Measurement of patient safety: a systematic review of the reliability and validity of adverse event detection with record review

Mirelle Hanskamp-Sebregts,1 Marieke Zegers,2 Charles Vincent,3 Petra J van Gurp,1 Henrica C W de Vet,3,4 Hub Wollersheim2

ABSTRACT

Objectives: Record review is the most used method to quantify patient safety. We systematically reviewed the reliability and validity of adverse event detection with record review.

Design: A systematic review of the literature.

Methods: We searched PubMed, EMBASE, CINAHL, PsycINFO and the Cochrane Library and from their inception through February 2015. We included all studies that aimed to describe the reliability and/or validity of record review. Two reviewers conducted data extraction. We pooled \( \kappa \) values (\( \kappa \)) and analysed the differences in subgroups according to number of reviewers, reviewer experience and training level, adjusted for the prevalence of adverse events.

Results: In 25 studies, the psychometric data of the Global Trigger Tool (GTT) and the Harvard Medical Practice Study (HMPS) were reported and 24 studies were included for statistical pooling. The inter-rater reliability of the GTT and HMPS showed a pooled \( \kappa \) of 0.65 and 0.55, respectively. The inter-rater agreement was statistically significantly higher when the group of reviewers carried out record review. We found no studies reporting on the validity of the GTT and HMPS.

Conclusions: The reliability of record review is moderate to substantial and improved when a small group of reviewers carried out record review. The validity of the record review method has never been evaluated, while clinical data registries, autopsy or direct observations of patient care are potential reference methods that can be used to test concurrent validity.

INTRODUCTION

Healthcare professionals are faced with the challenge of improving patient safety by detecting, preventing and mitigating the occurrence of adverse events (AEs).1–2 An AE is defined as an injury that is caused by healthcare management (rather than the underlying disease) and results in prolonged hospitalisation, disability at the time of discharge or even in patient’s death.3 Besides improving patient safety, transparency with reliable and valid data is necessary for accountability purposes.4,5 Non-valid or unreliable instruments for quantifying patient safety can lead to inadequate diagnosis of patient safety problems and subsequently to the implementation of inadequate patient safety improvement interventions.

Patient record review is the most thoroughly studied method used to measure the prevalence of AEs.6 Incidents, complaints and claims reporting systems are less suitable for counting AEs, because the amount of AEs strongly depends on the willingness of healthcare providers and patients to report them. Only 3–5% of the AEs detected in patient records are reported by healthcare providers in hospitals.7–11 In addition, the denominator, the related number of patients, is difficult to determine. These systems are therefore inadequate to count the actual number of incidents.12–14

Strengths and limitations of this study

- We have reviewed ~4000 articles across five databases on psychometric data regarding the record review as a method to detect adverse events.
- We evaluated the methodological quality of the included studies on measurement properties with the validated COSMIN checklist.
- Two instruments for record review, the Global Trigger Tool and the Harvard Medical Practice Study, were extensively tested on their reliability, but data regarding the validity of these instruments completely lack.
- The subgroup analyses were limited to the variables that were reported by the authors in the studies that were included in our systematic review.
Although record review is widely accepted as the
method for quantifying AEs, data about the psychometric
aspects of this method reported in previous literature
reviews are limited.\(^{12, 13, 15}\) or outdated.\(^{16}\) Therefore, we
systematically reviewed the reliability and validity of
record review and which factors are associated with these
psychometric measures. We assumed that the inter-
rater reliability of record review was higher for studies
with a small number of reviewers, more reviewer experi-
ence and a higher training level.

**METHODS**

**Search strategy and databases**

Our literature search strategy was prespecified and
aligned with recommendations outlined in the PRISMA
(Preferred Reporting Items for Systematic Reviews and
Meta-Analyses).\(^ {17}\) We included the study protocol in
online supplementary appendix 1.

We searched for full-text studies published until
October 2013 and updated our search in February 2015
using the following databases: PubMed (including
MEDLINE), EMBASE, CINAHL, PsycINFO and the
Cochrane Library. The references of the included
studies were manually checked, and the authors’
personal files and bibliographies of previously published
related reviews were searched to identify additional rele-
vant studies (snowballing). There were no language
restrictions. Online supplementary appendix 2 provides
a detailed listing of search strings.

**Selection criteria and process**

Two researchers (MH-S and MZ) independently screened
the titles and abstracts of all studies identified by the
search strategy for their eligibility. Studies were included
if (1) the record review method was described in detail,
(2) AEs were measured in a wide variety of patient groups
and (3) data about reliability and validity were reported.
Studies not available in full-text were excluded.

When the title and abstract did not clearly indicate
whether the inclusion criteria were met, the full text
(meaning the complete article) was obtained and
reviewed by two researchers (MH-S and MZ). The previ-
ously described inclusion criteria were applied again,
and a final set of studies was identified for data extrac-
tion. Disagreement about inclusion was solved by discus-
sion. When no consensus could be achieved, a third
researcher (HW) made the final decision.

**Terminology and definitions**

Different types of reliability and validity of measurement
instruments can be distinguished. Focus of our system-
atic review was on the inter-rater reliability, content
(face) validity and concurrent validity of record review.
Definitions are described in table 1.

**Quality assessment**

Assessment of the methodological quality of the selected
studies was carried out using the COSMIN checklist.\(^ {20}\)
The COSMIN checklist facilitates a separate judgement
of the methodological quality of the included studies
and their results.\(^ {21}\) The COSMIN checklist consists of
nine boxes with methodological standards for how each
measurement property should be assessed. Three of the
nine boxes were relevant for this systematic review
regarding inter-rater reliability, content validity and con-
current validity. There are no standards for assessing
face validity, because face validity requires a subjective
judgement of experts.\(^ {22}\) Each item in these relevant
boxes was scored on a four-point rating scale (ie, ‘poor’,
‘fair’, ‘good’ or ‘excellent’).\(^ {20, 21}\) An overall score for the
methodological quality of a study was determined by
taking the lowest rating of any of the items in a box. The
methodological quality of a study was assessed per meas-
urement property by MH-S, and 10% of the studies were
assessed independently by MZ. In cases of disagreement,
a third reviewer (HW) was consulted for a final decision.

**Data extraction**

Each article that met study eligibility criteria was independ-
ently abstracted by one reviewer (MH-S), and a second
reviewer (MZ) crosschecked the data extraction of the first
reviewer. Both reviewers used a standardised form, which
compromised a description of objectives, study population,

**Table 1 Definitions of reliability and validity in the context of record review**

<table>
<thead>
<tr>
<th>Terms</th>
<th>Definition (expressed by)</th>
<th>Comments relevant to record review</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inter-rater reliability(^ {18})</td>
<td>Measures consensus in the scores when different raters using the same measurement instrument in the same group of patients. Mostly expressed as a reliability measure (κ), or % agreement</td>
<td>Two independent reviewers assess patient records without discussion between the reviewers during the review process</td>
</tr>
<tr>
<td>Face validity(^ {18})</td>
<td>The degree to which the content of an instrument is an adequate reflection of the construct to be measured (descriptive, expert opinion)</td>
<td>Clinical data registries, autopsy or direct observations of patient care have the potential to be a criterion measure for record review</td>
</tr>
<tr>
<td>Concurrent validity(^ {19})</td>
<td>The extent to which scores on a new measure are related to scores from a criterion measure administered at the same time (Se, Sp, PPV and NPV)</td>
<td></td>
</tr>
</tbody>
</table>

NPV, negative predictive value; PPV, positive predictive value; Se, sensitivity; Sp, specificity;
design and methods used and the results of the analysis of the reliability and validity, including statistical parameters (see online supplementary appendix 1).

Data synthesis and analysis
We tabulated study characteristics and outcomes such as setting, number of records, percentage AEs and data about reliability and validity of record review. In some studies, percentage agreement was calculated from source data by MH-S and confirmed by MZ. To be able to rate the reliability of record review, we classified the $\kappa$ values as ‘slight’ ($\kappa=$0.00–0.20), ‘fair’ ($\kappa=$0.21–0.40), ‘moderate’ ($\kappa=$0.41–0.60), ‘substantial’ ($\kappa=$0.61–0.80) and ‘almost perfect’ ($\kappa=$0.81–1.00).23

We pooled the outcomes statistically by calculating the mean percentage agreement and the mean and pooled $\kappa$ on the presence of AEs to draw conclusions about the reliability of record review. We used the number of records on which the $\kappa$ value is calculated as weighing factor in the statistical pooling as a proxy for accuracy, since we missed information about the 95% CIs of the $\kappa$ values in the included studies.

To examine differences in $\kappa$ values depending on the number of reviewers, reviewer experience and reviewer training, we present descriptive statics per subgroup (mean with SD or median with IQR for non-normal distributions, minimum and maximum). In order to better interpret the results, we classified the number of reviewers per study, reviewer experience and reviewer training into three proportional classes: maximum 5 reviewers, >5–20 reviewers, >20 reviewers; <100 records per reviewer, 100–300 records per reviewer, >300 records per reviewer and <1 day training, 1 day training, >1 day training, respectively. We used the non-parametric Kruskal-Wallis test for the group characteristics, which are not normally distributed and an ANOVA for the group characteristics with a normal distribution. We checked whether the assumptions for ANCOVA were met. It was not possible to incorporate all variables (the number of reviewers, reviewer experience and reviewer training) in one ANCOVA, because the number of studies in our analyses was limited (n=20). Therefore, we performed three separate ANCOVAs, with prevalence of AE as covariate. We adjusted for prevalence of AEs, since a previous study of Lilford et al.27 showed correlation between prevalence and $\kappa$. Additionally, we studied the influence of the aim of the study and the type of instrument (Global Trigger Tool (GTT) vs Harvard Medical Practice Study (HMPS)) on $\kappa$ with two separate ANCOVAs adjusted for prevalence. A p value of <0.05 was regarded as statistically significant. Statistical software IBM SPSS V.22 was used for all statistical analyses and data processing.

RESULTS
Results of the literature search
Our literature study yielded 3915 citations (see online supplementary appendix 3, flow chart), of which 1790 were in PubMed, 1153 were in EMBASE, 515 were in CINAHL, 30 were in PsycINFO and 427 were in the Cochrane Library. After removing duplicates, 3415 studies remained, of which 148 were selected for full-text selection. A total of 137 studies were excluded after reading the full text, because these studies did not meet the inclusion criteria, including studies that did not focus on the reliability or validity of record review,24–26 did not have AEs as outcome27 or reported a different method than retrospective reviewing of medical records.28 29 We collected eight additional articles through manual searching of articles’ bibliographies. In February 2015, we updated our search and found six additional studies. The final set consisted of 25 record review studies; 24 studies were used for calculating the mean $\kappa$, and 20 studies were appropriate for the subgroup analysis. Five studies were excluded because only the intra-class correlation coefficient was calculated,23 the prevalence was an outlier,31 the prevalence was not reported32 33 or the number of reviewers was not reported.3

Description of the GTT and the HMPS
We found two record review instruments for detecting AEs, namely, the GTT and the HMPS. Both instruments use an implicit review style, meaning that the AE assessment relies on expert judgement instead of using well-defined criteria on a checklist (explicit review style).6 16 The GTT and the HMPS consist of a two-stage review process conducted by nurses and physicians (table 2). The GTT is primarily used as a quality improvement tool for clinical practice and for estimating and tracking AE rates over time in a hospital or a clinic. The HMPS is commonly used to measure the prevalence rate of AEs on a national level. The GTT is not meant to identify every single AE in a patient record, and, therefore, assessments have a time limit of 20 min per record.34 The GTT consists of 47–55 triggers to identify potential AEs. Reviewing the preventability of adverse events is originally no part of the GTT method, but has been recently included in the studies of Schildmeijer et al.35 Kennerly et al.,36 Najjar et al.37 and Hwang et al.38 In contrast, the HMPS consists of 16–18 screening criteria (triggers), 27 leading questions for AE detection, of which three questions are crucial for AE determination: injury present; resulting in prolongation of hospital stay, temporary or permanent disability or death and caused by healthcare management. Determination of preventability of AEs is standard within the HMPS method. The HMPS is more time-consuming and labour-intensive in assessing AEs (stage 2) than the GTT, due to the number of questions.

Characteristics and methodological quality of included studies
Most of the identified studies were carried out in the USA, UK, Canada, Europe and Australia (see online supplementary appendices 4 and 5). In these studies, the
Table 2: Description of the Global Trigger Tool and Harvard Medical Practice Study

<table>
<thead>
<tr>
<th>Instrument</th>
<th>Description</th>
<th>Safety outcomes</th>
<th>Conducted by</th>
<th>Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Global Trigger Tool&lt;sup&gt;34&lt;/sup&gt;</td>
<td>Two-stage retrospective record review&lt;br&gt;Stage 1: Screening records for the presence of triggers and determining the adverse event that caused harm to patients&lt;br&gt;Stage 2: Confirming or dismissing the occurrence and category of the adverse event</td>
<td>Triggers (mostly narrow)&lt;br&gt;Adverse events</td>
<td>Stage 1: Trained nurses or hospital pharmacists (primary reviewers, mostly two reviewers per records)&lt;br&gt;Stage 2: Trained physicians (second reviewers, mostly one reviewer)</td>
<td>Dichotomous: yes/no trigger&lt;br&gt;Dichotomous: yes/no AE&lt;br&gt;Definition of AE: Any unintended physical injury resulting from or contributed to by medical care that requires additional monitoring, treatment or hospitalisation or that results in death</td>
</tr>
<tr>
<td>Medical record review based on HMPS&lt;sup&gt;3&lt;/sup&gt;</td>
<td>Two-stage or three-stage retrospective record review&lt;br&gt;Stage 1: Screening records using criteria&lt;br&gt;Stage 2: Detailed review to confirm the presence of adverse events and their preventability</td>
<td>(Broad) Screening criteria (triggers)&lt;br&gt;Preventable&lt;br&gt;Adverse events</td>
<td>Stage 1: Trained nurses*&lt;br&gt;No time limit&lt;br&gt;Stage 2: Trained physicians (one or two reviewers per record)</td>
<td>Dichotomous: yes/no trigger&lt;br&gt;AE determination is based on three criteria:&lt;br&gt;1. Unintended injury to the patient (dichotomous: yes/no)&lt;br&gt;2. Resulted in prolongation of hospital stay, temporary or permanent disability or death (dichotomous: yes/no)&lt;br&gt;3. Caused by healthcare management (six-point scale)&lt;br&gt;Preventability: six-point scale&lt;br&gt;When criteria 1 and 2 are met and the score on criteria 3 is ≥4, then there has been an AE and an AE is preventable when the score on the preventability scale is ≥4</td>
</tr>
</tbody>
</table>

*With the exception of the study of Brennan et al.,<sup>32</sup> in which medical records were reviewed by medical-record-room administrators.<br>†In some studies, a third stage was used.<sup>33</sup>–<sup>42</sup>

AEs, adverse events; HMPS, Harvard Medical Practice Study.
GTT (n=10 studies) and HMPS (n=15 studies) were all tested in hospitals. The percentage AEs in GTT studies ranged from 7.2% to 27.0% (see online supplementary appendix 4). The total number of reviewers varied from 2 to 20 reviewers per study. Reviewers assessed 50 to 4043 records on average. The percentage AEs in HMPS studies ranged from 2.9% to 18.0%, and for preventable AEs they ranged from 1% to 8.6% (see online supplementary appendix 5). The total number of reviewers varied from 2 to 127 reviewers per study. Average records per reviewer ranged from 38 to 3872 records. The primary aim of most of the GTT studies included in this review was to examine the inter-rater reliability, whereas the primary aim of the HMPS studies reporting inter-rater reliability data was measuring AE rates.

The methodological quality of the included studies was good. In all these studies, the inter-rater reliability was evaluated. In one study, the face validity was evaluated.

Reliability of the GTT

The percentage agreement for reviewers of AE assessment was reported in four studies, ranging from 83% to 94% with a mean of 87.5% (SD 4.8%) (see online supplementary appendix 4). One study showed slight inter-rater reliability (κ=0.19),3 three studies showed moderate inter-rater reliability (κ=0.44–0.49)8 11 39 and three studies showed substantial inter-rater reliability (κ=0.69–0.76)40 49 51 for assessing preventable AEs. The mean κ and pooled κ of the HMPS for AE assessment are 0.54 (SD 0.10) and 0.55 (SD 0.07), respectively, and, for assessing preventability, they are 0.47 (SD 0.20) and 0.48 (SD 0.20), respectively. The inter-rater reliability of the HMPS is classified as moderate.

Subgroup analysis inter-rater reliability

The number of GTT studies (n=9) and HMPS studies (n=11) were too small to perform the subgroup analysis for the methods separately. Therefore, we used the κ statistics of all studies (n=20) to carry out the subgroup analysis. The assumptions for ANCOVA were met. Prevalence was not statistically significant associated with the κ values (p=0.069, p=0.189 and p=0.726, respectively). We found a statistically significant difference in the pooled κ values, p=0.006, among subgroups according to the number of reviewers (table 3). There were no differences in κ values between subgroups according to reviewer experience (p=0.062) and reviewer training.

<table>
<thead>
<tr>
<th>Group of reviewers</th>
<th>n</th>
<th>Pooled κ* (SD)</th>
<th>95% CI</th>
<th>p Value†</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max 5</td>
<td>7</td>
<td>0.80 (0.07)</td>
<td>0.66 to 0.94</td>
<td>0.006</td>
</tr>
<tr>
<td>&gt;5–20</td>
<td>7</td>
<td>0.52 (0.06)</td>
<td>0.40 to 0.64</td>
<td></td>
</tr>
<tr>
<td>&gt;20</td>
<td>6</td>
<td>0.54 (0.02)</td>
<td>0.50 to 0.59</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>20</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reviewer experience (records/reviewer)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;100</td>
<td>7</td>
<td>0.71 (0.06)</td>
<td>0.58 to 0.84</td>
<td>0.062</td>
</tr>
<tr>
<td>100–300</td>
<td>6</td>
<td>0.51 (0.04)</td>
<td>0.43 to 0.58</td>
<td></td>
</tr>
<tr>
<td>&gt;300</td>
<td>7</td>
<td>0.53 (0.04)</td>
<td>0.45 to 0.62</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>20</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Training</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;1 day</td>
<td>4</td>
<td>0.53 (0.07)</td>
<td>0.37 to 0.68</td>
<td>0.809</td>
</tr>
<tr>
<td>1 day</td>
<td>4</td>
<td>0.56 (0.14)</td>
<td>0.25 to 0.87</td>
<td></td>
</tr>
<tr>
<td>&gt;1 day</td>
<td>5</td>
<td>0.57 (0.05)</td>
<td>0.45 to 0.67</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>13</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*pooled κ weighted for the number of records on which the κ value is calculated.†p Values are obtained with the prevalence rate as covariate.

Table 3 Differences in pooled κ values (n=20) among subgroups according to number of reviewers, reviewer experience and reviewer training
The group of maximum five reviewers detected more AEs (average 17.1%) in comparison with the other two groups of reviewers (table 4). This group received the least training (median 6 hours) and assessed the largest number of records (median 213 records). There was no significant difference in the reviewer experience (p=0.351), the reviewer training (p=0.317) and the prevalence of AEs (p=0.480) between the three groups of reviewers (maximum 5 reviewers, >5–20 reviewers and >20 reviewers).

The number of studies that reported the $\kappa$ of preventable AEs (n=8) was too small for subgroup analysis. The aim of the study and the type of instrument (GTT vs HMPS) were not statistically significantly associated with $\kappa$ (p=0.572 and p=0.086, respectively).

**Validity**

The face validity of the HMPS was reported in one study as being a valid method to identify AEs.32 We found no studies in which the concurrent validity of the GTT or HMPS has been studied.

**DISCUSSION**

The inter-rater reliability of record review to detect AEs is moderate to substantial;23 with a pooled $\kappa$ of 0.65 and 0.55 for the GTT method and the HMPS method, respectively. The pooled $\kappa$ for preventability, measured with the HMPS method, is moderate, 0.48. The fact that there are no studies looking at concurrent validity is alarming, given the statements that record review is accepted worldwide as the ‘best’ means of measuring incidence rates of AEs (even called ‘the gold standard’).15 59 Even if the inter-rater reliability of record review is acceptable, there is no evidence that record review really detects AEs. Possible methods to test the concurrent validity of record review are clinical data registries, autopsy or direct observations of patient care. No single, even a small study experimented with above listed reference methods, although these methods capture valuable (real-time), accurate and precise patient data.13 60–63

We found statistically significant higher inter-rater reliability in subgroups in which the group of reviewers consisted of five reviewers or less. An explanation for this difference is that when the group of reviewers is small, the assessment of the presence of an AE becomes more standardised.40 64 Having a small group of reviewers stimulates (un)intentionally working closer together, resulting in less variation in the review methodology and more consensus about the definition of what constitutes harm in order to be counted as an AE. Additional advantages of having a small group of reviewers are that intensive review training can be organised, and the review process can be better monitored.30 In our review however, the group of maximum five reviewers received less training hours. Probably, they were better supervised (p=0.809). The group of maximum five reviewers detected more AEs (average 17.1%) in comparison with the other two groups of reviewers (table 4). This group received the least training (median 6 hours) and assessed the largest number of records (median 213 records). There was no significant difference in the reviewer experience (p=0.351), the reviewer training (p=0.317) and the prevalence of AEs (p=0.480) between the three groups of reviewers (maximum 5 reviewers, >5–20 reviewers and >20 reviewers).

The number of studies that reported the $\kappa$ of preventable AEs (n=8) was too small for subgroup analysis. The aim of the study and the type of instrument (GTT vs HMPS) were not statistically significantly associated with $\kappa$ (p=0.572 and p=0.086, respectively).

**Validity**

The face validity of the HMPS was reported in one study as being a valid method to identify AEs.32 We found no studies in which the concurrent validity of the GTT or HMPS has been studied.

**DISCUSSION**

The inter-rater reliability of record review to detect AEs is moderate to substantial;23 with a pooled $\kappa$ of 0.65 and 0.55 for the GTT method and the HMPS method, respectively. The pooled $\kappa$ for preventability, measured with the HMPS method, is moderate, 0.48. The fact that there are no studies looking at concurrent validity is alarming, given the statements that record review is accepted worldwide as the ‘best’ means of measuring incidence rates of AEs (even called ‘the gold standard’).15 59 Even if the inter-rater reliability of record review is acceptable, there is no evidence that record review really detects AEs. Possible methods to test the concurrent validity of record review are clinical data registries, autopsy or direct observations of patient care. No single, even a small study experimented with above listed reference methods, although these methods capture valuable (real-time), accurate and precise patient data.13 60–63

We found statistically significant higher inter-rater reliability in subgroups in which the group of reviewers consisted of five reviewers or less. An explanation for this difference is that when the group of reviewers is small, the assessment of the presence of an AE becomes more standardised.40 64 Having a small group of reviewers stimulates (un)intentionally working closer together, resulting in less variation in the review methodology and more consensus about the definition of what constitutes harm in order to be counted as an AE. Additional advantages of having a small group of reviewers are that intensive review training can be organised, and the review process can be better monitored.30 In our review however, the group of maximum five reviewers received less training hours. Probably, they were better supervised (p=0.809). The group of maximum five reviewers detected more AEs (average 17.1%) in comparison with the other two groups of reviewers (table 4). This group received the least training (median 6 hours) and assessed the largest number of records (median 213 records). There was no significant difference in the reviewer experience (p=0.351), the reviewer training (p=0.317) and the prevalence of AEs (p=0.480) between the three groups of reviewers (maximum 5 reviewers, >5–20 reviewers and >20 reviewers).

The number of studies that reported the $\kappa$ of preventable AEs (n=8) was too small for subgroup analysis. The aim of the study and the type of instrument (GTT vs HMPS) were not statistically significantly associated with $\kappa$ (p=0.572 and p=0.086, respectively).
or communicate better with each other during the study, which could increase the inter-rater agreement.

The inter-rater reliability was higher when reviewers assess a substantial number of records. We found no statistically significant differences between subgroups according to reviewer experience, despite the group of maximum five reviewers assessed a notable number of records compared to the groups of reviewers, which consist of 6–20 reviewers or more than 20 reviewers.

From other studies, we know that training improves the performance of review teams and the application of record review. We found no evidence for this in our review. In fact, the group of maximum five reviewers had half the training hours compared to the group of 6–20 reviewers but achieved a higher inter-rater agreement.

The systematic review of Lilford et al. showed that there was an association between $\kappa$ and the prevalence of AEs. We found no statistically significantly association between $\kappa$ and the prevalence of AEs. The smaller range of the prevalence rate (2.9–27.0%) in our review compared to the review of Lilford et al. (2.8–58.9%) could explain why we did not find an association between $\kappa$ and the prevalence of AEs.

Our systematic review has some strengths and limitations. First, the evidence of the results of the statistical pooling depends on the quality of the therein contained studies. We used the validated COSMIN tool to evaluate the methodological quality of the included studies. Second, it was not possible to formally estimate the pooled $\kappa$ statistics for the GTT and Medical Record Review (MRR) to assess between-study heterogeneity or to carry out analyses of the likelihood of publication bias, because CIs were lacking in approximately half of the reliability studies. Third, the subgroup analyses were limited to the variables that were reported by the authors in the included studies of our systematic review. Other factors that possibly influence the inter-rater agreement between reviewers, such as the level of cooperation between the reviewers during the review process, could therefore not be studied. Fourth, our review may have been influenced by publication bias, as studies reporting low reliability or validity may be less likely to be published than those with more positive results. Fifth, we statistically pooled the $\kappa$ values. However, specific agreement on the presence of AE, expressing the agreement separately for the positive and negative ratings, is recommended. After all, inter-rater reliability concerns when one reviewer finds an AE, and this AE is also found by a second reviewer. Unfortunately, in most of the studies, information about the number of records for which there was agreement, presented in a 2×2 cross table, was missing. Therefore, we could not perform a statistical pooling of the proportion of specific agreement.

In conclusion, users of the record review method to assess (preventable) AEs should be aware that the inter-rater agreement between reviewers is moderate to substantial and increases when using a smaller group of reviewers. More studies are needed to explore which factors increase the inter-rater reliability of record review. Most importantly, concurrent validity should be tested, otherwise it remains an imperfect, never evaluated method.

**Author affiliations**
1. Radboud University Medical Center, Institute of Quality Assurance and Patient Safety, Nijmegen, The Netherlands
2. Radboud University Medical Center, Radboud Institute for Health Sciences, IQ healthcare, Nijmegen, The Netherlands
4. Department of Epidemiology and Biostatistics, EMGO Institute for Health and Care Research, VU University Medical Center, Amsterdam, The Netherlands

**Acknowledgements** The authors thank Ir Reiner Akkermans, statistician, for his recommendations by the statistical pooling.

**Contributors** MZ and HW conceived the idea for the study. MH-S and MZ led the writing of the paper as well as analysed and interpreted the data. CV advised on study design and approach. HGWdV supervised the data analysis. HGWdV and CV contributed to the writing of the paper. PJvG and HW participated in revising this manuscript. All authors contributed substantially to the writing of the paper, and all reviewed and approved the final draft.

**Funding** This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

**Competing interests** None declared.

**Provenance and peer review** Not commissioned; externally peer reviewed.

**Data sharing statement** No additional data are available.

**Open Access** This is an Open Access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited and the use is non-commercial. See: http://creativecommons.org/licenses/by-nc/4.0/

**REFERENCES**

9. Christiansen-Dinglehoff I, Smits M, Zwaan L, et al. To what extent are adverse events found in patient records reported by patients and healthcare professionals via complaints, claims and incident reports? BMC Health Serv Res 2011;11:49.


