BMJ Open

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 Wollersheim[2]

To cite: Hanskamp-Sebregts M, Zegers M, Vincent C, *et al*. Measurement of patient safety: a systematic review of the reliability and validity of adverse event detection with record review. *BMJ Open* 2016;**6**:e011078. doi:10.1136/bmjopen-2016-011078

► Prepublication history and additional material is available. To view please visit the journal (http://dx.doi.org/10.1136/bmjopen-2016-011078).

Received 11 January 2016
Revised 22 July 2016
Accepted 29 July 2016

For numbered affiliations see end of article.

Correspondence to
Mirelle Hanskamp-Sebregts; Mirelle.Hanskamp-Sebregts@radboudumc.nl

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 κ values (κ) 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 κ of 0.65 and 0.55, respectively. The inter-rater agreement was statistically significantly higher when the group of reviewers within a study consisted of a maximum five reviewers. 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.

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.

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]

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 experience 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 relevant 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 previously described inclusion criteria were applied again, and a final set of studies was identified for data extraction. Disagreement about inclusion was solved by discussion. 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 systematic 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 concurrent 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 measurement 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 independently 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

Terms	Definition (expressed by)	Comments relevant to record review
Inter-rater reliability[18]	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	Two independent reviewers assess patient records without discussion between the reviewers during the review process
Face validity[18]	The degree to which the content of an instrument is an adequate reflection of the construct to be measured (descriptive, expert opinion)	
Concurrent validity[19]	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)	Clinical data registries, autopsy or direct observations of patient care have the potential to be a criterion measure for record review

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 κ values as 'slight' (κ=0.00–0.20), 'fair' (κ=0.21–0.40), 'moderate' (κ=0.41–0.60), 'substantial' (κ=0.61–0.80) and 'almost perfect' (κ=0.81–1.00).[23]

We pooled the outcomes statistically by calculating the mean percentage agreement and the mean and pooled κ on the presence of AEs to draw conclusions about the reliability of record review. We used the number of records on which the κ 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 κ values in the included studies.

To examine differences in κ 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*[16] showed correlation between prevalence and κ. 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 κ 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 outcome[27] 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 κ, and 20 studies were appropriate for the subgroup analysis. Five studies were excluded because only the intraclass correlation coefficient was calculated,[30] the prevalence was an outlier,[31] the prevalence was not reported[32 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

Instrument	Description	Safety outcomes	Conducted by	Scale
Global Trigger Tool[34]	Two-stage retrospective record review			
	Stage 1: Screening records for the presence of triggers and determining the adverse event that caused harm to patients	Triggers (mostly narrow)	Stage 1: Trained nurses or hospital pharmacists (primary reviewers, mostly two reviewers per records) Maximum 20 min per record	Dichotomous: yes/no trigger
	Stage 2: Confirming or dismissing the occurrence and category of the adverse event	Adverse events	Stage 2: Trained physicians (second reviewers, mostly one reviewer)	Dichotomous: yes/no AE 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
Medical record review based on HMPS[3]	Two-stage or three-stage retrospective record review			
	Stage 1: Screening records using criteria	(Broad) Screening criteria (triggers)	Stage 1: Trained nurses* No time limit	Dichotomous: yes/no trigger
	Stage 2: Detailed review to confirm the presence of adverse events and their preventability	(Preventable) Adverse events	Stage 2: Trained physicians (one or two reviewers per record)	AE determination is based on three criteria: 1. Unintended injury to the patient (dichotomous: yes/no) 2. Resulted in prolongation of hospital stay, temporary or permanent disability or death (dichotomous: yes/no) 3. Caused by healthcare management (six-point scale) Preventability: six-point scale 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
	Stage 3: Discussion or independently supervising review (consensus stage)†		Stage 3: Supervising physician	

*With the exception of the study of Brennan *et al*,[32] in which medical records were reviewed by medical-record-room administrators.
†In some studies, a third stage was used.[3 32 39–42]
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[3 11 30–33 35–58] was good. In all these studies, the inter-rater reliability was evaluated. In one study, the face validity was evaluated.[32]

Reliability of the GTT

The percentage agreement for reviewers of AE assessment was reported in four studies,[31 38 43 47] ranging from 83% to 94% with a mean of 87.5% (SD 4.8%) (see online supplementary appendix 4). One study showed fair inter-rater reliability (κ=0.34),[47] two studies showed moderate inter-rater reliability (κ=0.45),[35 43] five studies showed substantial inter-rater reliability (κ=0.62–0.74)[31 36 38 45 46] and two studies showed almost perfect inter-rater reliability (κ=0.85–0.89).[37 44] The mean κ and pooled κ are 0.65 (SD 0.19), meaning that the overall inter-rater reliability of the GTT is substantial.[23]

Reliability of the HMPS

The percentage agreement of AE assessment was reported in 10 studies and ranged from 73% to 91% with a mean of 83% (SD 6.1%);[3 11 39–42 49 50 52–54] percentage agreement for preventability of AE was assessed in six studies and ranged from 58% to 93% with a mean of 81% (SD 13%)[3 11 39 40 49 54] (see online supplementary appendix 5).

Ten studies showed moderate inter-rater reliability for AE detection (κ=0.40–0.57)[32 39 41 42 48–52 54] and in four studies the inter-rater reliability was substantial (κ=0.61–0.80).[3 11 40 49] In 10 studies, the κ for assessing preventable AEs was reported and ranged from 0.19 to 0.76.[3 11 32 39 40 48 49 51 53 54] One study showed slight inter-rater reliability (κ=0.19),[53] three studies showed fair inter-rater reliability (κ=0.24–0.34),[3 32 54] three studies showed moderate inter-rater reliability (κ=0.44–0.49)[11 39 48] 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.[23]

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 3 Differences in pooled κ values (n=20) among subgroups according to number of reviewers, reviewer experience and reviewer training

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

*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 4 The reviewer experience, reviewer training and the prevalence of AEs in the three groups of reviewers

	Max 5 reviewers		>5–20 reviewers		>20 reviewers		
	Median* (IQR)	Min–Max	Median* (IQR)	Min–Max	Median* (IQR)	Min–Max	p Value†
Reviewer experience (records/reviewer)	213 (60–1138)	50–4043	95 (39–317)	38–591	129 (109–616)	78–675	0.351
Training hours	6 (0–6)	0–12	16 (5–20)	2–24	8 (3–10)	2–16	0.317
	Mean* (SD)	Min–Max	Mean* (SD)	Min–Max	Mean* (SD)	Min–Max	p Value‡
Prevalence AEs (%)	17.1 (7.8)	7.2–27	13.5 (4.0)	7.5–21	12.7 (8.5)	2.9–25.1	0.480

*Unweighted statistics for reviewer experience, training and prevalence rate.
†p Values are obtained by the non-parametric Kruskal-Wallis test.
‡p Value is obtained with an ANOVA.
AEs, adverse events; Min, minimum; Max, maximum.

(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 κ 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 κ (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 κ of 0.65 and 0.55 for the GTT method and the HMPS method, respectively. The pooled κ 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.[40] In our review however, the group of maximum five reviewers received less training hours. Probably, they were better supervised

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.[40] 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.[65 66] 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*[16] showed that there was an association between κ and the prevalence of AEs. We found no statistically significantly association between κ 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*[16] (2.8–58.9%) could explain why we did not find an association between κ 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[20] to evaluate the methodological quality of the included studies. Second, it was not possible to formally estimate the pooled κ 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 κ values. However, specific agreement on the presence of AE, expressing the agreement separately for the positive and negative ratings, is recommended.[67] 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
[3]Department of Experimental Psychology, University of Oxford, Oxford, UK
[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 Reinier 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. HCWdV supervised the data analysis. HCWdV 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.

REFERENCES

1. Andermann A, Wu AW, Lashoher A, *et al*. Case studies of patient safety research classics to build research capacity in low- and middle-income countries. *Jt Comm J Qual Patient Saf* 2013;39:553–60.
2. Duckers M, Faber M, Cruijsberg J, *et al*. Safety and risk management interventions in hospitals: a systematic review of the literature. *Med Care Res Rev* 2009;66(6 Suppl):90S–119S.
3. Brennan TA, Leape LL, Laird NM, *et al*. Incidence of adverse events and negligence in hospitalized patients. Results of the Harvard Medical Practice Study I. *N Engl J Med* 1991;324:370–6.
4. Denis J-L. Accountability in healthcare organizations and systems. *Healthc Policy* 2014;10:8–11.
5. Werner RM, Asch DA. The unintended consequences of publicly reporting quality information. *JAMA* 2005;293:1239–44.
6. Weingart SN, Davis RB, Palmer RH, *et al*. Discrepancies between explicit and implicit review: physician and nurse assessments of complications and quality. *Health Serv Res* 2002;37:483–98.
7. Kennerly DA, Kudyakov R, da Graca B, *et al*. Characterization of adverse events detected in a large health care delivery system using an enhanced Global Trigger Tool over a five-year interval. *Health Serv Res* 2014;49:1407–25.
8. Rutberg H, Borgstedt Risberg M, Sjodahl R, *et al*. Characterisations of adverse events detected in a university hospital: a 4-year study using the Global Trigger Tool method. *BMJ Open* 2014;4:e004879.
9. Christiaans-Dingelhoff 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.
10. Classen DC, Resar R, Griffin F, *et al*. 'Global Trigger Tool' shows that adverse events in hospitals may be ten times greater than previously measured. *Health Aff (Millwood)* 2011;30:581–9.

11. Sari AB, Sheldon TA, Cracknell A, *et al*. Extent, nature and consequences of adverse events: results of a retrospective casenote review in a large NHS hospital. *Qual Saf Health Care* 2007;16:434–9.
12. Vincent C, Burnett S, Carthey J. *The measurement and monitoring of safety*. The Health Foundation, 2013.
13. Thomas EJ, Petersen LA. Measuring errors and adverse events in health care. *J Gen Intern Med* 2003;18:61–7.
14. Tsang C, Aylin P, Palmer W. Patient safety indicators: a systematic review of the literature. London, UK: Dr. Foster Unit, Imperial College. October 2008.
15. Murff HJ, Patel VL, Hripcsak G, *et al*. Detecting adverse events for patient safety research: a review of current methodologies. *J Biomed Inform* 2003;36:131–43.
16. Lilford R, Edwards A, Girling A, *et al*. Inter-rater reliability of case-note audit: a systematic review. *J Health Serv Res Policy* 2007;12:173–80.
17. Moher D, Liberati A, Tetzlaff J, *et al*. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: the PRISMA statement. *Ann Intern Med* 2009;151:264–9.
18. Mokkink LB, Terwee CB, Patrick DL, *et al*. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *J Clin Epidemiol* 2010;63:737–45.
19. Vet HCWd, Terwee CB, Mokkink LB, *et al*. *Measurement in medicine. A practical guide*. Cambridge: Cambridge University Press, 2011.
20. Cosmin. Secondary. http://www.cosmin.nl/ (accessed 4 Dec 2015).
21. Terwee CB, Mokkink LB, Knol DL, *et al*. Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. *Qual Life Res* 2012;21:651–7.
22. Mokkink LB, Terwee CB, Patrick DL, *et al*. *The COSMIN checklist manual*. Amsterdam: VU University Medical Centre, 2009.
23. Landis JR, Koch GG. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics* 1977;33:363–74.
24. Flynn EA, Barker KN, Pepper GA, *et al*. Comparison of methods for detecting medication errors in 36 hospitals and skilled-nursing facilities. *Am J Health Syst Pharm* 2002;59:436–46.
25. Forster AJ, Taljaard M, Bennett C, *et al*. Reliability of the peer-review process for adverse event rating. *PLoS One* 2012;7:e41239.
26. Forster AJ, O'Rourke K, Shojania KG, *et al*. Combining ratings from multiple physician reviewers helped to overcome the uncertainty associated with adverse event classification. *J Clin Epidemiol* 2007;60:892–901.
27. Nettleman MD, Nelson AP. Adverse occurrences during hospitalization on a general medicine service. *Clin Perform Qual Health Care* 1994;2:67–72.
28. Michel P, Quenon JL, Djihoud A, *et al*. French national survey of inpatient adverse events prospectively assessed with ward staff. *Qual Saf Health Care* 2007;16:369–77.
29. Michel P, Quenon JL, de Sarasqueta AM, *et al*. Comparison of three methods for estimating rates of adverse events and rates of preventable adverse events in acute care hospitals. *BMJ* 2004;328:199.
30. Hayward RA, Hofer TP. Estimating hospital deaths due to medical errors: preventability is in the eye of the reviewer. *JAMA* 2001;286:415–20.
31. Classen DC, Lloyd RC, Provost L, *et al*. Development and evaluation of the Institute for Healthcare Improvement Global Trigger Tool. *J Patient Saf* 2008;4:169–77.
32. Brennan TA, Localio RJ, Laird NL. Reliability and validity of judgments concerning adverse events suffered by hospitalized patients. *Med Care* 1989;27:1148–58.
33. Hofer TP, Bernstein SJ, DeMonner S, *et al*. Discussion between reviewers does not improve reliability of peer review of hospital quality. *Med Care* 2000;38:152–61.
34. GriffinFAResarRK I. *Global Trigger Tool for measuring adverse events*. 2nd edn. IHI Innovation Series white paper. Cambridge, MA: Institute for Healthcare Improvement, 2009.
35. Schildmeijer K, Nilsson L, Arestedt K, *et al*. Assessment of adverse events in medical care: lack of consistency between experienced teams using the Global Trigger Tool. *BMJ Qual Saf* 2012;21:307–14.
36. Kennerly DA, Saldana M, Kudyakov R, *et al*. Description and evaluation of adaptations to the Global Trigger Tool to enhance value to adverse event reduction efforts. *J Patient Saf* 2013;9:87–95.
37. Najjar S, Hamdan M, Euwema MC, *et al*. The Global Trigger Tool shows that one out of seven patients suffers harm in Palestinian hospitals: challenges for launching a strategic safety plan. *Int J Qual Health Care* 2013;25:640–7.
38. 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;20:27–35.
39. Baines RJ, Langelaan M, de Bruijne MC, *et al*. Changes in adverse event rates in hospitals over time: a longitudinal retrospective patient record review study. *BMJ Qual Saf* 2013;22:290–8.
40. Zegers M, de Bruijne MC, Wagner C, *et al*. The inter-rater agreement of retrospective assessments of adverse events does not improve with two reviewers per patient record. *J Clin Epidemiol* 2010;63:94–102.
41. Thomas EJ, Studdert DM, Burstin HR, *et al*. Incidence and types of adverse events and negligent care in Utah and Colorado. *Med Care* 2000;38:261–71.
42. Localio AR, Weaver SL, Landis JR, *et al*. Identifying adverse events caused by medical care: degree of physician agreement in a retrospective chart review. *Ann Intern Med* 1996;125:457–64.
43. Mattsson TO, Knudsen JL, Lauritsen J, *et al*. Assessment of the Global Trigger Tool to measure, monitor and evaluate patient safety in cancer patients: reliability concerns are raised. *BMJ Qual Saf* 2013;22:571–9.
44. Kirkendall ES, Kloppenborg E, Papp J, *et al*. Measuring adverse events and levels of harm in pediatric inpatients with the Global Trigger Tool. *Pediatrics* 2012;130:e1206–14.
45. Naessens JM, O'Byrne TJ, Johnson MG, *et al*. Measuring hospital adverse events: assessing inter-rater reliability and trigger performance of the Global Trigger Tool. *Int J Qual Health Care* 2010;22:266–74.
46. Sharek PJ, Parry G, Goldmann D, *et al*. Performance characteristics of a methodology to quantify adverse events over time in hospitalized patients. *Health Serv Res* 2011;46:654–78.
47. Matlow AG, Cronin CM, Flintoft V, *et al*. Description of the development and validation of the Canadian Paediatric Trigger Tool. *BMJ Qual Saf* 2011;20:416–23.
48. Hogan H, Healey F, Neale G, *et al*. Preventable deaths due to problems in care in English acute hospitals: a retrospective case record review study. *BMJ Qual Saf* 2012;21:737–45.
49. Soop M, Fryksmark U, Köster M, *et al*. The incidence of adverse events in Swedish hospitals: a retrospective medical record review study. *Int J Qual Health Care* 2009;21:285–91.
50. Forster AJ, Asmis TR, Clark HD, *et al*. Ottawa Hospital Patient Safety Study: incidence and timing of adverse events in patients admitted to a Canadian teaching hospital. *CMAJ* 2004;170:1235–40.
51. Baker GR, Norton PG, Flintoft V, *et al*. The Canadian Adverse Events Study: the incidence of adverse events among hospital patients in Canada. *CMAJ* 2004;170:1678–86.
52. Davis P, Lay-Yee R, Briant R, *et al*. Adverse events in New Zealand public hospitals: principal findings from a national survey. Occasional paper no 3. New Zealand: Ministry of Health, 2001.
53. Thomas EJ, Lipsitz SR, Studdert DM, *et al*. The reliability of medical record review for estimating adverse event rates. *Ann Intern Med* 2002;136:812–16.
54. Wilson RM, Runciman WB, Gibberd RW, *et al*. The quality in Australian health care study. *Med J Aust* 1995;163:458–71.
55. Langelaan M, Baines R, Broekens M, *et al*. *Monitor zorggerelateerde schade 2008. Dossieronderzoek in Nederlandse ziekenhuizen (Patient files study in Dutch hospitals)*. Report of EMGO Institute & VUmc/NIVEL, Amsterdam/Utrecht. 2010.
56. Zegers M, De Bruijne M, Wagner C, *et al*. Adverse events and potentially preventable deaths in Dutch hospitals: results of a retrospective patient record review study. *Qual Saf Health Care* 2009;18:297–302.
57. Davis P, Lay-Yee R, Briant R, *et al*. Adverse events in New Zealand public hospitals II: preventability and clinical context. *NZ Med J* 2003;116:U624.
58. Leape LL, Brennan TA, Laird N, *et al*. The nature of adverse events in hospitalized patients: results of the Harvard Medical Practice Study II. *N Engl J Med* 1991;324:377–84.
59. Zegers M, de Bruijne MC, Spreeuwenberg P, *et al*. Quality of patient record keeping: an indicator of the quality of care? *BMJ Qual Saf* 2011;20:314–18.
60. Association AH, Assiociation AS. Facts Clinical Registries. Secondary Facts Clinical Registries. http://www.heart.org/idc/groups/heart-public/@wcm/@adv/documents/downloadable/ucm_432451.pdf (accessed 4 Dec 2015).
61. Shojania KG, Burton EC, McDonald KM, *et al*. Changes in rates of autopsy-detected diagnostic errors over time: a systematic review. *JAMA* 2003;289:2849–56.

62. Michel P. *Strengths and weaknesses of available methods for assessing the nature and scale of harm caused by the health system: literature review*. World Health Organization, 2003.
63. Group WW. *Patient safety: rapid assessment methods for estimating hazards*. Report of the WHO Working Group meeting. Geneva, 17–19 December 2002.
64. Lilford RJ, Mohammed MA, Braunholtz D, *et al*. The measurement of active errors: methodological issues. *Qual Saf Health Care* 2003;12(Suppl 2):ii8–12.
65. von Plessen C, Kodal AM, Anhøj J. Experiences with Global Trigger Tool reviews in five Danish hospitals: an implementation study. *BMJ Open* 2012;2:e001324.
66. Schildmeijer K, Nilsson L, Perk J, *et al*. Strengths and weaknesses of working with the Global Trigger Tool method for retrospective record review: focus group interviews with team members. *BMJ Open* 2013;3:e003131.
67. de Vet HC, Mokkink LB, Terwee CB, *et al*. Clinicians are right not to like Cohen's κ. *BMJ* 2013;346:f2125.