How can Never Event data be used to reflect or improve hospital safety performance?

J. Olivarius-McAllister, M. Pandit, A. Sykes and J. J. Pandit

1 Foundation Year Trainee, 2 Chief Medical Officer, 3 Deputy Chief Medical Officer, 4 Consultant, Department of Respiratory Medicine, 5 Consultant, Nuffield Department of Anaesthesia, Oxford University Hospitals NHS Foundation Trust, Oxford, UK
6 Professor of Anaesthesia, University of Oxford, Oxford, UK

Summary
The absolute number of Never Events is used by UK regulators to help assess hospital safety performance, without account of hospital workload. We applied funnel plots, as an established means of taking workload into account, to published Never Event data for 151 acute Trusts in NHS England, matched to finished consultant episodes for 3 years, 2017–2020. Trusts with excess event rates should have the most Never Events if absolute number is a valid way to judge performance. The absolute number of Never Events was correlated with workload ($r^2 = 0.51$, $p < 0.001$), but the five Trusts above the upper 95% confidence limit did not have the highest number of Never Events. However, a limitation to interpretation was that the data were skewed; 12 out of 151 Trusts lay below the lower 95% limit. This skew probably arises because funnel plots pool all Never Events and workload data; whereas, ideally, different Never Events should use as denominator only the relevant workload actions that could cause them. We conclude that the manner in which Never Event data are currently used by regulators, in part to judge or rate hospitals, is mathematically invalid. The focus should shift from identifying ‘outlier’ hospitals to reducing the overall national mean Never Event rate through shared learning and an integrated system-wide approach.

Introduction
Never Events are serious incidents considered wholly preventable because there exist strong systemic protective barriers [1]. Their occurrence implies that the relevant barriers were not in place, or they were in place but not adhered to. Understandably, Never Events are used by regulators to help judge the ‘safety culture’ within a hospital. However, it is only their absolute number that is used: the more Never Events a hospital has in a year, the more concern there is about the underlying safety culture. The Care Quality Commission is unambiguous in its statement that “[even] a single never event can act as a red flag that an organisation’s systems may not be robust” [2]. There is some justification to this approach because a Never Event has the potential to cause serious patient harm or death (although fortunately only a small minority do) and might suggest a failure to apply or follow processes and policies. In this way, regulatory oversight might identify the consistently underperforming ‘outlier’ centres, where culture may be problematic, as a focus for intervention.

Yet, it is now well established that the number of Never Events occurring in an organisation is related primarily to its size (workload) and, at least for surgical Never Events, does not correlate with other measures of performance [3]. When
analysed statistically, the distribution of Never Events is approximately Poisson, implying they are random rare events [3, 4]. In other words, the bigger (busier) the hospital, the more Never Events it is likely to have. This is self-evident on first principles. A Never Event is a ‘binomial’ event; it either occurs or not; and can only occur if preceded by the necessary causative intervention. A nasogastric tube cannot be misplaced if never sited; a wrong side block cannot happen if there is no local anaesthetic injection.

Never Events have a detrimental effect on all those involved: Not just patients and their families, but also the staff and hospital Boards (who may be judged as failing in a duty to create safe systems). It is clear that the ‘number of Never Events’ is too crude a metric by which to judge a hospital’s safety performance, and the Care Quality Commission is undertaking a thematic review to consider different approaches [2].

In other spheres of analysing safety or performance, ‘funnel plots’ have become standard as a useful graphical way to identify outliers, especially in a way that takes account of workload [5–8]. The incidence of the outcome of interest (e.g. complications after hip surgery, expressed as % chance per operation) is plotted against the volume of activity (e.g. number of hip operations conducted by each team or centre). This yields a mean value for the incidence, which is plotted as a horizontal line, the individual results scattered each side of this line. Then, 95% (or 99%, or any other chosen limit) binomial confidence intervals are plotted using the mean value. This sets the statistical ‘expectation’ around the mean value and results in two curved lines, above and below the mean, very wide for low volumes of activity and closing towards the mean value as the volume of activity increases. Outlying values are easily identified ‘by eye’. Funnel plots are commonly used to assess clinical outcomes across teams or hospitals, but to our knowledge have never been published for Never Event data, and funnel plots or alternatives are not a focus of the Care Quality Commission thematic review.

We planned to assess what might be learned from applying funnel plots to Never Event data from English hospital Trusts, also remaining open to adapting or modifying these as necessary. If using absolute numbers was valid, we predicted that outlying Trusts in a funnel plot would also be those with the highest number of Never Events.

Methods

We analysed annual rates of Never Event (‘events’) and finished consultant episodes (‘episodes’) from 2017 to 2020 by acute hospital Trust, listed by NHS Digital (online Supporting Information Appendix S1) [9]. The number of Trusts varied from 145 to 153 due to their re-organisation during the 3-year period. We analysed events and episodes (to the end of February 2020) by Trust designation after re-organisation [10–13].

We analysed the correlation of the number of events with the number of episodes. We constructed funnel plots of events vs. episodes, for each year and the cumulative values to 3 years, with 95% and 99.8% CIs calculated from the mean value [5–8]. The primary mean value was calculated as the pooled estimate from a fixed-effect model, with a secondary mean calculated as the summed events divided by the summed episodes. The Clopper-Pearson method was used as the primary confidence interval calculation, with various secondary methods, including Agresti-Coull, asymptotic, Bayes, logit, probit and Wilson.

We assessed coverage of the confidence interval against a simulated sample, run for the number of reported episodes and events in the hospitals. In addition, we used Spiegelhalter’s method, which includes truncation of extreme values (we used bilateral 10% truncation) and random-effects (accommodating rate variance within Trusts) to calculate the mean value, and includes dispersed boundaries to identify outlying values [14]. We used R for analyses and plots, packages ‘binom’, ‘FunnelPlotR’, ‘ggplot2’, ‘meta’.

Results

We analysed events and episodes reported by 151 Trusts. There were 1140 events after 58,413,335 episodes, or 1 event per 51,240 episodes, which is a crude rate of 0.0000195, or a pooled rate of 0.0000222 (fixed-effect) or 0.0000210 (random-effects). The median (IQR [range]) numbers of events and episodes for three years were 6 (4–10 [0–29]) and 360,945 (237,370–507,244 [23,940–1,120,070]), respectively. More events happened in Trusts with more episodes, with the number of episodes ‘accounting’ for about half the linear variation in the number of Trust events, r² 0.51, p < 0.001 (Fig. 1). The cumulative event rate reported by 151 Trusts ranged from 0/376,415 episodes to a rate of 9/33,080 (0.000272) episodes, two extreme rates unlikely to be consistent with a pooled rate of 0.0000222 in a binomial distribution, two-sided p = 0.000047 and p = 0.000000013, respectively.

The number of Trusts reporting events in each year were 148, 148 and 144, with the annual median (IQR [range]) number of events and episodes 2 (1–4 [0–13]) and 122,920 (82,235–172,109 [6860–476,805]), respectively. The annual rates of events and episodes were 337/18,708,970 (0.0000180), 438/19,878,445 (0.0000220) and
365/19,825,920 (0.0000184). This variation is unlikely to be consistent with a common mean rate ($p = 0.0072$) but might be due to differences in contributing hospitals, as ‘year’ was not independently associated with event rate. Individual years had less power to determine whether the distribution of event rates was consistent with a binomial distribution, with 7/148 (5%), 8/148 (5%) and 6/144 (4%) outside the respective Clopper-Pearson 95%CI for their crude mean values, of which 9/21 (43%) were more than the upper limit and 12/21 (57%) were less than the lower limit (Fig. 2).

Cumulative events for 3 years were inconsistent with a standard binomial distribution. Of the 151 Trusts, 17 (11%) were outside the Clopper-Pearson 95%CI, with 5 high rates and 12 low rates, and 23 (17%) were outside Bayesian 95%CI. The excessive dispersion for events, noted for surgical Never Events by Moppett et al. [3], affected calculation of the mean rate and the identification of outliers. Bilateral truncation of extreme values reduced the mean rate to between 0.0000195 and 0.0000210. There were eight Trusts with cumulative event rates outside the 99.8% bounds of the dispersed distribution, two above and six below, which Spiegelhalter suggested would be suitable for further inspection to determine possible causes of extreme values (Fig. 3a)[14].

Figure 3b shows the five Trusts whose cumulative event rate and lower 95%CI exceed the national mean. The median (IQR [range]) number of events for these Trusts was 7 (5–9 [3–14]) vs. 6 (4–10 [0–29]) for the Trusts with rates within the 95%CI, $p = 0.91$. The median (IQR [range]) number of episodes for these five Trusts was 506,920 (351,640–727,980 [285,495–850,880]) vs. 360,863 (239,826–500,764 [23,940–1,120,070]) for the 134 Trusts within the 95%CI, $p = 0.13$. Figure 3c shows 12 Trusts whose event rates and upper 95%CIs are exceeded by the national mean. The median (IQR [range]) number of events for these 12 Trusts was 3 (1–5 [0–10]) vs. 6 (4–10 [0–29]), which was different to the 134 Trusts with rates within the 95%CIs, $p < 0.001$. The median (IQR [range]) number of episodes for these 12 Trusts were 492,495 (336,081–597,421 [192,975–878,145]), which was similar to the 134 Trusts with rates within the 95%CIs, $p = 0.10$. Figure 3d shows 12 example Trusts whose mean rates were above (n = 4) or below (n = 8) the national mean and whose 95% CIs encompass the national mean. Figure 4 plots the upper and lower 95%CI limits for a true event rate of 0.000022 binomial distribution as events vs. episodes.

**Discussion**

We think that funnel plots of event rate vs. case-load are a suitable way to display Never Events. The distribution of Never Events appears to be like other outcomes displayed by NHS funnel plots, being over-dispersed, inconsistent
Figure 2 Funnel plots of event rates vs. episodes reported by: (a) 148 Trusts, 2017–2018; (b) 148 Trusts, 2018–2019; (c) 144 Trusts, 2019–2020. The horizontal line is the mean value for each year, the black dashed lines are the Clopper-Pearson 95%CI, and the red dashed lines are the Bayesian 99.8%CI.
Figure 3 Funnel plots for event rates vs. finished consultant episodes reported for 3 years, with a mean rate of 0.0000222: (a) for all 151 Trusts, red circles are Trusts outside the 95%CI; (b) for five Trusts with rates and lower 95%CI greater than the mean; (c) for 12 Trusts with rates and upper 95%CI less than the mean; for 12 exemplar Trusts with 95%CI that include the mean, 4 with rates greater than the mean and 8 with rates less than the mean. The black dashed lines and red dashed lines are 95%CI and 99.8%CI, respectively.
with a binomial distribution and declaring outlying values only when they lie outside 99.8% CI limits, rather than outside 95% CI limits.

Safety performance of a hospital is, at least in large part, assessed by UK regulators on the absolute number of Never Events. Yet, there is a clear correlation of this absolute number with caseload, which should logically necessitate suitable adjustment of the data to take account of this. Conventional methods to do this include funnel or forest plots, and if using absolute number of events is valid, the prediction is that those hospitals that are outliers in these plots should also be those that have the most events. The fact that this is not the case invalidates any approach that relies on using Never Event data by absolute, unadjusted numbers.

We do not present the funnel plots as the ideal or recommended method to rank organisations in ‘league tables’ for safety performance. Rather, they are presented to illustrate the logical absurdity of using absolute Never Event numbers as a basis for making such judgements. Unless data are adjusted for case volume, then as reductio ad absurdum, a hospital performing just one procedure per year would be judged on the same basis as one performing several hundred thousand. One Never Event in the former would be, absurdly, as meaningful as one Never Event in the latter.

In fact, as explained further in online Supporting Information Figure S1, the NHS Never Event data do not conform to a binomial distribution around the calculated mean value, and therefore should not be subjected to unadjusted funnel plot representation. On one level, this is surprising. Never Events are by definition binomial; they either happen or not. However, on another level it is understandable: A Never Event is only binomial if related to the primary action that could plausibly cause it. The denominator needs to be appropriate. A problem with conventional funnel plot analysis is that it pools the Never Events as if a single category, and regards all the episodes as if equally likely to cause them. Yet, for example, it is only those attempts at local anaesthetic injection that could cause a wrong side block (insertion of nasogastric tube could not). In other words, the manner in which Never Event data are collected and presented [12] does not permit the adjustment of these by the relevant denominator data (i.e. the number of wrong side block should be adjusted by only the number of total blocks and not by any other statistic).
In mathematical terms, the distribution of each class of Never Event is almost certainly binomial but as a pooled group, all Never Events do not derive from the same binomial distribution.

The main message of the analysis is therefore not to present an alternative or ideal method of ranking Trusts, but rather the opposite: to stress that such league tables are inappropriate where Never Events are concerned (not least because each method of ranking will have its statistical shortcomings, as discussed above). Funnel and Forest plots may have a use in defending a Trust erroneously labelled ‘under-performing’ as based only on its absolute numbers of Never Events.

The notion of ‘bad apples’ has had some traction in the safety literature [15, 16]: the idea being that identifying poor performers at institutional [16, 17] or individual [18] level is a prerequisite to targeted intervention to raise overall standards. Eliminate or improve the poorly performing, leave the rest alone and, the theory goes, average performance will improve automatically. However, our data indicate that even eliminating Never Events from all the putative outlier hospitals will do little to change the overall national mean event rate because there are relatively few of them; the number of Never Events in these outlier centres is small, and they are all relatively small hospitals.

Therefore, a different, NHS-wide, rather than an individual Trust-focused, strategy will be necessary. Our data underline the reality that the national Never Event rate is not because of poor performance in just a few Trusts but arises because of broadly equivalent contributions by all Trusts.

Efforts directed to reducing the national mean Never Event rate might include a single, uniform way of undertaking root cause analyses with suitably anonymised conclusions shared across the NHS, that would help standardise practices and processes. This appears to be the philosophy underpinning the work of the recently formed Healthcare Safety Investigation Branch (https://www.hsib.org.uk). Its recent report into Never Events acknowledges that each is very different and pooling them to judge hospital performance may not work. This might also include a ‘learning teams’ approach; a ‘bottom-up’, supportive, blame-free method that focuses on improvement [19, 20]. This is especially important given increasing staff mobility [21], particularly trainees who rotate between hospitals and who might otherwise encounter varied or inconsistent approaches to safety. There should also be a strong link to the National Patient Safety Strategy, specifically the role of the Patient Safety Specialists. Finally, Trust Boards should actively engage in the safety and quality improvement agenda adopting the NHSI Just Culture Guide [22]. Our result is, therefore, important as it helps focus precious resources to the correct solution.

We noted that some Trusts can exhibit wide variation in performance across years (see online Supporting Information Figure S2), wherein some that lie above the confidence limit in one year can have zero Never Events in a later year. In part, this could be a result of reaction to a Never Event (e.g. a re-focus on preventing error after a recent event). Such ‘inoculative’ effect of a Never Event has been discussed previously [23], but it unlikely reflects a sustained shift in fundamental organisational safety culture. It is more likely simply reflective of the random distribution of Never Events, as has been previously modelled [3, 4].

It may be possible, using the data in online Supporting Information Appendix S1, to develop a method of ranking hospitals that does not have the shortcomings of funnel or Forest plots that we have discussed above. Verburg et al. [8] have described sophisticated techniques and analysing the NHS Never Event data by different distributions (e.g. hypergeometric, negative binomial, etc) [24] by methods like Monte Carlo simulation [25] may yield more accurate confidence limits that allow robust identification of outliers. Different methods of calculating binomial confidence intervals might also be assessed [26]. However, the magnitude of differences of confidence intervals between all these statistical distributions are generally modest [26]. Spiegelhalter has emphasised the benefits of simple graphical representation of performance, as opposed to detailed statistical analysis [5].

One of the limitations to any analysis is that it relies greatly on the primary reporting of both Never Events and also of caseload. The definition of Never Events was changed from February 2018 [12]. Trust mergers, closures or re-organisations clearly influence Trust caseload. We analysed the median change in episodes (2017–2019) as 4% (1%–6% [−4% to 247%]); the median and IQR range being relatively modest, but the larger range explained by known mergers. However, the same event and episode data are used widely across NHS strategy planning, so any deficiencies in this regard affect not just our study but numerous core NHS activities [27–29].

Another concern is the extent to which episodes represent the relevant ‘volume of activity’ [27–29]. An ideal metric would be to capture all interventions in patients attending for care, but this is impossible. An alternative is to use ‘number of operating theatres’, which has previously been attempted for surgical Never Events [3], but that would then miss many non-surgical Never Events. Another option is using whole time equivalent data as a surrogate for Trust
size; the assumption being that since interventions leading to Never Events are undertaken by staff, employee numbers might reflect the risk of a Never Event. However, whole time equivalent is influenced by recruitment and retention (staff turnover) and misleadingly includes non-clinical staff unlikely to be involved in a Never Event. In contrast, finished consultant episode is invariably used in the literature when activity volumes are of interest [27–29].

In conclusion, if using the absolute number of Never Events was a valid way of judging safety performance then it would have to take account of workload, given the correlation of absolute number with volume of activity. Yet, when conventional methods like funnel or forest plots are used to take account of workload, very few hospitals are outliers, and those that do not have the highest number of events. The proper focus should therefore shift from identifying outlier Trusts, to reducing the national mean Never Event rate through an integrated safety strategy. If hospital-level Never Event data are to be used at all to guide policy, case reporting needs greater refinement, especially with regard to identifying appropriate denominator data for each event.

Acknowledgements

JP is Chair of the Safe Anaesthesia Liaison Group (SALG) at the Royal College of Anaesthetists. The views expressed are personal, and not of those organisations. We thank Dr J. Carlisle, Consultant Anaesthetist, Torbay, for providing the analysis in online Supporting Information Figure S1. No external funding or other competing interests declared.

References

1. NHS England. Never Event Policy Framework. 2018. https://improvement.nhs.uk/documents/22655/Revised_Never_Events_policy_and_framework_FINAL.pdf (accessed 01/11/2020).
2. Care Quality Commission. Opening the Door to Change. 2018. https://www.cqc.org.uk/sites/default/files/20181224_openeingthedoor_report.pdf (accessed 01/11/2020).
3. Moppett IK, Moppett SH. Surgical caseload and the risk of surgical Never Events in England. Anaesthesia 2016; 71: 17–30.
4. Pandit JJ. Deaths by horsekick in the Prussian army - and other ‘Never Events’ in large organisations. Anaesthesia 2016; 71: 7–11.
5. Spiegelhalter DJ. Funnel plots for comparing institutional performance. Statistics in Medicine 2005; 24: 1185–202.
6. Mayer EK, Bottle A, Rao C, Darzi AW, Athanasiou T. Funnel plots and their emerging application in surgery. Annals of Surgery 2009; 249: 376–83.
7. Spiegelhalter D, Sherlaw-Johnson C, Bardsley M, Blunt I, Wood C, Grigg O. Statistical methods for healthcare regulation: rating, screening and surveillance. Journal of the Royal Statistical Society (A) 2012; 175: 1–47.
8. Verburg IW, Holman R, Peek N, Abu-Hanna A, de Keizer NF. Guidelines on constructing funnel plots for quality indicators: a case study on mortality in intensive care unit patients. Statistical Methods in Medical Research 2018; 27: 3350–66.
9. NHS Digital. NHS Trusts and CCGs in England by Health Education England (HEE) Region in September 2017. 2018. https://digital.nhs.uk/data-and-information/find-data-and-publications/supplementary-information/2018-supplementary-information-files/nhs-trusts-and-ccgs-in-england-by-health-education-england-nee-region-in-september-2017 (accessed 12/03/2021).
10. NHS Confederation Resources: NHS statistics, facts and figures. 2017. https://www.nhsconfed.org/resources/key-statistics-on-the-nhs (accessed 12/03/2021).
11. NHS Digital. NHS workforce statistics. https://digital.nhs.uk/data-and-information/publications/statistical/nhs-workforce-statistics (accessed 12/03/2021).
12. NHS Never Events data. https://www.england.nhs.uk/patient-safety/never-events-data (accessed 12/03/2021).
13. NHS Digital: Hospital Admitted Patient Care Activity. https://digital.nhs.uk/data-and-information/publications/statistical/hospital-admitted-patient-care-activity (accessed 12/03/2021).
14. Spiegelhalter DJ. Handling over-dispersion of performance indicators. Quality and Safety in Health Care 2005; 14: 347–51.
15. Shojania KG, Dixon-Woods M. ‘Bad apples’: time to redefine as a type of systems problem? British Medical Journal Quality and Safety 2013; 22: 528–31.
16. Dixon-Woods M, Yeung K, Bosk CL. Why is U.K. medicine no longer a self-regulating profession? The role of scandals involving “bad apple” doctors. Social Science Medicine 2011; 73: 1452–9.
17. Levitt P. Challenging the systems approach: why adverse event rates are not improving. BMJ Quality and Safety 2014; 23: 1051–2.
18. Dekker SW, Leveson NG. The bad apple theory won’t work: response to ‘Challenging the systems approach: why adverse event rates are not improving’ by Dr Levitt. BMJ Quality and Safety 2014; 23: 1050–1.
19. Wachter RM, Pronovost PJ. Balancing “no blame” with accountability in patient safety. New England Journal of Medicine 2009; 361: 1401–6.
20. Robbins T, Tipper S, King J, Ramachandran SK, Pandit JJ, Pandit M. Evaluation of learning teams versus root cause analysis for incident investigation in a large United Kingdom National Health Service hospital. Journal of Patient Safety 2020. https://doi.org/10.1097/PTS.0000000000000641. Epub 24 March.
21. Buchan J, Charlesworth A, Gershlick B, Seccombe I. A critical moment: NHS staffing trends, retention and attrition. London: The Health Foundation, 2019. https://www.health.org.uk/sites/default/files/upload/publications/2019/A%20Critical%20Moment_1.pdf (accessed 01/11/2020).
22. The NHS Patient Safety Strategy. https://www.england.nhs.uk/patient-safety/the-nhs-patient-safety-strategy (accessed 18/03/2021).
23. Hopping M, Merry AF, Pandit JJ. Exploring performance of, and attitudes to, Stop- and Mock-Before-You-Block in preventing wrong-side blocks. Anaesthesia 2018; 73: 421–7.
24. Poor HV. The maximum difference between the binomial and Poisson distributions. Statistics and Probability Letters 1991; 11: 103–6.
25. Xia L, He K, Li Y, Kalflesch J. Accounting for total variation and robustness in profiling health care providers. BioStatistics 2020; https://doi.org/10.1093/biostatistics/kkab024. Epub 3 June.
26. Newcombe RG. Two-sided confidence intervals for the single proportion: comparison of seven methods. Statistics in Medicine 1998; 17: 857–72.
27. Herbert A, Wijaars L, Zylbersztejn A, Cromwell D, Hardeled P. Data resource profile: hospital episode statistics admitted patient care (HES APC). International Journal of Epidemiology 2017; 46: 1093.
28. Aylin P, Williams S, Bottle A, Jarman B. Dr Foster’s case notes.
   Counting hospital activity: spells or episodes? British Medical Journal 2004; 329: 1207.
29. Leal J, Manetti S, Buchanan J. The impact of hospital costing methods on cost-effectiveness analysis: a case study.
   Pharmacoeconomics 2018; 36: 1263–72.

Supporting Information
Additional supporting information may be found online via the journal website.

Appendix S1. Original dataset.
Figure S1. Probability distributions for data.
Figure S2. Dynamic performance of example Trusts.