BMJ Open

Does increasing the size of bi-weekly samples of records influence results when using the Global Trigger Tool? An observational study of retrospective record reviews of two different sample sizes

Kjersti Mevik,[1] Frances A Griffin,[2] Tonje E Hansen,[3] Ellen T Deilkås,[4] Barthold Vonen[5]

To cite: Mevik K, Griffin FA, Hansen TE, *et al*. Does increasing the size of bi-weekly samples of records influence results when using the Global Trigger Tool? An observational study of retrospective record reviews of two different sample sizes. *BMJ Open* 2016;**6**:e010700. doi:10.1136/bmjopen-2015-010700

▸ Prepublication history and additional material is available. To view please visit the journal (http://dx.doi.org/10.1136/bmjopen-2015-010700).

Received 30 November 2015
Revised 4 March 2016
Accepted 5 April 2016

For numbered affiliations see end of article.

Correspondence to
Dr Kjersti Mevik;
kjersti.mevik@nlsh.no

ABSTRACT

Objectives: To investigate the impact of increasing sample of records reviewed bi-weekly with the Global Trigger Tool method to identify adverse events in hospitalised patients.

Design: Retrospective observational study.

Setting: A Norwegian 524-bed general hospital trust.

Participants: 1920 medical records selected from 1 January to 31 December 2010.

Primary outcomes: Rate, type and severity of adverse events identified in two different samples sizes of records selected as 10 and 70 records, bi-weekly.

Results: In the large sample, 1.45 (95% CI 1.07 to 1.97) times more adverse events per 1000 patient days (39.3 adverse events/1000 patient days) were identified than in the small sample (27.2 adverse events/1000 patient days). Hospital-acquired infections were the most common category of adverse events in both the samples, and the distributions of the other categories of adverse events did not differ significantly between the samples. The distribution of severity level of adverse events did not differ between the samples.

Conclusions: The findings suggest that while the distribution of categories and severity are not dependent on the sample size, the rate of adverse events is. Further studies are needed to conclude if the optimal sample size may need to be adjusted based on the hospital size in order to detect a more accurate rate of adverse events.

Strengths and limitations of this study

- The samples were similar in terms of age, sex and length of stay.
- Preventability of the adverse events was not assessed.
- Only two sample sizes were compared.
- Method for authentication of events differed slightly for each set of samples, however, high inter-rater reliability between the review teams indicates consistency and thus did not likely affect the results.

INTRODUCTION

For more than a decade, considerable efforts have been invested across healthcare to reduce adverse events, resulting in many efforts to identify reliable and valid tools to measure such events. The Institute for Healthcare Improvement (IHI) Global Trigger Tool is a widely used and considered an effective tool for measuring adverse events.[1–3] The method includes reviewing bi-weekly samples of 10 patient records selected randomly from the hospital discharge lists. Two non-physician reviewers search independently for predefined triggers that could indicate possible adverse events. A physician authenticates their consensus on the presence of adverse events and severity. The adverse events identified in the bi-weekly periods provide data for Statistical Process Control (SPC) charts used to analyse adverse events rates over time. However, concerns have been raised[2 4–8] about the method's ability to accurately detect rates of adverse events and changes in rates, due to the small sample size of 10 records bi-weekly recommended in the IHI method.

In Norway, all hospital trusts are required by the National Health Authority to use a translated version of the Global Trigger Tool to review a minimum of 10 records selected continuously and bi-weekly in order to monitor the rates of adverse events in each hospital trust and at a national level.[9] Good *et al*[10] suggest that sample size should be adjusted to hospital size and based on this, we increased the sample size at our trust to seven times greater than that required by the Health Authority, as we believed this would detect a more accurate rate of adverse events. Our rates of adverse events have been higher than other comparable trusts that are reviewing bi-weekly samples of 10 records, thus we sought to assess whether our higher rates were due to the larger sample size. The impact of sample size on adverse event rates has not been validated to our knowledge, thus demonstrating the need for this study.

Our aim was to obtain the rate, category and severity of the identified adverse events in two different sample sizes of records selected from the same population bi-weekly: one sample corresponding to the IHI recommendation and one sample seven times larger. We hypothesised that increasing the sample size would not yield a different rate of adverse events per 1000 patient days.

METHODS

Study design

The study is an observational cross-sectional study including retrospective record review of two samples of records, 1680 and 240, respectively (figure 1).

Setting

The study was performed in a 524-bed hospital trust at three geographical locations in Nordland County, North-Norway. Both the samples were selected from the same population discharged from 1 January to 31 December 2010. However, the large sample was first stratified according to discharges from the nine services in the trust and then 10 records were selected from 5 services and 5 records from 4 services, respectively, for a total of 70 records bi-weekly. The small sample included 10 records selected bi-weekly from the aggregated discharge lists of all the 9 services. Following the IHI guidelines, records were excluded in both samples for patients aged 17 years or younger, patients admitted primarily for psychiatric or rehabilitation care, or patients with a length of stay less than 24 h. The whole hospitalisation was reviewed including patient days at all services not only at the index service.

The study was approved by the data protection official in Nordland Hospital trust and by the Norwegian Regional Ethics Committee (ref 2012/1691).

Record review method

Training of the reviewers followed the IHI recommendations and included theory, practical review exercises, and debriefing sessions provided by experienced reviewers. The IHI definition of an adverse event was used, that is[1]: 'Unintended physical injury resulting from or contributed to by medical care that requires additional monitoring, treatment or hospitalisation, or that results in death'. Both adverse events associated with treatment given prior, during or after (within 30 days) to the index

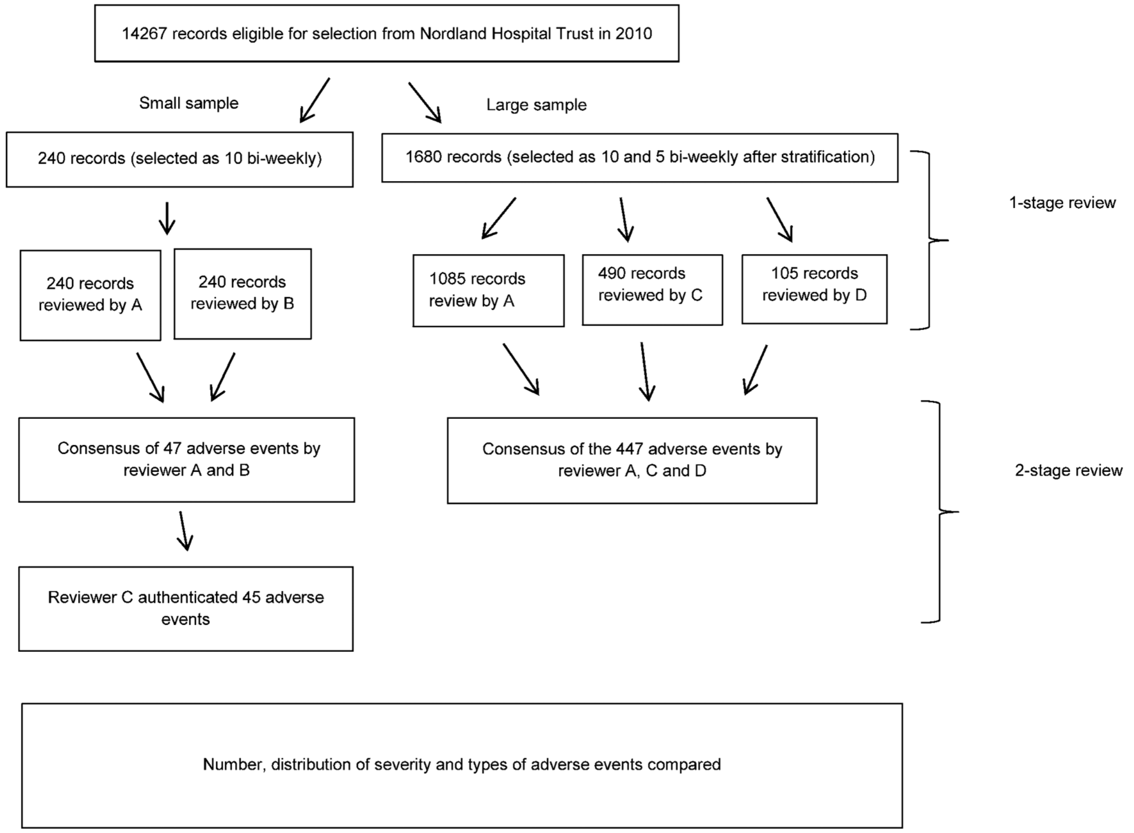
14267 records eligible for selection from Nordland Hospital Trust in 2010
Small sample
Large sample
240 records (selected as 10 bi-weekly)
1680 records (selected as 10 and 5 bi-weekly after stratification)
1-stage review
240 records reviewed by A
240 records reviewed by B
1085 records review by A
490 records reviewed by C
105 records reviewed by D
Consensus of 47 adverse events by reviewer A and B
Consensus of the 447 adverse events by reviewer A, C and D
2-stage review
Reviewer C authenticated 45 adverse events
Number, distribution of severity and types of adverse events compared

Figure 1 Overview of the study design. Non-physician reviewers; reviewer A and B, physician reviewers; reviewer C and D.

discharge (the discharge selected from the discharge lists of the services) were included to evaluate the total number of adverse events resulting from medical care. Preventability of the identified adverse events was not evaluated.

The identified adverse events were grouped into 23 categories derived from the Norwegian translation[11] of the IHI Global Trigger Tool. These categories were further aggregated into eight main categories (ie, hospital-acquired infections, surgical complications, bleeding/thrombosis, patient fall/fracture, medication harm, obstetric harm, pressure ulcer and other). The severity of adverse events was categorised into five levels (E–I) using definitions adapted from those of the National Coordinating Council for Medication Error Reporting and Prevention Index (NCC MERP):[12]

Category E: Temporary harm to the patient and required intervention
Category F: Temporary harm to the patient and required initial or prolonged hospitalisation
Category G: Permanent patient harm
Category H: Intervention required to sustain life
Category I: Patient death

The review process for both sets of samples followed the IHI method,[1] where reviewers checked each record for the presence of triggers from a standard list of triggers in the Norwegian translation of the Global Trigger Tool. When a trigger was identified, they checked for documentation indicating that an adverse event had occurred; for any adverse event detected, whether by a trigger or not, one of the above eight categories and a severity level was assigned. The process for authentication of adverse events differed slightly between the two sets of samples. For the small samples, two nurses (reviewer A and reviewer B) reviewed all records independently and then together reached consensus on presence, category and severity of adverse events. A physician (reviewer C) then authenticated their findings. The reviewing process of authentication with records from the large samples was slightly different in that each record was reviewed by one reviewer—either a nurse (reviewer A) or one of two physicians (reviewers C and D). The three reviewers discussed their findings and reached consensus of presence, category and severity of adverse events identified (figure 1). The modification with only one reviewer per record in the reviewing process for the large samples was due to limited resources available.

Statistical analysis

Demographic variables of the records were obtained. Categorical variables were compared between the samples with χ^2 test while continuous variables were compared using the independent t test.

SPC charts are used to evaluate variations between data points over time, which is a recommended approach for evaluating the rates of adverse events measured by the Global Trigger Tool.[1] [13] We used QI Macros in Excel 2013 to present the calculated rate of adverse events per 1000 patient days in U-charts and the calculated percentage of records with adverse events in a P-chart of both samples.[14] Test 1–3 of special cause variation (SCV) were applied in order to evaluate the rates. The tests are positive if data points are outside the control limits, eight or more data points are on the same side of the median or/and if six data points are either ascending or descending. We hypothesised that different rates of adverse events in the two samples would yield different results in terms of the tests and control limits.

To compare the calculated rates, proportions of severities and categories of adverse events between the samples, we used Poisson regression in generalised linear models to calculate the relative risk of adverse events between the samples as the risk ratio (RR). Poisson regression was chosen as it accounts for variations in the number of cases reviewed and variations in the length of stay. The number of adverse events was set as the dependent variable and log patient days as the offset variable (in the analysis of adverse events per patient day). When analysing adverse events per records and percentages of records with an adverse event, zero was set as the fixed value. A p value of <0.05 was defined as statistically significant. We also adjusted for services and variables associated with the index service. Associations between adverse events and demographic variables were explored using Pearson's correlation and logistic regression. To assess the inter-rater reliability between the review teams of the two samples, we used κ and weighted κ statistics. The following interpretations from Landis and Koch was used for the Cohen κ coefficient: poor (<0.0), slight (0.00–0.20), fair (0.21–0.40), moderate (0.41–0.60), substantial (0.61–0.80) and almost perfect (0.81–1.00).[15] We used SPSS (V.22.0; SPSS Chicago, Illinois, USA) for statistical analyses.

RESULTS

Demographics characteristics

A total of 1920 records were reviewed in the study using the Global Trigger Tool. Demographic characteristics in the samples and the overall population from which the samples were drawn from are shown in table 1. A total of 12% of the overall population (14 267 discharges) was reviewed in the large samples, while 2% was reviewed in the small sample. Length of stay, age and sex were derived for the whole hospitalisation and these did not differ between the large and the small sample. Patients in the large sample were different to the overall population in terms of sex and length of stay while patients in the small sample did not differ from the overall population. Type of admissions (acute or planned), case mix (discharge diagnose), services (functional units), case mix index, admission to surgery and numbers of transfers were derived from the index discharge (source of the random selection) and adjusted for.

Table 1 Demographic characteristics of the two samples and the overall population

	Samples			p Value		
	Large sample	Small sample	Overall population	Large vs small sample	Large vs overall population	Small vs overall population
n	1680	240	14 267			
Length of stay (days)*	6.8 (7.5)	6.9 (11.1)	6.3 (6.9)	0.852	0.014	0.400†
Average age (years)*	62 (21)	61 (21)	62 (21)	0.487	0.592	0.344†
Sex (percent women)‡	62	59	57	0.446	<0.001	0.410§

n.s=non-significant=p value>0.05.
*Values presented as mean with SDs.
†t test.
‡Values presented as percent.
§χ^2 test.

Comparison of adverse events

In the large sample of 1680 records comprising 11 367 patient days, we identified 447 adverse events in 347 discharges. This corresponds to a rate of 39.3 adverse events per 1000 patient days (95% CI 35.8 to 43.1, SE=1.86) or 26.6 adverse events per 100 discharges (95% CI 24.3 to 29.2, SE=1.26). The percentage of patients with an adverse event was 20.5% in the large sample. In the small sample of 240 records comprising 1657 patient days, we identified 45 adverse events in 30 discharges. This corresponds to a rate of 27.2 adverse events per 1000 patient days (95% CI 20.3 to 36.4, SE=4.05) or 18.8 adverse events per 100 discharges (95% CI 14.0 to 25.1, SE=2.80). The percentage of patients experiencing an adverse event was 12.5%. Some patients experienced more than one adverse event. Patients experiencing adverse events had longer hospital stays (large sample r^2=0.21, p<0.001 and small sample r^2=0.46, p<0.001) than patients without experiencing adverse events. In the large sample age correlated (r^2=0.03, p<0.001) with number of adverse events, while in the small sample age did not correlate with number of adverse events (r^2=−0.003, p=0.54).

The rate of adverse events per 1000 patient days was 45% higher in the large sample than in the small sample (RR=1.45, 95% CI 1.07 to 1.97; p=0.02). Likewise, the rate of adverse events per record was 42% higher in the large sample than in the small sample (RR=1.42, 95% CI 1.04 to 1.93, p=0.03). The percentage of records including an adverse event was 65% higher in the large sample than in the small sample (RR=1.65, 95% CI 1.14 to 2.34, p=0.008). In figure 2, the rates of adverse events per 1000 patient days in both samples are presented in control U-charts and percentages of records with adverse events in control P-charts over the 24 bi-weekly periods in 2010. In both charts, the control limits are much wider in the small sample than in the large sample. SCVs (positivity of tests 1) were identified only for the small sample. This is marked with a black dot in the U-chart. None of the other tests were positive for either of the samples.

To adjust for the stratification made before selection of records to the large sample, we adjusted for the variables that were associated from the index discharge. The primary results did not alter as the RR was 1.83 (95% CI 1.32 to 2.54, p<0.001) of identifying an adverse event per 1000 patient days in the large sample compared with the small sample when adjusting for these variables.

The inter-rater reliability of the two teams that reviewed the different sets of samples was obtained to

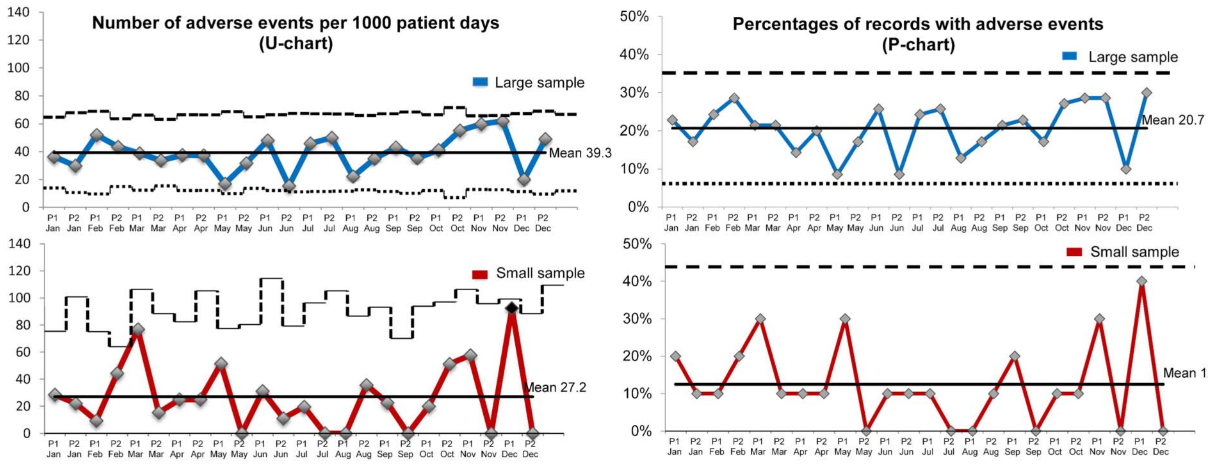

Figure 2 Comparison of statistical process control charts (U-chart) and (P-chart) between large and small samples. Dashed line = upper control limits; dotted line = lower control limits.

assess for possible impact from the different authentication processes. The two review teams reviewed a set of 50 patient records, and agreement regarding the presence of adverse events (κ =0.75), number of adverse events (κ=0.68) and severity level (κ=0.69) was substantial.

Hospital-acquired infections were the most frequent category of identified adverse events in both samples. There were no significant differences between the estimated proportions of identified adverse events between the samples for the six main categories of adverse events; hospital-acquired infections (RR=1.52, 95% CI 0.94 to 2.47, p=0.09), surgical complications (RR=1.28, 95% CI 0.67 to 2.47, p=0.46), bleeding/thrombosis (RR=1.44, 95% CI 0.70 to 2.98, p=0.33), medication harm (RR=1.68, 95% CI 0.60 to 4.66, p=0.32), patient fall (RR=0.83, 95% CI 0.24 to 2.82, p=0.76) and pressure ulcers (RR=0.73, 95% CI 0.16 to 3.33, p=0.68) (see online supplementary file 1). For the categories obstetric harm and other, no adverse events were identified in the small sample and a comparison was not performed.

The least severe adverse events (category E) accounted for more than half of the adverse events identified in both samples. Severity level including prolonged stay accounted for the same amount (30–40%) in both samples. No significant differences were found between the rate of adverse events per 1000 patient days between the samples, when adverse events were analysed separately according to severity of the adverse events: E (RR=1.50, 95% CI 1.00 to 2.26, p=0.05) and F (RR=1.68, 95% CI 0.99 to 2.85, p=0.05) and F, G, H and I (RR=0.47, 95% CI 0.17 to 1.27, P=0.14) and G, H and I (RR=1.38, 95% CI 0.87 to 2.18, p=0.17).

DISCUSSION

The rate of adverse events was 1.45 higher in the large sample than in the small sample. Our findings indicate that the sample size may influence the rate of identified adverse events. The differences in CI and SE indicate that increasing the sample size decreases the variation, as expected. We believe that the higher rate of adverse events detected was due to the use of a larger sample and may be more reflective of the total population given the size of the hospital. Since the distribution of severity level and types of adverse events were the same in both sample sizes, we suggest that these distributions are unaffected by sample size.

While evaluations of the Global Trigger Tool have reported high sensitivity[3] and acceptable reliability,[16 17] the impact of the sample size in determining the level of adverse events has hardly been discussed. We believe this is the first attempt to assess the impact of the sample size to the rate of adverse events identified with the Global Trigger Tool. Good *et al*[10] adjusted the sample size to the hospital sizes without further comparisons between different sample sizes selected in the same time period. We wanted to evaluate whether a larger sample of records reviewed bi-weekly could yield higher rates of adverse events than a sample of 10 records reviewed bi-weekly. Our trust had increased our bi-weekly samples to correspond to 12% of the total number of discharges and found higher rates of adverse events than comparable Norwegian trusts that reviewed samples of 10 records bi-weekly. Thus we determined it legitimate, necessary and original to assess whether using the Global Trigger Tool with different sample sizes would produce different results.

While our findings may challenge the sensitivity of the recommended small sample size in order to identify an accurate rate of adverse events, they also underline the ability of that sample size to reflect distribution of severities and categories of adverse events accurately. Our results in terms of this, corresponds well with other studies.[18 19] In the small sample, no adverse events of category I were identified. This is most likely due to the fact that the Global Trigger Tool is not designed to identify all such cases (category I). Owing to their infrequent occurrence, other methods should be used to monitor these specific types of events, for example, investigating all hospital deaths.[20 21] Thus, we compared the rate of adverse events in category I along with the rate of adverse events in other categories (category F, G and H).

Several factors could explain the differences in the rate of adverse events identified in the two samples. First, the authentication processes differed slightly for the two samples. To assess for possible bias, we evaluated the inter-rater reliability of the two teams that reviewed the different samples. We found substantial agreement between the two review teams regarding presence, number and severity level of adverse events, thus conclude that the difference in adverse event rates between the samples are most likely not due to bias from the minor difference in authentication processes. These findings are supported by the work of Zegers *et al*.[22] Second, the Simpson paradox, implying that statistical results from aggregated data could give a different result from a group-level analysis.[23] A skewness regarding the variables associated with the index discharges could be present in our study, as the large sample was stratified according to the services before sampling and the small sample was not. However, the primary results did not differ when adjusting for these variables. Neither did the demographic characteristics sex, age and length of stay differ between the large and the small sample. Third, the study was undertaken for only 1 year of discharges comprising 240 records in the small sample. A meta-analysis of different sample sizes showed that the variation of adverse event rates decreases as the sample size increases,[4] thus underlining the importance of having a large enough sample size in order to obtain valid results.

CONCLUSION

We believe the findings in this study could challenge the appropriateness of the sampling methods commonly used as the rate of adverse events increased when the number of records reviewed bi-weekly was increased, though limitations of the study must be considered. The distributions of adverse event categories and severity level did not differ between the samples and only the rate of adverse events appeared to be influenced by the sample size. Further studies are needed to determine whether there is an optimal sample size and if it should be based on hospital size, especially as reviewing larger sample sizes requires more resources. Until further studies, we suggest using a relative increase in sample size to 8–10% of total number of discharges.

Author affiliations

[1]Regional Patient Safety Resource Center, Nordland Hospital Trust, Bodø, Norway
[2]Fran Griffin & Associates, LLC, Neptune, New Jersey, USA
[3]Nordland Hospital Trust, Bodø, Norway
[4]Center for Health Service Research, Akershus University Hospital, Lørenskog, Norway
[5]CMO, Nordland Hospital Trust, Bodø, Norway
[6]Institute for community medicine, The Artic University of Norway, Tromsø, Norway

Acknowledgements The authors would like to thank Birger Hveding and Inger Lise Øvre who conducted the reviews together with the coauthors TEH and KM. Frank Federico and Carol Haraden for guidance on the development of the study protocol. Tom Wilsgaard for advice on statistical analysis, and Alexander Ringdal and Marina Mineeva for help with data processing.

Contributors KM and BV designed the study. TEH and KM reviewed the records. KM conducted the data analysis and wrote the first draft of the manuscript and rewrote the draft of the manuscript after all coauthors had reviewed and revised. KM performed the statistical analyses and produced the graphs.

Funding KM received a grant from the Northern Norway Regional Health Authority.

Competing interests None declared.

Ethics approval Data protection official in Nordland Hospital trust. Norwegian Regional Ethics Committee (ref 2012/1691).

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement No additional data are available.

REFERENCES

1. Griffin F, Resar R. IHI Global Trigger Tool for measuring adverse events. *IHI Innov Ser white Pap* 2007:1–44. http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:IHI+ Global+ Trigger+ Tool+ for + Measuring+ Adverse+ Events#0 (accessed 26 Nov 2014).
2. Commission S, Zealand N. *The Global Trigger Tool: a review of the evidence*. Wellington: 2013. http://www.hqsc.govt.nz
3. Classen DC, Resar R, Griffin F, *et al*. 'Global trigger tool' shows that adverse events in hospitals may be ten times greater than previously measured. *Health Aff (Millwood)* 2011;30:581–9.
4. Lessing C, Schmitz A, Albers B, *et al*. Impact of sample size on variation of adverse events and preventable adverse events: systematic review on epidemiology and contributing factors. *Qual Saf Health Care* 2010;19:e24.
5. 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.
6. 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.
7. James J. A new, evidence-based estimate of patient harms associated with hospital care. *J Patient Saf* 2013;9:122–8.
8. Shojania KG, Thomas EJ. Trends in adverse events over time: why are we not improving? *BMJ Qual Saf* 2013;22:273–7.
9. Deilkås E, Bukholm G, Lindstrøm J, *et al*. Monitoring adverse events in Norwegian hospitals from 2010 to 2013. *BMJ Open* 2015;5: e008576. http://dx.doi.org/ 10.1136/bmjopen-2015-008576
10. Good VS, Saldaña M, Gilder R, *et al*. Large-scale deployment of the Global Trigger Tool across a large hospital system: refinements for the characterisation of adverse events to support patient safety learning opportunities. *BMJ Qual Saf* 2011;20:25–30.
11. Strukturert journalundersøkelse, ved bruk av Global Trigger Tool for å identifisere og måle forekomst av skader i helsetjenesten. 2010. Den nasjonale pasientsikkerhetskampanjen, Nasjonal enhet for pasientsikkerhet, Nasjonalt kunnskapssenter for helsetjenesten. Oslo
12. Hartwig SC, Denger SD, Schneider PJ. Severity-indexed, incident report-based medication error-reporting program. *Am J Hosp Pharm* 1991;48:2611–16. http://www.nccmerp.org/types-medication-errors
13. Benneyan JC, Lloyd RC, Plsek PE. Statistical process control as a tool for research and healthcare improvement. *Qual Saf Health Care* 2003;12:458–64.
14. Provost LP, Murray S. *The Health Care Data Guide: Learning from Data for Improvement*. 2011. http://books.google.com/books?hl=no&lr=&id=pRLcaOkswQsC&pgis=1 (accessed 7 Jan2015).
15. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;33:159–74.
16. Sharek PJ, Parry G, Goldmann D, *et al*. Performance characteristics of a methodology to quantify adverse events over time in hospitalized patients. *Health Serv Res* 2011;46:654–78.
17. Naessens JM, O'Byrne TJ, Johnson MG, *et al*. Measuring hospital adverse events: assessing inter-rater reliability and trigger performance of the Global Trigger Tool. *Int J Qual Health Care* 2010;22:266–74.
18. Kennerly DA, Saldaña M, Kudyakov R, *et al*. Description and evaluation of adaptations to the global trigger tool to enhance value to adverse event reduction efforts. *J Patient Saf* 2013;9:87–95.
19. Rutberg H, Borgstedt Risberg M, Sjödahl R, *et al*. Characterisations of adverse events detected in a university hospital: a 4-year study using the Global Trigger Tool method. *BMJ Open* 2014;4: e004879.
20. Lau H, Litman KC. Saving lives by studying deaths: Using standardized mortality reviews to improve inpatient safety. *Jt Comm J Qual Patient Saf* 2011;37:400–8.
21. Move your dot: Measuring, Evaluating, and Reducing Hospital Mortality Rates (part 1) IHI Innovation Series white paper. Boston: 2003. http://scholar.google.no/scholar?q=Move+ Your+ DotE284A23A&btnG=&hl=no&as_sdt=0%2C5#1 (accessed 5 Jun 2015).
22. 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.
23. Blyth CR. On Simpson's Paradox and the sure-thing principle. *J Am Stat Assoc* 1972;67:364–6.