PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (**http://bmjopen.bmj.com/site/about/resources/checklist.pdf**) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

TITLE (PROVISIONAL)	Exploring similarities and differences in hospital adverse event rates between Norway and Sweden using Global Trigger Tool
AUTHORS	Deilkås, Ellen; Risberg, Madeleine; Haugen, Marion; Lindstrøm, Jonas; Nylen, Urban; Rutberg, Hans; Soop, Michael

VERSION 1 - REVIEW

REVIEWER	Natasha Rafter Royal College of Surgeons in Ireland Ireland
REVIEW RETURNED	13-May-2016

GENERAL COMMENTS	This is a very good article and the comparison of national rates between countries represents a major step forward in the field of quality and adverse event measurement. The paper however, needs to be be significantly improved with additional detail about the study and GTT methodology, statistical analysis, previous national research and critique of the comparison. Introduction A paragraph detailing the GTT method should be included so that the paper is understandable to all readers. Previous relevant national AE research should be referred to here when providing background and in the discussion as a comparison– in particular Deilkas et al. Monitoring adverse events in Norwegian hospitals from 2010 to 2013. BMJ Open 2015;5:e008576 doi:10.1136/bmjopen-2015-008576 and Soop et al. The incidence of adverse events in Swedish hospitals: a retrospective medical record review study. Int J Qual Health Care 2009;21:285–91.) Why did you not use the Harvard Medical Practice Study methodology as employed by Soop et al in Sweden? Please provide references for the two patient safety campaigns. Information about the healthcare systems in Sweden and Norway and their similarities/differences is required (see comment in conclusion). Methods It should be made clearer that the data collections and methods were separate in the two countries then brought together for comparison (rather than the research designed as one study in two countries). Please explicitly state your objective in the main body of the article as well as in the abstract. Sampling Define somatic care. How did you include/exclude admissions? Did you use ICD codes? Were there differences in how these criteria

	were applied between the countries? Please provide more detail on how the hospitals randomly selected their samples. Settings Please justify the inclusion of data from two planned care hospitals in Norway. Were their results similar to the rest of the hospitals? What is planned care? The Norwegian figures therefore include data from both acute care and planned care hospitals but the Swedish hospitals are all acute care. It appears you are not comparing the same type of hospitals? I note that the numbers of reviewers in stage one and two of the review process were different between the countries. The implications of this should be discussed in the discussion. For example, Norway had two RNs in stage one whereas Sweden used one. Would having two increase the number of triggered admissions (and therefore increase the likelihood of an AE being determined)? Furthermore in stage two it appears that Norway used one physician but Sweden used at least one RN and one physician. Again would more reviewers mean that more or less AEs were determined? In the results section you should present a flowchart of the review process in each country displaying the number of charts that went to stage two review and any differences/similarities should be commented on in the discussion. Translation and validation Please reference the "original GTT white paper" (line 14). The sentence "It was then slightly modified regarding a few triggers" is insufficient, and similarly for the Swedish adaptation about omitted triggers. More detail is required so that the reader could replicate your study methodology. You may wish to list the changes as an appendix and/or refer the reader to table 3 which presumably describes the full list of triggers used for each country (however, I note that the Swedish triggers in the table totals 26 rather than the 27 given on line24). In the paragraph about the Swedish methodology you mention the triggers were reformulated and expanded to assess severity and preventability. This should be described or available as an appendix. Were the same or similar reformulations applied in the Norwegian methodology? Please provide an indication of how similar/difference the Norwegian and Swedish translations and protocols were. Training and standardization It appears that separate training was given in the two countries. This should be explicitly stated and details provided of the differences/similarities. More detail is required about the coordination of the training, who provided the phone support, and team meetings/reviews. All this is important information for readers who may wish to collect national data and compare with other countries. Were any analyses of inter rater reliability conducted between reviewers/teams/hospitals/countries? Definitions Please reference the source of the AE definition. Your division of the events into higher severity (F-I) and all AEs (E-I) should be defined and justified here (have other studies used this categorization?). Line 11 should be mutually exclusive. Line 31 please reference the report by SALAR regarding the findings

	of the record review. Could an admission have more than one AE associated with it? If so then how did you adjust and analyse for this. You state that in Norway the total number of admissions that the sample was selected from was reported. Did this occur in Sweden also? Please state to whom (which organisation) these reports were sent. For example, who is the “campaign secretariat”? Ethics Please state explicitly whether research ethics committee review was obtained or not required. Line 14 id should read identification. Statistical analysis Please describe how the p values were calculated. Was inter-rater reliability calculated with Cohen’s kappa statistic? Did you do a power calculation? Results A flowchart of the number of reviews at each stage in both countries would be helpful in assessing differences between countries in use of the GTT methodology. I am not sure about the journal formal but should thousands etc be separated by commas? (1,000 instead of 1.000). Please state the total number of AEs collected. Please state exact p values. You provide the range of AE rates “between GTT teams in the individual hospitals”. Do you mean between hospitals or between teams within the hospitals? If it is the latter did all the hospitals have several teams working on the reviews? Please add % signs for these AE rates. Are these mean AE rates? If so please state. These results should be discussed in the discussion, in particular how could a team have a 0% AE rate for AEs in the category F-I? Were you able to compare rates of preventability between the countries? Discussion Your findings should be compared with your previous national studies as mentioned above. Line 20 should read “No previous studies…compared national rates of hospital AEs between countries based on the GTT.” Why are your AE rates lower than American hospitals? Please expand on this. Line 30. I do not agree with your statement that the similarity of results shows how robust GTT is. Because you get the same result does not necessarily indicate a robust method unless you can justify that the underlying systems are the same. Line 51. Please expand on the differences between the Norwegian and Swedish patient safety campaigns – are you referring to different methods or areas targeted or results? Conclusions You conclude that the results were expected considering the similarities between the two healthcare systems. It would be helpful for the reader to include some information regarding these similarities, i.e. information about the respective systems in the introduction. References If possible please provide an web address for the references that are

	reports, e.g. numbers 2, 5, 8, 10 etc.

REVIEWER	Paulo Sousa National School of Public Health, Universidade Nova de Lisboa
REVIEW RETURNED	19-May-2016

GENERAL COMMENTS	Very interesting and up-to-date topic. Robust methodology and clear reporting of results and discussion. I have three comments and one concern: i) In page 4 the authors mention “Plotted in control charts” but in page 6 regarding similar data the mention “…in run charts”. Is this correct or is there a mistake? As you know each chart is used to show different information, for different aims. ii) Regarding table 2, 3 and 4, there is too much information in the titles. My recommendation is to separate the title from other information, namely level of significance. iii) In general the English is clear, but there are a couple of “minor mistakes” such as: page 7 paragraph 2 “(…) from for the team” and “from for all teams”; page 10 paragraph one repetition of “in”; page 13 in conclusion “”health care system” and should be plural “systems”. iv) My concern is that no hospital and patient characteristics are provided in this comparison. It would be interesting to have some information regarding hospital dimension, specificity/complexity; case-mix/patient characteristics/or some information on risk adjustment.

REVIEWER	Zegers, Marieke Radboudumc, the Netherlands
REVIEW RETURNED	19-May-2016

GENERAL COMMENTS	Measuring adverse events by health care professionals is, I believe, the most important tool to increase awareness for and to improve patient safety. My major concern is that it is unclear whether the GTT teams exists of people/health care professionals working in the hospitals under review or are independent (from another institute/hospital). If not, there is a risk for bias because professionals are screening the records of their colleagues or even their own records. For quality improvement, this is okay. But for measuring national or local rates of adverse events over time or to compare rates between countries, valid and reliable measurements are required. No psychometric figures were reported in the manuscript. The variation between teams in individual hospitals was large: for me an indication that reliability of the measurements is low. How can you conclude that comparison of adverse event rates based on these results can be done with the GTT? No reliability measures (kappa or agreement scores) were done/reported. And were the adverse event rates corrected for patient characteristics? Were the samples of both countries comparable? I miss a table with the characteristics of the study participants in the

	manuscript. Minor points: Abstract: The rates of adverse events were not reported in the abstract, while the aim of this study is to compare the rates between Norway and Sweden (see also the title of the manuscript). Strengths and limitations: What do you mean with 'The samples are limited regarding detailed information.' Introduction: You state that Norway and Sweden are the only countries that have used GTT as part of a national government policy. In the Netherlands, the GTT tool is also used in the national patient safety program. See: http://www.vmszorg.nl/_page/vms_inline?nodeid=4641&subjectid=6615. This website is in Dutch, but refers to the IHI global trigger tool. Methods- translation and validation: I miss a description of the triggers. And which four triggers were added to the Swedish trigger tool (and which were removed)? Methods – definitions, categorizing and reporting of data: 'In Norway 23 types of AEs were specified; in Sweden 27 types' Give an overview of these types. Methods – which statistical test was used to compare adverse event rates?

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1
Reviewer Name: Natasha Rafter
Institution and Country: Royal College of Surgeons in Ireland, Ireland
Competing Interests: None declared

This is a very good article and the comparison of national rates between countries represents a major step forward in the field of quality and adverse event measurement.
Thanks for the comment.
The paper however, needs to be significantly improved with additional detail about the study and GTT methodology, statistical analysis, previous national research and critique of the comparison.
Introduction
1. A paragraph detailing the GTT method should be included so that the paper is understandable to all readers.
Thanks for the suggestion. We have included a paragraph about the GTT method under "Methods".
2. Previous relevant national AE research should be referred to here when providing background and in the discussion as a comparison– in particular Deilkas et al. Monitoring adverse events in Norwegian hospitals from 2010 to 2013. BMJ Open 2015;5:e008576 doi:10.1136/bmjopen-2015-008576 and Soop et al. The incidence of adverse events in Swedish hospitals: a retrospective medical record review study. Int J Qual Health Care 2009;21:285–91.) Why did you not use the Harvard Medical Practice Study methodology as employed by Soop et al in Sweden?
Thanks for the suggestions. We have included the suggested references and elaborated to answer your question regarding choosing the GTT as a method.

3. Please provide references for the two patient safety campaigns.
Thanks for the suggestions. References are provided.
4. Information about the healthcare systems in Sweden and Norway and their similarities/differences is required (see comment in conclusion).
Thanks for the suggestion. We have included information and a reference regarding the healthcare and welfare systems in the introduction.
Methods
5. It should be made clearer that the data collections and methods were separate in the two countries then brought together for comparison (rather than the research designed as one study in two countries).
Thanks for the suggestion. We have added a phrase about this in the beginning of the “Sampling” paragraph.
6. Please explicitly state your objective in the main body of the article as well as in the abstract.
Thanks for the correction. The objective is now specified in the Introduction.

Sampling
7. Define somatic care.
To clarify, we have removed the word “somatic” and rephrased the whole sentence.
8. How did you include/exclude admissions?
In Norwegian hospitals eligible admissions were selected according to the criteria, using the hospitals own electronically patient administrative systems.
9. Did you use ICD codes?
ICD codes are used by reviewers as part of the record review, but are not reported nationally.
10. Were there differences in how these criteria were applied between the countries?
No.
11. Please provide more detail on how the hospitals randomly selected their samples.
Thanks for the suggestion. We have included more information about this.
Settings
12. Please justify the inclusion of data from two planned care hospitals in Norway.
Thanks for pointing out the need to clarify this. In Norway and Sweden acute care hospitals also do planned care. In fact most planned hospital care is done in acute care hospitals. The two planned care hospitals represent less than one per cent (0.3 % and 0.5 % respectively) of the total population of the eligible Norwegian admissions included in the study, and could in size well have been part of the planned activity in one or several acute care hospitals.
13. Were their results similar to the rest of the hospitals?
Samples from each hospital are too small to compare results between hospitals, so we have not done that. The bimonthly samples are just large enough to make time series that can be analysed with statistical process control.
14. What is planned care?
Planned care refers to admissions that are not acute. In order to avoid confusion the sentence is rephrased.
15. The Norwegian figures therefore include data from both acute care and planned care hospitals but the Swedish hospitals are all acute care. It appears you are not comparing the same type of hospitals?
See clarification at point 11. The hospitals results are weighted according to their small populations (0.8% of total population) in the national estimates and have equally small influence on the estimates.
16. I note that the numbers of reviewers in stage one and two of the review process were different between the countries. The implications of this should be discussed in the discussion. For example, Norway had two RNs in stage one whereas Sweden used one.
Would having two increase the number of triggered admissions (and therefore increase the likelihood of an AE being determined)?

We have included comments on this in the discussion.

17. Furthermore in stage two it appears that Norway used one physician but Sweden used at least one RN and one physician. Again would more reviewers mean that more or less AEs were determined?
We doubt it since the GTT emphasizes that it is the physician who validates the results.

18. In the results section you should present a flowchart of the review process in each country displaying the number of charts that went to stage two review and any differences/similarities should be commented on in the discussion.
This is not possible since information about the charts in stage one was not reported to national level. Only information about AEs was reported, with supplementary information regarding type, severity, and numbering according to chart, so that AE's identified in the same chart could be aggregated. In Sweden supplementary information was more extensive, with age, gender and length of stay.
Translation and validation
19. Please reference the "original GTT white paper" (line 14).
Thanks for the suggestion. It is done.
20. The sentence "It was then slightly modified regarding a few triggers" is insufficient,
We have included a reference to where the slight Norwegian modifications are described.
21. … and similarly for the Swedish adaptation about omitted triggers. More detail is required so that the reader could replicate your study methodology.
We have added a reference to the document which describes the Swedish triggers.
22. You may wish to list the changes as an appendix and/or refer the reader to table 3 which presumably describes the full list of triggers used for each country (however, I note that the Swedish triggers in the table totals 26 rather than the 27 given on line24).
The categories in Table 3 refer to adverse events types, not triggers. We have corrected the number to 26.
23. In the paragraph about the Swedish methodology you mention the triggers were reformulated and expanded to assess severity
The descriptions of the triggers (not the triggers themselves) were reformulated and expanded to facilitate assessment of the severity… This should be described or available as an appendix.
Since the extended Swedish trigger descriptions constitute a 44 page long document we suggest to refer to the document rather than include it as an appendix.
24. ….and preventability...
The element of preventability was added only in Sweden, and we have therefore left it out of the manuscript, since it does not apply to the comparison between countries.
25. Were the same or similar reformulations applied in the Norwegian methodology?
No.
26. Please provide an indication of how similar/difference the Norwegian and Swedish translations and protocols were.
We have added comments on this in the discussion.

Training and standardization
27. It appears that separate training was given in the two countries. This should be explicitly stated and details provided of the differences/similarities. More detail is required about the coordination of the training, who provided the phone support, and team meetings/reviews. All this is important information for readers who may wish to collect national data and compare with other countries.
Thanks for the suggestions. We have elaborated on that.

28. Were any analyses of inter rater reliability conducted between reviewers/teams/hospitals/countries?
Lack of resources did not allow to conduct inter rater reliability analyses between reviewer teams

across hospitals or countries.
Definitions
29. Please reference the source of the AE definition.
Thanks for the suggestion. We have done that.
30. Your division of the events into higher severity (F-I) and all AEs (E-I) should be defined and justified here (have other studies used this categorization?).
Thanks for the suggestions. The categories are now defined and references to other publications have been made.
31. Line 11 should be mutually exclusive.
Thanks for the correction.
32. Line 31 please reference the report by SALAR regarding the findings of the record review.
That is done.
33. Could an admission have more than one AE associated with it? If so then how did you adjust and analyse for this.
Thanks for pointing out need for clarity. The point is now clarified.
34. You state that in Norway the total number of admissions that the sample was selected from was reported. …Please state to whom (which organisation) these reports were sent. For example, who is the “campaign secretariat”?
It is now specified.
35. Did this (the total number of admissions that the sample was selected from was reported. …) occur in Sweden also?
Yes, the hospitals in Sweden reported the total number of admissions that the investigated records had been randomly selected from. These numbers have been used to weight the hospital results, when making national estimates. National AE rates have been calculated as a weighted average of individual mean AE rates for 45 GTT teams in Norway and as a weighted average of individual mean AE rates for 63 hospitals in Sweden. This is now clarified in the manuscript.
Ethics
36. Please state explicitly whether research ethics committee review was obtained or not required.
This is now specified.

37. Line 14 id should read identification.
It is now corrected.

Statistical analysis
38. Please describe how the p values were calculated.
We do not calculate exact p-values. Inferences of the difference in mean AE rates between Sweden and Norway have been made from 95% bootstrap confidence intervals based on 10.000 simulations. If zero is not contained within the 95% non-parametric confidence interval, then the probability that the difference in mean AE rates between the two countries is zero is less than 5%.
As we describe in the beginning of the statistical analysis paragraph, national AE rates with associated 95% confidence intervals are divided into types and severities in our cross-sectional analysis. We have calculated bootstrap confidence intervals since AEs according to type and severity do not follow a symmetrical distribution, i.e. there are few observations for some combinations of type and severity. This gives a positively skewed distribution and a t-confidence interval is not adequate.
39. Was inter-rater reliability calculated with Cohen’s kappa statistic?
Inter-rater reliability was not studied due to lack of funding available for the study.
40. Did you do a power calculation?
We did not do a power calculation. Our study is based on all available data and it would not be possible to gather more data as the data is collected as part of the hospitals quality control process. Furthermore, as we have written in the methods chapter in our manuscript, our study was planned and designed after the data collection was finished.

Results

41. A flowchart of the number of reviews at each stage in both countries would be helpful in assessing differences between countries in use of the GTT methodology.
Number of review stages was the same in both countries. We do not have information about how many records that had triggers compared to how many records that had AEs for every bimonthly review. We are therefore not able to make a flowchart for all stages in the review.

42. I am not sure about the journal formal but should thousands etc be separated by commas? (1,000 instead of 1.000).

In a previous article in this journal, neither comma nor period was required[4].

43. Please state the total number of AEs collected.
In the samples a total of 1672 AEs across all severity categories were identified in Norway, and 3217 in Sweden. In the paper we present hospital admissions with at least one AE. We think it may be confusing for the reader if we state these in the manuscript. But we have included a flowchart
44. Please state exact p values.
See line of reasoning at point 38.
45. You provide the range of AE rates “between GTT teams in the individual hospitals”. Do you mean between hospitals or between teams within the hospitals?
In Norway we have weighted individual mean AE rates for 45 GTT teams but in Sweden we have weighted individual mean AE rates for 63 hospitals. The range of AE rates is therefore between GTT teams in Norway and between hospitals in Sweden. This is now clarified in the manuscript.
46. If it is the latter did all the hospitals have several teams working on the reviews?
No, only some hospitals had several teams working on the review. Four hospitals in Norway had more than one team(i.e. two hospitals had two teams and two hospitals had seven teams). In Sweden three hospitals had more than one team.
47. Please add % signs for these AE rates.
Thanks for the suggestion. It is done.
48. Are these mean AE rates? If so please state.
Yes, the range of AE rates shows the minimum and maximum of the individual mean AE rates for 45 GTT teams in Norway and 63 hospitals in Sweden, respectively. These are based on cross-sectional analysis of a sample between 240 and 480 charts per team/hospital, where at least ten charts have been randomly selected bimonthly during the whole year.
49. These results should be discussed in the discussion, in particular how could a team have a 0% AE rate for AEs in the category F-I?
The small sample sizes probably contribute to the large difference in AE rates between hospitals in addition to differences between the characteristics of the hospitals patient populations and activities.
50. Were you able to compare rates of preventability between the countries?
No, since this element is not part of the original GTT manual, and not part of the Norwegian GTT procedure.

Discussion

51. Your findings should be compared with your previous national studies as mentioned above.
We have elaborated on that.
52. Line 20 should read “No previous studies…compared national rates of hospital AEs between countries based on the GTT.”
Thanks for pointing out need for this clarification. It is now corrected.
53. Why are your AE rates lower than American hospitals? Please expand on this.
Thanks for the question. We have chosen to remove the sentence that compares our results with the American GTT studies since the American samples were either drawn from a different time period

(2002 – 2007) or from a more elderly population (65 years and above). Both factors could influence levels of risk that make comparison not justifiable.
54. Line 30. I do not agree with your statement that the similarity of results shows how robust GTT is. Because you get the same result does not necessarily indicate a robust method unless you can justify that the underlying systems are the same.
Thanks for the comment. We have now justified our conclusion by describing conditions and context for Norwegian and Swedish healthcare showing their similarities, in a paragraph on page 4. We have also adjusted our statement on page 12 regarding how GTT can be useful for cross country comparison.

55. Line 51. Please expand on the differences between the Norwegian and Swedish patient safety campaigns – are you referring to different methods or areas targeted or results?
We have removed the sentence since the target areas of the patient safety campaigns were similar.

Conclusions
56. You conclude that the results were expected considering the similarities between the two healthcare systems. It would be helpful for the reader to include some information regarding these similarities, i.e. information about the respective systems in the introduction.
We have included this information in the introduction and discussion and adjusted the conclusion to comply with adjustments.
References
57. If possible please provide a web address for the references that are reports, e.g. numbers 2, 5, 8, 10 etc.
The list is as follows:
2 OECD. Health at a Glance 2013: OECD Indicators. Health at a glance, 2013. - https://www.oecd.org/els/health-systems/Health-at-a-Glance-2013.pdf
4 Nordisk Ministerråd. Nordisk kvalitetsmåling i sundhedsvæsenet. København, Denmark, 2010:119-23. - https://sundhedsstyrelsen.dk/da/sundhed-og-livsstil/tandpleje/~/media/5AB21C0326F342CEBE76D577D763D191.ashx
5 Working group on Patient Safety under Nordic Ministry Council. A report on patient safety. Copenhagen: Styrelsen for Patientsikkerhed, 2016. - https://stps.dk/da/nyheder/2016/~/media/2B522420CECE4EDBA2B273B0B6518F2A.ashx
6 Griffin F, Resar R. IHI Global Trigger Tool for Measuring Adverse Events IHI Innovation series. 2nd ed. Cambridge, MA, 2009. - http://www.ihi.org/resources/pages/ihiwhitepapers/ihiglobaltriggertoolwhitepaper.aspx
8 Health Foundation. Evidence scan: Global trigger tools. London, United Kingdom: The Health Foundation, 2010. – http://www.health.org.uk/sites/health/files/EvidenceScanGlobalTriggerTools.pdf
10 Health Quality & Safety Commission. The Global Trigger Tool: A Review of the Evidence - Report for the Health Quality & Safety Commission New Zealand. Wellington: Health Quality & Safety Commission, 2013. - http://www.hqsc.govt.nz/assets/GTT/PR/GTT-evidence-review-Jan-2016.pdf
18 Patientsäkerhetssatsning 2011 överenskommelse mellan staten och Sveriges Kommuner och Landsting om förbättrad patientsäkerhet. In: Government TS, ed. Stockholm, 2011. - http://www.regeringen.se/contentassets/7b321cef1a814194babb11a788f385ef/bemyndigande-att-underteckna-en-overenskommelse-om-patientsakerhetssatsning-2011
19 Den nasjonale pasientsikkerhetskampanjen. Strukturert journalundersøkelse, ved bruk av Global Trigger Tool for å identifisere og måle forekomst av skader i helsetjenesten, 2010. - http://www.pasientsikkerhetsprogrammet.no/no/M%C3%A5linger/Materiell/_attachment/249?_ts=135a0601909
20 Sveriges Kommuner och Landsting. Markörbaserad journalgranskning - för att identifiera och mäta skador i vården. LTAB: Sveriges Kommuner och Landsting, 2012. - http://webbutik.skl.se/bilder/artiklar/pdf/7164-847-1.pdf?issuusl=ignore
21 Sveriges Kommuner och Landsting. Markörer med definitioner Stockholm: Sveriges Kommune og

Landsting, 2012. - http://webbutik.skl.se/internt/artiklar/7164-847-1/Rapport_markorer_och_definitioner.pdf?issuusl=ignore
22 Levinson DR. Adverse events in hospitals. National incidence among Medicare beneficiaries.: Department of Health and Human Services, Office of Inspector General, 2010. doi:OEI-06-09-00090 - https://oig.hhs.gov/oei/reports/oei-06-09-00090.pdf
23 Sveriges Kommuner och Landsting. Vårdskador VAD TRODDE VI DÅ – VAD VET VI NU? Stockholm: Sveriges Kommuner och Landsting, 2016. - http://skl.se/download/18.6ee610e515505b6272d35cbd/1465398354772/Vad+trodde+vi+d%C3%A5+-+vad+vet+vi+nu_+SKL.pdf

1 Deilkås E, Bukholm G, Lindstrøm JC, et al. Monitoring adverse events in Norwegian hospitals from 2010 to 2013 BMJ open 2015.

Reviewer 2
Reviewer Name: Paulo Sousa
Institution and Country: National School of Public Health, Universidade Nova de Lisboa, Portugal.
Competing Interests: None declared
1. Very interesting and up-to-date topic. Robust methodology and clear reporting of results and discussion.
Thanks.
I have three comments and one concern:
2. In page 4 the authors mention “Plotted in control charts” but in page 6 regarding similar data the mention “…in run charts”. Is this correct or is there a mistake? As you know each chart is used to show different information, for different aims.
Thanks for pointing this out. We have provided both run charts and control charts for different purposes related to analysing the GTT data locally. To simplify I have revised the manuscript to only refer to “run” charts.
3. Regarding table 2, 3 and 4, there is too much information in the titles. My recommendation is to separate the title from other information, namely level of significance.
Thanks for the suggestion. We have followed your advice.
4. In general the English is clear, but there are a couple of “minor mistakes” such as: page 7 paragraph 2 “(…) from for the team” and “from for all teams”; page 10 paragraph one repetition of “in”; page 13 in conclusion “”health care system” and should be plural “systems”.
Thanks for pointing out the mistakes. They are now corrected.
5. My concern is that no hospital and patient characteristics are provided in this comparison. It would be interesting to have some information regarding hospital dimension, specificity/complexity; case-mix/patient characteristics/or some information on risk adjustment.
We agree that this information would be very interesting to have. Data protection regulation did not allow us to collect data from the hospitals that could identify individual patients. Since the random samples have been drawn from all somatic hospital discharges, we consider the risk for a skewed sample to be minute regarding demographic characteristics.

Reviewer 3
Reviewer: 3
Reviewer Name: M Zegers
Institution and Country: Radboudumc, the Netherlands
Competing Interests: None declared

1. Measuring adverse events by health care professionals is, I believe, the most important tool to increase awareness for and to improve patient safety.
Thanks. We agree.

2. My major concern is that it is unclear whether the GTT teams exists of people/health care professionals working in the hospitals under review or are independent (from another institute/hospital). If not, there is a risk for bias because professionals are screening the records of their colleagues or even their own records.
Thanks for the important point. In our study the reviewers are employed in the hospitals where they work. We recommend that reviewers are clinically experienced, and find that most reviewers are, although many work in the quality department, where they at the time do not do clinical work. Our strategy leans on a study where internal GTT teams found more AEs than external teams[3].
1. For quality improvement, this is okay. But for measuring national or local rates of adverse events over time or to compare rates between countries, valid and reliable measurements are required. No psychometric figures were reported in the manuscript.
We agree that more research is needed to explore both inter-rater reliability across countries, and validity related to other patient safety measurements. We have added comments on that in the discussion.
2. The variation between teams in individual hospitals was large: for me an indication that reliability of the measurements is low.
We have added the following comment on the variation between teams in individual hospitals in the discussion.
“Between hospitals in each country variation in total AE rates was large. That is expected since rates are based on cross sectional analyses of small samples allowing large random variation. In addition there are differences between characteristics of hospitals’ patient populations and activities. For this reason we do not use GTT results for comparison between hospitals.”
3. How can you conclude that comparison of adverse event rates based on these results can be done with the GTT? No reliability measures (kappa or agreement scores) were done/reported.
We agree that this study is explorative and not conclusive. We have adjusted the manuscript accordingly.
4. And were the adverse event rates corrected for patient characteristics?
Data protection regulation did not allow us to collect data from the hospitals that could identify individual patients.
5. Were the samples of both countries comparable?
Since the random samples have been drawn from all somatic hospital discharges in both countries, we consider the risk for a skewed sample to be minute regarding demographic characteristics.
6. I miss a table with the characteristics of the study participants in the manuscript.
Unfortunately we do not have this information for reasons explained in point 4 and 5. We agree that this information would be very interesting to have.
Minor points:
Abstract:

7. The rates of adverse events were not reported in the abstract, while the aim of this study is to compare the rates between Norway and Sweden (see also the title of the manuscript).
We have adjusted the title according to point 3 and included the overall rates in the abstract.
Strengths and limitations:
8. What do you mean with ‘The samples are limited regarding detailed information.’
Thanks for pointing out need for clarification. We have replaced the sentence as follows:
“The study does not include demographic characteristics.”

Introduction:
9. You state that Norway and Sweden are the only countries that have used GTT as part of a national government policy. In the Netherlands, the GTT tool is also used in the national patient safety program. See: http://www.vmszorg.nl/_page/vms_inline?nodeid=4641&subjectid=6615. This website is in Dutch, but refers to the IHI global trigger tool.
Thanks for the correction. We have adjusted the manuscript accordingly.

10. Methods- translation and validation: I miss a description of the triggers. And which four triggers were added to the Swedish trigger tool (and which were removed)?
The extended Swedish trigger description constitutes a 44 page long document. We therefore suggest to refer to the document rather than include it in the manuscript or as an appendix.
The Norwegian trigger description constitutes 11 pages and is directly translated from the original English version with minor contextual adjustments specified in an article which we refer to.
11. Methods – definitions, categorizing and reporting of data: 'In Norway 23 types of AEs were specified; in Sweden 27 types' Give an overview of these types.
Thanks for pointing out need for clarification. We have added that the types are specified in Table 3.
12. Methods – which statistical test was used to compare adverse event rates?
Inferences of the difference in mean AE rates between Sweden and Norway have been made from 95% bootstrap confidence intervals based on 10.000 simulations. If zero is not contained within the 95% non-parametric confidence interval, then the probability that the difference in mean AE rates between the two countries is zero is less than 5%. We do not calculate exact p-values.
As we describe in the beginning of the statistical analysis paragraph, national AE rates with associated 95% confidence intervals are divided into types and severities in our cross-sectional analysis. We have calculated bootstrap confidence intervals since AEs according to type and severity do not follow a symmetrical distribution, i.e. there are few observations for some combinations of type and severity. This gives a positively skewed distribution and a t-confidence interval is not adequate.

VERSION 2 – REVIEW

REVIEWER	Natasha Rafter Royal College of Surgeons in Ireland Ireland
REVIEW RETURNED	16-Sep-2016

GENERAL COMMENTS	This paper has been significantly improved with the revisions. I have only a few minor comments: 1. GTT and/or medical record review should be included in the title. 2. Because this is the first comparison between countries using the GTT and others are likely to follow, statistical peer review is needed to confirm the validity of the method to combine data from teams/sites with a ten fold difference in AE rates (eg in Norway: 2.1% - 21.7%) and the validity of the statistically significant differences in multiple comparisons of event types. 3. The confidence intervals around the estimates of 13.0% and 14.4% should be given in the abstract. 4. The acknowledgements section refers to a flowchart but one is not present in the paper?

REVIEWER	Paulo Sousa National School of Public Health - New University of Lisbon Portugal
REVIEW RETURNED	12-Sep-2016

GENERAL COMMENTS	1- I would like to reinforce the relevance of the topic and the internal robustness of the paper as a whole. 2- In my opinion, the authors have made a great effort to improve the paper and include the reviewers suggestions/comments. For these reasons, my opinion is that the paper is fit for publishing.

REVIEWER	Zegers, Marieke Radboudumc, Nijmegen, The Netherlands
REVIEW RETURNED	13-Sep-2016

GENERAL COMMENTS	The paper is certainly improved. I have, however, still concerns about the validity of the conclusion 'the level of patient safety was essentially the same in both countries'. Can you state this without adjusting the adverse event rates for case-mix correction? In the letter to the editor you write 'the objective of our patient safety research is to study relations between structures, processes and results in healthcare.' To explore this, you certainly have to correct for case-mix and adjust for clustering of data in hospitals (see: http://www.ncbi.nlm.nih.gov/pubmed/21227956). You gave answers on several comments, but you did not change the text in the paper. Almost all Dutch hospitals are reviewing patient records to identify adverse events, obligated by the government. In the introduction section you still write 'Norway and Sweden which, to our knowledge, are the only countries that have used GTT in all hospitals as part of a national government policy. ' In your response you write 'data protection regulation did not allow us to collect data from the hospitals that could identify individual patients'. An important limitation that is not mentioned in the discussion section. No text is included about the independency of the reviewers (reviewing the records of their own hospital). Your answer on this comment is excellent. Please, incorporate this in the text of the paper. You state 'more research on issues like inter-rater reliability is however needed to explore validity and reliability……' You may refer to: http://www.ncbi.nlm.nih.gov/pubmed/27550650 You did not include a (short) description of the used triggers. Readers cannot repeat your study and copy your tool for quality improvement in their institution.

REVIEWER	Catherine Bresee, MS Cedars-Sinai Medical Center Biostatistics & Bioinformatics Research Center 8700 Beverly Blvd PACT Building, Suite 900c Los Angeles, CA 90048
REVIEW RETURNED	31-Oct-2016

GENERAL COMMENTS	Overall I found the manuscript as presented by Deilkas et all to be well executed and well presented. As a Biostatistician, I suggest that the following items be addressed prior to publication: 1) (Major Comment) Most readers will expect to see p-values associated with the bootstrap 95% Confidence Intervals presented which should be able to be calculated. See the white paper from Wludyka and Smotherman. "Using SAS to Create a p-value Resampling Distribution for a Statistical Test.". Also see Li, Jialiang, Bee Choo Tai, and David J. Nott. "Confidence interval for the bootstrap P-value and sample size calculation of the bootstrap test." Journal of Nonparametric Statistics 21.5 (2009): 649-661.

	2) (Major Comment) In the Statistical Analysis methods, it is unclear what is the sampling frame. Is the analysis performed on the absolute number of records sampled, or the rates within each GTT team? For example in Table 2, in the Norway column, what is the Total N. Is it N=10986 records, or N=45 hospitals? 3) In the Abstract, Line 14, the term “somatic care” is unclear. Please revise. 4) In the Abstract Line 34, change “When excluding AEs of least severity…” suggest rewrite as “In sub-analysis of only serious and severe AEs…” 5) In the Methods Section, Line 55, state that the AE’s are rated on a 5-point severity scale and cite Table 1. 6) In the Statistical Analysis Methods Section, please indicate what R-package was used to perform the bootstrap sampling. 7) Results Section, Lines 9 & 12, cite the number of Norwegian and Swedish hospitals sampled. 8) Tables 2, 3 & 4 – Remove double asterisks, and replace with single. A 95% confidence interval either contains zero or does not. “Almost” doesn’t count. Including a p-value will assist readers to evaluate the magnitude of the effect.

VERSION 2 – AUTHOR RESPONSE

Reviewer: 1

Reviewer Name: Natasha Rafter

Institution and Country: Royal College of Surgeons in Ireland, Ireland Competing Interests: None declared

This paper has been significantly improved with the revisions. I have only a few minor comments:

1. GTT and/or medical record review should be included in the title.

The title is changed accordingly.

2. Because this is the first comparison between countries using the GTT and others are likely to follow, statistical peer review is needed to confirm the validity of the method to combine data from teams/sites with a ten fold difference in AE rates (eg in Norway: 2.1% - 21.7%) and the validity of the statistically significant differences in multiple comparisons of event types.

We are grateful for the statistical peer review.

3. The confidence intervals around the estimates of 13.0% and 14.4% should be given in the abstract.

The confidence intervals are now included in the abstract.

4. The acknowledgements section refers to a flowchart but one is not present in the paper?

Please excuse our mistake. The reference is now removed.

Reviewer: 2

Reviewer Name: Paulo Sousa

Institution and Country: National School of Public Health - New University of Lisbon, Portugal
Competing Interests: None to declare

1- I would like to reinforce the relevance of the topic and the internal robustness of the paper as a whole.

Thanks. We appreciate your comment.

2- In my opinion, the authors have made a great effort to improve the paper and include the reviewers suggestions/comments.

For these reasons, my opinion is that the paper is fit for publishing.

We are pleased to note this. Thanks.

Reviewer: 3

Reviewer Name: M Zegers

Institution and Country: Radboudumc, Nijmegen, The Netherlands Competing Interests: None declared

1. The paper is certainly improved. I have, however, still concerns about the validity of the conclusion 'the level of patient safety was essentially the same in both countries'. Can you state this without adjusting the adverse event rates for case-mix correction? In the letter to the editor you write 'the objective of our patient safety research is to study relations between structures, processes and results in healthcare.' To explore this, you certainly have to correct for case-mix and adjust for clustering of data in hospitals (see: http://www.ncbi.nlm.nih.gov/pubmed/21227956).

The concept of case mix is interesting and would have been especially important if we had reason to believe that requirements for healthcare were different in the two countries. As described in the third paragraph, we do not suspect this since the characteristics of the two neighbouring countries are similar. The concept of case mix would also have been relevant if we were comparing hospitals or departments within countries, and not all somatic hospital admissions between two similar countries.

2. You gave answers on several comments, but you did not change the text in the paper. Almost all Dutch hospitals are reviewing patient records to identify adverse events, obligated by the government. In the introduction section you still write 'Norway and Sweden which, to our knowledge, are the only countries that have used GTT in all hospitals as part of a national government policy. '

We have changed the statement as follows:

"are among the few countries that have required use of GTT in all hospitals as part of a national government policy. "

3. In your response you write 'data protection regulation did not allow us to collect data from the hospitals that could identify individual patients'. An important limitation that is not mentioned in the discussion section.

Thanks for the suggestion. We have now commented on this in the discussion:

"Data protection regulations did not allow us to collect individual demographic patient data from the hospitals."

4. No text is included about the independency of the reviewers (reviewing the records of their own hospital). Your answer on this comment is excellent. Please, incorporate this in the text of the paper.

Thanks for the suggestion. We have now included a comment on this:

"Reviewers in both countries were employed in the hospitals where they worked clinically. They were all clinically experienced although some worked in the quality department, where they sometimes at the time of the review did not do clinical work. Our strategy leans on a study where internal GTT teams found more AEs than external teams [9]."

You state 'more research on issues like inter-rater reliability is however needed to explore validity and reliability……' You may refer to: http://www.ncbi.nlm.nih.gov/pubmed/27550650

Thanks for the good idea. We have referred to the article in the introduction and in the discussion.

5. You did not include a (short) description of the used triggers.

- We have provided references to the Norwegian and Swedish trigger tools which are publically available online free of charge.
- The Swedish trigger description constitutes a 44 page long document.
- The Norwegian trigger description constitutes 11 pages. It is directly translated from the original English version with minor contextual adjustments specified in an article which we refer to.
- Since the lists of triggers are too long it is not possible to include them in the manuscript or as an appendix.

6. Readers cannot repeat your study and copy your tool for quality improvement in their institution.

The sources for the tools are referred to in the paper, and they are available online, free of charge, enabling readers to repeat the study.

Reviewer: 4

Reviewer Name: Catherine Bresee, MS

Institution and Country: Cedars-Sinai Medical Center, Biostatistics & Bioinformatics Research Center, USA Competing Interests: None declared

Overall I found the manuscript as presented by Deilkas et all to be well executed and well presented. As a Biostatistician, I suggest that the following items be addressed prior to publication:

1) (Major Comment) Most readers will expect to see p-values associated with the bootstrap 95% Confidence Intervals presented which should be able to be calculated. See the white paper from Wludyka and Smotherman. "Using SAS to Create a p-value Resampling Distribution for a Statistical Test.". Also see Li, Jialiang, Bee Choo Tai, and David J. Nott. "Confidence interval for the bootstrap P-value and sample size calculation of the bootstrap test." Journal of Nonparametric Statistics 21.5 (2009): 649-661.

We do not understand this comment; either p-values or confidence intervals may be used to determine statistical significance. It should be sufficient to only include confidence intervals since they also contain more information than p-values. Along with the statistical significance, confidence intervals permit statements about the direction or size of the difference between the AE rates in the two countries. This is particularly useful when the results are not significant. In addition, the variability of the estimation can be assessed by the width of the confidence interval; the narrower the confidence interval, the more precise the estimation.

We did not expect to find large differences in this study, due to the similar structural conditions and contexts for healthcare in Norway and Sweden. The similarities are described from line number 40 in the introduction. This supports our use of confidence intervals.

Our analyses were explorative rather than confirmatory. The study was designed after the collection of GTT data from 2013. Therefore, we do not test a priori hypotheses that are made before the measurement phase begins. We explore similarities and differences in hospital AE rates between the two countries by examining GTT data from 2013.

2) (Major Comment) In the Statistical Analysis methods, it is unclear what is the sampling frame. Is the analysis performed on the absolute number of records sampled, or the rates within each GTT team? For example in Table 2, in the Norway column, what is the Total N. Is it N=10986 records, or N=45 hospitals?

Under statistical analysis we have written that the basis for the bootstrap procedure is the individual means for all GTT teams in Norway and all hospitals in Sweden, i.e. it was done on the level of each GTT team since each hospital in Sweden has its own review team. We thought that it was clear enough, so thank you for the feedback. We have now clarified that N=45 for Norway and N=63 for Sweden under statistical analysis:

“The bootstrap was performed by drawing randomly with replacement from the individual means for all 45 GTT teams in Norway and all 63 hospitals in Sweden (each with its own review team).”

3) In the Abstract, Line 14, the term “somatic care” is unclear. Please revise.

Thanks for the comment. It is now rephrased as follows:

“…undergoing care with an in-hospital stay of at least 24 hours, excluding psychiatric care and rehabilitation.”

4) In the Abstract Line 34, change “When excluding AEs of least severity…” suggest rewrite as “In sub-analysis of only serious and severe AEs…”

Thanks for the suggestion. We have rephrased as follows:

“In sub-analysis of more severe AEs, …”

5) In the Methods Section, Line 55, state that the AE’s are rated on a 5-point severity scale and cite Table 1.

Thanks for the suggestion. We have done that.

6) In the Statistical Analysis Methods Section, please indicate what R-package was used to perform the bootstrap sampling.

We have not applied any R-package to perform the bootstrap sampling. We have programmed the bootstrap ourselves in R.

7) Results Section, Lines 9 & 12, cite the number of Norwegian and Swedish hospitals sampled.

Thanks for the suggestion. It is done. Together with these numbers, we have also specified the number of GTT teams in the two countries to make sure that the sample size is clear. The sampling is done on the level of each GTT team, where N=45 in Norway and N=63 in Sweden.

8) Tables 2, 3 & 4 – Remove double asterisks, and replace with single. A 95% confidence interval either contains zero or does not. “Almost” doesn’t count. Including a p-value will assist readers to evaluate the magnitude of the effect.

With the single and double asterisks we wanted to make a distinction between clear statistical significance (one asterisk) and statistical significance where zero is near to the limits of the confidence interval (two asterisks). We have now deleted both the single and double asterisks in the tables 2-4.

Regarding p-values please see comment on 1).

VERSION 3 – REVIEW

REVIEWER	Zegers, Marieke Radboudumc, Nijmegen, The Netherlands
REVIEW RETURNED	06-Dec-2016

GENERAL COMMENTS	It is still a concern that the authors do not mention the lack of information about the study population. How can you draw conclusions that AE rates are equal between two countries without knowing (and correcting for) differences in characteristics of the population studied? A remark about this, and it’s possible implications for the validity of the study results and conclusions, is lacking in the discussion section and abstract of the paper and should be included before the paper is acceptable for publication.

REVIEWER	Catherine Bresee, MS Biostatistics & Bioinformatics Research Center Cedars-Sinai Medical Center Los Angeles, CA

REVIEW RETURNED	20-Dec-2016

GENERAL COMMENTS	With the edits I find the revision satisfactory.

VERSION 3 – AUTHOR RESPONSE

Thanks for the following comment from the reviewers:

"It is still a concern that the authors do not mention the lack of information about the study population. How can you draw conclusions that AE rates are equal between two countries without knowing (and correcting for) differences in characteristics of the population studied? A remark about this, and its possible implications for the validity of the study results and conclusions, is lacking in the discussion section and abstract of the paper and should be included before the paper is acceptable for publication."

We have complied with the reviewers concern by elaborating further on the point, and adding the following to

1. The abstract:
"Similar contexts for healthcare and similar socio- economic status and demographic characteristics have inspired the Nordic countries to exchange experiences from measuring and monitoring quality and patient safety in healthcare."

2. “Strengths and limitations of this study”:
• The study does not include demographic data or other patient characteristics.

3. The introduction
“Norway and Sweden have similar structural conditions and contexts for healthcare as tax based funding, similar socio- economic status, demographic characteristics, publically funded education of healthcare employees and democratic policies pursuing equal access”

4. The discussion:
"In this study the numbers of records reviewed are probably the largest random samples of admissions drawn from all hospitals in two countries that have been reviewed with GTT. That should with reason be sufficient for comparing overall results as well as specific AE type rates. Unfortunately, data protection regulations did not allow us to collect individual demographic or other patient data, which would have allowed us to identify and correct for possible differences in characteristics between the populations of the two countries which could have influenced the results. However, the risk for such differences is probably minor as the general demographics and other population characteristics of the two countries are very similar."

We hope the adjustments meet your expectations.