Is inter-rater reliability of Global Trigger Tool results altered when members of the review team are replaced? Running title: An observational study comparing results of the Global Trigger Tool between teams with varying degrees of replacements of the reviewers

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

[1] Nordland Hospital, Post box 1480, 8092 Bodø, Norway, [2] Fran Griffin & Associates, LLC, 318 Sea Spray Lane Neptune, NJ USA 07753, [3] Akershus University Hospital, Post box 1000, 1478 Lørenskog, Norway, [4] Nordland Hospital Trust, Post box 1480, 8092 Bodø, Norway and Institute for Community Medicine, and [5] The Artic University of Norway, Post box 6050, Langnes 9037 Tromsø, Norway

Abstract:

Objective. To evaluate the inter-rater reliability of results from Global Trigger Tool (GTT) reviews when one of the three reviewers remains consistent while one or two reviewers rotate

Design. Comparison of results from retrospective record review performed as a cross-sectional study with three review teams each consisting of two non-physicians and one physician; Team I (three consistent reviewers), Team II (one of the two non-physician reviewers or/and the physician from Team I are replaced for different review periods) and Team III (three consistent reviewers different from reviewers in Team I and Team II).

Setting. Medium sized hospital trust in Northern Norway.

Participants. 120 records selected as bi-weekly samples of ten from discharge lists between July 1st and December 31st 2010 for a threefold review.

Main Outcome Measure(s). Inter-rater reliability assessed with the Cohen Kappa coefficient between different teams regarding presence and severity level of adverse events.

Results. Substantial inter-rater reliability regarding presence and severity level of adverse events was obtained between Team I and Team II while moderate inter-rater reliability was obtained between Team I and Team III.

Conclusions. Replacement of reviewers did not influence the results provided that one of the non-physician reviewers remains consistent. The experience of the consistent reviewer can result in continued consistency in interpretation with the new reviewer through discussion of events. These findings could encourage more hospital to rotate reviewers in order to optimize resources when using the GTT.

Keywords: inter-rater reliability, Global trigger tool, adverse events, quality measurement, incident reporting and analysis, medical errors, drug errors

INTRODUCTION

Identifying and measuring adverse events is challenging both in terms of which method to use and how to ensure valid results. Record reviews have identified a prevalence of adverse events in 9-16 % of hospitalized patients in the Nordic countries[1,2]. The Institute for Healthcare Improvement (IHI) Global Trigger Tool is a method for retrospective review of continuous random samples of inpatient records to identify adverse events that is widely used and has demonstrated a high sensitivity and specificity in identifying adverse events compared to other commonly used methods such as voluntary incident reporting or safety indicators from administrative data[3–7]. The method involves a two-step review process where two non-physician, clinical reviewers independently review the records for pre-defined triggers that could indicate that an adverse event has occurred. These reviewers determine whether an adverse event is indeed present, and if so, categorize the severity level. A physician authenticates the consensus of the findings by the non-physician reviewers and may change or overturn the determinations based on assessment of documentation in the record.

The agreement between reviewers and between different teams as measured by inter-rater reliability has been reported from fair to substantial[8,9]. The Global Trigger Tool procedure recommends that the review team of three reviewers should be kept consistent as much as possible to ensure consistency of interpretations and high inter-rater reliability[3]. However, replacement of reviewers does occur in clinical work environments due to various reasons, such as medical leave or job changes, and can result in replacement of one, two or all reviewers. In addition to these practical reasons to replace reviewers, the resources necessary for review could also lead to frequent replacement of reviewers.

Thus it is necessary to assess whether replacement of one or two of the reviewers affects the level of agreement as much as replacement of all three does. To our knowledge no studies have evaluated the agreement when one of the non-physician reviewers is kept consistent while the rest of the reviewers are replaced. The aim of this study is to evaluate the agreement of teams with varying replacement of reviewers regarding presence and severity of identified adverse with the Global Trigger Tool.

METHOD

Setting:

The study was carried out at Nordland Hospital trust, a 524-bed trust with hospitals in three different geographic sites in Northern Norway. The hospitals had a total of 7087 discharges fulfilling the study's inclusion criteria's with 43750 patient days in the period from July to December 2010. A total of 120 inpatient records were obtained by selecting ten records randomly from the hospital discharge lists bi-weekly for the period of July 1th to December 31th, 2010. Due to resources available we found that 120 records selected from a six month time period were sufficient to obtain valid results. Others who have assessed inter-rater reliability have included both lower and higher number of cases [4,9] Patients excluded from the samples were as per the IHI method: length of stay less than 24 hours (to avoid any patients for observation) and less than 18 years of age, or admitted to psychiatric and/or rehab units as the

triggers in the tool were designed for adult, medical-surgical, acute care only patients. The study was approved by the Data protection official in Nordland Hospital trust and by the Norwegian Regional Ethics Committee (ref 2012/1691).

Review process:

The record review method described in the Global Trigger Tool[3] was applied with the adapted 57 triggers in the Norwegian translation (See Appendix 1)[10] using a two-stage review process. In the first stage, the two non-physician reviewers (nurses) reviewed the records independently to identify triggers that could represent possible adverse events for a maximum of 20 minutes per record. Examples of triggers included a given procedure, a lab result or a medication administration. After the independent review a consensus was reached for each record as to the adverse events identified and the severity level for each. In the second stage, the consensus findings were authenticated by the physician. The physician did not systematically review the entire record, just the sections with documentation indicating or supporting the presence of the suspected adverse event.

The definition of an adverse event used by IHI[3]: *"unintended physical injury resulting from or contributed to by medical care that requires additional monitoring, treatment or hospitalization, or that results in death"* was applied. Preventability of the adverse events was not assessed. The severity levels were adapted by IHI from the National Coordinating Council for Medication Error Reporting and Prevention index (NCC MERP)[11] and applied in the study with five severity levels:

E: Temporary harm requiring intervention

F: Temporary harm requiring initial or prolonged hospitalization

G: Permanent harm

H: Intervention required to sustain life

I: Harm contributing to death

Selection and training of reviewers:

Five non-physician reviewers (A-E) and three physician reviewers (1-3) participated in the study. All reviewers had received the same training in the Global Trigger Tool method. The training included theory, identical practical review exercises and debrief sessions as recommended by IHI[3]. The training period was performed before the reviewers were included in the study as all reviewers were reviewers on a regular basis and internal to the trust. They were experienced with the Global Trigger Tool method, having previously used the Global Trigger Tool for at least 2 years. No additional training was done just prior to study start or during the study period. All reviewers were instructed in the study design, ensuring similar reviewing procedures among the reviewers. The areas of clinical practice and years of experience for the reviewers are shown in table 1. The mean number of experience of Team I was 18 years, Team II 17 years and Team III 21 years.

Study design:

The records were reviewed using the hospitals electronic patient journal system in sets of ten records from each bi-weekly period. To account for the replacement of reviewers that occur in a clinical work environment three different review teams were assembled; Team I (three consistent reviewers), Team II (one of the non-physician reviewers or/and the physician from Team I are replaced for different review periods) and Team III (three consistent reviewers different from reviewers in Team I and Team II) to evaluate the agreement of teams regarding presence and severity level of adverse events identified by the Global Trigger Tool method .

Statistical analysis

To describe characteristics of the records descriptive statistics were used presented as frequencies, means, medians and ranges. The level of agreement between Team I and Team II and between Team I and Team III in terms of inter-rater reliability was assessed using kappa statistic for nominal data (agreement on presence or absence

of adverse events) and weighted kappa for ordinal data (number of adverse events and severity levels). The following interpretations from Landis and Koch was used for the Cohen Kappa 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) [12]. All analysis was performed using SPSS (version 22.0; including extension of weighted kappa, SPSS Chicago, IL).

RESULTS

Demographic characteristics

Of the 120 reviewed records 49 (41 %) of the patients were men and the mean age was 61.6 years (SD 20.7, range 19-102). Total number of patient days analyzed was 761, corresponding to a mean length of stay of 6.3 days (SD 7.2, range 2-64). 3037 (43 %) of the patients in the overall population from where the records were selected were men, mean length of stay was 6.2 days (SD 6.4, range 2-113) and mean age was 61.9 years (SD 20.7, range 18-102).

Adverse events identified

Altogether the teams identified 34 unique adverse events (figure 1). Team I identified a total of 23 adverse events corresponding to a rate of 30.2 adverse events per 1000 patient days. Team II identified 20 adverse events for a rate of 26.3 adverse events per 1000 patient days and Team III identified 18 adverse events corresponding to a rate of 23.7 adverse events per 1000 patient days. The level of severity assigned by each team in each cases of adverse events identified is included in table 2. In table 3 the agreement and disagreement according to Team I’s findings are listed. There was disagreement in 4 records between Team I and Team II and in seven records between Team I and Team III. Three of five records with pneumonia identified by Team I were missed by Team II as well as two records with surgical complications. Team III missed six of six records identified with a medication event by Team I as well as three records identified with pneumonia by Team I.

Inter-rater reliability

Agreements were substantial on presences of adverse events between Team I and Team II and moderate between Team I and Team III (table 4). The agreement in terms of number of adverse events and severity levels was substantial between Team I and Team II and moderate between Team I and Team III.

DISCUSSION

To our knowledge this is the first attempt to assess inter-rater reliability between review teams experiencing replacement of reviewers in varying degrees. We found that if one of the non-physician reviewers was consistent while one or both of the other reviewers were changed (Team I vs Team II), the agreement in terms of presence of adverse events and severity levels was substantial compared to moderate agreement when all reviewers were different (Team I vs Team III). This indicates that the level of agreement between two teams with completely different reviewers is lower than between teams where at least one of the reviewers remains consistent. The results in our study indicate that keeping at least one of the non-physician reviewers consistent when other reviewers must be changed is better than changing all reviewers. In this way the interpretation of adverse events will be more consistent over time than if all reviewers are replaced[9]. Rotation of non-physician reviewers was used in one study and the level of agreement did not change, which is in accordance with our results[8].

This study has some potential limitations. First, the study was performed without giving the reviewers additional training before or under the study. Others have also conducted studies without further training[9]. In our setting we did not consider this as relevant as we assumed that using regular reviewers ensured a similar level of experience. However, all reviewers were instructed in the study design ensuring that the record reviews were conducted in similar fashion. Second, we did not replace both non-physicians from Team I in Team II in neither of the bi-weekly review periods. We assume that some continuity is needed to ensure that the non-physician reviewers represent some consistency as they perform the primary reviews. Third

since the definition of the types of adverse events depend on a subjectively assignment we chose not to include the level of agreement of the types of adverse events. We therefore only evaluated the level of agreement of the presence of an adverse event and its severity level. .

As this is a methodological study of the record review method described by the IHI, the results is generalizability to other users of the IHI Global Trigger Tool. The results are in accordance to other studies regarding the rate of adverse events and severity assigned. However, these results would not be applicable in settings other than adult, acute care hospitals.

Conclusion

We found substantial agreement in terms of adverse events and their severity level when at least one of the non-physician reviewers was consistent while other reviewers in the team were replaced. This is in contrast to only moderate agreement between two teams with all different reviewers. Our findings indicate that hospitals can rely on rotating reviewers to optimize resources. Hospitals are encouraged to perform record review even with frequent replacement of reviewers as this can be done without the risk of biasing the results as long as one reviewer remains consistent.

Acknowledgments: The authors wants to thank Birger Hveding, Inger Lise Øvre, Berit Enoksen, Unn Mari Dahle, Ida Bakke, Anita F. Jensen and Kåre Nordland who conducted the reviews together with the co-author TEH. Frank Federico and Carol Haraden for the development of the study design and Alexander Ringdal for help with data organising.

Funding: This work was supported by the North Norwegian health Authority as a PhD grant to the author K.M.

References

1 Doupi P, Svaar H, Bjørn B, *et al.* Use of the Global Trigger Tool in patient safety improvement efforts: Nordic experiences. *Cogn Technol Work* 2014;**17**:45–54. doi:10.1007/s10111-014-0302-2

2 Deilkås E, Bukholm G, Lindstrøm J, *et al.* Monitoring adverse events in Norwegian hospitals from 2010 to 2013. *BMJ Open* Published Online First: 2015.http://bmjopen.bmj.com/content/5/12/e008576.short (accessed 4 Mar2016).

3 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 Nov2014).

4 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. doi:10.1377/hlthaff.2011.0190

5 Naessens JM, Campbell CR, Huddleston JM, *et al.* A comparison of hospital adverse events identified by three widely used detection methods. *Int J Qual Health Care* 2009;**21**:301–7. doi:10.1093/intqhc/mzp027

6 Maass C, Kuske S, Lessing C, *et al.* Are administrative data valid when measuring patient safety in hospitals? A comparison of data collection methods using a chart review and administrative data. *Int J Qual Health Care* 2015;**27**:305–13. doi:10.1093/intqhc/mzv045

7 Najjar S, Hamdan M, Euwema MC, *et al.* The Global Trigger Tool shows that one out of seven patients suffers harm in Palestinian hospitals: challenges for launching a strategic safety plan. *Int J Qual Health Care* 2013;**25**:640–7. doi:10.1093/intqhc/mzt066

8 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. doi:10.1093/intqhc/mzq026

9 Schildmeijer K, Nilsson L, Arestedt K, *et al.* Assessment of adverse events in medical care: lack of consistency between experienced teams using the global trigger tool. BMJ Qual. Saf.

2012;**21**:307–14. doi:10.1136/bmjqs-2011-000279

10 Strukturert journalundersøkelse, ved bruk av Global Trigger Tool for å identifisere og måle forekomst av skader i helsetjenesten. Oslo: 2010.

11 Hartwig SC, Denger SD, Schneider PJ. Severity-indexed, incident report-based medication error-reporting program. *Am J Hosp Pharm* 1991;**48**:2611–6.http://www.nccmerp.org/types-medication-errors

12 Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;**33**:159–74.http://www.ncbi.nlm.nih.gov/pubmed/843571 (accessed 20 Jul2014).

Table 1 Area of clinical practice of the reviewers and years of clinical experience

	Reviewers	**Area of clinical practice**	**Years of clinical experience**
Primary reviewers (nurses)	A	Cardiac Intensive care	25
	B	Neurology	22
	C	Neurology	15
	D	Anesthesiology	29
	E	Orthopedics	28
Secondary reviewers (physicians)	1	Neurology	7
	2	Surgery	13
	3	Pediatrics	7

Table 2 Severity level of each adverse events identified by the teams respectively

Severity category	Team I	Team II	Team III
E	11	10	10
F	12	10	7
G			
H			1
I			
Total	23	20	18

E: Temporary harm requiring intervention, F: Temporary harm requiring initial or prolonged hospitalization, G: Permanent harm, H: Intervention required to sustain life, I: Harm contributing to death

Table 3 Agreement and disagreement to Team I's identified adverse events

Team I	Agreement	Disagreement
Pressure ulcer	Team II	Team III (postoperative bleeding)
Other infection		
Pneumonia		
Fracture		Team III (postoperative bleeding) Team II (medication event)
Medication event	Team II	
Pneumonia	Team II	Team III (urinary tract infection)
Medication event	Team II	
Pneumonia	Team II Team III	

Other infection		
Other surgical complication		Team III (other infection)
Reoperation	Team II	Team III (postoperative infection)
Medication event	Team II	
Urinary tract infection	Team III	Team II (patient fall)
Reoperation		Team II (urinary tract infection) Team III (urinary tract infection)
Medication event	Team II	
Other surgical complication		
Patient fall		
Postoperative bleeding	Team II	Team III (fracture)
Medication event	Team II	Team III (pneumonia)
Medication event		Team II (deterioration of chronic disease)
Pressure ulcer	Team II	
Pneumonia	Team III	
Pneumonia		

Table 4 The level of agreement between Team I and Team II and between Team I and Team III in terms of adverse events and severity level

	Team I vs Team II (kappa coefficient, 95 % CI)	**Team I vs Team III (kappa coefficient, 95 % CI)**
Presence of adverse events*	0.640 (0.434-0.846)	0.468 (0.232-0.703)
Number of adverse events**	0.661 (0.479-0.842)	0.468 (0.278-0.694)
Severity level**	0.652 (0.469-0.836)	0.442 (0.260-0.624)

*Unweighted kappa analysis, **Weighted kappa analysis

Figure legends:

Figure 1 Venn diagram of number of adverse events identified by Team I, II and III