Temporal trends in patient safety in the Netherlands: reductions in preventable adverse events or the end of adverse events as a useful metric?

Kaveh G Shojania,[1,2] Perla J Marang-van de Mheen[3]

[1]Department of Medicine, Sunnybrook Health Sciences Centre and the University of Toronto, Toronto, Ontario, Canada
[2]University of Toronto Centre for Quality Improvement and Patient Safety, Toronto, Ontario, Canada
[3]Department of Medical Decision Making, Leiden University Medical Centre, Leiden, The Netherlands

Correspondence to
Dr Kaveh G Shojania, Department of Medicine, Sunnybrook Health Sciences Centre and the University of Toronto, Room H468, 2075 Bayview Avenue Toronto, Ontario, Canada M4N 3M5; kaveh.shojania@sunnybrook.ca

Accepted 3 June 2015
Published Online First 6 July 2015

▸ http://dx.doi.org/10.1136/bmjqs-2014-003702
▸ http://dx.doi.org/10.1136/bmjqs-2015-004403

To cite: Shojania KG, Marang-van de Mheen PJ. *BMJ Qual Saf* 2015;**24**:541–544.

Two years ago, *BMJ Quality & Safety* published the first example of a longitudinal national adverse event (AE) study.[1] That study included 400 admissions from each of 21 randomly selected hospitals in the Netherlands in 2004 and 200 admissions from 20 hospitals in 2008. The authors reported an increase in AEs (ie, harm from medical care) from 4.1% in 2004 to 6.2% in 2008. Reassuringly, the preventable AE rate did not change, leaving one to wonder if the increase in non-preventable AE rates reflected better documentation in medical records (or just a chance finding). The lack of improvement in patient safety over time in the Netherlands mirrored the results of a US study that showed no improvement in preventable AEs from 2002 to 2007.[2]

Commenting on this lack of improvement over time, an editorial in *BMJ Quality & Safety* (including one of us as an author) suggested that, while the results at least partially reflect the paucity of effective patient safety interventions, they may also highlight limitations of AEs as a metric of improvement.[3] AEs represent a conceptually simple but practically heterogeneous category, including medication problems, healthcare-acquired infections, postoperative complications, delayed diagnoses, fall-related injuries, pressure ulcers, and many other errors and complications. This heterogeneity of AE types presents measurement problems because a broad effort to look at all AEs will probably not capture all events within a given category of interest.

Suppose institutions have generally targeted, say, surgical complications (with checklists), a few specific healthcare-associated infections (eg, catheter-associated bloodstream infections with the central line bundle) and medication-ordering errors (with clinical pharmacists and/or computerised order systems). Then, it makes more sense to capture these outcomes comprehensively than to partially capture all types of harm from medical care, including ones for which we have not implemented any effective interventions. With AEs as the metric, random error from incomplete data capture for specific outcomes of interest limits our ability to document improvements even if they have occurred, especially if reductions in one category of AE have been counterbalanced by increases in another.

Interestingly, Dutch investigators have now added a third time point to their previous study[1] and report a substantial albeit non-significant reduction in preventable AEs.[4] After adjustment for oversampling of deceased patients and patient characteristics, the preventable AE rate fell by 30% from 2008 to 2012 ($p=0.10$). Despite this encouraging signal of improvement, the editorial by Vincent and Amalberti[5] accompanying this latest study again calls for a move away from focusing on AEs and the use of more granular measurement, focusing on outcomes that capture the impacts of specific interventions. We agree. However, it may seem strange that a paper reporting possible improvements in preventable AEs should elicit critical reflections on the utility of AEs as a metric similar to those made in response to previous studies[1 2] that showed no improvement.

The previous study[1] showed zero evidence of improvement, so it made sense to wonder if the tool for measuring change might be inadequate. However, now we have a study showing a substantial reduction in preventable AEs. Even if not statistically significant, this signal of possible improvement surely shows that changes in AEs can be detected. It seems like the use of AEs receives criticism when rates do not improve, but also when they do. We discuss the case for abandoning AE rates as a measure of improvement over time. However, first we examine in more detail this latest study of AEs in the Netherlands and how confident we can be that the non-significant 30% reduction in preventable AEs relates to patient safety interventions implemented in the Netherlands in recent years.

REVIEWER AGREEMENT: THE ACHILLES' HEEL OF AE STUDIES

In the first Dutch AE study,[6] investigators used the standard method for measuring AEs, namely record review that begins with triggers or flags for possible quality-of-care problems (eg, unexpected death, unplanned readmission, unexpected admission to intensive care, adverse drug reactions, dissatisfaction with care documented in the medical record). Physicians then reviewed records with at least one trigger for the presence of AEs and made judgements about the preventability of any AEs they identified. Reviewers indicated the probability of prevention using a 6-point Likert scale, ranging from 1 (virtually no evidence for preventability) to 6 (certain evidence of preventability), with values of 3 and 4 capturing the transition from probably not preventable (less than 50/50 chance, but 'close call') to probably preventable (more than 50/50, but 'close call'). The main results in this and other such studies typically classify scores of 1–3 as non-preventable and 4–6 as preventable.[7–11]

While reviewers in such studies often identify the same AEs, they frequently disagree about preventability (or similar judgements about the presence of errors or negligence). AE studies frequently document the level of agreement between reviewers using the κ coefficient, which measures agreement beyond that expected on the basis of chance alone. A κ value of zero does not mean zero agreement. It means no more agreement than would occur from the reviewers flipping coins to make their judgements. In the original Harvard Medical Practice Study,[7] reviewers agreed on the characterisation of an event as an AE with κ=0.61 ('substantial agreement' beyond chance according to commonly used labels), whereas negligence had a κ of only 0.24 ('fair' agreement). Subsequent studies have used reviewer training or more structured review forms to increase agreement between reviewers, achieving κ scores in the 0.4–0.6 range even for the more difficult judgement of preventability of AEs.[6 11] However, even this level of persistent disagreement remains somewhat disturbing when it involves the 'gold standard' outcome for a field.

Given this problem with agreement between reviewers, it is notable that the two subsequent Dutch studies (including AEs from 2008[1] and 2012[4]) abandoned double review for identifying AEs and characterising preventability. The authors justified this methodological departure because they obtained acceptable agreement between pairs of reviewers in the 2004 study and because discussion between the reviewers working together did not improve overall agreement. Physicians who reviewed records as a pair showed substantial agreement in identifying AEs (κ of 0.64), but agreement between pairs of reviewers was only fair (with a κ of 0.25).[12]

Other investigators have also shown better agreement between reviewers who work together but poor agreement with other reviewers. In one study,[13] discussion between physicians who worked as a pair improved their agreement over time, but different pairs of reviewers showed particularly poor agreement (κ of 0.14), which barely improved with discussion (κ of 0.17). This study also showed better agreement between reviewers working together even before any discussion took place. It seems therefore that, after reviewing charts together and discussing disagreements encountered, reviewers adjust their perspectives about the presence of AEs and their preventability. This unconscious harmonisation of judgements masks the degree to which other reviewers (eg, other pairs of reviewers), even similarly expert ones, will continue to make different judgements about the same events.

Furthermore, changes in evidence over time (eg, between the different Dutch studies) may increase disagreement between reviewers about preventability of AEs. As mentioned by Vincent and Almaberti, rising standards of care may result in some AEs crossing the transition from less than 50/50 to more than 50/50 chance of being preventable. If reviewers agree, this will increase the proportion of preventable AEs. However, if reviewers differ in the extent to which they regard that new evidence has turned some AEs into preventable AEs (eg, central-line-associated infections or hospital-acquired delirium), they will disagree. The rate of preventable AEs obtained from studies with single reviewers then depends on which reviewers made the judgements.

It is tempting to think that the overall preventable AE rate might not change much as a result of using single review. One reviewer might have identified different preventable AEs, but the proportion of patients who experienced a preventable AE might remain similar. This may be true. However, it is also hard to know what to make of the preventability of an AE (already a collapsed, graded judgement on a scale) when another reviewer might not have even called it an AE in the first place. Simply put, there is an important error bar surrounding any estimated preventable AE rate.

HOW PLAUSIBLE IS A REDUCTION IN PREVENTABLE AEs FROM 2008 TO 2012?

Even if the lack of double review introduces an element of measurement error, the investigators report a fairly large reduction in preventable AEs of 30% between 2008 and 2012. While not statistically significant (p=0.10), this 30% reduction in preventable AEs could still reflect a true improvement. Maybe, therefore, we should consider a Bayesian perspective. If the hypothesis that no change in preventable AEs has occurred is sufficiently unlikely, then p=0.1 might provide adequate grounds for rejecting the null hypothesis. Unfortunately, using the Bayesian approach, the probability of the null hypothesis has to be less than about 17% in order for an observed p value of 0.1 to generate a final (or Bayesian posterior) probability of 5% or less that the results are due to chance alone.[14]

Why is this technical point about Bayesian inference useful to consider? Because it forces one to ask the question: do we really think, before seeing the data from the study, that the chance that preventable AEs had gone down in the Netherlands as the result of the national safety programme was at least 83%? This seems far too high. For one thing, very few patient safety interventions have shown significant improvements in patient outcomes. However, let us consider the plausibility of the specific improvements seen in the present study.

Most of the reductions in preventable AEs occurred in surgical patients and patients over 80. The Netherlands was the site of a major study of a surgical safety programme, including checklists at several stages of the surgical process.[15] This study reported a small but significant absolute reduction in mortality of 0.7% (95% CI 0.2% to 1.2%), and the proportion of patients with one or more complications decreased from 15.4% to 10.6% (p<0.001). These effects are of comparable magnitude to the 30% reduction in preventable AEs. The question is how likely is it that this programme was successfully implemented in a much larger group of Dutch hospitals over a 2–3-year period.

The current AE study does not report measures of implementation for any of the programme elements. A technical report[16] provides some data on implementation but, as mentioned by Baines *et al*,[4] about 19 hospitals participated in this evaluation study for each of the 10 themes, with few hospitals providing data for all themes. We cannot know from these self-reported data what stage of implementation each hospital really achieved, with what fidelity they replicated the original programme, or how such hospital-level implementation data relate to AE rates, as these data could not be linked. Published evaluations of implementation efforts for surgical checklists do not encourage the notion that hospitals will routinely reduce AEs. One recent study of surgical checklists from a province in Canada where the surgical checklist has been mandated showed no significant improvements in mortality or morbidity.[17] Another study of a more intensive effort to implement the surgical checklist as intended by its proponents[18] showed no reduction in postoperative complications, surgical site infections, or 30-day mortality.

Aside from the practical obstacles to implementing any intervention successfully, surgical checklists face the additional problem that the active ingredient of the intervention remains unclear: is it the checklist itself, changes in team interactions and safety culture, or some combination of the three?[19–22] Some institutions may focus on the checklist itself. Others may address the changes in teamwork intended by many proponents to accompany the checklist. Still others may choose to improve teamwork in operating room settings in alternative ways. These varying options highlight the complexity of interpreting possible improvements in surgical safety without more detailed information about processes of care that really changed in the participating institutions.

For elderly patients, no study has shown such a marked reduction in the common AEs that befall frail elderly patients. Furthermore, the authors acknowledge that the goals were not met for the part of the national safety programme involving elderly patients for falls, poor nutrition, physical limitations and delirium. This raises the question whether these reductions should be attributed to the national programme or have another explanation, such as chance. In that context it is also noteworthy that diagnostic errors showed a substantial reduction even though none of the components of the national programme in the Netherlands targeted this type of AE.

Thus, the signal of a 30% reduction in preventable AEs (p=0.1) may well be a chance finding. This would be the traditional interpretation of this p value. Furthermore, even with a more Bayesian view of the evidence, taking into account the prior probability that change has occurred, we do not have good reason to reject the hypothesis that no significant improvement occurred. Looking at the supplementary material provided by Baines *et al* (appendix 2 of that article),[4] their multilevel models explained only 10% of the variance in AE and 12% of the variance in preventable AEs, suggesting that many other factors influence these outcomes. It thus leaves a lot of room for the possibility that random variation explains the non-significant reduction in preventable AEs from 2008 to 2012.

ABANDONING AES AS THE GOLD STANDARD MEASURE OF IMPROVED PATIENT SAFETY

The fact that preventable AEs may not have decreased does not represent a failure of this latest study.[4] The authors have conducted an impressive study—what amounts to three national AE studies (2004, 2008 and 2012)—an unprecedented accomplishment in the field of patient safety. They also reach appropriately tentative conclusions about the impact of the national programme in the Netherlands. Furthermore, interestingly, they call for the same movement away from AEs as a measure of

improvement in patient safety as do Vincent and Almaberti.[5] Baines *et al* do so on the grounds that the sample size required to turn p=0.1 into a more significant p value is prohibitively high. This is probably true.

It is also true, as Vincent and Amalberti write,[5] that AEs provide a very general sense of the 'burden of disease'—the degree to which safety problems cause measurable impacts on morbidity and mortality, and, as with any disease, one eventually wants more specific measures, especially when it comes to evaluating treatments. AE studies still make sense when a new clinical area is being investigated. For instance, most major AE studies have not included paediatrics. So, to characterise the approximate burden of the problem and the main categories of patient safety problems in paediatrics, it made sense to conduct a paediatric AE study.[23] Similarly for home care, the overall burden of patient safety problems in this setting was not known, so it made sense to start with a broad measurement of AEs.[24] However, to show progress in any of these settings once we have a general sense of the burden and types of patient safety problems, studies will need to capture specific AEs that measure the impact of implemented interventions, rather than continuing to rely on broad heterogeneous measures such as AEs, as they will dilute real effects that may have occurred. For instance, if hospitals have invested implemented safety strategies for frail elderly patients, measurement must comprehensively capture fall-related injuries, delirium that develops after admission, aspiration events, or whatever other outcomes the strategies targeted. We cannot expect to detect improvements by partially capturing all possible harms that elderly patients experience in hospital.

Fifteen years into the field of patient safety, we would of course like to say that we finally have a study showing substantial reductions in preventable AEs on a large national scale. Such a finding would indeed constitute a milestone for maturation of the field. For now, though, we may have to settle for the milestone consisting of moving on to better metrics of improvement than the broad measure of harm that established the field in the first place.

Contributors KGS and PJM-vdM both contributed to conception of the paper; they both critically read and modified subsequent drafts and approved the final version. They are both editors at *BMJ Quality & Safety*.

Competing interests None declared.

Provenance and peer review Not commissioned; internally peer reviewed.

REFERENCES

1 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.
2 Landrigan CP, Parry GJ, Bones CB, *et al*. Temporal trends in rates of patient harm resulting from medical care. *N Engl J Med* 2010;363:2124–34.
3 Shojania KG, Thomas EJ. Trends in adverse events over time: why are we not improving? *BMJ Qual Saf* 2013;22:273–7.
4 Baines RJ, *et al*. How effective are patient safety initiatives? A retrospective patient record review study of changes to patient safety over time. *BMJ Qual Saf* 2015;24:561–71.
5 Vincent C, Amalberti R. Safety in healthcare is a moving target. *BMJ Qual Saf* 2015;24:539–40.
6 Zegers M, de Bruijne MC, 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.
7 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.
8 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.
9 Vincent C, Neale G, Woloshynowych M. Adverse events in British hospitals: preliminary retrospective record review. *BMJ* 2001;322:517–19.
10 Davis P, Lay-Yee R, Briant R, *et al*. Adverse events in New Zealand public hospitals I: occurrence and impact. *N Z Med J* 2002;115:U271.
11 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.
12 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.
13 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.
14 Goodman SN. Toward evidence-based medical statistics. 2: The Bayes factor. *Ann Intern Med* 1999;130:1005–13.
15 de Vries EN, Prins HA, Crolla RM, *et al*. Effect of a comprehensive surgical safety system on patient outcomes. *N Engl J Med* 2010;363:1928–37.
16 de Blok C, Koster E, Schilp J, *et al*. Implementatie VMS Veiligheidsprogramma. Evaluatieonderzoek in Nederlandse ziekenhuizen. Utrecht: NIVEL, 2013.
17 Urbach DR, Govindarajan A, Saskin R, *et al*. Introduction of surgical safety checklists in Ontario, Canada. *N Engl J Med* 2014;370:1029–38.
18 Reames BN, Krell RW, Campbell DA Jr, *et al*. A checklist-based intervention to improve surgical outcomes in Michigan: evaluation of the keystone surgery program. *JAMA Surg* 2015;150:208–15.
19 Chopra V, Shojania KG. Recipes for checklists and bundles: one part active ingredient, two parts measurement. *BMJ Qual Saf* 2013;22:93–6.
20 Dixon-Woods M, Leslie M, Tarrant C, *et al*. Explaining Matching Michigan: an ethnographic study of a patient safety program. *Implement Sci* 2013;8:70.
21 Davidoff F, Dixon-Woods M, Leviton L, *et al*. Demystifying theory and its use in improvement. *BMJ Qual Saf* 2015;24:228–38.
22 Catchpole K, Russ S. The problem with checklists. *BMJ Qual Saf* 2015;24:545–9.
23 Matlow AG, Baker GR, Flintoft V, *et al*. Adverse events among children in Canadian hospitals: the Canadian Paediatric Adverse Events Study. *CMAJ* 2012;184:E709–18.
24 Blais R, Sears NA, Doran D, *et al*. Assessing adverse events among home care clients in three Canadian provinces using chart review. *BMJ Qual Saf* 2013;22:989–97.