WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

How Reliable Is the Severity Recorded in Hospital Incident Reports? Comparing Frontline and Expert Grading in an Indian Tertiary-Care Hospital

Dr. Deepak Dongre[a,*], Purva Agrawal[b], Ms. Pooja Tulara[c], Dr. Amit Kumar Singh[d]

[a] Medical Administrator and Quality Head, Nagpur, Maharashtra, India
[b] Pharmacist and MBA (Hospital Management), India

A B S T R A C T :

Background: Every incident report carries a severity grade, and that grade decides whether the event gets a full investigation. Few hospitals ever check whether the grade is right.

Objective: To compare the severity chosen by the person who reports an incident with the Severity Assessment Code (SAC) given by expert reviewers, and to see how well two expert reviewers agree with each other.

Methods: We studied all 1,083 incident reports made between August 2024 and October 2026 in a private tertiary-care hospital in Nagpur, India. For 1,075 reports, the reporter's grade (minor, moderate or severe) was compared with the first expert's SAC consequence. A second expert review was available for 726 reports. We used percentage agreement, Cohen's kappa and logistic regression, and we also checked basic data quality.

Results: The reporter and the expert gave the same grade in 446 reports (41.5%; kappa 0.13). Reporters graded 467 incidents (43.4%) lower than the expert and 162 (15.1%) higher. Falls (odds ratio 4.42) and injuries (3.11) were the incidents most often graded too low. The two experts seemed to agree well on consequence (kappa 0.65), but 411 of the 726 second reviews (56.6%) simply endorsed the first one. When the second expert gave his or her own reasons, kappa fell to 0.41 for consequence and 0.14 for likelihood. Nearly a third of occurrence times were rounded to the hour, and 17 of 22 sentinel or never events were marked as causing no or minor harm.

Conclusion: The reporter's grade was a poor guide to the expert's view, and expert agreement looked better than it really was. Any hospital can repeat this simple self-audit with its own data before it trusts its severity figures.

Keywords: patient safety; incident reporting; severity assessment code; reliability; kappa; data quality; India

1. Introduction

Hospitals learn about harm mainly through incident reports (Macrae, 2016; Mitchell et al., 2016). A nurse, doctor or other staff member writes down what went wrong and says how serious it was. That choice matters. A report marked "minor" may be closed on the ward the same day. A report marked "severe" may lead to a root cause analysis, a meeting with management and a change in practice (Reason, 2000).

Many hospitals add a second step. A head of department (HOD) reviews each report and gives it a Severity Assessment Code, or SAC. The SAC was first used in the United States Veterans Affairs health system (Bagian et al., 2002). It combines two judgements: how bad the outcome was or could have been (consequence), and how likely the event is to happen again (likelihood). Together they give a score from 1 to 3, and a score of 3 calls for a full root cause analysis (U.S. Department of Veterans Affairs, n.d.). In India, hospitals accredited by the National Accreditation Board for Hospitals and Healthcare Providers (NABH) are expected to run such a system and to act on what it shows (NABH, 2020).

Studies that used written case stories found that staff do not apply harm scales in the same way (Williams et al., 2015; Abbasi et al., 2015). Training improved agreement only a little (Lee & Shin, 2023). Other work shows that incident reports pick up only a small share of the harm found when patient records are checked (Sari et al., 2007; Westbrook et al., 2015; Noble & Pronovost, 2010). Very few studies, however, have used real reports to set the reporter's grade beside the expert's grade. Fewer still have asked whether a second expert truly forms an own opinion or just agrees with the first. We did not find any such study from India.

We therefore asked three simple questions. How often does the reporter's grade match the expert's? Do two experts agree when each one judges the report independently? And is the basic data in the reporting system good enough to rely on?

2. Methods

2.1 Study design and setting

This was a retrospective study of routine hospital data. We followed the Guidelines for Reporting Reliability and Agreement Studies, known as GRRAS (Kottner et al., 2011). The study hospital is a private tertiary-care multispecialty hospital in Nagpur, Maharashtra. It offers cardiac, cancer, neuro, critical care, kidney and dialysis, and mother and child services. Any staff member can report an incident through an electronic system, and reporting is voluntary.

2.2 How an incident report is graded

Each report passes through four steps (Figure 1). The reporter fills in the form and picks minor, moderate or severe. The reporter also ticks a harm level and an event type, such as near miss or sentinel event. The HOD of the department concerned then reviews the report. The HOD chooses a consequence (minor, moderate, major or severe) and a likelihood (remote, uncommon, occasional or frequent), and the software turns these into a SAC score of 1, 2 or 3. When another department is involved, its HOD adds a second review with a separate SAC. The second reviewer can see the first review. Last, the quality team and management check the actions taken and close the report.

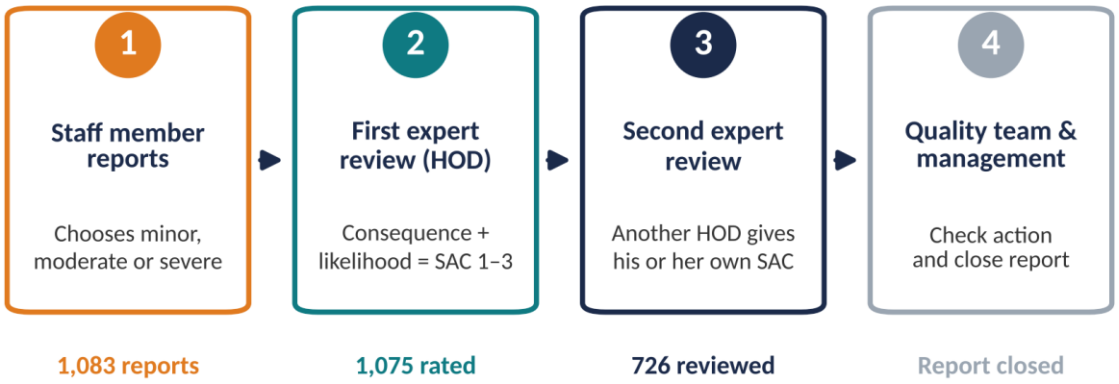


Figure 1. How an incident report moves through grading at the study hospital, with the number of reports at each step.

2.3 Data source and sample

We exported every report with an occurrence date from 27 August 2024 to 25 October 2026. Seventeen reports had no occurrence time; we kept them because the rest of each record was complete. In all, 1,083 reports were included, and no sampling was done. Eight reports had no expert SAC and were left out of the agreement analysis, which left 1,075. A second expert review was present in 726 reports. Before analysis, all patient names, staff names, hospital numbers and bed numbers were removed.

2.4 What we compared

The reporter's scale has three levels and the SAC scale has four. To compare them, we joined the expert's "major" and "severe" into one level called "severe". A report was called **under-graded** when the expert's level was higher than the reporter's, and **over-graded** when it was lower. We repeated the comparison in three other ways to see whether this choice changed the result (Table 2).

A second review was called an **endorsement** when its explanation did no more than agree with the first review, using words such as "satisfactory", "as per above" or "as explained by" the first reviewer. All other second reviews were called **independent**. We also counted simple data problems: times recorded exactly on the hour, a missing department, harm labels that contradict the event type, and incident types left as "Others". Reporting delay was the time from the event to the report.

2.5 Statistical analysis

We worked out percentage agreement with Wilson 95% confidence intervals, or CI (Newcombe, 1998). We also used Cohen's kappa (Cohen, 1960) and weighted kappa (Cohen, 1968). Kappa shows how much two raters agree beyond what chance alone would give; 0 means agreement no better than chance and 1 means perfect agreement. We described kappa with the Landis and Koch (1977) bands: 0.20 or less is slight, 0.21 to 0.40 fair, 0.41 to 0.60 moderate and 0.61 to 0.80 substantial. Because kappa can be low even when raw agreement is high (Feinstein & Cicchetti, 1990), we also show the full counts.

The CI for each kappa came from 2,000 bootstrap samples. Logistic regression showed which kinds of reports were more likely to be under-graded. Delays were compared with the Kruskal–Wallis test. A p value below 0.05 was taken as significant. All analysis was done in Python 3.

2.6 Ethical considerations

The study used routine, anonymised data collected for quality improvement. No patient was contacted and no care was changed. No names or other identifiers appear in the data set or in this paper.

3. Results

3.1 Profile of the reports

There were 1,083 reports over 26 months. Most of them (82.5%) involved in-patients. Nearly half (48.9%) had the incident type left as "Others"; medication errors made up 23.1% and injuries 13.4% (Figure 2A). Reporters chose "minor" for 47.8% of reports, while experts chose "minor" for only 27.7% and "major" or "severe" for 32.4% (Figure 2B, Table 1).

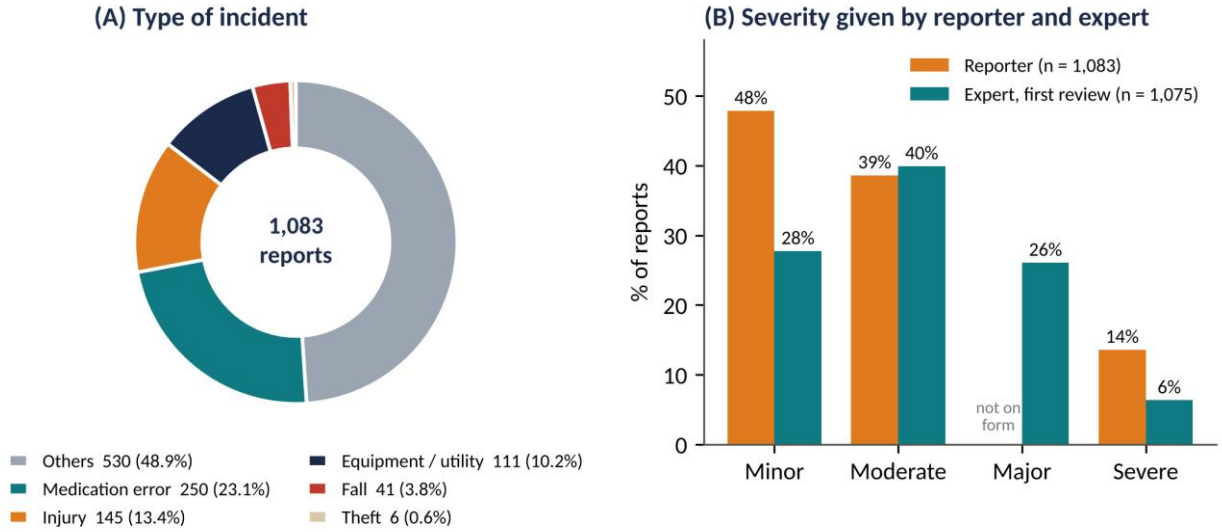


Figure 2. (A) Types of incident reported (n = 1,083). (B) Severity chosen by reporters and by the first expert reviewer.

Table 1. Profile of the incident reports

Characteristic	Group	n (%)
Event related to (n = 1,083)	In-patient	893 (82.5)
	Other, staff, out-patient, facility, visitor	190 (17.5)
Event type (reporter)	Incident	824 (76.1)
	Near miss	121 (11.2)
	No harm	84 (7.8)
	Adverse event	32 (3.0)
	Sentinel or never event	22 (2.0)
Reporter grade (n = 1,083)	Minor / Moderate / Severe	518 (47.8) / 418 (38.6) / 147 (13.6)
Expert consequence (n = 1,075)	Minor / Moderate	298 (27.7) / 429 (39.9)
	Major / Severe	280 (26.0) / 68 (6.3)
SAC score (n = 1,075)	1 / 2 / 3	616 (57.3) / 366 (34.0) / 93 (8.7)

3.2 Reporter versus expert

The reporter and the expert gave the same grade in 446 of 1,075 reports (41.5%; 95% CI 38.6–44.5). Kappa was 0.13 (95% CI 0.09–0.17), which is slight agreement; weighted kappa was 0.16. The disagreement mostly ran one way. Reporters graded 467 incidents (43.4%) lower than the expert and only 162 (15.1%) higher (Figure 3A).

Of the 512 events that reporters called minor, experts rated 323 (63.1%) as moderate or worse and 131 (25.6%) as major or severe (Figure 3B). Of the 147 events that reporters called severe, the expert agreed in 73 (49.7%). The result stayed much the same when the two scales were joined in other ways (Table 2).

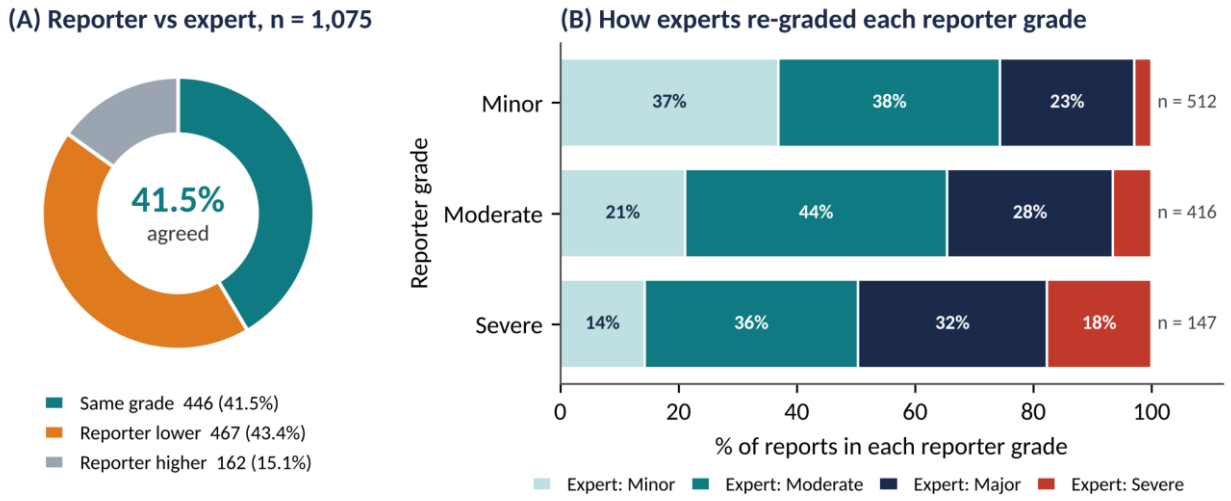


Figure 3. (A) Agreement between reporter and expert (n = 1,075). (B) How the expert graded the reports in each reporter grade.

Table 2. Agreement between the reporter's grade and the expert review, tested in four ways

Comparison	n	Same grade, % (95% CI)	Kappa (95% CI)	Graded lower / higher
Main analysis: expert major + severe joined	1,075	41.5 (38.6–44.5)	0.13 (0.09–0.17)	467 / 162
Reporter grade against SAC score (1–3)	1,075	47.5 (44.6–50.5)	0.10 (0.05–0.15)	220 / 344
All four expert levels kept	1,075	37.1 (34.3–40.0)	0.11 (0.07–0.14)	467 / 209
Only reports with a valid event time	1,058	41.5 (38.6–44.5)	0.12 (0.08–0.16)	457 / 162

Linear weighted kappa for the main analysis: 0.16 (0.11–0.20). CI = confidence interval.

3.3 Which incidents were under-graded

Falls were graded too low in 77.5% of cases (31 of 40) and injuries in 68.3% (99 of 145), against 43.4% for all reports (Figure 4A). After allowing for the other factors, a fall had more than four times the odds of being under-graded (odds ratio 4.42; 95% CI 2.05–9.54) and an injury about three times (3.11; 2.09–4.64). Medication errors (0.65; 0.47–0.91) and events not involving in-patients (0.46; 0.30–0.71) were less likely to be graded too low (Figure 4B). Event type, reporting delay and conflicting harm labels made no clear difference. The model's area under the curve was 0.66, so these factors explain only part of the problem.

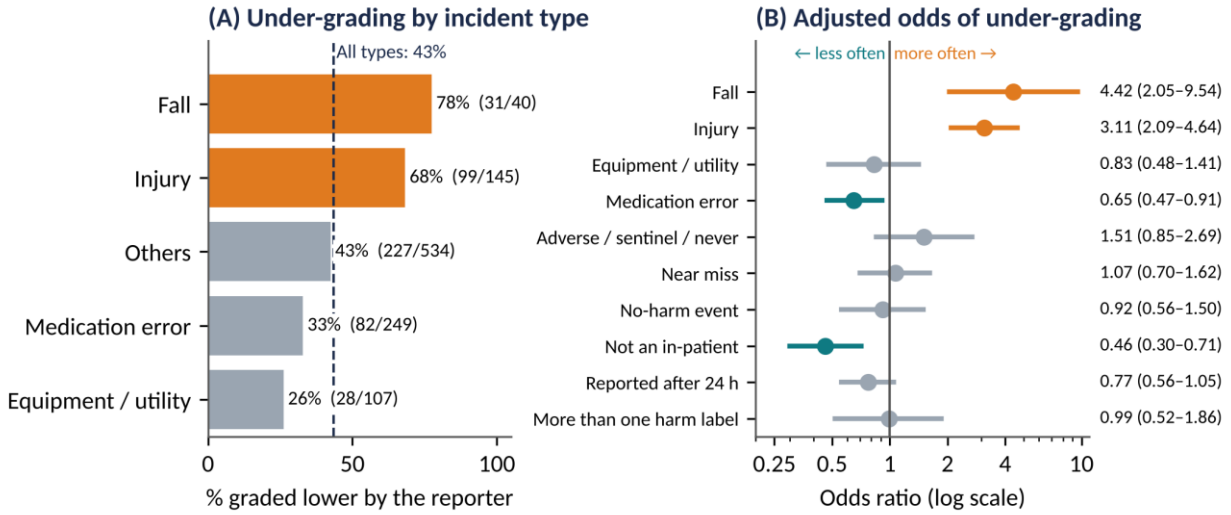


Figure 4. (A) Share of reports graded lower by the reporter, by incident type. (B) Adjusted odds ratios for under-grading from the logistic regression model (n = 1,075).

3.4 Agreement between two experts

Looked at as a whole, the two experts agreed substantially on consequence (kappa 0.65) and fairly on likelihood (kappa 0.37). This picture changed once the type of second review was taken into account. Of the 726 second reviews, 411 (56.6%) only endorsed the first review (Figure 5A). In these, agreement was naturally very high (kappa 0.83 for consequence).

In the 315 independent second reviews, kappa fell to 0.41 (95% CI 0.34–0.48) for consequence, 0.14 (0.07–0.22) for likelihood and 0.33 (0.24–0.43) for the SAC score (Figure 5B). Two independent experts chose the same consequence in 58.4% of reports and the same likelihood in only 42.5%. Reports that received a second review were also more serious than the rest (36.0% versus 24.9% major or severe; $p < 0.001$).

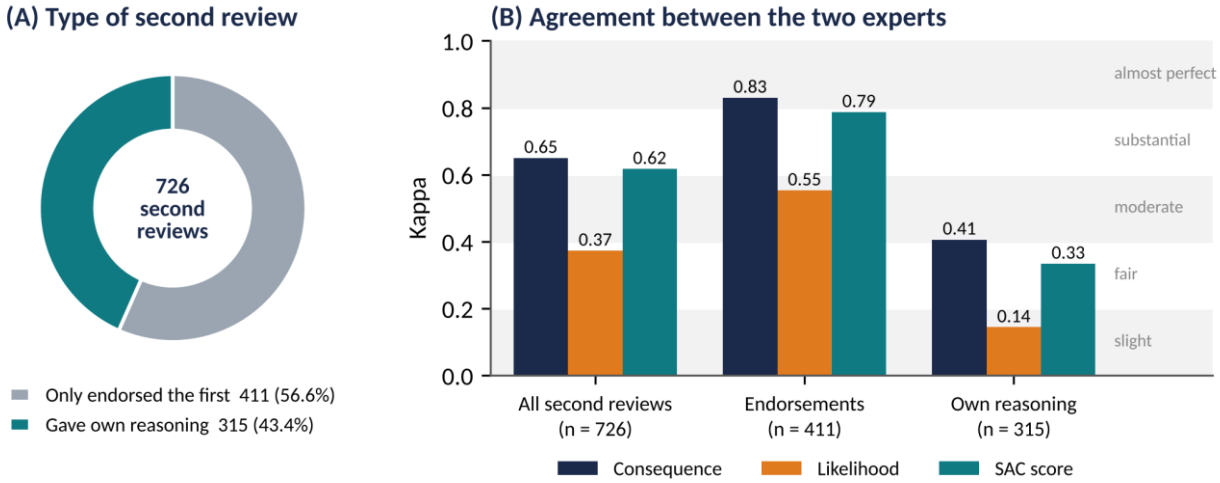


Figure 5. (A) Second reviews that only endorsed the first versus those that gave their own reasons. (B) Kappa between the first and second expert. Shaded bands show the Landis and Koch levels.

3.5 Data quality and reporting delay

Out of 1,066 reports with a time, 310 (29.1%) had the event time exactly on the hour, against about 1.7% expected by chance; 55 were set at exactly 12:00. The reporter's department was "undefined" in 177 reports (16.3%). Of the 22 events labelled sentinel or never events, 17 (77.3%) were marked as causing no or minor harm; examples included a broken oxygen flowmeter and a wall cabinet that fell. Twenty-two of 121 near misses (18.2%) showed harm, although a near miss by definition does not reach the patient (Figure 6A).

The median time from event to report was 3.2 hours (IQR 0.3–24.2), and 74.3% of reports were made within 24 hours. Delay rose with the reporter's grade: 1.3 hours for minor, 4.9 hours for moderate and 14.6 hours for severe events ($p < 0.001$; Figure 6B). It did not rise steadily with the expert's grade (1.4, 4.4, 4.4 and 2.1 hours from minor to severe). Management closed 986 of 1,033 reviewed reports (95.5%) with the same standard words, "root cause satisfied, solution satisfied".

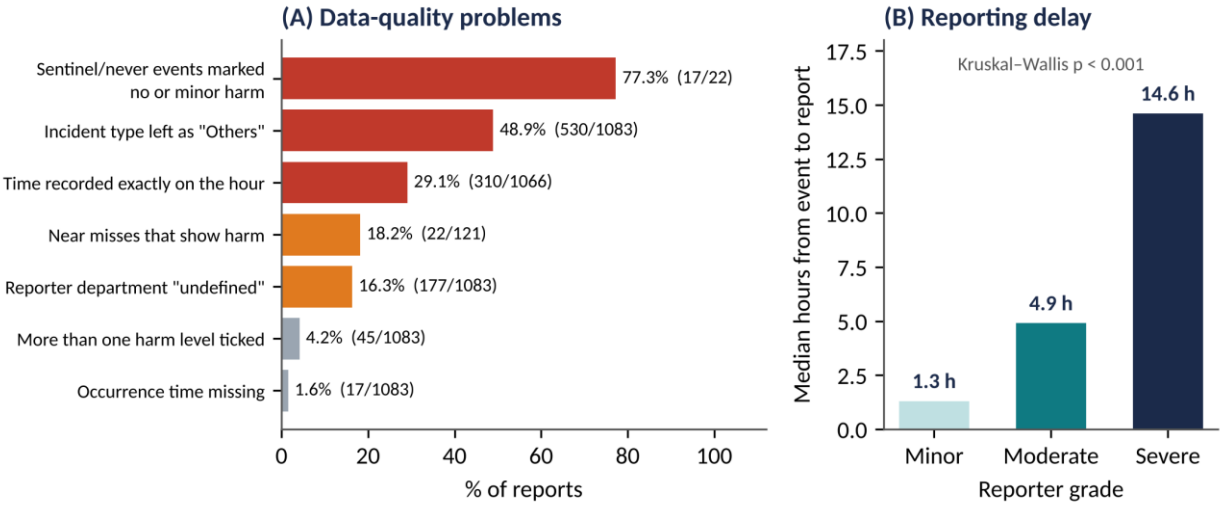


Figure 6. (A) Data-quality problems in the reporting system. (B) Median hours from event to report, by the reporter's grade.

4. Discussion

In this hospital, the reporter and the expert agreed on severity in only two of every five reports, and kappa was only a little above chance. Reporters usually rated events lower than the experts did. On its own, then, the reporter's grade cannot be used to decide which events need an investigation.

Some of this gap is expected. The reporter writes what he or she sees at the bedside, often within minutes. The expert reviews later, knows more, and is asked to judge what could have happened, not only what did happen. Falls and injuries show this clearly, because the harm often appears hours later. In our data, one fall in the dialysis unit was reported as causing minimal harm; an X-ray later showed a hip fracture. The two scales also differ in size (three levels against four), which by itself invites confusion. Fear of blame may play a part as well, though our data cannot show this.

Our results fit with earlier work. Trained safety officers reached only modest agreement when grading written cases (Williams et al., 2015), nurses agreed poorly (Abbasi et al., 2015), and training raised nurses' kappa only from 0.23 to 0.31 (Lee & Shin, 2023). Our kappa of 0.13 on real reports is lower still. This is not surprising: written cases give every rater the same facts, while in daily work the reporter and the expert see different facts at different times.

The expert finding is just as important. A headline kappa of 0.65 between two experts looks reassuring, but most of it came from endorsements. When the second expert thought the case through independently, agreement on likelihood was close to chance (kappa 0.14). Likelihood feeds straight into the SAC score, so two experts could easily reach different answers on whether a root cause analysis is needed. Rounded times, missing departments, contradictory labels and standard closing words all point the same way: the severity data may be weaker than a dashboard suggests (Macrae, 2016).

4.1 What other hospitals can do

These methods can be reused. Most hospitals that use a frontline scale and a SAC or risk matrix, including those preparing for NABH or Joint Commission International accreditation, already hold the same fields. Any of them can repeat this study as a self-audit using a routine export and a spreadsheet (Figure 7). Table 3 links each finding to a practical action and to an indicator that can be tracked.

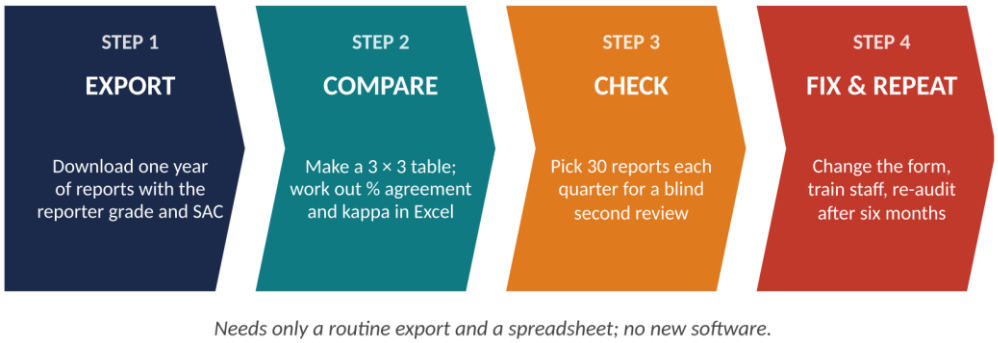


Figure 7. A four-step self-audit that other hospitals can adapt (proposed; not yet tested).

Table 3. Turning the findings into action

Finding in this study	What a hospital can do	Indicator to track
Reporter and expert scales differ (3 vs 4 levels)	Use one shared scale with written definitions and examples	% same grade; kappa
Falls and injuries graded too low	Add grading tips for falls and injuries; re-grade after the X-ray or scan	% falls and injuries under-graded
Most second reviews only endorse the first	Blind double review of a random sample each quarter; ask for a reason when grades differ	Kappa between independent reviewers
Sentinel events labelled "no harm"	Single-choice harm field; automatic alert for entries that contradict each other	% reports with conflicting labels
Times rounded; department missing	No default time; compulsory department drop-down list	% times on the hour; % undefined
Severe events reported later	Encourage early reporting with a provisional grade	Median delay by grade

Most of these actions are changes to settings and training, so they need no new software. That makes them practical for Indian hospitals with limited resources. Since our results come from one hospital, each hospital should first measure its own starting point. The proposed actions should then be tested in a before-and-after or multicentre study.

4.2 Strengths and limitations

The study covered every report over more than two years, used real reports rather than written cases, and tested whether the experts truly judged independently. It also has limits. It comes from a single hospital. The experts were not blinded and could see the reporter's grade. Endorsement was judged from the wording of the review. The reports that received a second review were not a random sample. Rounded times suggest, but do not prove, that times were entered from memory. Finally, neither the reporter's grade nor the expert's grade is a true gold standard.

5. Conclusion

The severity recorded by the person reporting an incident was a weak guide to the expert's judgement, and agreement between experts was made to look better by endorsements. Hospitals should use one shared severity scale, check reviewer agreement independently from time to time, and fix basic data quality before they use incident severity to set priorities.

Declarations

Conflict of interest: The authors declare no conflict of interest.

Funding: No external funding was received for this study.

Data availability: The anonymized, coded data set is available from the corresponding author on reasonable request, subject to approval by the hospital.

Author contributions: DD: idea, data access, analysis and writing. PA: data preparation, literature review and editing. Both authors approved the final version. Please confirm.

Use of AI tools: The authors take full responsibility for the manuscript's content.

Acknowledgements: The authors thank the staff and heads of department who reported and reviewed incidents.

REFERENCES

- Abbasi, T., Adornetto-Garcia, D., Johnston, P. A., Segovia, J. H., & Summers, B. (2015). Accuracy of harm scores entered into an event reporting system. *Journal of Nursing Administration, 45*(4), 218–225. https://doi.org/10.1097/NNA.0000000000000188
- Bagian, J. P., Gosbee, J., Lee, C. Z., Williams, L., McKnight, S. D., & Mannos, D. M. (2002). The Veterans Affairs root cause analysis system in action. *Joint Commission Journal on Quality Improvement, 28*(10), 531–545.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46. https://doi.org/10.1177/001316446002000104
- Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin, 70*(4), 213–220. https://doi.org/10.1037/h0026256
- Feinstein, A. R., & Cicchetti, D. V. (1990). High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology, 43*(6), 543–549. https://doi.org/10.1016/0895-4356(90)90158-L
- Kottner, J., Audigé, L., Brorson, S., Donner, A., Gajewski, B. J., Hróbjartsson, A., Roberts, C., Shoukri, M., & Streiner, D. L. (2011). Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. *Journal of Clinical Epidemiology, 64*(1), 96–106. https://doi.org/10.1016/j.jclinepi.2010.03.002
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174. https://doi.org/10.2307/2529310
- Lee, K., & Shin, E. (2023). Comparison of perception differences among nurses based on harm assessment education in the patient safety classification system. *The Open Nursing Journal, 17*, e187443462308170. https://doi.org/10.2174/18744346-v17-230927-2023-5
- Macrae, C. (2016). The problem with incident reporting. *BMJ Quality & Safety, 25*(2), 71–75. https://doi.org/10.1136/bmjqs-2015-004732
- Mitchell, I., Schuster, A., Smith, K., Pronovost, P., & Wu, A. (2016). Patient safety incident reporting: A qualitative study of thoughts and perceptions of experts 15 years after 'To Err is Human'. *BMJ Quality & Safety, 25*(2), 92–99. https://doi.org/10.1136/bmjqs-2015-004405
- National Accreditation Board for Hospitals & Healthcare Providers. (2020). *Accreditation standards for hospitals* (5th ed.). Quality Council of India.
- Newcombe, R. G. (1998). Two-sided confidence intervals for the single proportion: Comparison of seven methods. *Statistics in Medicine, 17*(8), 857–872.
- Noble, D. J., & Pronovost, P. J. (2010). Underreporting of patient safety incidents reduces health care's ability to quantify and accurately measure harm reduction. *Journal of Patient Safety, 6*(4), 247–250. https://doi.org/10.1097/PTS.0b013e3181fd1697
- Reason, J. (2000). Human error: Models and management. *BMJ, 320*(7237), 768–770. https://doi.org/10.1136/bmj.320.7237.768
- Sari, A. B., Sheldon, T. A., Cracknell, A., & Turnbull, A. (2007). Sensitivity of routine system for reporting patient safety incidents in an NHS hospital: Retrospective patient case note review. *BMJ, 334*(7584), 79. https://doi.org/10.1136/bmj.39031.507153.AE

- U.S. Department of Veterans Affairs, National Center for Patient Safety. (n.d.). *Safety Assessment Code (SAC) matrix*. Retrieved October 27, 2026, from https://www.patientsafety.va.gov/professionals/publications/matrix.asp
- Westbrook, J. I., Li, L., Lehnbom, E. C., Baysari, M. T., Braithwaite, J., Burke, R., Conn, C., & Day, R. O. (2015). What are incident reports telling us? A comparative study at two Australian hospitals of medication errors identified at audit, detected by staff and reported to an incident system. *International Journal for Quality in Health Care, 27*(1), 1–9. https://doi.org/10.1093/intqhc/mzu098
- Williams, T. L., Szekendi, M. K., Pavkovic, S., Clevenger, W., & Cerese, J. (2015). The reliability of AHRQ Common Format Harm Scales in rating patient safety events. *Journal of Patient Safety, 11*(1), 52–59. https://doi.org/10.1097/PTS.0b013e3182948ef9