. Hebert et al., Identifying adverse events caused by medical care: degree of physician agreement in a retrospective chart review, Ann Intern Med 125 (1996), pp. 457–464.

[8] E.J. Thomas, D.M. Studdert, H.R. Burstin, E.J. Orav, T. Zeena and E.J. Williams et al., Incidence and types of adverse events and negligent care in Utah and Colorado, Med Care 38 (2000), pp. 261–271.

[9] E.J. Thomas, D.M. Studdert and T.A. Brennan, The reliability of medical record review for estimating adverse event rates, Ann Intern Med 136 (2002), pp. 812–816.

[10] R.M. Wilson, W.B. Runciman, R.W. Gibberd, B.T. Harrison, L. Newby and J.D. Hamilton, The quality in Australian Health Care Study, Med J Aust 163 (1995), pp. 458–471.

[11] G.R. Baker, Commentary. Harvard medical Practice Study, Qual Saf Health Care 13 (2004), pp. 151–152. [12] D.L. Streiner and G.R. Norman, Health measurement scales: a practical guide to their development and use, Oxford University Press, Oxford (1999).

[13] H.C.W de Vet, C.B. Terwee and L.M. Bouter, Current challenges in clinimetrics, J Clin Epidemiol 56 (2003), pp. 1137–1141

[14] M. Zegers, M.C. de Bruijne, C. Wagner, P.P. Groenewegen, R. Waaijman and G. van der Wal, Design of a retrospective patient record study on the occurrence of adverse events among patients in Dutch hospitals, BMC Health Serv Res 7 (2007).

[15] L.L. Leape, T.A. Brennan, N. Laird, A.G. Lawthers, A.R. Localio and B.A. Barnes et al., The nature of adverse events in hospitalized patients. Results of the Harvard Medical Practice Study II, N Engl J Med 324 (1991), pp. 377–384.

[16] P. Michel, J.L. Quenon, A.M. de Sarasqueta and O. Scemama, Comparison of three methods for estimating rates of adverse events and rates of preventable adverse events in acute care hospitals, Br Med J 328 (2004), p. 199.

[17] T. Schioler, H. Lipczak, B.L. Pedersen, T.S. Mogensen, K.B. Bech and A. Stockmarr et al., Incidence of adverse events in hospitals. A retrospective study of medical records, Ugeskr Laeger 163 (2001), pp. 5370–5378.

[18] P. Davis, R. Lay-Yee, R. Briant, W. Ali, A. Scott and S. Schug, Adverse events in New Zealand public hospitals II: preventability and clinical context, N Z Med J 116 (2003), p. U624.

[19] C. Vincent, G. Neale and M. Woloshynowych, Adverse events in British hospitals: preliminary retrospective record review, Br Med J 322 (2001), pp. 517–519.
[20] M. Zegers, M.C. de Bruijne, C. Wagner, L.H.F. Hoonhout, R. Waaijman and M. Smits et al., Adverse events and potentially preventable deaths in Dutch hospitals. Results of a retrospective patient record review study, Qual Safety Health Care (2009) [in press].
[21] J.R. Landis and G.G. Koch, The measurement of observer agreement for categorical data, Biometrics 33 (1977), pp. 159–174.
[22] T.P. Hofer, S.J. Bernstein, S. Demonner and R.A. Hayward, Discussion between reviewers does not improve reliability of peer review of hospital quality, Med Care 38 (2000), pp. 152–161.
[23] C.P. Friedman, G.G. Gatti, T.M. Franz, G.C. Murphy, F.M. Wolf and P.S. Heckerling et al., Do physicians know when their diagnoses are correct? Implications for decision support and error reduction, J Gen Intern Med 20 (2005), pp. 334–339.
[24] D.L. Kunac, D.M. Reith, J. Kennedy, N.C. Austin and S.M. Williams, Inter- and intra-rater reliability for classification of medication related events in paediatric inpatients, Qual Saf Health Care 15 (2006), pp. 196–201.
[25] R. Lilford, A. Edwards, A. Girling, T. Hofer, G.L.D. Tanna and J. Petty et al., Inter-rater reliability of case-note audit: a systematic review, J Health Serv Res Policy 12 (2007), pp. 173–180. [26] D.G. Altman, Practical statistics for medical research, Chapman and Hall, London (1991).
[27] A.R. Feinstein and D.V. Cicchetti, High agreement but low kappa: I. The problems of two paradoxes, J Clin Epidemiol 43 (1990), pp. 543–549.

[APPENDIX A]

FIGURES, TABLES AND APPENDIX

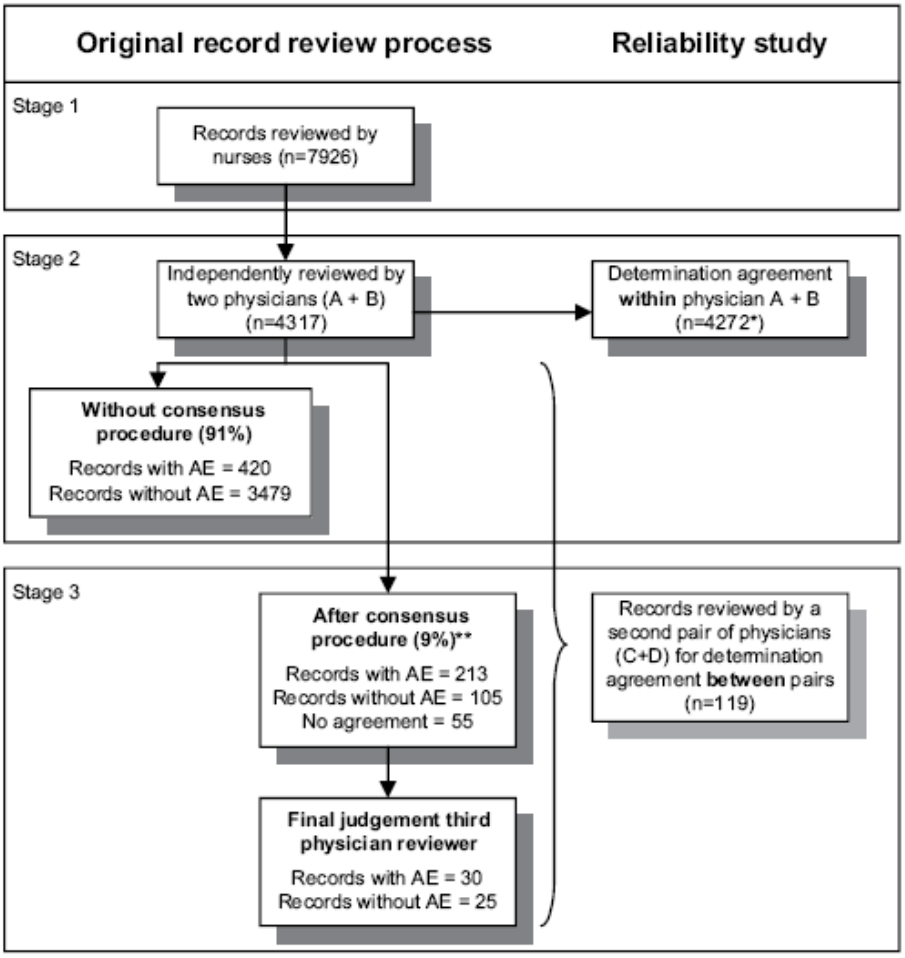

Fig. 1. Flow chart of the record review process. *Forty-five records were excluded, because the AEs had occurred outside the participating hospitals (thus in another hospital) or were not related to the sampled admission. **Consensus procedure about the presence of an AE and/or preventability.

Table 1
Sample selection to evaluate the review process *within* and *between* pairs of physicians

Strata	Result review process physician A and B	Evaluation agreement *within* pairs: all stage 2 records ($n = 4,272$)	Evaluation agreement *between* pairs: stratified sample ($n = 119$)
AE− C−	No AE was found; direct agreement (%)	81.4	38.6
AE+ C−	AE was found; direct agreement	9.8	28.6
AE− C+	No AE was found; consensus procedure[a] (and if applicable third review) was necessary	3.1	9.2
AE+ C+	AE was found; consensus procedure[a] (and if applicable third review) was necessary	5.7	23.5

Abbreviation: AE, adverse event.
[a] Consensus procedure between physician A and B about the presence of an adverse event.

Table 2
Agreement and kappa statistic of duplicate review process *within* pairs of physicians (physician A+B)

	Physician B					
	AE			Preventable AE		
Physician A	Absent	Present	Total	Absent	Present	Total
Absent	3479	201	3680	158	38	196
Present	172	420[a]	592	20	201	221
Total	3651	621	4272	178	239	417[a]
Agreement (%)			**91.3**			**86.1**
Kappa statistic (95% CI)			**0.64 (0.61, 0.68)**			**0.72 (0.66, 0.79)**

Abbreviations: AE, adverse event; CI, confidence interval.
[a] The inter-rater agreement for preventability was estimated in records in which both physicians found an AE. For three cases the preventability score was missing.

Table 3
Agreement and kappa statistic of duplicate review process *between* pairs of physicians (physician A+B and C+D)

	Physician C+D					
	AE			Preventable AE		
Physician A+B	Absent	Present	Total	Absent	Present	Total
Absent	45	12	57	7	2	9
Present	34	28[a]	62	6	12	18
Total	79	40	119	13	14	27[a]
Agreement (%)			**75.6[b]**			**70.4**
Kappa statistic (95% CI)			**0.25 (0.05, 0.45)[b]**			**0.40 (0.07, 0.73)**

Abbreviations: AE, adverse event; CI, confidence interval.
[a] The inter-rater agreement for preventability was estimated in records in which both pairs of physicians found an AE. For one case the preventability score was missing.
[b] Adjusted for sampling frame.

Table 4
Inter-rater agreement *within* pairs of physicians for AE determination for subgroups of records ($n = 4,272$)

Subgroup of records	*N*	Kappa (95% CI)	Agreement (%)
Records of discharged patients	1,515	0.68 (0.62, 0.73)	92.3
Records of deceased patients	2,757	0.62 (0.58, 0.67)	90.7
Records reviewed by 2 internists ($n = 25$)	2,757	0.64 (0.59, 0.68)	91.7
Records reviewed by 2 surgeons ($n = 20$)	1,011	0.63 (0.57, 0.70)	88.9
Records reviewed by 2 neurologists ($n = 5$)	372	0.77 (0.65, 0.88)[a]	96.0
Records reviewed by 2 pediatricians ($n = 5$)	59	0.25 (−0.19, 0.69)	91.5
Records reviewed by 2 mix of physicians	73	0.56 (0.32, 0.78)	84.9
Records reviewed by 2 physicians who reviewed ≥143 records[b]	2,738	0.68 (0.64, 0.73)	92.7
Records reviewed by 2 physicians who reviewed <143 records[b]	377	0.63 (0.51, 0.75)	91.8

Abbreviations: AE, adverse event; CI, confidence interval.
[a] The kappa value within pairs of neurologists significantly differs with kappa values of other specialists ($P < 0.05$).
[b] The median number of reviewed records per physicians was 143; 1,157 records were excluded from this analysis because they were reviewed by a pair of physicians who reviewed more and less than 143 records.

Appendix A

Definitions and outcome measures [14]

Description of 18 screening criteria for potential AEs

1. Unplanned admission before index admission (admission reasons are related to the index admission)
2. Unplanned readmission after discharge from index admission
3. Hospital-incurred patient injury (permanent or temporary injury obtained (acquired) during index admission)
4. Adverse drug reaction
5. Unplanned transfer from general care to (an) intensive care (unit)
6. Unplanned transfer to another acute care hospital (after unexpected deterioration of the patient)
7. Unplanned return to the operating room
8. Unplanned removal, injury, or repair of an organ during surgery
9. Hospital-acquired infection or sepsis
10. Other patient complications
11. Development of neurologic deficit not present on admission
12. Unexpected death
13. Cardiac or respiratory arrest
14. Injury related to abortion or delivery
15. Inappropriate discharge to home
16. Dissatisfaction with care documented in the medical record
17. Documentation or correspondence indicating litigation
18. Any other undesirable outcome not covered above

The determination of an AE was based on three criteria:

1. an unintended (physical and/or mental) *injury* which
2. resulted in temporary or permanent *disability*, death, or prolongation of hospital stay, and is
3. *caused by health care management* rather than the patient's disease

To determine whether the injury was caused by health care management or the disease process a 6-point scale was used:

1. (Virtually) no evidence for management causation
2. Slight to modest evidence of management causation
3. Management causation not likely (less than 50/50, but "close call")
4. Management causation more likely (more than 50/50, but "close call")
5. Moderate to strong evidence of management causation
6. (Virtually) certain evidence of management causation

The degree of preventability of the AEs was measured on a 6-point scale:

1. (Virtually) no evidence for preventability
2. Slight to modest evidence of preventability
3. Preventability not quite likely (less than 50/50, but "close call")
4. Preventability more than likely (more than 50/50, but "close call")
5. Strong evidence of preventability
6. (Virtually) certain evidence of preventability