Identifying adverse events: reflections on an imperfect gold standard after 20 years of patient safety research

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

[1]Medicine, University of Toronto Faculty of Medicine, Toronto, Ontario, Canada
[2]Department of Biomedical Data Sciences, J10-S, Leids Universitair Medisch Centrum, Leiden, The Netherlands

Correspondence to
Dr Kaveh G Shojania, Sunnybrook Health Sciences Centre, Room H468, 2075 Bayview Avenue, Toronto, ON M4N 3M5, Canada; kaveh.shojania@sunnybrook.ca

Accepted 30 January 2020
Published Online First 17 February 2020

Linked

► http://dx.doi.org/10.1136/bmjqs-2018-008664
► http://dx.doi.org/10.1136/bmjqs-2019-009824

To cite: Shojania KG, Marang-van de Mheen PJ. *BMJ Qual Saf* 2020;**29**:265–270.

In ancient Roman religion, Janus was the god of gates and doorways, but also beginnings, endings, transitions, passages, time and duality. Usually depicted as having two faces, Janus looks at the past with one face and to the future with the other. Why mention Janus in an editorial about patient safety? Partly because the 20-year anniversary of *To Err is Human*[1] marks a transition—from the beginnings of patient safety as a fledgling field to a more mature research endeavour.

Beyond this symbolism of a transition period, Janus's past and future looking faces bear another connection to patient safety. The 'gold standard' research method in patient safety, record review to look for 'adverse events' (AEs), defined as harms from medical care, has taken two forms. The more common method, famously used in the Harvard Medical Practice Study (HMPS)[2] and other studies which have emulated it,[3–9] involves retrospective ('backwards looking') record review. An initial review looks for signs of possible harms from medical care, which, when present, trigger more detailed review to adjudicate the presence of AEs and judge the degree to which adhering to accepted standards of care could have prevented them.

More recently, some investigators have conducted prospective ('forward looking') surveillance to identify AEs in near-real time.[10–12] These forward-looking and backward-looking AE studies have succeeded in showing the scope of many safety problems. But, after 20 years of research, can we continue to use the same metric for both measuring safety and monitoring its improvement over time?

IDENTIFYING ADVERSE EVENTS THROUGH RETROSPECTIVE RECORD REVIEW

Though widely attributed to the HMPS,[2] retrospective record review to identify AEs originally came from a much less well-known study.[13] Carried out by the California Medical Association and California Hospital Association, this study sought to explore alternate models for compensating patients harmed by their medical care. This was also the main motivation for the HMPS. Showing the extent of preventable harm caused by medical care would potentially provide the basis for changing from the traditional malpractice system to a 'no fault' compensation system, as seen in Denmark, Sweden, Finland and New Zealand.[14]

In a preliminary publication, the HMPS authors wrote "In the California Medical Insurance Feasibility Study, certain screening criteria—such as death, transfer to a special care unit, an undesirable outcome and readmission to the hospital—were found to be associated with an increased likelihood of medical injury. In the absence of such criteria, AEs were generally not found. We eliminated several of the California criteria that we found redundant and added two…"[15] Such adding and eliminating has been the story of this methodology over the subsequent decades. Various national AE studies[3–9] have added new screening criteria or 'triggers' (eg, to detect safety problems of relevance to particular clinical settings or patient populations[16–19]) and abandoned others (eg, excessive length of stay) that, in practice, detected adverse outcomes from patients'

underlying illnesses rather than harms due to their medical care (ie, non-AEs).

Efforts to refine the AE methodology have also aimed at improving the review process to increase agreement between reviewers about key judgements. Reviewers have generally exhibited moderate to good agreement when it comes to distinguishing AEs from harms not caused by medical care, but only fair agreement when it comes to judging preventability or errors.[20] Some studies have reported better agreement about preventable AEs, achieving kappa values in the 0.4–0.6 range.[6 8] Even this improved agreement falls short of what one would expect for a field's 'gold standard' measure.

Beyond reviewer disagreement about the presence of AEs and their preventability, these retrospective studies suffer from the complete reliance on documentation practices. As with incident reporting, more events can simply mean more reporting, not worse safety.[21] When investigators in the Netherlands conducted a second national AE study, they found an AE rate in 2008 of 6.2%, higher than the 4.1% found in 2004.[22] Since preventable AEs did not increase, the accompanying editorial[23] suggested that more frequent non-preventable AEs reflected better documentation as a result of the growing interest in patient safety.

More commonly, though, researchers have worried about the converse situation: harms that go undocumented. For instance, members of the surgical team caring for a patient after a bowel resection that has gone well may not document a brief adverse drug event. And, even if someone mentions the event, lack of relevant details will hamper judgments about preventability. Moreover, some events do not cause harm immediately and thus do not appear to warrant documentation when they occur.

Detection of AEs using triggers for retrospective record review launched the field of patient safety, but the method has clear shortcomings. Many AEs go undocumented in medical records and reviewers often disagree about those that are documented. Enter the prospective version of the traditional AE detection method.

PROSPECTIVE APPLICATION OF THE TRIGGER TOOL METHOD FOR DETECTING ADVERSE EVENTS

Prospective AE surveillance still relies heavily (but not exclusively) on the use of triggers—signs of possible quality of care problems such as unexpected death, unplanned admission to intensive care, documented of patient dissatisfaction with care, as well as signs of specific events of interest—for instance, a laboratory test positive for *Clostridium difficile* infection. Importantly, though, trigger detection occurs in near-real time (usually within 48 hours) as opposed to months or years later. And, a trained observer integrated in the clinical environment supplements record review with debriefs of front-line staff, obtaining relevant details not noted in medical records. Observers can also learn of possible AEs from observing wards rounds, which can identify many events not captured in medical records,[24 25] reviewing incident reporting systems and direct communication from front line staff.[10–12]

This intensified prospective surveillance strategy aims to detect more candidate AEs and obtain key details relevant to judgments about harm and preventability. Moreover, involving staff from the clinical unit in identifying possible harms (and also in weekly conferences for reviewing the identified cases) may engage them in efforts to improve patient safety in a way that learning about AEs affecting patients who received care years ago might not.

COULD PROSPECTIVE SURVEILLANCE ENHANCE THE VALUE OF AES AS A PERFORMANCE MEASURE?

Regardless of whether any method for identifying AEs can also inform improvement efforts, many might argue that we need at least one 'gold standard' measure for tracking progress and/or comparing different healthcare organisations. Just as we compare hospitals using risk-adjusted mortality and readmission rates, maybe we could compare hospitals using a robust measure of patient safety. Existing methods for comparing performance on safety measures tend to use administrative data and have limited validity along with poor positive predictive values when compared with clinical data.[26–28] Maybe prospective surveillance, with its likely enhanced detection of preventable AEs, can provide such a comparative performance measure for patient safety.

In this issue of *BMJ Quality & Safety,* Forster *et al* report on their use of prospective AE surveillance at five hospitals in two Canadian provinces.[29] They sought to determine the degree to which observed variations in rates of (preventable) AEs likely reflect true differences in safety versus variations in the measurement method, including observer and reviewer behaviours. To help characterise the contribution of measurement issues to apparent differences in rates of AEs, Forster and colleagues added the elegant methodological feature of rotating observers between hospitals during the study. And, they restricted the study to general medicine wards to avoid another potential source of variation, as units within hospitals can show greater variation than seen across hospitals.[30]

The five hospitals consisted of four academic centres offering tertiary and quaternary services and one large urban community hospital. The percentage of hospital admissions with at least one AE ranged from a low of 9.9% (at the community hospital) to a high of 35.8% at one of the academic hospitals, with an overall AE risk per hospitalisation of 22% across the five hospitals. Admissions with at least one preventable AE ranged from 9.9% (again, at the community hospital) to 29.7% (at the same academic hospital with the highest AE risk). These risks for AEs and preventable AEs generally exceed those seen in the previous

study using retrospective AE record review conducted in Canada.[6] That study reported an overall risk for AEs of 7.5% (10.9% in teaching hospitals) and 2.8% for preventable AEs (3.3% in teaching hospitals). The higher rates of AEs in teaching hospitals likely reflect differences in documentation (more clinicians tend to enter notes on a given patient) and/or differences in case-mix, including transfers of particularly complex patients from non-teaching hospitals.

Regardless of hospital type, the focus of this latest prospective AE study lay in determining the degree to which this method allows identification of true differences in safety between hospitals as opposed to variations intrinsic to the AE detection method. Forster and colleagues reported large variation between the trained observers detecting triggers within the same hospital and also that the magnitude of this observer effect was highly correlated with the hospital. For instance, there was a twofold variation between observers in the hospital with the lowest risk of AEs and a smaller variation in the hospital with the highest risk. The subsequent physician review process somewhat dampened this variation in observer behaviour. But, as in retrospective record review studies, physician reviewers exhibited only modest agreement for judging preventability, with a kappa score of 0.55 (95% CI 0.41 to 0.69).

Even with the ability to detect AEs not captured in the medical record and a greater likelihood of obtaining information relevant to judging preventability, the prospective surveillance method does not appear to solve the issue of variation in the measurement method for detecting AEs. The rates at which observers identify triggers for more detailed record review and persistent limitations in reviewer agreement about key judgments prevent distinguishing true differences in safety between hospitals from measurement variation intrinsic to the AE detection method.

HETEROGENEITY AS THE FUNDAMENTAL CHALLENGE TO USING AES AS A METRIC

Past discussions of problems with AE studies have focused on issues such as reviewer behaviour, properties of the triggers and, now with prospective surveillance, the behaviours of the trained reviewers. More fundamental, though, is the problem that the AE rate is a composite indicator comprising multiple heterogeneous components. Composite performance measures, such as the Overall Hospital Quality Star Ratings in the USA[31 32] or the NHS England Overall Patient Experience Score,[33] combine multiple indicators of care quality into a single score. Such composites offer two advantages—the simplicity of a single overall measure and increased statistical power from having more eligible events. But, these composites can create problems,[34 35] just as with composite outcomes in clinical trials.[36 37] Composite outcomes pose particular problems when the components vary substantially in terms of their frequency and/or severity and when an intervention exerts differential effects on the various components.

AEs have these problems to a far greater extent than most composite outcomes. Instead of a small number of component submeasures, the AE rate encompasses all possible injuries from medical care: adverse drug events, complications of surgery and other invasive procedures, hospital acquired infections, non-infectious hazards of hospitalisation (eg, fall-related injuries, pressure ulcers, venous thromboembolism, delirium, malnutrition), diagnostic delays and so on. Each of these major categories is itself heterogeneous (figure 1). For instance, the category of preventable adverse drug events includes harms caused at the time of ordering medications, harms arising during drug dispensing and others from medication administration. Computerised provider order entry may prevent some adverse drug events at the ordering stage, but it will do little to reduce harms arising at the stages of dispensing or medication administration. Similarly, hospital acquired infections include central line associated bloodstream infections,[38 39] catheter-associated urinary tract infections,[40–42] *C. difficile*[42] and so on, each with different interventions to reduce these events. So, the AE rate at a given hospital at a given time represents a composite comprising a very long list of distinct event types, ranging from common to very infrequent harms, and with very different potentials for improvement from a given safety intervention or even multiple interventions.

Incomplete capture for many specific types of AEs further compounds the measurement problem. The trigger tool methodology—whether retrospective or prospective—incompletely captures specific categories of AEs,[43 44] as many will not produce death, transfer to an intensive care unit, readmission within 30 days or other triggers for record review. Thus, we have a composite outcome (AEs) comprising dozens of component categories (distinct types of AEs), most of which will include few cases in a given sample and many of which suffer from underdetection. The noise of chance variations in the small numbers constituting the numerous components of the AE composite will overwhelm any true signal of, for instance, specific adverse drug events reduced by computerised decision support. Heterogeneity both across and within categories of AEs combined with the small numbers for each means that measurements of AE rates have poor signal-to-noise ratio, thus preventing robust comparisons across hospitals. This same signal-to-noise problem will bedevil efforts to monitor progress over time. Fluctuations in (preventable) AEs at a single institution are at least as likely to reflect measurement variation (ie, noise) as they are true changes in safety.

LOOKING TO THE FUTURE

Returning to Janus and looking back on 20 years of patient safety research, the AE as a metric in various studies, both retrospective and prospective, has served

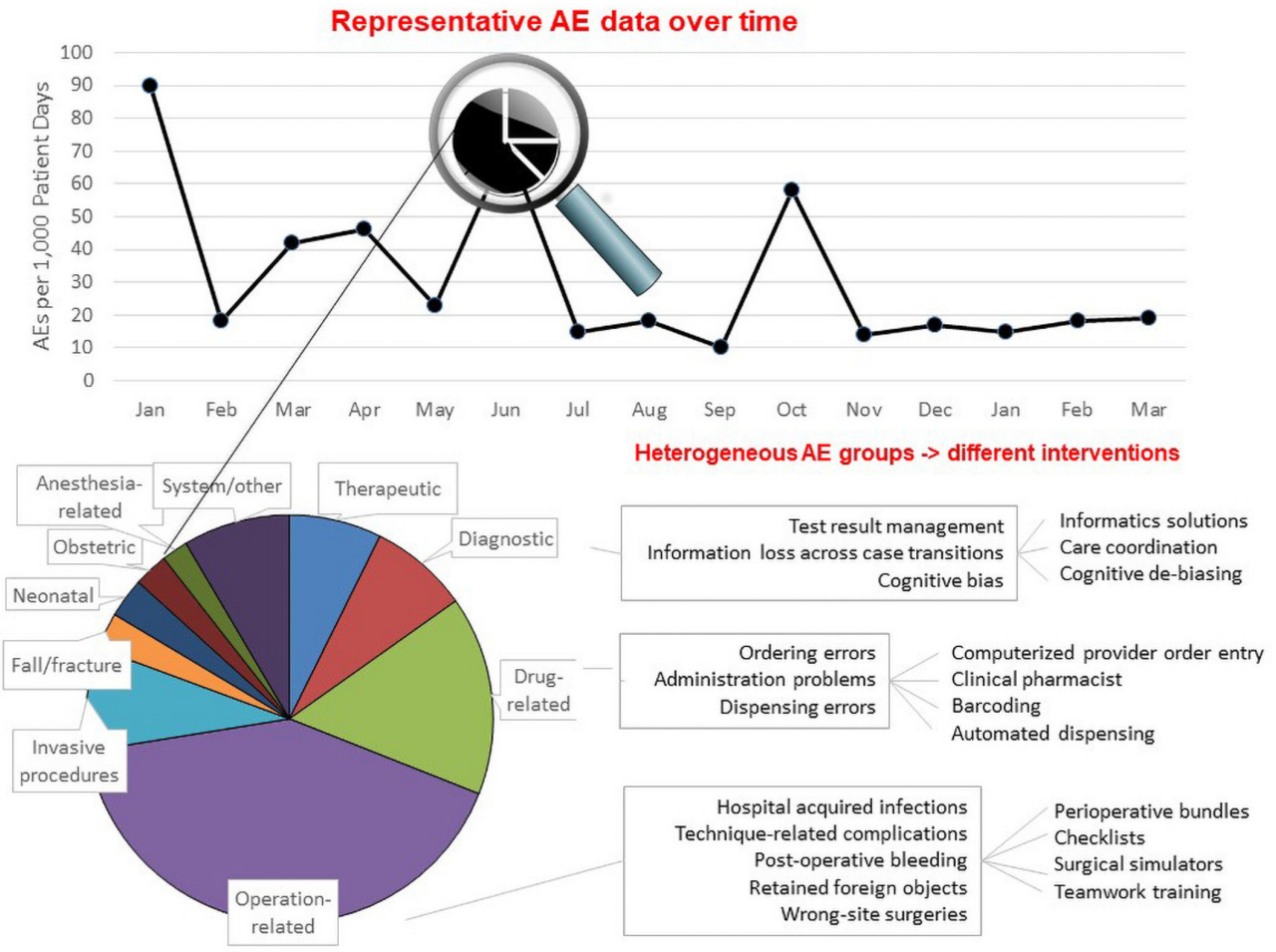

Figure 1 Depiction of the intrinsic heterogeneity associated with AE rates. The categories of AEs and their distribution come from a systematic review of retrospective record review studies.[20] The point of the figure lies in illustrating the deceptive degree of heterogeneity associated with the label 'adverse event', not the specific categories or their relative sizes. Definitions reflect those used in most individual studies, although some studies varied in the names and definitions of certain categories. 'Therapeutic' refers to AEs involving inappropriate or delayed treatment despite a correct diagnosis. System/other: includes AEs that cannot be attributed to an individual or specific source (eg, lack of/defective equipment or supplies, inadequate reporting or communication, inadequate staffing/training/supervision, no protocol/failure to implement protocol). AE, adverse event.

to demonstrate the scope of the problem and to engage clinicians, managers, researchers and policy makers. But, looking forward, such a broad, omnibus metric will not detect important differences in safety between institutions or track progress over time. For these tasks, we need to measure specific events of interest. These include established measures for capturing common healthcare-acquired infections, prospective registries for capturing outcomes of surgery,[45–47] validated text mining algorithms applied to electronic health records to capture specific care-related injuries,[48] methods to track missed diagnoses leading to harm,[49 50] and so on.

Generating broad interest in patient safety required an easily understood measure to demonstrate the scope of preventable harms caused by the healthcare system. Few would question that scope now. To make progress in this now well-established field, we need measures tailored to specific patient harms. No other field attempts to measure progress in the form of an omnibus measure. To assess progress in cardiovascular health, for instance, one looks at trends in the incidence and prognosis over time for common cardiovascular diseases, such as myocardial infarction and stroke, not an omnibus measure of all possible harms from 'heart and blood vessel disease'. Similarly, we will show progress in patient safety by tracking common, well-defined patient safety problems, not some general measure of all possible harms from medical care, the nature of which will inevitably change over time.[51]

The figure of Janus looking to the past and to the future captures the possibility of measuring AEs using both retrospective and prospective methods. But Janus also represents transitions. After 20 years of active research in patient safety, the time has come to put away the imperfect gold standard of AE rates and transition to more specific measures of important safety problems.

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.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent for publication Not required.

Provenance and peer review Commissioned; internally peer reviewed.

ORCID iD
Perla J Marang-van de Mheen http://orcid.org/0000-0003-1439-0989

REFERENCES

1 Kohn LT, Corrigan J, Donaldson MS. *To err is human: building a safer health system*. Washington, DC: National Academy Press, 2000.
2 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.
3 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.
4 Vincent C, Neale G, Woloshynowych M. Adverse events in British hospitals: preliminary retrospective record review. *BMJ* 2001;322:517–9.
5 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.
6 Baker GR*et al*. The Canadian adverse events study: the incidence of adverse events among hospital patients in Canada. *Can Med Assoc J* 2004;170:1678–86.
7 Wilson RM, Runciman WB, Gibberd RW, *et al*. The quality in Australian health care study. *Med J Aust* 1995;163:458–71.
8 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.
9 Rafter N, Hickey A, Conroy RM, *et al*. The Irish national adverse events study (INAES): the frequency and nature of adverse events in Irish hospitals—a retrospective record review study. *BMJ Qual Saf* 2017;26:111–9.
10 Forster AJ, Worthington JR, Hawken S, *et al*. Using prospective clinical surveillance to identify adverse events in hospital. *BMJ Qual Saf* 2011;20:756–63.
11 Forster AJ, Fung I, Caughey S, *et al*. Adverse events detected by clinical surveillance on an obstetric service. *Obstet Gynecol* 2006;108:1073–83.
12 Wong BM, Dyal S, Etchells EE, *et al*. Application of a trigger tool in near real time to inform quality improvement activities: a prospective study in a general medicine ward. *BMJ Qual Saf* 2015;24:272–81.
13 Mills DH. Medical insurance feasibility study. A technical summary. *West J Med* 1978;128:360–5.
14 Studdert DM, Brennan TA. No-fault compensation for medical injuries: the prospect for error prevention. *JAMA* 2001;286:217–23.
15 Hiatt HH, Barnes BA, Brennan TA, *et al*. A study of medical injury and medical malpractice. *N Engl J Med* 1989;321:480–4.
16 de Wet C, Black C, Luty S, *et al*. Implementation of the trigger review method in Scottish general practices: patient safety outcomes and potential for quality improvement. *BMJ Qual Saf* 2017;26:335-342.
17 Lindblad M, Schildmeijer K, Nilsson L, *et al*. Development of a trigger tool to identify adverse events and no-harm incidents that affect patients admitted to home healthcare. *BMJ Qual Saf* 2018;27:502–11.
18 Matlow AG, Cronin CMG, Flintoft V, *et al*. Description of the development and validation of the Canadian paediatric trigger tool. *BMJ Qual Saf* 2011;20:416–23.
19 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.
20 de Vries EN, Ramrattan MA, Smorenburg SM, *et al*. The incidence and nature of in-hospital adverse events: a systematic review. *Qual Saf Health Care* 2008;17:216–23.
21 Shojania KG. The frustrating case of incident-reporting systems. *Qual Saf Health Care* 2008;17:400–2.
22 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.
23 Shojania KG, Thomas EJ. Trends in adverse events over time: why are we not improving? *BMJ Qual Saf* 2013;22:273–7.
24 Andrews LB, Stocking C, Krizek T, *et al*. An alternative strategy for studying adverse events in medical care. *Lancet* 1997;349:309–13.
25 Lamba AR, Linn K, Fletcher KE. Identifying patient safety problems during team rounds: an ethnographic study: Table 1. *BMJ Qual Saf* 2014;23:667–9.
26 McIsaac DI, Hamilton GM, Abdulla K, *et al*. Validation of new ICD-10-based patient safety indicators for identification of in-hospital complications in surgical patients: a study of diagnostic accuracy. *BMJ Qual Saf* 2020;29:209–16.
27 Borzecki AM, Rosen AK. Is there a 'best measure' of patient safety? *BMJ Qual Saf* 2020;29:185–8.
28 Winters BD, Bharmal A, Wilson RF, *et al*. Validity of the agency for health care research and quality patient safety indicators and the centers for Medicare and Medicaid hospital-acquired conditions. *Med Care* 2016;54:1105–11.
29 Forster AJ, Huang A, Lee TC, *et al*. Study of a multisite prospective adverse event surveillance system. *BMJ Qual Saf* 2020;29:277–85.
30 Baines R, Langelaan M, de Bruijne M, *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.
31 Chatterjee P, Joynt Maddox K. Patterns of performance and improvement in US Medicare's Hospital StAR ratings, 2016-2017. *BMJ Qual Saf* 2019;28:486–94.
32 Figueroa J, Feyman Y, Blumenthal D, *et al*. Do the stars align? distribution of high-quality ratings of healthcare sectors across US markets. *BMJ Qual Saf* 2018;27:287–92.
33 NHS England. Methods, Reasoning and scope statement of methodology for the overall patient experience scores, 2018. Available: https://www.england.nhs.uk/statistics/wp-content/uploads/sites/2/2018/11/Methods-statement_2018Nov_Update.pdf [Accessed 27 Jan 2020].
34 Barclay M, Dixon-Woods M, Lyratzopoulos G. The problem with composite indicators. *BMJ Qual Saf* 2019;28:338–44.
35 Friebel R, Steventon A. Composite measures of healthcare quality: sensible in theory, problematic in practice. *BMJ Qual Saf* 2019;28:85–8.
36 Ferreira-González I, Permanyer-Miralda G, Domingo-Salvany A, Busse JW, Heels-Ansdell D, *et al*. Problems with use of

composite end points in cardiovascular trials: systematic review of randomised controlled trials. *BMJ* 2007;334:786.
37 Montori VM, Permanyer-Miralda G, Ferreira-González I, *et al*. Validity of composite end points in clinical trials. *BMJ* 2005;330:594–6.
38 Dandoy CE, Hausfeld J, Flesch L, *et al*. Rapid cycle development of a multifactorial intervention achieved sustained reductions in central line-associated bloodstream infections in haematology oncology units at a children's hospital: a time series analysis. *BMJ Qual Saf* 2016;25:633–43.
39 Marang-van de Mheen PJ, van Bodegom-Vos L. Meta-analysis of the central line bundle for preventing catheter-related infections: a case study in appraising the evidence in quality improvement. *BMJ Qual Saf* 2016;25:118–29.
40 Meddings J, Greene MT, Ratz D, *et al*. Multistate programme to reduce catheter-associated infections in intensive care units with elevated infection rates. *BMJ Qual Saf* 2020:bmjqs-2019-009330.
41 Meddings J, Rogers MAM, Krein SL, *et al*. Reducing unnecessary urinary catheter use and other strategies to prevent catheter-associated urinary tract infection: an integrative review. *BMJ Qual Saf* 2014;23:277–89.
42 Janzen J, Buurman BM, Spanjaard L, *et al*. Reduction of unnecessary use of indwelling urinary catheters. *BMJ Qual Saf* 2013;22:984–8.
43 Levtzion-Korach O, Frankel A, Alcalai H, *et al*. Integrating incident data from five reporting systems to assess patient safety: making sense of the elephant. *The Joint J Qual Patient Saf* 2010;36:402–18.
44 Shojania KG. The elephant of patient safety: what you see depends on how you look. *Jt Comm J Qual Patient Saf* 2010;36:399–AP3.
45 Etzioni DA, Wasif N, Dueck AC, *et al*. Association of hospital participation in a surgical outcomes monitoring program with inpatient complications and mortality. *JAMA* 2015;313:505–11.
46 Maggard-Gibbons M. The use of report cards and outcome measurements to improve the safety of surgical care: the American College of surgeons national surgical quality improvement program. *BMJ Qual Saf* 2014;23:589–99.
47 Sadeghi M, Leis JA, Laflamme C, *et al*. Standardisation of perioperative urinary catheter use to reduce postsurgical urinary tract infection: an interrupted time series study. *BMJ Qual Saf* 2019;28:32–8.
48 Reeson M, Forster A, van Walraven C. Incidence and trends of central line associated pneumothorax using radiograph report text search versus administrative database codes. *BMJ Qual Saf* 2018;27:982–8.
49 Liberman AL, Newman-Toker DE. Symptom-Disease pair analysis of diagnostic error (spade): a conceptual framework and methodological approach for unearthing misdiagnosis-related harms using big data. *BMJ Qual Saf* 2018;27:557–66.
50 Mane KK, Rubenstein KB, Nassery N, *et al*. Diagnostic performance dashboards: tracking diagnostic errors using big data. *BMJ Qual Saf* 2018;27:567–70.
51 Vincent C, Amalberti R. Safety in healthcare is a moving target. *BMJ Qual Saf* 2015;24:539–40.