Faculty of Health Science

Department of Clinical Medicine

Identifying and measuring patient harms

A study of measuring adverse events in hospitalised patients by the Global Trigger Tool record review method

—

Kjersti Mevik

A dissertation for the degree of Philosophiae Doctor – February 2019

For all future patients

Cover: Illustration by Elin Karlsnes. Reprinted from The Journal of the Norwegian medical Association issue no.23, 9 December 2014 with kind permission from the Journal of the Norwegian Medical Association.

CONTENT

LIST OF TABLES

LIST OF FIGURES

Acknowledgements

First, my greatest gratitude goes to my main supervisor professor *Barthold Vonen*. I appreciate that you believed in me when I knocked on your door in 2003 as a young medical student looking for some academic work for the summer. Our first project was on fecal incontinence and you showed me how to fight for and help patients that might not be prioritized in our healthcare system. I am grateful that you involved me in your department and introduced me to other research fellows at the Department of Digestive Surgery in Tromsø. In Bodø, once again you involved me in your work, where you introduced me and the rest of Nordland Hospital Trust to the concept of patient safety. The PhD travel with you has been both fun and exciting. Thank you for caring for me and my girls and for your friendly concerns of my well-being and support in life challenges.

I am grateful to my co-supervisor *Ellen Deilkås* for your knowledge in the field of patient safety field and your valuable discussions. Thank you for teaching me and (many more) the GTT method.

The contribution from my co-writers *Tonje Hansen*, *Fran Griffin* and *Alexander Ringdal* was priceless. Dear Tonje, thanks for your valuable feedbacks, conducting the record reviews in the papers and for all the good times! Dear Fran, I am so grateful that you share your knowledge concerning the record review method with me. I really appreciate your hospitality when opening your home in Jersey for me and my girls, so we could work on the research. Dear Alexander, thank you for all excellent support understanding the intricacies of excel. The databases have not been the same without your contribution! I am forever grateful that you always give me IT support from wherever you are. Thanks to the reviewers *Kåre Nordland, Unn Marit Dahle, Ida Bakke, Berit Enoksen, Anita Jensen, Inger Lise Øvre*, and *Birger Hveding* conducting the reviews in Paper I and Paper II.

I wish to express my gratitude to the people at the Institute for Healthcare Improvement in Boston for welcoming me to their institute, especially thanks to *Frank Federico* for introducing me to wonderful people who have shared their knowledge and experiences regarding record reviews with me.

The enthusiastic patient safety colleagues at the Regional Patient Safety Resource Center at Nordland Hospital: I really appreciate your support and valuable discussions creating an atmosphere for learning and improvement. I am also grateful to my co-PhD fellow *Ellinor Haukland* for our valuable discussions and creating the PhD path together.

Thanks for practical help: To *Tonje Braaten* og *Bjarne Jacobsen* for discussions regarding the study designs, to my friend *Laila Bjølgerud* for making my chaotic sketches into the most beautiful and informative illustrations, to *Elisabeth Mentzoni* who always help finding the right numbers and patient lists and to *Tom Wilsgaard* for statistical support, teaching me statistics and ensuring that my assumptions were fulfilled.

My colleagues at BRENDO in Tromsø; *Frimann, Marit, Vegard, Amund* and *Ingvild* - you enlighten the work days in Tromsø! Thank you for being open-minded to my inputs of patient safety issues and that you find my ideas valuable for our patients. I am grateful that you encourage research as a part of our daily work and teaching me surgical skills.

Thanks to the Health Authorities of North Norway for funding this project and to the Surgical Department in Bodø for giving me leave to conduct the PhD.

I am grateful to my dearest parents *Kate* and *Fritz* for always listening to me despite that you sometimes do not know what I talk and write about. I really appreciate your ongoing feedback! Thank you for reading through manuscripts, babysitting the girls and caring for Christian. You handed my faith and taught me that hard work is necessary in this world. I love you.

To my favorite supergirls: *Alva, Tuva* and *Eva*, my true love and treasures. I am so proud of you. Thank you for your unconditional love, your companionship on our travels and your curiosity. You three bring so much joy in our life! Finally, to my dearest husband *Christian*, thanks for your patience during the PhD work and my everlasting talks of the Global Trigger Tool. Thank you for literally following me around the world bringing the girls along. Thanks

for your humor and support and that you take so good care of the girls and me (and the house and the boat and the cabin). I love you and I need you.

Bodø 01.02.19

SUMMARY

Patient harms, or adverse events (AEs) which is the term used in this PhD thesis, is a major global health problem. They cause suffering for patients, are stressful for involved health personnel and costly for the healthcare services. Acknowledging that such events happen is necessary in order to improve patient safety. The Global Trigger Tool (GTT) has been used to track AEs over time in Norwegian hospitals from 2011. The method involves a review team who screens randomly selected patient records for predefined triggers (situations) that could indicate that an AE has happened. A trigger can be use of blood products, an infection, abrupt medication stop or a readmission. If one or more of such triggers are present, a more in-depth review is performed to decide if the trigger represent an AE. The GTT method has demonstrated high sensitivity in comparison to other methods, such as voluntary incident reporting, quality indicators from administrative data and claims for compensation. However, the GTT method is criticized because of the sampling strategy, low agreement between review teams and that the method is time consuming to perform.

This PhD evaluated if increasing the number of records to be reviewed (Paper I), changes of reviewers (Paper II) and automatically identification of triggers (Paper III) improved the reliability and validity of the GTT method.

The results showed that increasing the number of reviewed records seven times increased the rate of identified AEs by 45 %. The confidence interval was narrower in a large sample compared to a small sample. Review teams with at least one identical reviewer demonstrated substantial agreement compared to moderate agreement between review teams with no identical reviewers. Automatic identification of triggers saved review time and use of this tool identified equal rates of AEs comparable to the original GTT method with manual trigger identification.

In conclusion, these studies showed that if the number of reviewed records is increased, at least one reviewer is consistent and automatic trigger identification is used, the method's reliability and validity are improved and the review time reduced.

SAMMENDRAG (summary in Norwegian)

Pasientskader, eller uønskede hendelser som er begrepet brukt i denne ph.d. avhandlingen, er et betydelig globalt helseproblem. De forårsaker lidelse hos pasienter, er belastende for involvert helsepersonell og kostbare for helsevesenet. Anerkjenning av at slike hendelser skjer er nødvendig for å kunne bedre pasientsikkerheten. Metoden Global Trigger Tool (GTT) ble derfor innført ved alle norske sykehus fra 2011 med det formål å følge antall uønskede hendelser over tid. Metoden går ut på at ett granskningsteam gransker et tilfeldig utvalg av pasientopphold etter forhåndsdefinerte triggere (situasjoner) som kan indikere at en uønsket hendelse kan ha skjedd. En trigger kan være bruk av blodprodukter, en infeksjon, plutselig seponering av ett medikament eller en reinnleggelse. Hvis en eller flere slike triggere er tilstede, gjøres en mer grundig gjennomgang for å finne ut om triggeren er assosiert med en uønsket hendelse. GTT metoden har høy sensitivitet i forhold til andre metoder som avviksmeldinger, kvalitetsindikatorer basert på administrative data og klagesaker. Imidlertid er GTT metoden kritisert fordi den baseres på granskning av små utvalg av pasientopphold, har dårlig samsvar mellom forskjellige granskningsteam og at metoden er tidskrevende å gjennomføre.

Denne doktorgradsavhandlingen evaluerte om økning av antall pasientopphold som granskes (Artikkel I), utskifting av granskere (Artikkel II) og automatisk identifisering av triggere (Artikkel III) bedret metodens reliabilitet (pålitelighet) og validitet (gyldighet).

Resultatene viste at ved å øke utvalget av granskede pasientopphold sju ganger, økte raten av antall identifiserte uønskede hendelser med 45 %. Konfidensintervallet var smalere i et stort utvalg sammenlignet med ett lite utvalg. Granskingsteam som hadde minst en lik gransker viste godt samsvar sammenlignet med team som ikke hadde noen like granskere. Automatisk identifisering av triggere sparer granskningstid, og bruk av dette verktøyet identifiserte samme rate av uønskede hendelser som ved bruk av den original GTT metoden med manuell trigger identifisering. Oppsummert viser studien at hvis man gransker større utvalg av pasientopphold, beholder minst en gransker stabil i granskingsteamet og bruker automatisk identifisering av triggere, vil metodens reliabilitet og validitet forbedres og tidsbruken reduseres.

LISTS OF PAPERS

This thesis is based upon three publications, referenced in the text by their respective roman numerals:

I. **Mevik K**, Griffin F, Hansen T, 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.
BMJ open, 2016, 6.4: e010700.

II. **Mevik K**, Griffin FA, Hansen TE, et al.
Is inter-rater reliability of Global Trigger Tool results altered when members of the review team are replaced?
International Journal for Quality in Health Care 28.4 2016: 492-496

III. **Mevik K**, Hansen TE, Deilkås EC, et al.
Is a modified Global Trigger Tool method using automatic trigger identification valid when measuring adverse events? A comparison of review methods using automatic and manual trigger identification.
International Journal for Quality in Health Care, 2018.

ABBREVIATIONS

GTT	Global Trigger Tool
AE	Adverse events
IHI	Institute for Healthcare Improvement
COSMIN	Consensus-based Standards for the selection of health status Measurement Instruments
NPE	Norsk pasientskadeerstatning
WHO	World Health Organization
HMPS	Harvard Medical Practice Study
SPC	Statistical Process Control
EHR	Electronic health records
INR	International normalized ratio
QI	Quality indicator
PSI	Patient safety indicator
PROM	Patient reported outcome measure
CI	Confidence interval
SE	Standard error
RR	Risk ratio

"To err is human; to cover up is unforgivable; and to fail to learn is inexcusable."

Sir Liam Donaldson at the launch of the World Alliance for Patient Safety Oct 2004

1 INTRODUCTION

1.1 Background

Patient harms, or adverse events due to medical care, is a major global health problem as they cause suffering for patients and are stressful for involved healthcare professionals [1]. In addition they are costly for the healthcare services [2]. Acknowledging that such events happen and measuring them, are necessary for improving health care and increasing patient safety [3].

The common methods (i.e.; incident reporting, quality indicators, processes for dealing with complaints and mortality & morbidity conferences) of reporting and analysing adverse events are unfortunately inappropriate for measuring adverse events mostly due to reporting bias [3]. These systems depend on either the patients, their relatives or health personnel voluntary reporting the adverse events.

Review of patient records for specific triggers (situations) such as use of blood products, abrupt stop in medication or readmissions, is an alternative method to identify and measure adverse events. Such method has demonstrated high sensitivity in comparison to the referred methods above [4]. The widely used method for identifying and measuring adverse events is the Global Trigger Tool (GTT), developed by the Institute of Healthcare Improvement (IHI) in Cambridge, USA [5]. Frequent use of the GTT method has demonstrated that adverse events are far more common than first assumed [6], [7]. Estimates show that adverse events happen as frequent as up to 30 % of the inpatient population [6].

However, the GTT has some practical disadvantages. It is rather resource intensive due to time and personnel required. The sampling approach, reviewing only small samples of records, together with frequent replacement of reviewers question the reliability and validity of the method [8], [9]. This thesis examined the effect on the results of identified adverse

events by increasing the number of reviewed records and changing the reviewers. Use of automatic identification of triggers was also evaluated. As the GTT is used in all Norwegian hospitals the aim of the thesis was to make the GTT method a more efficient, valid and reliable strategy to identify and measure adverse events in hospitalised patients.

1.2 Adverse events

1.2.1 Definitions

Several different terms describing adverse outcomes of medical care are used (table 1). Inconsistent use of terms, which appear both in the literature and in the clinical settings, complicates the understanding of adverse outcomes due to medical care [10].

Table 1 Terms describing adverse outcomes

Term	Definition	Pros	Cons
Errors	a failure to carry out a planned action as intended or application of an incorrect plan [11]	Identify failures	Promotes blaming Inhibit system approach
Injuries	damage to tissues caused by an agent or event [11]		Only severe events
Patient harms	an outcome that negatively affects a patient's health and/or quality of life [12]	Already in use	Used differently whatever considered a severe event, a claim or adverse outcomes
Adverse events	unintended physical injury resulting from or contributed to by medical care that requires additional monitoring, treatment or hospitalisation, or that results in death [5]	System approach Promotes a no blame culture Promotes interventions to reduce them	New term
Complications	an unfavourable evolution or consequence of a disease, a health condition or a therapy [13]	Already in use	Acceptance of the incidence of the events
Healthcare-associated harm	harm arising from or associated with plans or actions taken during the provision of healthcare, rather than from an underlying disease or injury [10]	No doubt that the harm is due to the healthcare given	Too complicated for everyday use

Identification and measurement of adverse outcomes from medical care depend on a common definition of what constitutes this term, in order to increase the understanding of such events [14]. Consistent use of patient safety terms is also necessary for making comparison between facilities possible and to track the trends over time [10]. A group, initiated from the World Health Organisation (WHO), agreed upon 48 concepts aiming for that this agreement could pave the way for a common understanding of the concepts of patient safety [10]. Common definitions would probably increase the focus on these events promoting implementation of interventions to prevent them. However, deciding the contribution of medical intervention in regard to the underlying disease to an event, is often difficult. For example; an unplanned unit of blood is infused to an anaemic patient after an operation. It is not always obvious if the anaemia is due to the medical condition or due to the operation. The type of medical condition is important to consider when deciding if the event was due to the condition. A definition including criteria for defining it as an adverse outcome due to medical care, would make it easier to decide. A discussion concerning when to use and not to use the different terms follow, as well as their suitability as measures of adverse outcomes.

Using the term *error* for the adverse outcome often brings up the question of whom is to blame. The blame perspective makes the culture for analysis the event difficult. A *"*just" culture promotes a system approach, rather than blaming and shaming on individuals [3], [15], [16]. Most errors are committed by good hardworking people and identifying who's to blame is a distraction. It is far more productive to identify the situations that caused the error and implementing systems that will prevent them from happening again [17]. However, the fact that all errors do not result in adverse outcomes and all adverse outcomes are not necessary a result of errors, makes measuring errors not suitable as a measure [18].

The terms *injury* or *harm* do not distinguish between injuries as adverse outcomes due to medical care or due to injuries caused by the patients' disease or by an accident. In the clinical setting the term *patient harm* has traditionally been used when a patient suffers a harm due to a severe and highly unexpected event caused by the medical care given. This unresolved understanding of the term *patient harm* was not considered when the Norwegian Patient Safety Campaign (later defined as program) "In safe hands" ("I trygge hender") was launched

in 2011. They chose to use the term *patient harm* (pasientskade) for all events when implementing the GTT to measure adverse outcomes due to medical care [19]. The manual of the original GTT define such events as adverse events and do not use the term patient harm. "Patient harm" used in the Norwegian campaign included both minor events, such as catheter based urinary infections, and more severe events, such as injury to the ureter during a laparotomy. This "new" use of the term *patient harm* was not immediately adapted by the clinical health personnel in Norwegian hospitals as they have reserved this term for the severe events and events that could qualify for compensation through the Norwegian System of Patient Injury Compensation (NPE) [20]. According to the Act on Patient Injury Compensation [21] three criteria must be fulfilled before a claim for compensation is accepted. It must have been a failure in treatment (with some exceptions), economic loss of more than 10000 NOK and the injury could not be more than three years old when applying. The patient harms measured by the GTT method is mostly less severe than the events traditionally defined as patient harm by the clinical health personnel.

The term *complication* does neither distinguish between events caused by the patients' underlying disease, or by medical care. However, complications are often agreed as foreseeable unintended events due to medical care. If an event is considered foreseeable it is often a silent acceptance that they happen from time to time. Accepting that such events happen could act as an obstacle to identify, measure and prevent them. The Norwegian Patient Safety Program wanted to include events that were defined as complications as well as events that were previously not considered a patient harm (i.e.: urinary tract infection due to catheter), addressing all these events as patient harms.

The original GTT defined the adverse outcomes due to medical care as *adverse events* (uønskede hendelser) with the definition described in table 1. As described previously, unplanned and unintended events have traditionally been defined as complications, if acknowledged at all by the clinical health personnel. The authors of the GTT focused on the events that harm the patients rather than errors that easily promote a perspective of whom to blame. *Adverse events* has been used in the literature for decades, but first used in relation to patient harms in the mid 80's [22].

In this thesis we will investigate how the GTT's ability to identify and measure adverse outcomes could be improved. We therefore decided to use the term *adverse event* in this thesis. We argue that this term includes most of the relevant events due to medical care; whether considered a complication, a preventable event, an error or a failure of systems.

1.2.2 Identification

Table 2 shows the different systems that are used for reporting or measuring adverse events in hospitals [23]. These are unlike the methods that are used for dealing with adverse events such as root cause analysis, mortality & morbidity conferences, malpractice claims and compensation systems which all are inappropriate to use as measurement methods due to reporting bias. Also, selection bias, confounding bias, information bias or hindsight bias could influence the reporting of adverse events in the different measurement methods referred to in table 2. Selection bias could occur when patients are seemingly selected non-randomly, but for whatever reason still selected due to a specific variable such as their age, sex, department admitted to or selected because of the adverse event. Confounding bias can occur if an alternative explanation of the adverse event (which is not accounted for) is present, such as age. For example, if age is not adjusted for, the adverse event rates could be explained by that the selected patients are mainly above a certain age. If there is an error concerning the measurement method, it is defined as information bias. This could be present if there is something wrong with the measurement method. Hindsight bias could be due to that the outcome is known for the reviewer when determining if adverse events are present.

Voluntary incident reports and patient reported outcome measures, rely on the commitment of health personnel and patients to report adverse events. These systems are therefore subject to reporting bias. The patient voice is an emerging part in the patient safety field, but there is so far no tradition to include patient reports in the measurement of adverse events [24]. Patients mostly identify problems related to doctor-patient-relationship (lack of respect, time pressure, rudeness, break of confidence), coordination, access (long waiting time, no appointments available) and communication (between doctor and patient, among health care professionals) [24], [25]. Medical record review is the method with highest correlation with patient reported events, in contrast to incident reporting by staff with no or low concordance with patient reported events [26]–[29]. The few studies performed suggest that patient reported outcomes can be included in the hospitals measurement of adverse events, but the risk of both overestimating and underestimating due to inconsistent use of terms must be accounted for [28].

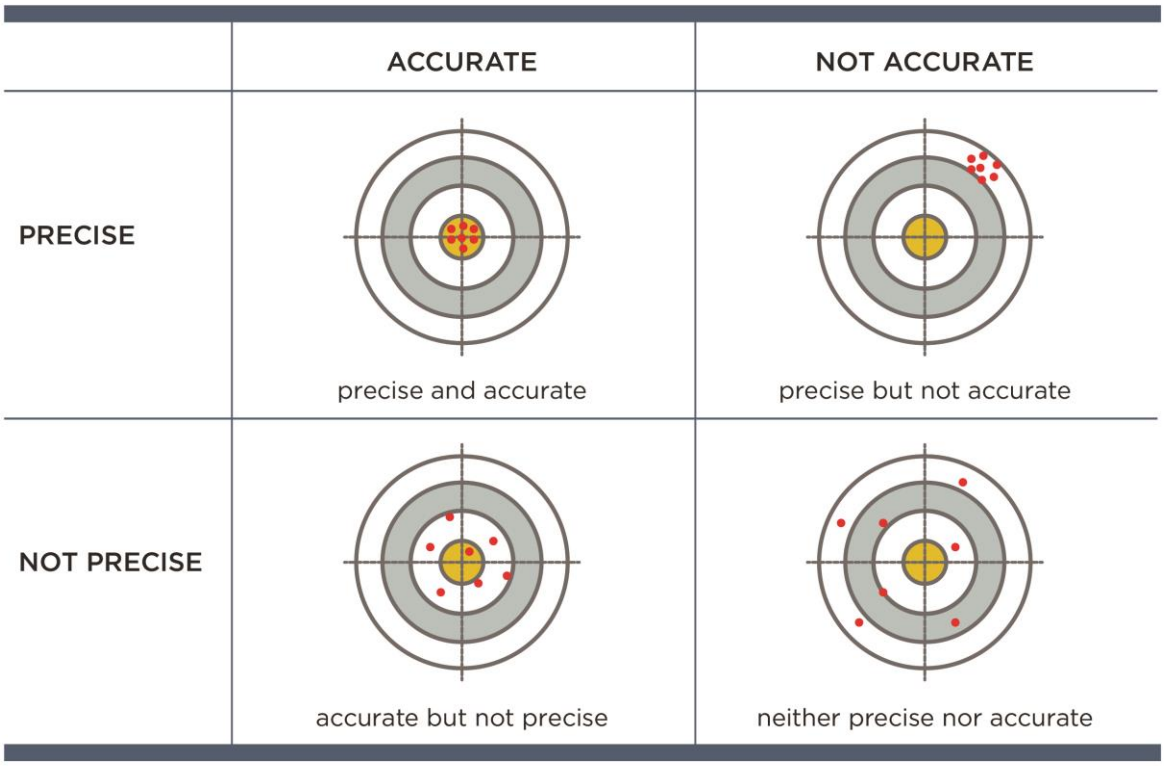
Table 2 Strengths and limitations of common methods to identify adverse events

Methods	Strengths	Limitations
Administrative data (e.g.: ICD 10 codes)	Few resources required Inexpensive Utilize readily available data	Low sensitivity- many false positives Requires correct diagnosis, procedures
Quality indicators (QI) (e.g.: readmission after 30 days)	No clinical resources needed for computerized systems Objective measure Inexpensive to run when first developed	Low sensitivity- many false positives Requires correct documentation of the data
Patient safety indicators (PSIs) (e.g.: decubitus ulcer)	Do not rely on clinical judgment Identifies adverse events directly Comprehensive Screening tool Inexpensive to run when first developed	Requires technology development Depend on the accuracy of the ICD-10 coding Some indicators are just indicators of adverse events, and not just an adverse event by itself Narrow range of adverse events Administrative data lack information about the severity
Voluntary reporting (e.g.: incident reporting)	Inexpensive Can detect latent events (near-misses)	Relies on awareness and willingness of staff to volunteer submit event notification Requires a no blame culture Reporting bias Hindsight bias
Trigger tools	Sample based	Rely on documentation in the health record
Manual (e.g.: GTT, HPMS)	Commonly used No technology development required Works in paper records	Requires extensive clinical resources Inter-rater reliability can vary Hindsight bias
Automatic (e.g.: automatic trigger identification)	Inexpensive when first developed Efficient Objective identification of triggers Integrates multiple data sources	Technology development required Manual review required of the triggered records
Full chart review	Works in paper records Commonly used Gold standard?	Incomplete medical records Judgment of adverse events are subject to reviewers decision Expensive Resource intensive Hindsight bias
Patient reported outcome measure (PROM)	Reflects the patients view of adverse events No technology development required	Inconsistent reporting routine No standard definition of an adverse events
Clinical surveillance (e.g.: EKG of all post-operative patients)	Accurate and precise Limited to specific interventions	Costly as all patient in a cohort are screened
Observation of patient care (e.g.: videotaping or observation)	Direct observation	Confidentiality concerns (punishments) Hawthorne effect (people do not act “normal” when observed) Evaluates a specific situation Resource intensive training of observers

1.2.3 Evaluations of measures

Measures should be of high precision and with high accuracy. Precision refers to if the measure consistently provides the same results if it is repeated. The accuracy refers to whether the measure measures exactly what it is supposed to measure [30]. The precision describes the difference between repeated measures of the same value and the accuracy reflects the difference between the measured and the true value (figure 1).

Figure 1 Precision and accuracy (Illustration by Laila Bjølgerud)



The confidence interval (CI) is calculated from the observed data based on the standard error (SE). The confidence level is usually set to 95 %. The accuracy regarding the CI defines if the interval contains the true population mean while the precision refers to the width of the CI. To increase accuracy the confidence level is increased which will widens the CI. But if the width of the CI increases the precision goes down.

Methodological quality in studies on measurement properties can be assessed by using the Consensus-based Standards for the selection of health status Measurement Instruments (COSMIN) checklist [31]. The checklist include the measurement properties internal consistency, reliability, measurement error, content validity, structural validity, hypotheses testing, cross-cultural validity, criterion validity, responsiveness and interpretability [31]. The measurement properties used in this thesis is further discussed.

For academic use the term *reliability* describes how reliable and precise the results from a measure are. *Reliability* refers to the consistency of a measure with the types: test-retest reliability, internal consistency and inter-rater reliability. Test-retest reliability is administering a test to a group of individuals, re-administering the same test to the same group at some later time, correlating the first set of scores with the second in a scatterplot computing Pearson's *r* [14]. Inter-rater reliability is the correlation of scores between two or more reviewers who scores the same item. This is typically measured by the Cohen's Kappa coefficient where kappa is the "true" agreement when accounting for agreement by chance [32]. This method could also be used to evaluate the agreement of repeated administration of a test performed by one rater (intra-rater reliability). Internal consistency is the correlation between different items on the same test measured by Cronbach's alpha [33].

Validity is not defined by one definition [34]. It could be explained as the degree of which a concept measure what it is supposed to measure and how valid and accurate the results from the measure are. It could be evaluated by comparing the results of the measure to the results of another measure (referred to as gold standard) [35]. Content validity evaluate if the content of an instrument is an adequate reflection of the item to be measured. If this is obtained by expert opinions as a descriptive evaluation without any statistically analysis, it is called face validity. Construct validity evaluates if the measure measures what it is supposed to measure [36]. Criterion validity is how good the measure correlates with or predicts another valid and observable variable at the same time (concurrent validity) or later (predictive validity). For example, if the adverse event urinary tract infection is related to the rate of indwelling urine catheter used [37]. Validity is also divided in internal and external validity. Internal validity refers to whether the findings relate or are caused by the phenomena under investigation. For

example, if the adverse event identified, really is caused by the intervention given in the actual admission. External validity is the extent to which the results can be generalized for other patient groups [38].

A measure needs to have high reliability and high validity, but low validity is considered more critical than low reliability. If the measure measures some other variable and not the one we think it measures or if the measure is systematically wrong, a larger sample will not help, it will rather do more harm [36]. For example, if the method used for measuring adverse events have low validity, the events measured might not be true adverse events. Low reliability could be improved by increasing the sample size.

1.2.4 Types

A brief description, prevalence and source of the main types of adverse events referred to in the literature are presented in table 3.

Table 3 Overview of the common types of adverse events

Type	Including	Incidence in hospitalised patients	Source
Infections	Healthcare associated infections, hospital acquired infections, iatrogenic infections and nosocomial infections such as • Ventilator associated pneumonia • Pneumonia • Central line associated bloodstream infections • Catheter associated urinary tract infections • Surgical site infections • Gastrointestinal illness • Blood stream infections	5 %	Surveillance Prevalence study Trigger tools Record reviews QIs Administrative data Chart review
Surgical	• Surgical site infections • Hematoma/Bleeding • Postoperative thromboembolism • Wrong site surgery • Retained foreign objects	2 %	Surveillance PSIs Chart review

	• Medical device related harms (gas/air embolism, burning, stent thrombosis)		
Obstetric/perinatal	• Foetal asphyxia • Anal sphincter tear • Infections • Shoulder dystocia • Injury of intestines or urinary tract • Uterine rupture • Thromboembolism	0.3 %	Surveillance Claims for compensation systems PSIs Chart review
Falls		20 %	Surveillance Voluntary reporting Chart review
Pressure ulcer	• Bedsores • Decubitus ulcer • Pressure sores	14 %	Surveillance PSIs Chart review
Medications	• Adverse drug event • Adverse blood infusion event • Adverse infusions events (vaccines)	20 %	Surveillance Trigger tools Chart review
Diagnostics	• Misdiagnosis • Missed diagnosis • Delayed diagnosis	Unknown	PROMs Malpractice claims/Compensation system

Infections

Infections associated with medical care has been named healthcare associated infections, hospital acquired infections, iatrogenic infections or nosocomial infections as opposed to community-acquired infections. The terms are mostly used interchangeably, but “healthcare associated infection” are recommended to use when the patient recently has been hospitalised, had haemodialysis, received intravenous chemotherapy or resided in a long-term care facility in contrast to “hospital acquired” infection where the patient received the infection diagnose within 72 hours of admittance to hospital or developed the infection within 10 days of discharge from the hospital [39]. The percentage of patients experiencing at least one healthcare associated infection is approximately 4 % in the US [40], 5.7 % in Europe and 4.9 % in Norway [41] making this one of the most common type of adverse event.

Adverse events following surgery

According to the WHO almost half of the identified adverse events (48%) are related to surgical procedures [42]. The most frequent adverse surgical events are blood loss, surgical site infections and postoperative venous thromboembolism. Surgical site infections increase mortality, length of stay, readmissions and use of health-care services [43]. Postoperative venous thromboembolism is a common adverse event, occurring in 7 % of hospitalised patients [44] and is associated with reduced survival and substantial health-care costs [45].

Wrong site surgery could be defined as surgery on the wrong person, on the wrong body part or at the wrong side of the patient body [46]. Wrong site surgery and retained foreign objects are rare but receive major attention if they occur. Risk factors are emergency operations, unusual time pressures to start or complete a procedure or the involvement of different surgeons [47].

Manufacturer-related errors, user errors and design errors of medical devices can cause adverse events such as gas emboli after laparoscopy/hysteroscopy, air embolism after infusions, stent thrombosis and burning scar after diathermic procedures [43]. In some cases it is difficult to identify these as the cause of the adverse event [48].

Obstetric and perinatal adverse events

Worldwide the maternal and infant mortality rates are high mostly due to lack of access to medical facilities and adequate medical care [43]. The rate of obstetric related adverse events has been reported to less than 1 % in developed countries [49]. However, despite their infrequencies, obstetric events are one of the ten most common cases for claims for compensation in the Norwegian System of Compensation to patients (e.g.: fetal asphyxia, anal sphincter tear, infections, shoulder dystocia, injury of intestines or urinary tract, uterine rupture and thromboembolism) [50].

Fall with injury and pressure ulcer

Patient fall is the most common reported adverse event in the voluntary reporting systems [43]. The overall rate of patient fall is estimated to 5-9 per 1000 patient days and 30 % of the

events lead to harms. Negative outcome of a fall frequently includes hip fractures with prolonged hospitalisation. The prevalence of pressure ulcer in hospitals is estimated from 10 % to 15 % of admitted patients and the risk factors includes immobility, friction, incontinence, cognitive impairment and poor nutritional status [42].

Adverse drug events

An adverse drug event can be caused both by drugs, blood products or fluid infusion. The adverse events related to drug treatment are one of the most common adverse events in developed countries. The adverse events relate mostly to prescribing, monitoring and administering medicines with look-alike labelling, wrong use of medication or failure to recognize drug interactions [43]. The consequences of an adverse drug event could be substantial, and it is estimated that it occurs in 1 of 16 hospitalised patients, with huge financial impacts [51]. Injections are one of the most common healthcare procedures with 16 billion injections annually in developed countries including immunizations, local anaesthetics and contraceptives. Adverse events concerning injections are mostly related to devices that could transmit infections and not to the drug itself [43].

Diagnostics challenges

Diagnostics challenges include missed diagnosis, misdiagnosis and delayed diagnosis. This is an unexplored perspective of patient safety but is rarely registered as a type of adverse events on its own. This could be due to the difficulty studying the problem and the complex causes of it [43]. Many of the claims in the Norwegian compensation system for patient harm are related to delays in diagnosis or delayed or missed follow-ups. Andreasen et al found considerable variations of experts' evaluations regarding the claims after alleged birth complications demonstrating the difficulty of studying the issues related to diagnostics challenges [52].

1.2.5 Incidence

Measuring number of patients being harmed while hospitalised was first referred by the Tort system of medical malpractice in the U.S [53], [54]. Later, the Harvard Medical Practical

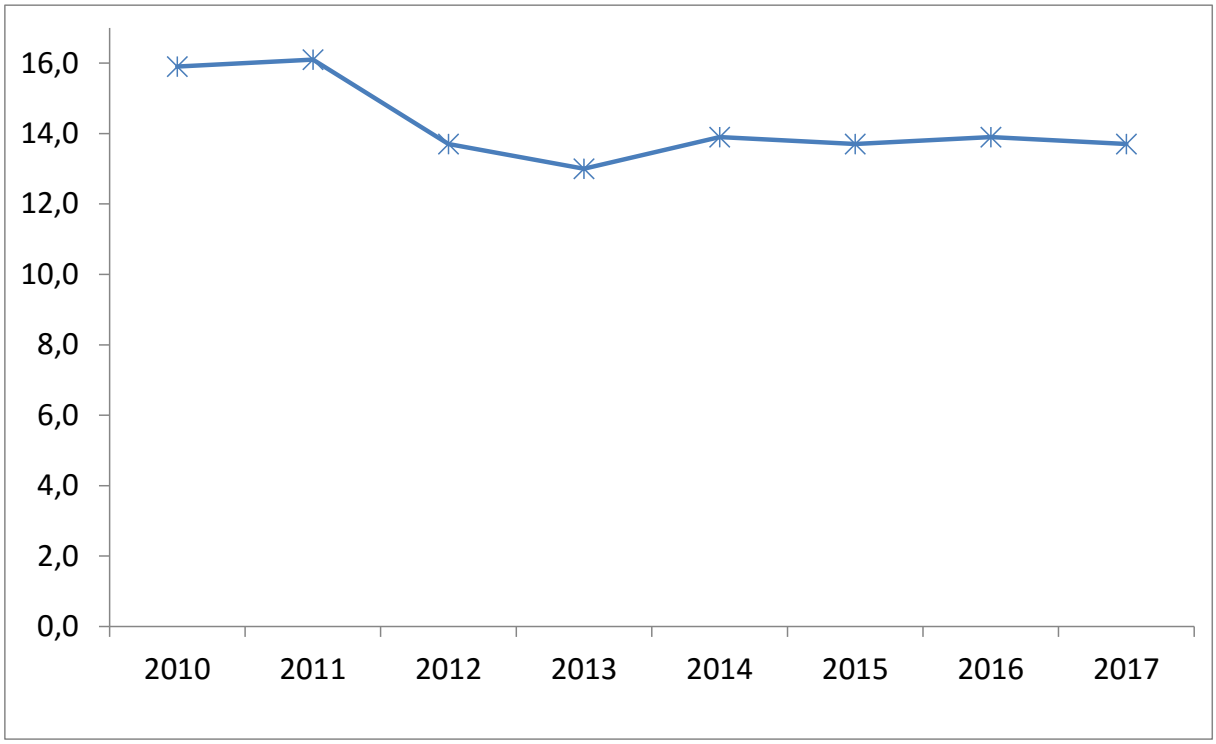
Study (HMPS) measured adverse events and negligence in hospitalised patients by reviewing patient records [22], [55]. The definition of an adverse event "as an injury that was caused by medical management (rather than the underlying disease) that prolonged the hospitalisation, produced a disability at the time of discharge, or both", was applied. They estimated that adverse events occurred in 3.7 % of the hospitalised patients. The Institute of Medicine's report "To Err is Human" brought the issue of measuring adverse events to national and international attention as they estimated that 98,000 Americans died as a results of medical errors every year [56]. This made measuring adverse events in hospitalised patients a growing focus for quality and safety in healthcare worldwide [57].

Several studies followed, demonstrating that the level of adverse events was higher than first estimated [58]–[62]. However, comparing the results between the studies were challenging as the studies applied different definitions of what they had measured [63]. Although no gold standard to identify the true level of adverse events exists, it is a common agreement that adverse events is a major global health problem [1], [63]. Valid and reliable methods that measure adverse events are demanded. The existing systems, such as the GTT, are inadequate to count the actual number of events, [38] but are used for estimating the rate of adverse events.

The WHO estimated a total of 47.7 million events when including seven different types of adverse events occurring annually in patients across the world [64]. In Norway, commissioning documents from the Ministry of Health have instructed the hospitals since 2011 to perform the GTT to measure adverse events yearly. The most common identified adverse events in Norwegian hospitals during the period 2010-2015 were hospital-acquired infections and medication related harms [65]. Interventions to reduce adverse events were initiated and implemented in the hospitals as part of the Norwegian Patient Safety Program ("I trygge hender"). In the period from 2010 to 2017 the rate of adverse events have slowly decreased from 16 % to 14 % of the admissions (figure 2) [65], [66]. This rate is below the rate of adverse events reported in international studies [6]. The reduction of the rate in Norway could reflect a true reduction of rate, or it could be due to random variability. Even though the total rate remains unchanged, the rate of the different types of adverse events could

have changed [67]. Many have argued that the rate of adverse events is still persistently high, despite the many different interventions implemented to reduce the rate of adverse events [67], [68].

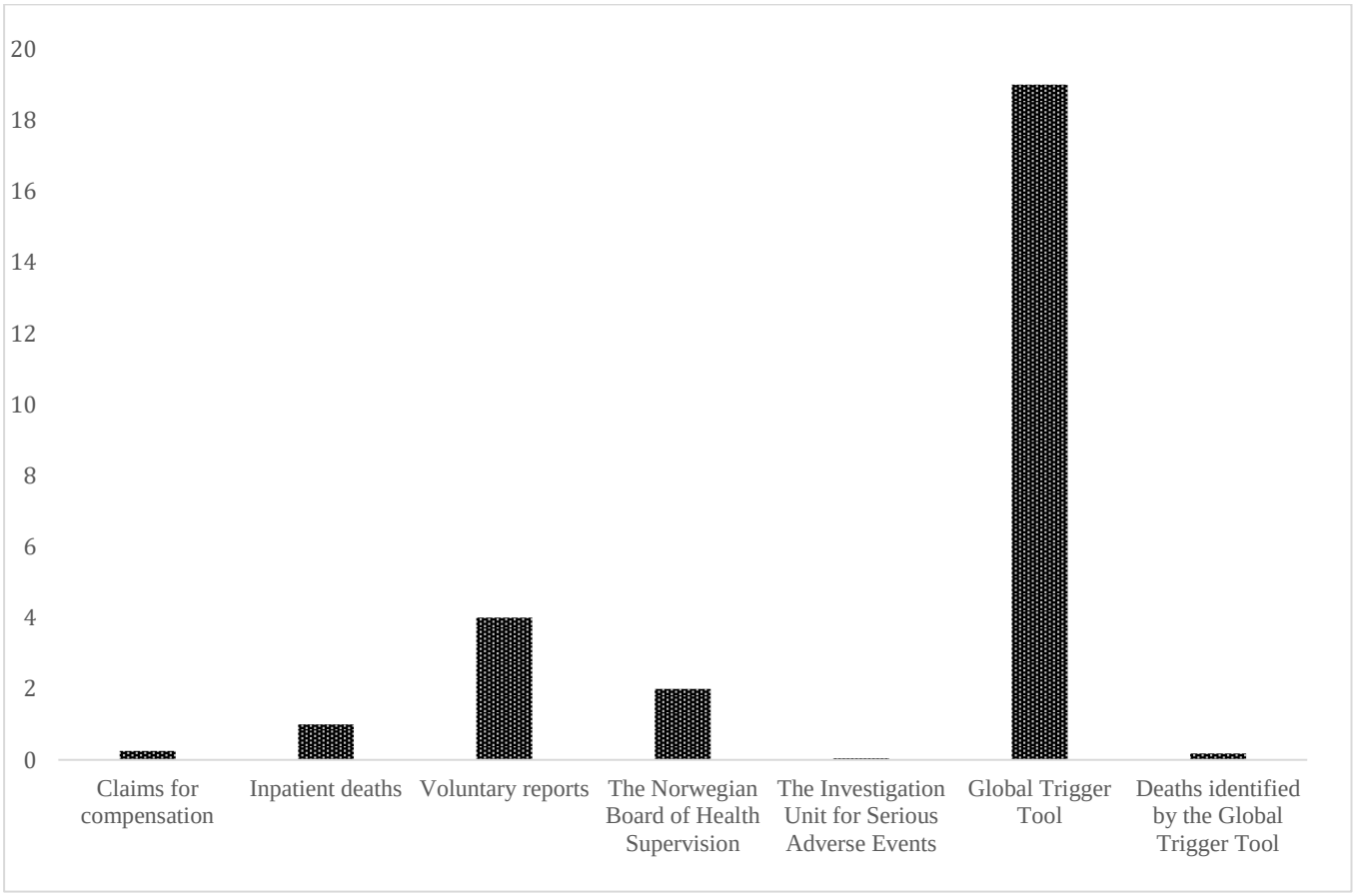
Figure 2 Percent of admissions with adverse events in Norwegian hospitals measured by the GTT



As described previously, the results from the systems for dealing with and reporting adverse events can only estimate the number of adverse events. However, when reporting systems are used for estimating how many patients who are harmed, the results of this are often misleading. The Norwegian claims for compensation system are based on patients' own claims, voluntary reports rely on health personnel to report, and severe events are investigated by the health supervision only if someone report the events. To illustrate how many events the different systems handle, we compared the reported adverse events per 100 admissions between the existing systems in our trust for 2013. Unfortunately, patient reported outcome measures are not included. The results are illustrated in figure 3. The data were collected from

the NPE, the trust's system for voluntary reporting of adverse events, the Norwegian Board of Health supervision and the GTT results. The GTT identified four times more adverse events than the other systems. We argue that this demonstrates that the GTT is the most appropriate system to quantify the number of adverse events. However, in most cases the events were reported only by one of the methods. Others have found similar results with no overlap of the identified events between the methods [69]. According to these findings, different methods might be used to reveal as many adverse events as possible.

Figure 3 Adverse events/reported events in 2013 per 100 admissions by the different systems in Nordland Hospital



1.3 The Global Trigger Tool (GTT)

1.3.1 Background

Identification of triggers in patient records to measure adverse events was first introduced by Jick in 1974 [70]. Classen et al developed the method further to be used for identifying adverse drug events [71]. Later, these trigger tools were introduced to measure adverse events in surgical departments, intensive care departments and children's departments [4], [71]–[74]. The trigger tools represented an alternative approach to measure adverse events [55]. The IHI developed the GTT initially for reviewing randomly selected paper patient records to identify triggers that could represent that an adverse event had occurred [5]. The GTT has successfully been advocated with the aim to monitor adverse events in adult inpatients demonstrated by widespread adoption [61], [75]–[78].

The intention of the GTT was to develop an easy-to-use approach for the hospitals to identify and measure adverse events [5]. The results were not intended for benchmarking between hospitals as they have different demographic background of the patients, they treat different conditions, the number of inpatients differ, and the functions of the hospitals differ. These issues make comparing GTT results between different hospitals challenging. The developers of the GTT argued that the results should be used within the hospital to acknowledge the rate and severity of the adverse events. Once the adverse events are identified, interventions that can prevent them from happening should be implemented. The effect of the interventions can be evaluated by the use of the GTT following the rate trends over time [5] .

Reviewing all inpatient records manually is impossible except in very small hospitals, hence the sample strategy. To obtain consistent results regarding the rate of adverse events, the sampling methodology needs to be truly random as the numbers of records selected must be identical in every sampling period from the same discharge lists. The recommended sample size in the GTT is ten closed inpatient records for every bi-weekly period. The patients eligible for selection must be 18 years or older, admitted for more than 24 hours and not be admitted for rehabilitation or psychiatric care since the triggers are not developed for these

areas of care. The triggers in the GTT are neither developed for children and teenagers or for outpatients.

1.3.2 Implementation

The GTT is a two-step method with manual retrospective review of records: Two primary reviewers individually review the records for 53 specific triggers (see appendices) and determine if the triggers represent any adverse events, before reaching consensus (step 1). A secondary reviewer, a physician, authenticates their findings (step 2) [5]. The two primary reviewers, either nurses or other health personnel with clinical background, review the records independently in a predefined order; discharge codes (particularly infections, complications, or certain diagnoses), discharge summary, medications administration record, laboratory results, prescriber orders, operative record, nursing notes, physician progress note and last if time permits; history, consult notes and emergency department notes. The reviewers look for any of the triggers and possible concurrent adverse events within a maximum 20-minute review time limit per record. The intention of reviewing for triggers is that this provides a more efficient and focused review of the records to identify adverse events instead of reviewing the records in their entirety. This approach help select the records in the sample that are more likely to have documented an adverse event. The triggers are classified according to the care that is provided in addition a medication module:

- General Patient care
- Surgical care
- Perinatal care
- Intensive care
- Care given in the emergency department

If a trigger is identified, the reviewer checks the relevant documentation to determine if the trigger is related to an adverse event according to the GTT definition: “unintended physical injury resulting from or contributed to by medical care that requires additional monitoring, treatment or hospitalization, or that results in death” [5]. For example, a venous thrombosis in the leg after a hip replacement is an unintended outcome, while the permanent scar from the

surgery is an intended outcome. The former is an adverse event and the latter is not. With this approach, all unintended events presented as signs, symptoms and diseases and that requires intervention, are considered an adverse event. To help the reviewers to determine whether an event is an adverse event, the following questions should be asked [5]:

- "Would I be happy if it happened to me?"
- "Was it a natural progression of the underlying disease?"
- "Was it an intended result of care?"

If the answers are no in all three questions, it is likely an adverse event. With these questions the method focus on how the patient perceives the event and stress that the patient's perspective should be emphasized when deciding if the event is an adverse event or not.

In some cases, it can be difficult to distinguish between consequences of medical care and the natural progression of the underlying disease as referred earlier. For example, if the patient suffers from a brain tumour and is treated with an operation and the patient receives blood transfusion after the operation- is the blood transfusion a result of an adverse event (e.g. the patient experienced unexpected or excessive blood loss) or was it due to the disease? In this case there was no reaction to the transfusion, but the transfusion was not a planned event. In such cases the event could be defined as an adverse event. Another example of an adverse event is if a patient develops a urinary tract infection while or after having an indwelling urine catheter. In the last case the infection is obviously due to the use of the catheter. Determining that this is an adverse event should be straightforward. The former described case with the blood transfusion is more difficult. Hence, the determination is to some extent a matter of the subjectivity of the reviewers although the common definition and guidelines should be used.

After an adverse event is identified, the reviewer determines the severity level of the event. The grading of the severity is based on a modification from the National Coordinating Council for Medication Error Reporting and Prevention Index with categories ranging from A to I (NCC MERP) [79]. The categories A-D concern events that do not reach or cause any harm to the patient (near-misses): Category A is circumstances or events that have the capacity to cause adverse events, while category B is adverse events that do not reach the patient. Category C is adverse events that reach the patient but do not cause harm, and

category D is adverse events that reach the patients and monitoring to confirm that no harm occurred is required. The few events reported through the voluntary reporting system are often near-misses. The category A-D is not included in the GTT definition as only events that cause harm to the patient are classified as adverse events in the GTT:

- Temporary harm to the patient that required intervention (Category E)
- Temporary harm to the patient that required initial or prolonged hospitalisation (Category F)
- Permanent patient harm (Category G)
- Intervention required to sustain life (Category H)
- Patient death (Category I)

The adverse events are often classified according to their type. Classification of types it not a part of the original GTT, but included in the Norwegian translation of the GTT [19] (see appendices).

The results of the reviewed bi-weekly data are then presented in three ways:

- Adverse events per 1,000 patient days
- Adverse events per 100 admissions
- Percent of admissions with an adverse event

"Adverse events per 1,000 patient days" is the recommended measure to apply when evaluating the rate of adverse events, since this measure accounts for the different length of stay in the records. Longer length of stay is associated with adverse events [80]. The "Percent of admissions with adverse events" is more easily understood by non-clinical staff and is recommended to use when the results are shared public [5]. This measure does not include that some patients experience more than one adverse event or the variability of length of stay [5].

To visualize how the rate of adverse events change over time, continual data plotting in a run chart enables to uncover either upwards or downwards trends. The data series are plotted in a time sequence. Special cause variations are identified by looking for trends (six consecutive

jumps above or over the mean/median), shifts (eight or more point above/over the central line), patterns (pattern that reoccur) and last looking for outliers that lie far from the central line. A more advanced version is the control chart in Statistical process control (SPC) which includes the upper and lower control limits which detect special cause variation quicker and more accurate [81]. Random variations are synonym with common causes that are causes that cannot be eliminated or determined. If sample size increases, random variation decreases. The SPC is used to identify special causes, or systematic errors, that might influence the process [82].

When identifying adverse events according to the definition given in the GTT, the preventability of the events is not considered. The authors of the GTT explain that this is not included as the definition of what is preventable constantly change. Events considered unpreventable today can quickly change to preventable when new innovations are introduced. When evaluating the adverse events over time, categorization of preventable versus unpreventable adverse events will be meaningless over time [5]. In Sweden, the assessment of preventability of adverse events has been evaluated by a grading system from 1-6; where 1-3 are considered non-preventable and 4-6 are considered preventable [9]. Schildmeijer et al found great differences in the assessments of preventability and doubt the benefit of including this aspect as there are no standard of how to decide preventability. They argue, as other also have [6], that all adverse events should be considered preventable.

Also, when using the GTT to identify adverse events, events due to omission is excluded as the definition only includes events due to medical care given. For example, if the patient does not receive his antithrombotic medication when indicated, and a cardiac attack occur, this type of adverse event is not included in the GTT. Such cases are often due to missed diagnoses which is difficult to reveal as discussed in 1.2.4.

Hanskamp-Sebregts et al reviewed the literature concerning validity and reliability of the record review methods using the COSMIN checklist [83]. They evaluated the studies in regards to face validity and concurrent validity and they found no reference that the validity of the GTT were evaluated [38]. The inter-rater reliability between different review teams

have been reported moderate to substantial [38]. However, the face validity of the GTT is evaluated to some extent by Schildmeijer et al [84]. They found that the GTT was a useful method to identify adverse events.

Further discussions regarding challenges with identifying and measuring adverse events with the GTT method will be described in the next chapter.

1.3.3 Challenges

There are some issues to consider when using the GTT as a measure of adverse events. First, critics argue that the GTT is too resource intensive due to time and labour required [7], [85]–[87]. The GTT is based on a 20-minute maximum review time per record per primary reviewer which equates a maximum of six hours per reviewer per month if 10 records are reviewed bi-weekly. In addition, the time used of the authenticator is estimated to one to two hours per month [5]. Also, the method requires trained personnel to perform the review. The training is a recurring event every time a reviewer or authenticator is replaced.

Second, the results of the GTT are used to make estimates of the rate of adverse events which are based on reviewing a small sample of records. The authors of the GTT explain that if the same sampling strategy is used, the method is reliable for evaluating if the rate of adverse events is reducing or increasing [74]. The results are less accurate when a small number of records is used for estimating the rate, and make it less valid as a measure of the total number of adverse events [88]. The number of identified adverse events are used to estimate the total incidence of adverse events by extrapolation. Extrapolation is a statistical method estimating a value (e.g.: expected rate of adverse events) based on extending a known sequence of values beyond the area that is certainly known [89].

Third, identification of the individual triggers varies between reviewers as triggers based on indexed variables (i.e.; blood transfusion and dialysis) have higher agreement than triggers based on free text (i.e.; pressure ulcers, patient fall) [87]. The results are to some extent subject to the reviewers subjectivity as inter-rater reliability between reviewers and review teams have been reported from low to moderate [74], [90].

Schildmeijer et al addressed strength and limitations from the GTT reviewer's perspective. They interviewed the reviewers concerning the usefulness and application of the GTT, preventability of the adverse events, review teams and dependence of the documentation provided in the health records [84]. They concluded that changing the approach of the method could influence the GTT results. They also meant that the reviewers should be more focused at looking at the patient's perspective when deciding if an adverse event had happened.

These issues are further discussed in a review of the GTT which found widespread adoption with different modification demonstrating its flexibility [91]. With these concerns Hibbert et all proposed that "the GTT should be reframed as an opportunity to identify adverse events, raise awareness of these within hospitals and to describe the most frequent type of adverse events to prioritize quality improvement", rather than an exclusively measuring method [91]. This demonstrate that the GTT could be modified in order to act as a method both for acknowledging and measuring adverse events.

Forster et al demonstrated that triggers were identified in 19-56 % of the records suggesting that half of the records are excessively reviewed when manual review for triggers are performed [92]. With automatic identification of triggers, manual reviews are only needed in records where triggers are identified in order to determine if the trigger is associated with any adverse events [87]. This reduce the number of records needed to be reviewed as the first part of the review (trigger identification) is done automatically. Such approach has been demonstrated to identify adverse drug events and adverse paediatric events with promising results [93], [94].

Accounting for these challenges we initiated our studies to evaluate the GTT method regarding sample size, inter-rater reliability and automatic identification of triggers.

2 AIMS OF THE THESIS

Overall aim

The general aim of this thesis was to evaluate the GTT method regarding sample size, changes of reviewers and automatic trigger identification to improve the method's reliability and validity and to reduce the resources required.

Specific aims:

Paper I

To investigate the influence on the results of increasing the sample of reviewed records by the GTT.

Paper II

To evaluate the inter-rater reliability when reviewers are replaced when identifying adverse events by the GTT.

Paper III

To evaluate a modified GTT method with automatic trigger identification to the original GTT method with manual trigger identification.

3 MATERIAL AND METHODS

3.1 Setting

The GTT was implemented in Norwegian hospitals in 2011 as a part of the National Patient Safety Program “In safe hands” launched in 2010. All hospitals were required to review ten closed inpatients records randomly selected every bi-weekly period. Our trust, Nordland Hospital trust, chose to multiply the recommended sample size times seven. This was done partly because we wanted to measure adverse events separately for our seven main units, but we also thought that ten records reviewed bi-weekly were too small for reliable results. The trust implemented seven different GTT review teams corresponding to the seven different units. The seven review teams reviewed records discharged from their department respectively. The reviewers in the studies were recruited from the GTT review teams in the trust and had the same basic training with the GTT.

The electronic health record (EHR) system was implemented in the trust in 1992 (DIPS, ASA). The EHRs include both free text (i.e.: discharge summaries, operative reports, pathology reports, radiology results, transfer of service notes, admission notes, medical progress notes and notes from other healthcare professionals) and indexed variables (i.e.: laboratory results, admissions and discharge data, diagnosis and procedure codes). In Norwegian hospitals medication administration, prescriber orders and vital parameters are still hand-written and scanned into the EHRs but are currently being digitalized and indexed.

The first national Norwegian GTT results from all Norwegian hospital were used to estimate the number of deaths and harms caused by medical treatment. These calculations were made by extrapolations from the rate of the identified adverse events and contributed to major resistance and objections from health personnel against the GTT when published [20], [95]. The critics from the health personnel were mainly concerning the small sample size. Also, the definition of the adverse event defined as patient harm were not necessarily acknowledged by the clinical staff. Last, the GTT required resources which were considered unmanageable by the clinical staff. We designed these studies to examine the arguments from the critics.

3.2 Study design

All records included are selected from the discharge lists in Nordland Hospital Trust (figure 4 and figure 5). A total of 3153 different admissions were included altogether. Exclusion criteria were; patients aged 17 years or younger, patients admitted primarily for psychiatric or rehabilitation care, or patients with a length of stay less than 24 hours. The exclusion criteria were adapted from the GTT as the triggers are developed for adult somatic inpatients only [5].

Anonymous bi-weekly discharge lists were obtained from the hospital administrative system. Included records were randomly selected as described in the Norwegian GTT [19]. The discharge lists included information regarding type of admission (acute or planned), diagnoses, services which the patient was admitted to, case mix index (the value is dependent on diagnosis and the allocation of resources to care for and/or treatment included in the admission), wherever the patient underwent surgery, sex, age and length of stay.

Figure 4 Flowchart of the study populations in Paper I and Paper II

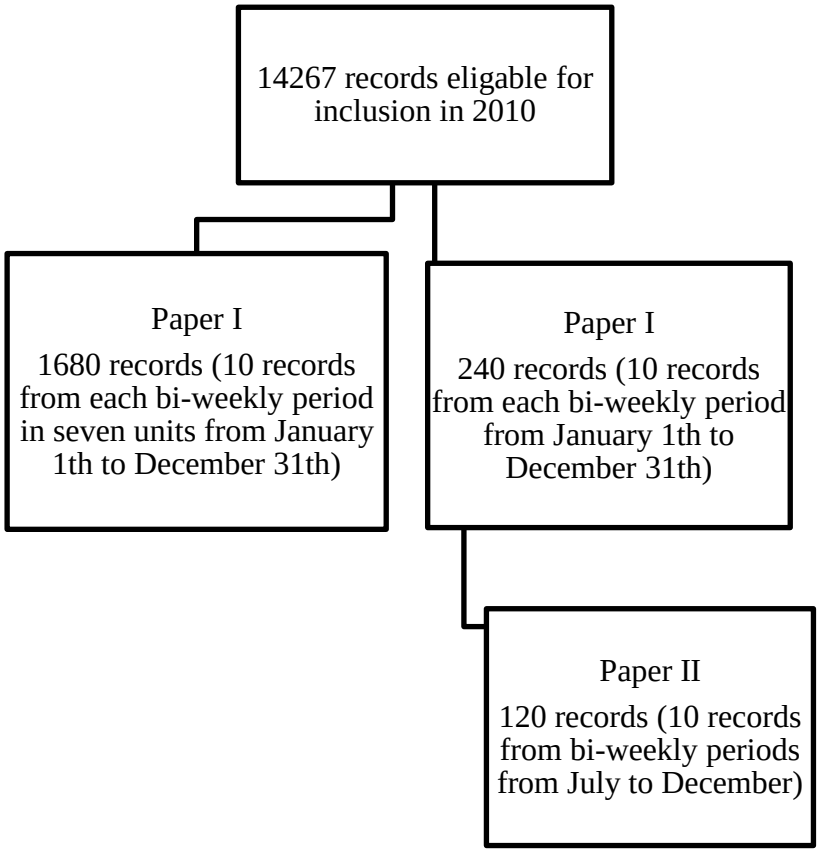
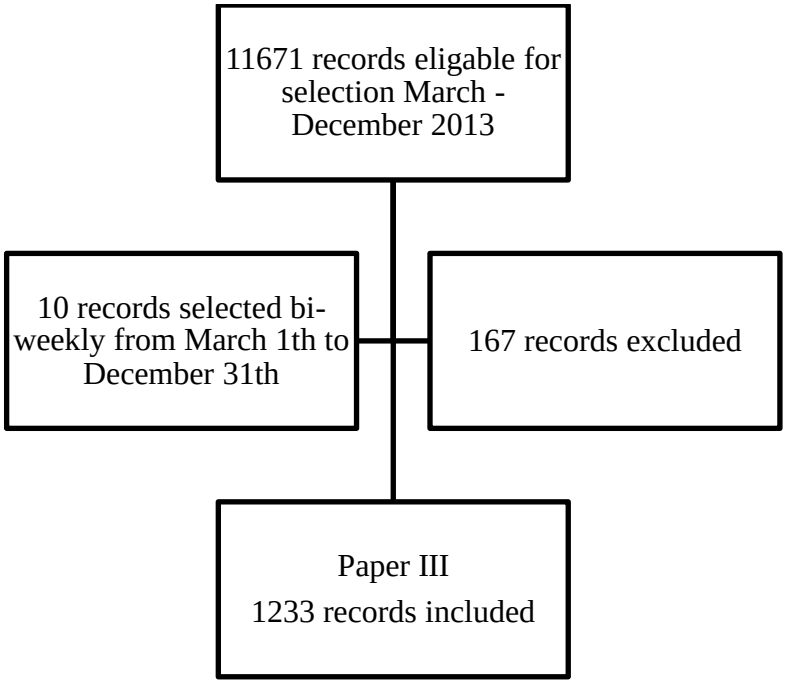


Figure 5 Flow chart of study population in Paper III



.

In all three papers collection of data was done retrospectively. All three studies are observational studies. Observational studies are either prospective, retrospective or cross-sectional studies. In prospective studies a sample of study objects are classified in some way and then followed over a period to see if they develop a condition. In retrospective studies the cases of the condition have already occurred at study initiation, and the study investigates if the subjects were exposed to any risk factors. In cross-sectional studies the samples are randomly selected at a specific point in time and cross-classified if they have the condition or not. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE)-checklist [96] was used to include relevant information in the papers.

In Paper I, the study was designed as an observational retrospective study. We chose this design as the data were collected retrospectively and the rate of identified adverse events in two different samples were compared. In the paper we have defined it as a cross-sectional study. This is not correct, as patient days were accounted for. We used the appropriate

statistical measurements and evaluated the study as a retrospective cohort study. The hypothesis was: will increasing number of records reviewed affect the results of identified adverse events? This was examined by comparing the rate, type and severity of adverse events identified by the GTT in two different sample sizes; one small and one large by obtaining the risk ratio (RR). Altogether 1920 records selected from the bi-weekly discharge lists were included. The large sample included records selected as 10 records bi-weekly from the seven units discharge lists (n=1680) while the small sample included ten records selected bi-weekly from the trust´s discharge lists (n=240) (figure 4). The manual review to identify triggers and adverse events differed in some way between the two samples. The records in the large sample were reviewed by one of three primary reviewers (two physicians and one nurse) and all three reached consensuses regarding the adverse events they all had identified separately. The records in the small sample were reviewed as described in the GTT; two primary reviewers (nurses) individually reviewed the records and reached consensus regarding the adverse events before a secondary reviewer (a physician) authenticated their common findings [5]. The reviewers of both samples were the same, except that one of the physicians from the small sample was replaced by a nurse in the large sample.

Paper II was designed as an observational cross-sectional study including 120 records (figure 4). We did not consider an alternative design as the study compares agreement between different reviewers regarding the prevalence of identified adverse events within a sample. The study evaluated the reproducibility of the method. The length of patient stay in the different records do not affect the results as the review teams review the same 120 records. Three review teams review the records as described in the GTT with two primary reviewers and one secondary reviewer [5]; Team I (three consistent reviewers- two primary reviewers and one secondary reviewer), Team II (one of the two primary reviewers or/and the secondary reviewer from Team I are replaced for different review periods) and Team III (no identical reviewers with Team I or Team II). The presence, type and severity of the adverse events identified by the three review teams were compared to assess the inter-rater reliability between the teams.

Paper III describes an observational cross-sectional study including 1233 records. The study evaluates two different methods to identify and measure adverse events, the modified GTT method versus the original GTT method. As in Paper II we did not consider an alternative design. 70 records were selected bi-weekly from the discharge lists from March 1th to December 31th 2013 (figure 5), but 167 records were excluded as data for these records were missing in the automatic trigger system. A modified GTT method, including automatic identification of triggers with manual review of the triggered records performed by a physician, was compared to the original GTT method [5]. The original GTT method included manual review of triggers and possible corresponding adverse events by two primary reviewers and authentication of their findings regarding adverse events was performed by a secondary reviewer. The identified adverse events by the modified GTT method were compared to the adverse events identified by the original GTT method. The concurrent validity of the modified GTT was evaluated by obtaining sensitivity, specificity, precision and reliability.

3.3 Intervention

In all three papers we evaluated the use of the GTT. The definition of an adverse event adopted from the GTT was applied in all three papers [5]. The training of the reviewers included the following understanding of how to determine if an adverse event was present: If the patient had experienced an unplanned event that led to either treatment, prolonged stay, permanent injury, immediate treatment to sustain life or death, the event was defined as an adverse event. The perspective of the patient was assessed by asking the questions as addressed in the GTT, mentioned in the chapter 1.3.2. [5]. Near misses that did not lead to the above criteria were not counted as adverse events and preventability of the adverse events was not evaluated.

The reviewers followed the approach as described in 1.3.2 except from the review team who reviewed the records of the large sample in Paper I. The triggers identified by the review teams and by the automatic trigger identification system was recorded in the databases. If the review team identified an adverse event, the type and severity was decided and recorded in

the databases. The categories of type of adverse events was adopted from the Norwegian GTT manual and sub classified in these main categories [19]:

- Surgical complications
- Bleeding/thrombosis
- Medication harm
- Patient fall
- Pressure ulcers
- Obstetric harm
- Other

3.4 Methodological consideration

Possible bias, presented in 1.2.2, could be present in all three studies. In Paper I selection and confounding bias are most likely to occur as the selection of records differed between the two samples. In Paper II hindsight bias could occur as the reviewers reviewed the records with different reviewers. The subjectivity of the reviewers could influence the results as the reviewers decided to some extent by themselves if the event was an adverse event or not. Selection bias is less likely in Paper II and Paper III as the reviewers reviewed the same records. In all three papers information bias could be present, as the findings of the adverse events rely on documentation in the records. Also, identification of triggers relies on that the information needed to identify a trigger is documented in the patient records. We consider that the results could be generalized for patient populations elsewhere.

3.5 Statistical analyses

The size of the small sample size in Paper I and Paper II are equal to the recommended sample size in the GTT with ten records selected bi-weekly. The size of the large sample in Paper I and Paper III with 70 records selected bi-weekly is the same as the total sample size used in our trust for the GTT. Power estimates of the sample sizes in Paper I was done with 80 % power. We assumed that the incidence of adverse events was 20 %. We then needed at least 7 % difference in the rate of identified adverse events between the samples for

significant results. In Paper II and Paper III kappa statistic was used. Power estimates was not performed, but if the CI is narrow then the power is considered good. The primary endpoint in all three papers was the rate of identified adverse events. Secondary endpoints were severity and types of adverse events.

To examine the means of the adverse event rates in the different samples, SPC charts were applied using QI Macros for Excel. The SPC charts include control limits, which are calculated based on the values presented and are ± 3 standard deviations from the central line (average). Any variation between the control limits are common causes of variation, while variation above or under the control limits are due to special cause variation.

Poisson regressions in Generalized linear models were applied to compare the rates of adverse events, severity level and categories of types of adverse events between the different sample sizes. Poisson regression was selected as it accounts for variation of number of cases and length of stay. Adjustments of demographical variables were done by including these as covariates. RR was obtained.

We used Cohen's kappa to determine the inter-rater reliability. For nominal data (value of 0 if no agreement and value of 1 if perfect agreement) kappa statistic was applied. For ordinal data (values from 1 to usual not more than 4 or 5 is applied for the different categories) weighted kappa was applied. Weighed kappa was applied when comparing severity level as we decided that it was less agreement if the event was rated category E by one reviewer and category H by another reviewer than rated category E versus category F. If no adverse events were identified the value of 1 was applied, if the adverse event was severity category E value 2 was applied, severity category F value 3 was applied, severity category G value 4 was applied, severity category H value 5 was applied and severity category I value 6 was applied. The 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) [97]. The inter-rater reliability has been used to evaluate measures of adverse events by many [98].

Statistical association of categorical variables was assessed by Chi-square test while continuous data were compared using independent t-tests. To compare the number of adverse events identified by different methods or by different teams, Paired t-tests were used. When evaluating the performance of the modified GTT method, the original GTT method was set as gold standard. We calculated sensitivity (recall as used in Paper I), positive predictive value (or precision) (PPV) and specificity with their respective 95 % confidence intervals (CI) to evaluate the validity of the modified GTT method [99]:

$$Sensitivity = \frac{No.of\ correct\ positive\ records\ identified\ by\ the\ modified\ GTT\ method}{No.of\ positive\ records\ identified\ by\ gold\ standard}$$

$$PPV = \frac{No.of\ correct\ positive\ records\ identified\ by\ the\ modified\ GTT\ method}{Total\ no.of\ positive\ records\ identified\ by\ the\ modified\ GTT\ method}$$

$$Specificity = \frac{No.of\ correct\ negative\ records\ identified\ by\ the\ modified\ GTT\ method}{No.of\ negative\ records\ identified\ by\ gold\ standard}$$

The CI for sensitivity, PPV and specificity was calculated using the Wilson score method [100]. CI for Cohen's Kappa was calculated as $\kappa \pm 1.96*SE$.

For all analyses, we used two-sided tests. The significance level was set at 5 % and 95 % CI were reported if relevant. The statistical analyses were performed using the SPSS statistical package, version 22.0 (SPSS Chicago, IL, USA).

When performing a statistical test there is always a chance of committing Type I error (incorrectly rejecting a true null hypothesis). The maximum probability of Type I error equals the specified significance level (5%) and we reject the null hypothesis whenever the p-value is lower than 0.05. In situations where we have multiple tests the Type I error probability in at least one test would be higher than 5 %. One way of reducing this error probability is to reduce the significance level proportionally to the number of tests (Bonferroni adjustments). Regardless of significance level, adjustment for other variables (age, length of stay etc.) increase the validity of our results.

3.6 Ethical consideration

The studies were performed in accordance to the Helsinki Declaration of 1975, and approved by the Norwegian Regional Committee for Medical and Health Research Ethics (Protocol ID: 2012/1691) and the Data Protection Office at the Nordland Hospital trust.

The information from the patient records were anonymised when extracted from the hospital administrative system and included in databases. The databases were hosted within an encrypted environment restricting the access to granted personnel only.

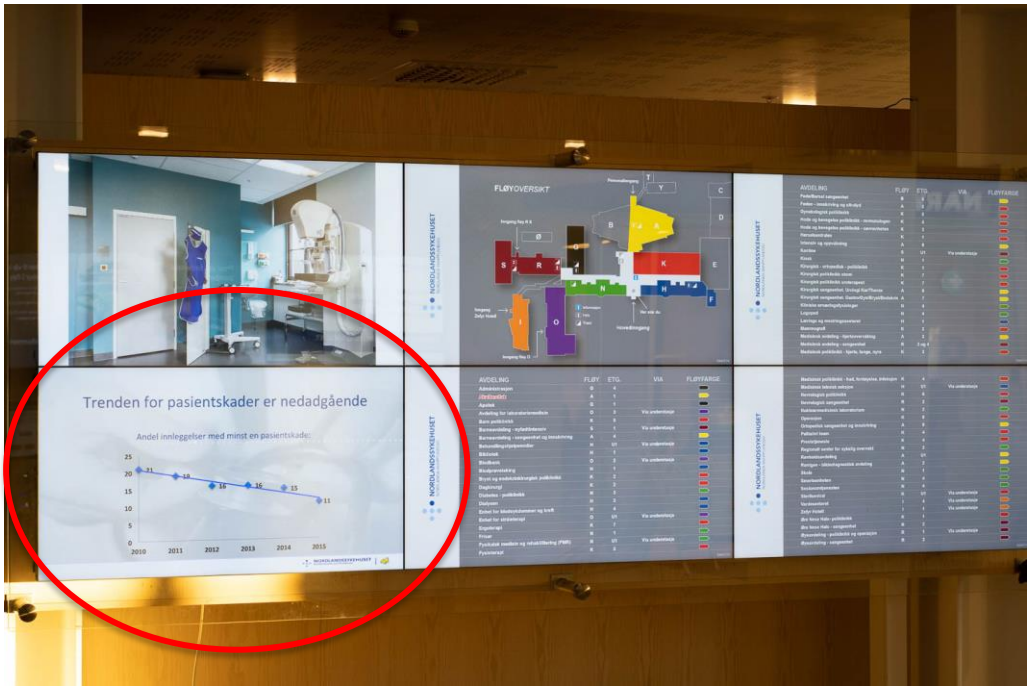
Information from medical records can be obtained for the purpose of internal control and quality assurance according to the Norwegian Health Professional Act. The trusts are required to obtain information and develop statistics about unintended events involving patients according to the regulations of the Norwegian Health Specialized Service Act. The health information can be obtained without consent in such cases. According to the same act the health services are also obligated to report severe adverse events to the Norwegian Board of Health Supervision.

The need for patient consent was waived for the records included in Paper III and for the largest sample in Paper I on the basis of :1) the records had already been selected and reviewed as a part of the trust's measurement of adverse events; 2) retrospectively collecting informed consent from patients or relatives of deceased patients would be costly with respect to time and money and might be considered a burden or inconvenience for the patients/relatives; 3) that the risk of being included and disadvantages of not being informed are considered minimal. This is in accordance to the criteria for waiving consent by Baker et al [101]. In Paper I we included a sample of records (n=240) that were not already selected for the trust's measurements of adverse events. 120 of the 240 records used in Paper I were also used in Paper II. We argued that these patients should be contacted and asked for consent when we applied for approval of the study. 26 denied consent or did not respond to the consent letter. To include the correct amount of records (ten records bi-weekly), replacements of records were performed with random selection from the same discharge lists where the

patients had denied consent or not responded. The “new” included patients were also asked for consent. We included information in the consent letter that the patients could contact the study leader upon questions and that they at any time could withdraw their consent. Retrospectively we considered that asking these patients for consent was not necessary in concordance with the referred criteria.

There is great variability in the interpretation of research issues related to patient safety and quality. Research committees and national legislations practice consent waiving differently [102]. The WHO recommend that when in doubt, all projects should be submitted to the ethic committees before study start, to determine if consent is needed [103]. The ethic committees can waive the usual requirement of individual informed consent when the research involves minimal risks and obtaining consent would be impracticable [104]. An alternative when formal consent is waived, is to provide information of the studies being performed either by posters, leaflets or as a part of the general patient information [103]. In our trust we provide information in the lobby regarding that patient records are being reviewed to identify adverse events, along with the identified rate of adverse events (figure 6).

Figure 6 The lobby in Nordland Hospital Trust, Bodø



4. RESULTS

The thesis examined different methodological aspects of the GTT record review method. In the first part of the study, the results suggested that increasing the sample sizes narrowed the CI thereby giving more precise results that can be extrapolated to institutional levels. The rate of identified adverse events was higher in a large sample compared to a small sample. The second part of the study evaluated how a larger sample of records could be efficient reviewed with valid results. The review of the triggered records should be done with consistent reviewers and automatic trigger identification enabling increasing the sample size without increasing the resources needed.

4.1 Patient characteristics

Demographic variables for the included 3153 patients compared to all records eligible for selection in 2010 and 2013, are presented in table 4. Median age was 64 (range 18-84) years. Most of the patients were women (60 %). Adverse events were identified in 655 (21 %) of the records included in the studies.

Table 4 Demographic characteristic

	Included records (n=3153)	**Mean (SD)**	**All records (n=25938)**	**Mean (SD)**
Hospital				
	N (%)		N (%)	
Bodø	2301 (73)		17153 (66)	
Lofoten	395 (13)		4201(16)	
Vesterålen	457 (15)		4584 (18)	
Age				
≤ 65	1632 (52)	60.1	12335 (48)	62.1 (20.7)
> 65	1521 (48)	(21.3)	13603 (52)	
Type of admission				
Acute	2240 (71)		19092 (74)	
Planned	913 (29)		6846 (27)	
Sex				
Male	1258 (40)		11189 (43)	
Female	1895 (60)		14749 (57)	
Number of patient days				
3≤	1421 (45)	6.1	11082 (43)	6.1 (6.5)
3>	1732 (55)	(7.3)	14856 (57)	
Adverse event present*				
Yes	655 (21)			
No	2498 (79)			

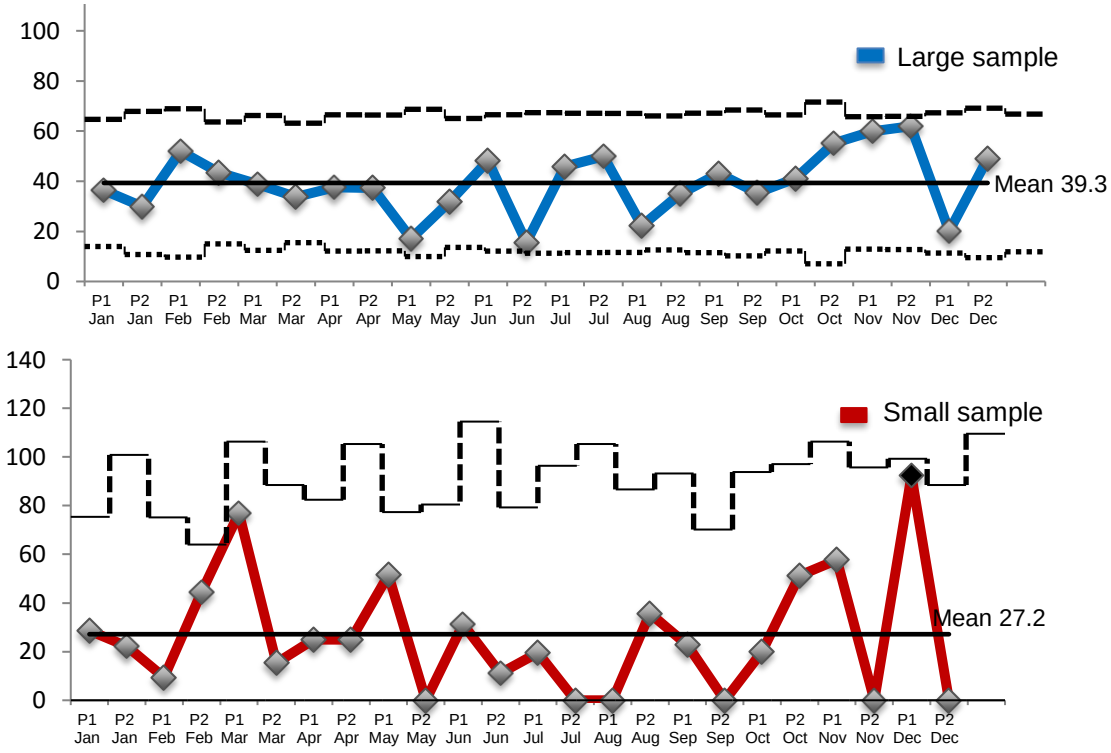
*Adverse events identified in at least one of the studies

4.2 Paper I

We found that a large sample size of 70 records selected bi-weekly identified 45 % (RR: 1.45 CI: 1.07-1.97) more adverse events per 1000 patient days, than a smaller sample size of ten records selected bi-weekly. In the large sample 39.3 adverse events per 1000 patient days (CI: 35.8-43.1, SE: 1.86) were identified while in the small sample 27.2 adverse events per 1000 patient days (CI: 20.3-36.4, SE: 4.05) were identified. The difference was significant (p=0.02, CI: 1.04-1.93). As expected, the CI was narrower and the SE was lower in the large sample than in the small sample. However, there was no difference regarding variation over time between the samples. This is in accordance with the main purpose of the GTT; to monitor the rate of adverse events over time. There was no significant difference between the samples regarding length of stay, average age or sex. When adjusting for services, diagnosis, case mix index, surgical treatment, acute or planned admission and numbers of transfers related to the index hospitalisation, the overall results were not altered.

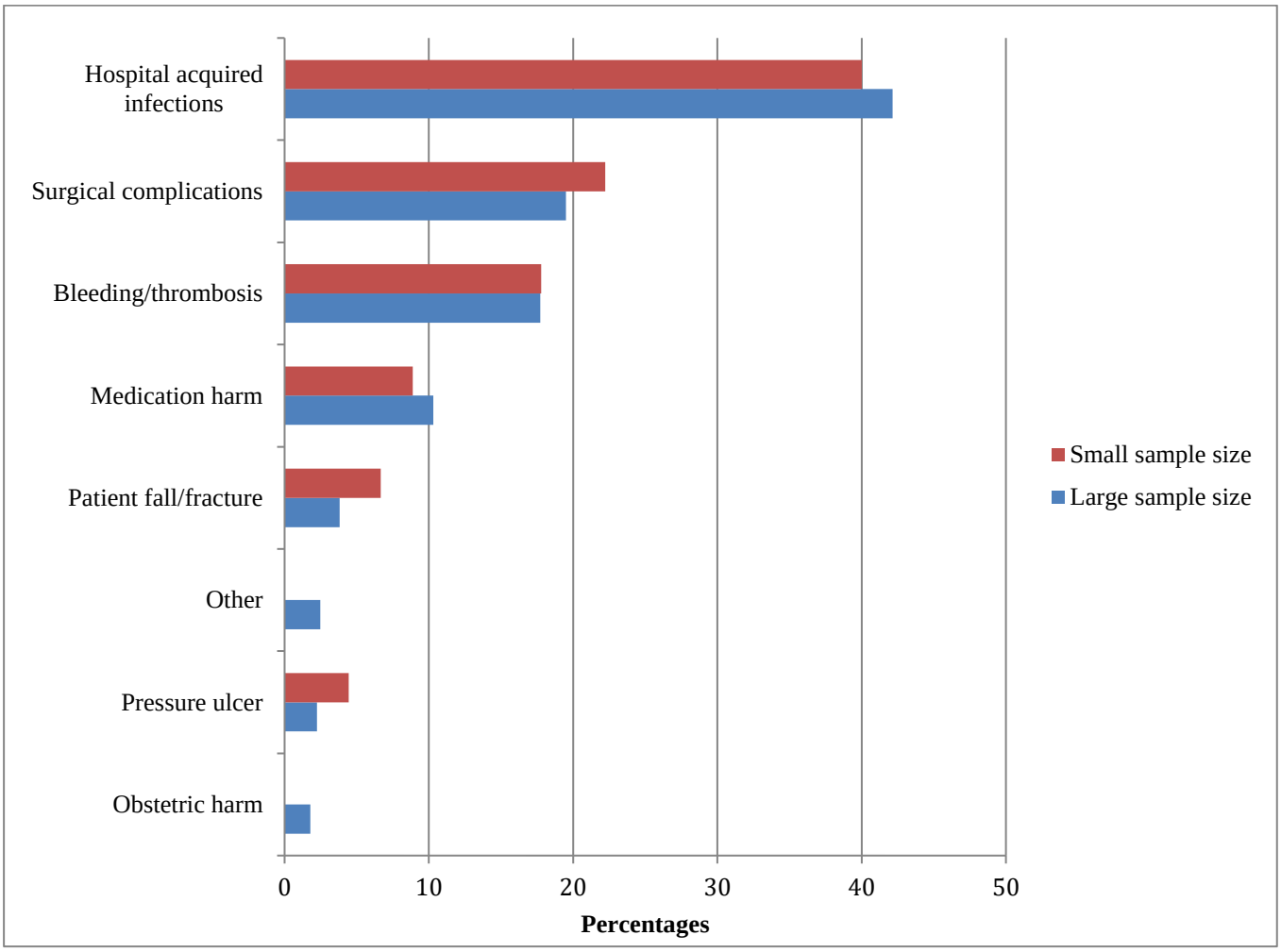
SPC charts were applied to compare the mean rate of adverse events over time to examine if any of the tests of special causes were positive. In the small sample test 1 was positive (i.e.; data points outside the control limits). None of the tests were positive for the large sample.

Figure 7 Number of adverse events per 1000 patient days in SPC U-chart



Hospital acquired infection was the most frequent type of adverse event in both samples followed by surgical related harms, medication harms, bleeding/thromboembolism, patient falls, pressure ulcer and obstetric harms (figure 8). No significant difference between the samples regarding the types of adverse events or the severity level was identified. 57 % of the adverse events identified in the large sample were defined as category E (harms requiring interventions) compared to 56 % in the small sample (RR: 1.5 p= 0.054, CI; 0.99-2.26). Respectively 39 % and 33 % of the adverse events were category F (RR: 1.69 p=0.051, CI: 1.00-2.86) and 3 % and 11 % were defined as severe adverse events (category G, H or I) (RR: 0.47 p=0.14, CI: 0.17-1.27).

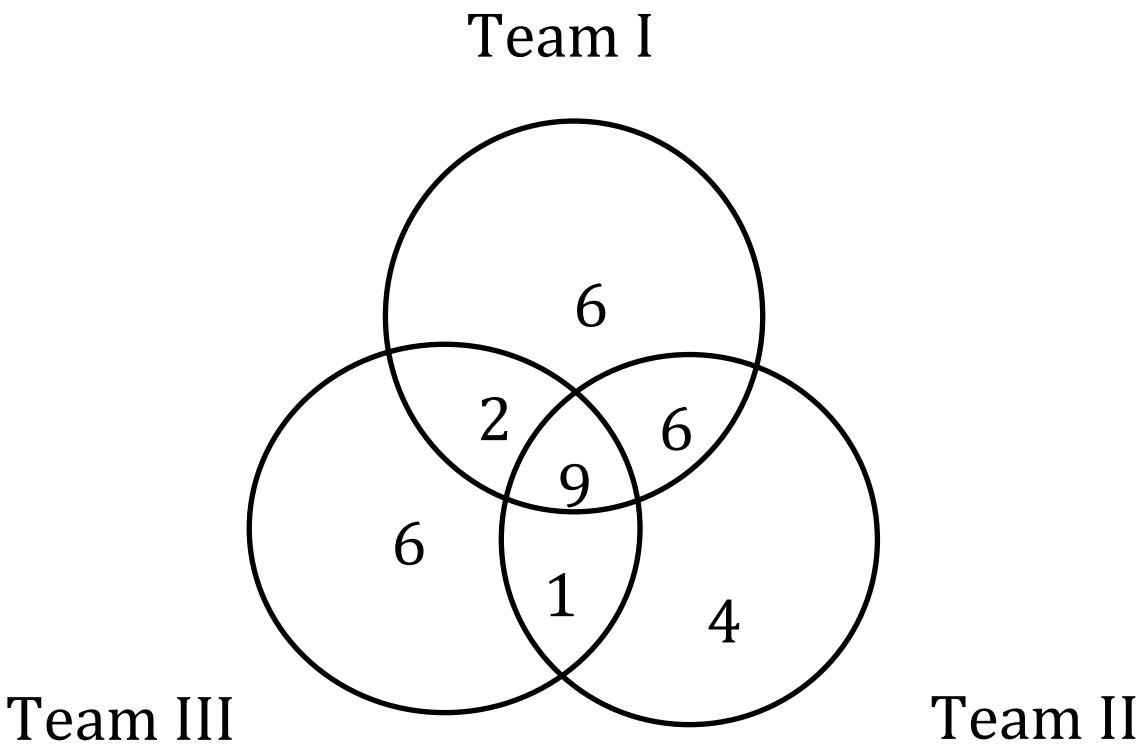
Figure 8 Types of adverse events identified



4.3 Paper II

120 records were reviewed by three review teams; Team I, Team II and Team III. Team I and Team II had one or two identical reviewers throughout the review of the 120 records. Team III had none identical reviewers with Team I and Team II. Team I identified 23 adverse events, Team II identified 20 adverse events and Team III identified 18 adverse events (figure 9). Team I and Team II identified six identical adverse events. The same six adverse events were not identified by Team III. In seven records Team III disagreed with Team I in regard of type of adverse event, while Team II disagreed to Team I in three records.

Figure 9 Number of identified adverse events by the three teams



We found that the agreement in regards of presence of adverse events was substantial (κ=0.64) when one or two of the reviewers were identical (Team I versus Team II) compared to moderate (κ=0.47) when none reviewers were identical (Team I versus Team III). Regarding the number of adverse events and categorizing of the severity level, the agreement

was substantial between Team I and Team II compared to between Team I and Team III (table 5).

Table 5 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

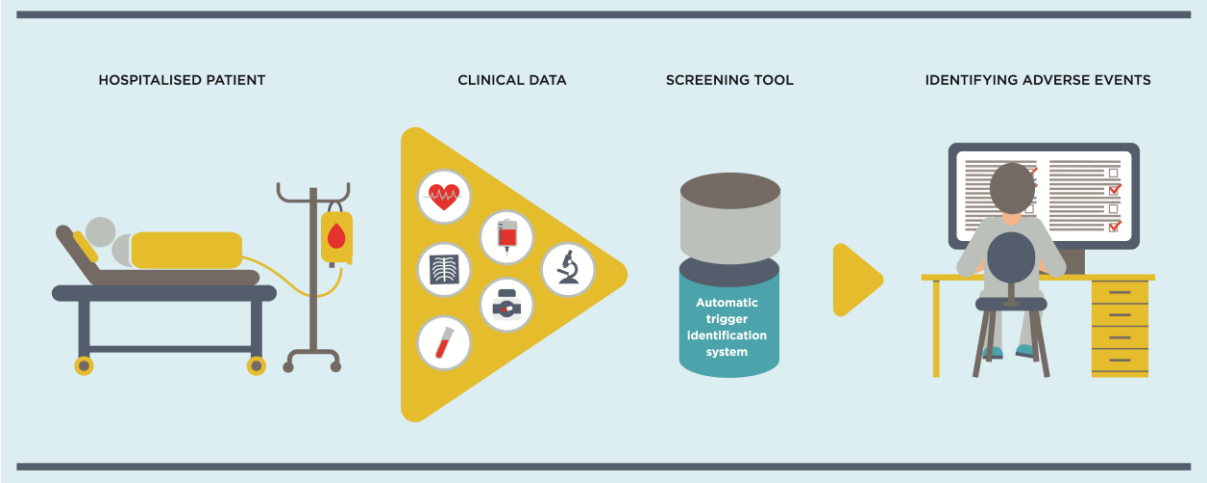
	Team I vs Team II (kappa coefficient, 95 % CI)	**Team I vs Team III (kappa coefficient, 95 % CI)**
Presence of adverse events*	0.640 (0.434-0.846)	0.468 (0.232-0.703)
Number of adverse events**	0.661 (0.479-0.842)	0.468 (0.278-0.694)
Severity level**	0.652 (0.469-0.836)	0.442 (0.260-0.624)

*Unweighted kappa analysis, **Weighted kappa analysis

4.4 Paper III

We evaluated the performance of a modified GTT method (figure 10). The modified GTT method included manual reviews for adverse events in 658 records identified with triggers by an automatic trigger identification system. The automatic trigger system screened 1233 records. The results were compared to the original GTT method which included manual review of all 1233 records to identify both triggers and adverse events.

Figure 10 The modified GTT method (Illustrated by Laila Bjølgerud)

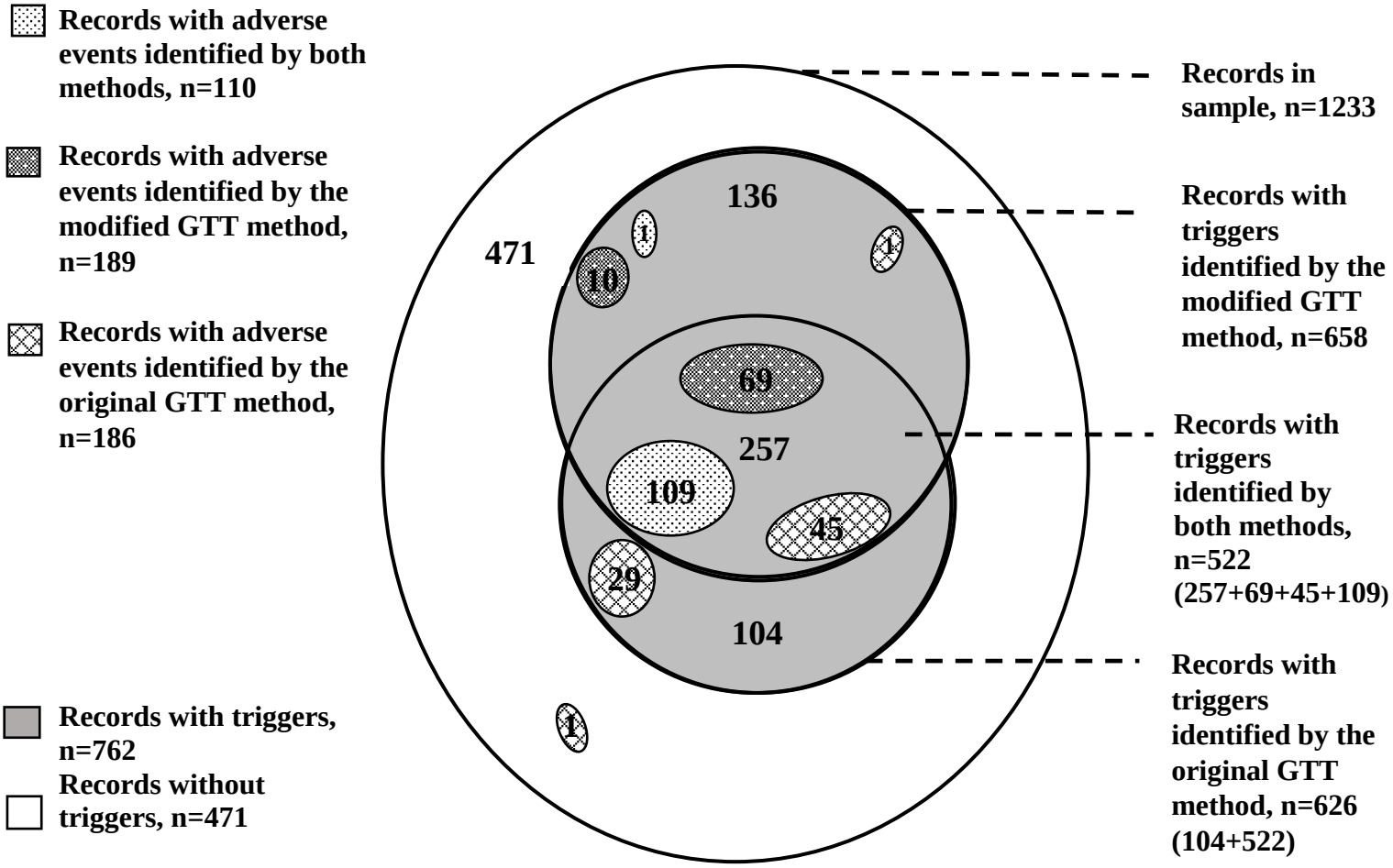


The modified GTT method identified the same rate of adverse events as the original GTT method; 35 adverse events per 1000 patient days. Sensitivity, PPV, specificity and reliability for records identified with adverse events were respectively 0.59, 0.58, 0.92 and 0.51 for the modified GTT method in respect to the original GTT method as gold standard. The total manual review time in the modified GTT method was 23 hours, while the manual review time using the original GTT method was 411 hours.

Number of records identified with adverse events (15.3 % versus 15.1 % of the total number of records, p=0.81, CI; -0.02-0.02) and number of identified adverse events (p=0.90, CI; -0.03-0.03) did not differ significantly between the modified GTT method and the original

GTT method. The modified GTT method reduced the number of records needed to be manual reviewed by 50 % (figure 11).

Figure 11 Number of records identified with triggers and adverse events by the modified GTT method and the original GTT method



5. DISCUSSION

5.1 Summary of strength and weaknesses

A summary of the strengths and weaknesses of the studies included in this thesis are presented in Table 6.

Table 6 A summary of strength and weaknesses

Weaknesses	Strengths
Study design	
Only two sample size (10 vs 70 records bi-weekly) compared	Large number of records reviewed from bi-weekly periods
Different sampling methods	Adjusted for different sample methods
	Records were randomly selected
No power estimates were performed	The sample size recommended in the GTT was applied
Some of the triggers were not possible to identify automatically	Automatic identification of 42 triggers
Small samples increase the risk of type 2 error	
Information bias due to retrospective collection of data	
No validation to an external patient cohort	The results correspond with other studies
Cross-sectional studies require large samples	Observational study design allows for different variables
Reviews	
	All reviews were performed by expert reviewers
	Reviewers underwent the same training program
Data rely on documentation in the EHR	
Identification of adverse events rely on triggers identified	A common definition of adverse events was applied in all paper
Manual review demanded time and methodical skills to assembly	
Data analysis	
	Generalized method was used accounting for different length of stay and different sample sizes
	Performance of a modified GTT method demonstrated a valid method to measure adverse events
Manual review is difficult to reproduce and compare between studies	Inter-rater reliability between reviewers was obtained when applicable and was substantial
A minimum P-value has increased type 1 error (false positive) and difficult to compare across studies	A minimum P-value approach is appropriate for exploratory studies (reducing type 2)

5.2 Paper I

Our study demonstrated that increasing the sample size affected the rate of adverse events, while the type and severity of identified adverse events were not influenced. 1.45 more adverse events per 1000 patient days were identified in the large sample (n=1680 records) than in the small sample (n=240). We argue that the rate of adverse events identified in the large sample is more representative and precise than the rate of adverse events identified in the small sample due to the narrow CI in the large sample.

A narrow CI makes the results more precise, and extrapolation with such results makes estimates of the total number of adverse events less uncertain. Many have debated the difficulty with estimating rates of adverse events based on small samples, and our results demonstrate this challenge [63], [105], [106]. Also the infrequent severe adverse events are often missed when sampling approaches are used [107]. No severe adverse events (category I) were identified in the small sample. Due to the infrequently occurrence of severe adverse events, other methods should be used to monitor these specific types of events, for example, investigating all hospital deaths [108], [109].

In the small sample there were an outliner with excessive patient days. We tested if exclusion of 10 % of the patient days form each sample altered the results. In the large sample 24 records with a total length of stay of 1150 patient days were excluded. In the small sample two records were excluded with a length of stay of 197 patient days. The result was not altered as the RR was 1.55 (CI: 1.1-2.1). In the large sample the rate was 39.2 adverse events per 1000 patient days while in the small sample the rate was 25.3 adverse events per 1000 patient days. As an explanation of the difference in rate, we therefore argue that this is most likely due to random variation as other explanations were adjusted for.

Several studies refer to high sensitivity [6] and acceptable reliability [7], [110] of the GTT, but the impact of the sample size has been discussed only by few. To our knowledge this study is the first attempt to assess the impact of the sample size on the results identified with the GTT. The large sample in our study represented approximately 12 % of the overall patient

population in our trust. Kennerly et al, among others, proposed that the sample size of records to be reviewed should be adjusted to the hospital size [6], [75], [111].

It is important to consider the Simpsons paradox when evaluating the results. This is implying that statistical results from aggregated data could give a different result when extracting the results to a group-level analysis [112]. This is important to be aware of when using the statistics for causal interpretations. We therefore adjusted for the variations such as case mix, if surgery was performed, case mix index, hospital locations, units and type of admissions, which were correlated to the index discharges (the sources of the selection of the records). When adjusting for these variations the results were not changed; the rate of adverse events was still significantly different between the two samples while the type and severity were not. The results did also not differ when adjusting for demographic variables such as gender, age and total length of stay. We do therefore not consider that the Simpsons paradox is a relevant problem in this study.

Another factor is that we do not have any information of the patient records in the small sample where patients had denied consent or refused to answer. These patients could have experienced an adverse event and would not participate because of a bad experience with the trust. This could bias the result from the small sample as the aim was to compare the rate of adverse events between the samples.

The SE of the mean represents the degree of the variability of the mean. The SE is low in the large sample while the small sample has a higher SE. The means of the rate of adverse events identified in the large sample has less random variability. With these assumptions we consider that a larger sample include more trustworthy results. The smaller sample is less resilient for outliners as there are too few records included.

The review process of the two samples differed slightly as described in the study design. To adjust for this possible bias, we assessed the agreement between the different authentication processes and found the agreement to be substantial. Zegers et al concluded that different

authentication processes did not impact the results [113]. We argue that the difference in review process between the samples did not influence the results in our study.

The power estimate of the large sample size was based on a difference of 7 % between the samples with 80 % power. The difference in the identified adverse events rate between the samples was 45 %. We therefore assume that the sample size was large enough. A larger sample size could reflect the population more accurate than a smaller sample and the rate of adverse events that were identified in a larger sample could be more reliable [114]. Further, we could have included more records by enhancing the study length period. Variation of number of patients and medical care given differ more between the different parts of the year than between two different years. We assumed that inclusion of records from one year was enough to obtain reliable results.

Other limitations when interpreting the results, is the categorization of types of adverse events which are not mutually exclusive. The determination of type of adverse event is based on the subjectivity of the reviewers as no common definitions of which type of adverse events to include in the different categories exists.

The length of stay in the records, which is the denominator in the estimated rate of adverse events per 1000 patient days, must be accounted for when comparing the means of the rate of adverse events. We therefore applied the Poisson regression in the generalized linear models as it is appropriate for rate data when the dependent variable is a count of events divided by some measure of that unit's exposure, i.e. number of adverse events per 1,000 patient days. The difference in number of records included in the two samples is also being accounted for when using this statistical test. The RR could then by obtained. The wide control limits in the SPC chart of the mean rates in the small sample demonstrated that these rates did vary more than the rates identified in the large sample.

Our results imply that the recommend sample size of ten records reviewed bi-weekly is too uncertain. Hence, further studies are needed to determine whether there is an optimal sample size. For example, if the sample size should be based on hospital size, especially as reviewing

larger sample sizes requires more resources. Until further studies, we have suggested using a relative increase in sample size to 8–10% of total number of discharges when using the GTT to achieve a narrow CI and hence more precise results. The increase in sample size requires a more effective strategy to review the records which is evaluated in Paper II and Paper III.

5.3 Paper II

In this study we evaluated the inter-rater reliability as we compared the results from different review teams who reviewed the same records. Others have examined the inter-rater reliability and have found at best a moderate to substantial agreement [98]. However, in this study we demonstrated substantial inter-rater reliability between review teams where at least one of the reviewers were identical. Moderate inter-rater reliability was found between review teams with no identical reviewers.

Members in the review teams performing the GTT are often replaced due to practical issues such as relocation of work place, sick leave and maternity/paternity leave. To our knowledge, this is the first attempt to assess inter-rater reliability between review teams experiencing replacement of reviewers to varying degrees. Evaluating the inter-rater reliability between all different teams, as replacement of all reviewers, have been described previously and reported to be poor [68]. We therefore evaluated how the results are affected when review members are changed except from one of the primary reviewers. We chose to keep one of the primary reviewers consistent as the GTT recommend that the primary reviewers are the ones who conduct the first screening of the records and therefore most important to keep consistent [5]. We considered replacement of both primary reviewers as equal to replacement of all review members as the primary reviewers perform the initial review. The secondary reviewer only authenticates the findings without accessing the records routinely. Unlike our assumptions O'Leary et al highlighted that the variation was higher between confirmation of adverse events than for identification of potentials adverse events [115].

The variables concerning the reviewers such as review experience, clinical background and years of experience could influence the results. The mean years of clinical experience was 18.3 years (range 7-29) of the reviewers and the total mean years of review experience of the three teams were 2 years. To evaluate the agreement of identified adverse events between the teams, the kappa statistics was used. This analysis is not able to adjust for clinical experience between the teams. We have therefore not discussed any influence such as psychology or social influence of the consistent reviewer. This viewpoint would be more of a study of group

dynamics which was not the intention of this study. We intended to evaluate a practical solution with a pool of reviewers performing the GTT at different times without influencing the results.

Our findings indicate that hospitals can rely on rotating reviewers from a consistent pool of reviewers in order to optimize resources. With this approach hospitals are encouraged to perform the GTT even if they experience frequent replacement of reviewers. However, the CI is wide which indicate that the sample size might not be large enough. Our results must therefore be interpreted with some caution.

5.4 Paper III

Identifying and measuring adverse events in hospitalised patients is challenging. So far, we consider the GTT the most robust method to measure adverse events in comparison to most other existing methods. The practical disadvantage with the method being resource intensive, can to a certain extent be addressed by automating the trigger identification. We developed an automatic trigger identification system to automate 42 of the GTT triggers. The study demonstrated that the modified GTT method using automatic trigger identification is a valid measure in respect to the original GTT method. To our best knowledge such study has not been performed previously.

Since the late 90's Classen et al along with others, have demonstrated computerized surveillance of adverse drug event by automated detection of triggers that could represent possible adverse events [116]–[121]. When triggers are automatically identified, only the records with triggers are reviewed manually to determine if the trigger represents an adverse event. This approach has showed promising results [93], [94].

The "gold standard" of determination of an adverse event has traditionally been the judgment of clinicians [122]. Automatic identification of adverse events based on administrative data have showed disappointing results [123]. With such approach the positive predictive value are reported to be low, ranging from 12-30 % [124]–[136]. We consider that a manual review is still needed to determine if the triggers automatically identified, represent an adverse event according to the GTT definition applied in the study. However, machine learning is slowly integrated in medical decisions, such as radiology imagination and treatment outcomes [137], [138]. In the future it is therefore possible that adverse events can be identified automatically. We consider that the automatic trigger identification system could be further developed to a system that can predict which patients who are at risk to experience an adverse event enabling the clinicians to act in real-time to prevent adverse event.

The modified GTT method, with manual review of only records with automatic identified triggers, demonstrated a more resource efficient method than the original GTT method. The

number of manually reviewed records were reduced by 50 % with the modified GTT method (n=658) compared to the records manually reviewed in the original GTT method (n=1233). This is because the original GTT method demands the reviewers to screen all records to identify any triggers and then do a more in-depth review when triggers are identified to find any possible corresponding adverse events. In the modified GTT method the automatic trigger identification system performed the screening of triggers and manual review was only performed in the triggered records. The time used with the modified GTT method was only 6 % of the time used with the original GTT method. We consider this an exceptional result. Others have showed that the time using computerized strategies is 20 % compared with the time used with manual strategies [107], [139].

We found good agreement between the two methods with regards to the records identified with adverse events (κ=0.51 CI: 0.44–0.57). Our results demonstrated better agreement between the automated method versus all manual methods, compared to other studies, who have found only up to 12 % agreement [115], [139]. The modified GTT method identified 59 % of the records identified with adverse events by the original GTT method (110 of 186 records). The variation between the methods concerning the difference of number of records identified with adverse events could be explained by using different review team. The automatic triggered records were reviewed by one physician. The original GTT method was performed as described in the GTT manual [5]. We have argued that using different reviewers may affect the results as demonstrated by for example O'Leary et al [115]. However, we concluded that this did not bias the results in this study.

A recent review by Hibbert et al found that the GTT identified adverse events in 7-40 % with a cluster around 20-29 % of the reviewed records [91]. O'Leary et al found that 22- 26 % of records identified by automatic system were confirmed with adverse event [115]. The modified GTT method confirmed adverse events in 33 % of the triggered records. To examine how many of the total number of records are triggered, we ran the automatic trigger identification system for all admissions eligible for inclusion in the GTT in 2017. The automatic trigger identification system identified at least one trigger in 62 % (n=10807 records) of the records. The modified GTT method identified adverse events approximately in

30 % of the triggered records in the study, constituting 15 % of the original record sample. If we apply this result to the aggregated numbers of 2017, the estimated number of records with adverse events would be 3242 records; or one of five hospitalised patients are harmed due to medical care. This emphasize the modified GTT method as a valid method to measure adverse events [6].

We have not considered the financial aspect of the automatic tool as this was beyond the scope of the study. The GTT method is criticized because it is resource intensive due to time and personnel required. We have therefore evaluated how to reduce resources in regard to personnel and time needed for reviewing the records.

Our results recommend that the modified GTT should be preferred rather than the original GTT method, as the modified GTT method is less resource intensive. The resources saved by using the modified GTT method is considerable, enabling increasing the sample size as proposed in Paper I and reviewing the records with consistent reviewers as demonstrated in Paper II.

6. CONCLUSION

In Paper I we found that increasing the sample size provides a narrower CI, reduce the random variation and increase the precision of the results. The rate of identified adverse events was higher in a large sample than in a small sample. We argue that a large sample should be preferred as this is a more reliable source for extrapolation of rates when calculating the total number of adverse events. In Paper II we demonstrated that keeping one reviewer consistent provide more reliable results. Using the modified GTT method, as demonstrated in Paper III to identify and measure adverse events, is a time-effective strategy. We suggest that our findings can guide hospitals to identify and measure adverse events more effectively and that using such approaches would gain valid and reliable results.

7. IMPLICATION FOR FUTURE RESEARCH

The results of these studies lead to suggestions of some changes in the practical use of the GTT. First, we suggest that the sample size should be increased, second, we argue the importance of keeping one of the primary reviewers consistent and finally we introduced automatic trigger identification as a successful alternative approach to the manual GTT. These implications could facilitate more widespread adoption of the GTT as a method to identify and measure adverse events.

Future research could include a comparison of review of the automated triggered records by two different review teams or by two different reviewers. This could be done as a cross-sectional study comparing the rate of identified adverse events but also comparing the findings in each record by the two review teams/reviewers. If the agreement is substantial it could demonstrate that automatic trigger identification increase the agreement with an objective screening of the records.

Also, an automatic trigger identification system could be developed further to a prospective approach. Sammer et al presented a system allowing for real-time bedside intervention, real-time trend analysis and continued learning about harm measurement using a sociotechnical approach of people, process and technology [107]. Their framework emphasizes the framework of Donabedian [140] to assess quality of care; structure, process and outcome. We believe that moving to a prospective system, to identify patient at risk, would be beneficial for the clinical health personnel as it allows them to prevent adverse events from happening to the actual patient. Novel technologies such as identifying risk factor for developing adverse events must be integrated in the EHR. This could be performed either as a cohort study or as a cross- sectional study depending on the study questions. Such prospective system could be used to improve clinical outcome, optimize treatment, reduce the financial burden of patient harm, reduce the burden for involved health personnel and most importantly; reduce the suffering for patients due to adverse events.

8. REFERENCES

[1] M. A. Makary and M. Daniel, “Medical error—the third leading cause of death in the US,” *BMJ*, vol. 353, 2016.

[2] L. Adler *et al.*, “Impact of Inpatient Harms on Hospital Finances and Patient Clinical Outcomes.,” *J. Patient Saf.*, vol. 14, no. 2, pp. 67–73, Jun. 2018.

[3] W. Martinez, L. S. Lehmann, Y.-Y. Hu, S. P. Desai, and J. Shapiro, “Processes for Identifying and Reviewing Adverse Events and Near Misses at an Academic Medical Center.,” *Jt. Comm. J. Qual. patient Saf.*, vol. 43, no. 1, pp. 5–15, Jan. 2017.

[4] R. K. Resar, J. D. Rozich, and D. Classen, “Methodology and rationale for the measurement of harm with trigger tools.,” *Qual. Saf. Health Care*, vol. 12 Suppl 2, pp. ii39-i45, 2003.

[5] F. Griffin and R. Resar, “IHI Global Trigger Tool for measuring adverse events (Second Edition),” *IHI Innov. Ser. white Pap. Cambridge, Massachusetts Inst. Healthc. Imrovement;*, pp. 1–44, 2009.

[6] D. C. Classen *et al.*, “‘Global trigger tool’ shows that adverse events in hospitals may be ten times greater than previously measured.,” *Health Aff. (Millwood).*, vol. 30, no. 4, pp. 581–9, Apr. 2011.

[7] J. M. Naessens *et al.*, “A comparison of hospital adverse events identified by three widely used detection methods.,” *Int. J. Qual. Health Care*, vol. 21, no. 4, pp. 301–7, Aug. 2009.

[8] T. O. Mattsson, J. L. Knudsen, J. Lauritsen, K. Brixen, and J. Herrstedt, “Assessment of the global trigger tool to measure, monitor and evaluate patient safety in cancer patients: reliability concerns are raised.,” *BMJ Qual. Saf.*, vol. 22, no. 7, pp. 571–9, Jul. 2013.

[9] K. Schildmeijer, L. Nilsson, K. Arestedt, and J. Perk, “Assessment of adverse events in medical care: lack of consistency between experienced teams using the global trigger tool,” *BMJ Quality & Safety*, vol. 21. pp. 307–314, 2012.

[10] W. Runciman, P. Hibbert, R. Thomson, T. Van Der Schaaf, H. Sherman, and P.

Lewalle, “Towards an International Classification for Patient Safety: key concepts and terms,” *Int. J. Qual. Heal. Care*, vol. 21, no. 1, pp. 18–26, Feb. 2009.

[11] W. Runciman, “Shared meanings: preferred terms and definitions for safety and quality concepts,” *Med. J. Aust.*, 2006.

[12] G. Parry, A. Cline, and D. Goldmann, “Deciphering harm measurement.,” *JAMA*, vol. 307, no. 20, pp. 2155–6, May 2012.

[13] A. Gawande, “Complications: A surgeon’s notes on an imperfect science,” 2010.

[14] K. Walshe, “Adverse events in health care: issues in measurement,” *Qual. Heal. Care*, 2000.

[15] R. M. Wachter and P. J. Pronovost, “Balancing ‘No Blame’ with Accountability in Patient Safety,” *N. Engl. J. Med.*, vol. 361, no. 14, pp. 1401–1406, Oct. 2009.

[16] A. S. Frankel, M. W. Leonard, and C. R. Denham, “Fair and just culture, team behavior, and leadership engagement: The tools to achieve high reliability,” *Health Services Research*, vol. 41, no. 4 II. pp. 1690–1709, 2006.

[17] R. M. Wachter, *Understanding patient safety*. McGraw Hill Medical, 2012.

[18] E. Thomas, D. Studdert, H. Burstin, and E. Orav, “Incidence and types of adverse events and negligent care in Utah and Colorado,” *Med. Care*, 2000.

[19] E. Deilkås, “Gjennomføring av journalundersøkelse med Global Trigger Tool (GTT) i den norske pasientsikkerhetskampanjen,” 2011.

[20] “Mener 200 sykehus-dødsfall kunne vært unngått - Forskning - Dagens Medisin.” [Online]. Available: https://www.dagensmedisin.no/artikler/2014/10/09/-ikke-4500-men-200-sykehus-dodsfall-kunne-vart-unngatt/. [Accessed: 22-Nov-2018].

[21] “Lov om erstatning ved pasientskader mv. (pasientskadeloven) - Lovdata.” [Online]. Available: https://lovdata.no/dokument/NL/lov/2001-06-15-53. [Accessed: 22-Nov-2018].

[22] T. A. Brennan *et al.*, “Incidence of adverse events and negligence in hospitalized patients. Results of the Harvard Medical Practice Study I,” *N Engl J Med*, vol. 324, pp.

370–376, 1991.

[23] E. J. Thomas and L. A. Petersen, “Measuring errors and adverse events in health care,” *J. Gen. Intern. Med.*, vol. 18, pp. 61–67, 2003.

[24] R. Harrison *et al.*, “The missing evidence: a systematic review of patients’ experiences of adverse events in health care,” *Int. J. Qual. Heal. Care*, vol. 27, no. 6, pp. 424–442, Dec. 2015.

[25] S. Lang, M. V. Garrido, and C. Heintze, “Patients’ views of adverse events in primary and ambulatory care: A systematic review to assess methods and the content of what patients consider to be adverse events,” *BMC Fam. Pract.*, vol. 17, no. 1, 2016.

[26] S. N. Weingart *et al.*, “What can hospitalized patients tell us about adverse events? Learning from patient-reported incidents.,” *J. Gen. Intern. Med.*, vol. 20, no. 9, pp. 830–6, Sep. 2005.

[27] J. S. Weissman *et al.*, “Comparing patient-reported hospital adverse events with medical record review: do patients know something that hospitals do not?,” *Ann. Intern. Med.*, vol. 149, no. 2, pp. 100–8, Jul. 2008.

[28] J. K. Ward and G. Armitage, “Can patients report patient safety incidents in a hospital setting? A systematic review,” *BMJ Qual. Saf.*, vol. 21, no. 8, pp. 685–699, 2012.

[29] O. Bjertnaes, E. T. Deilkås, K. E. Skudal, H. H. Iversen, and A. M. Bjerkan, “The association between patient-reported incidents in hospitals and estimated rates of patient harm.,” *Int. J. Qual. Health Care*, p. mzu087-, Nov. 2014.

[30] B. Röhrig, J.-B. du Prel, and M. Blettner, “Study design in medical research: part 2 of a series on the evaluation of scientific publications.,” *Dtsch. Arztebl. Int.*, vol. 106, no. 11, pp. 184–9, Mar. 2009.

[31] L. B. Mokkink *et al.*, “The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study.,” *Qual. Life Res.*, vol. 19, no. 4, pp. 539–49, May 2010.

[32] J. Sim and C. C. Wright, “The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements,” *Phys. Ther.*, vol. 85, no. 3, pp. 257–

268, Mar. 2005.

[33] L. J. Cronbach, “Coefficient alpha and the internal structure of tests,” *Psychometrika*, vol. 16, pp. 297–334, 1951.

[34] G. Winter and T. Q. Report, “A Comparative Discussion of the Notion of ‘Validity’ in Qualitative and Quantitative Research.pdf,” *Qual. Rep.*, vol. 4, pp. 1–12, 2000.

[35] R. D. Moen, T. W. Nolan, and L. P. Provost, *Quality improvement through planned experimentation*. McGraw-Hill Professional, 2012.

[36] A. H. Pripp, “Validitet,” *Tidsskr Nor Legeforen 2018 doi 10.4045/tidsskr.18.0398*.

[37] R. Fletcher, S. Fletcher, and G. Fletcher, “Clinical epidemiology: the essentials,” 2012.

[38] M. Hanskamp-Sebregts, M. Zegers, C. Vincent, P. J. van Gurp, H. C. W. de Vet, and H. Wollersheim, “Measurement of patient safety: a systematic review of the reliability and validity of adverse event detection with record review.,” *BMJ Open*, vol. 6, no. 8, p. e011078, Aug. 2016.

[39] M. Venditti, M. LA, S. Corrao, G. Licata, and P. Serra, “Outcomes of Patients Hospitalized With Community-Acquired, Health Care–Associated, and Hospital-Acquired Pneumonia,” *Ann. Intern. Med.*, vol. 150, no. 1, p. 19, Jan. 2009.

[40] S. S. Magill *et al.*, “Multistate Point-Prevalence Survey of Health Care–Associated Infections,” *N. Engl. J. Med.*, vol. 370, no. 13, pp. 1198–1208, Mar. 2014.

[41] “Sykehus: Infeksjoner og bruk av antibiotika høsten 2016 - FHI.” [Online]. Available: https://fhi.no/hn/helseregistre-og-registre/nois/resultater/resultater-sykehus/. [Accessed: 10-Mar-2017].

[42] “World Alliance for Patient Safety: Forward program,” 2006.

[43] “Summary of the evidence on patient safety: implications for research,” 2008.

[44] N. Kucher *et al.*, “Electronic Alerts to Prevent Venous Thromboembolism among Hospitalized Patients,” *N. Engl. J. Med.*, vol. 352, no. 10, pp. 969–977, Mar. 2005.

[45] J. A. Heit, F. A. Spencer, and R. H. White, “The epidemiology of venous thromboembolism,” *J. Thromb. Thrombolysis*, vol. 41, no. 1, pp. 3–14, Jan. 2016.

[46] M. R. Kwaan, D. M. Studdert, M. J. Zinner, and A. A. Gawande, “Incidence, patterns, and prevention of wrong-site surgery.,” *Arch. Surg.*, vol. 141, no. 4, pp. 353-7; discussion 357-8, 2006.

[47] P. J. Pronovost, I. Joint Commission Resources, and Joint Commission International., *Safe surgery guide*. Joint Commission Resources, 2010.

[48] “Avisa Nordland - Tragediene i Bodø har ført til stort medisinsk funn: - Dette var ingen av oss forberedt på.” [Online]. Available: https://www.an.no/nordlandssykehuset/forskning/bodo/tragediene-i-bodo-har-fort-til-stort-medisinsk-funn-dette-var-ingen-av-oss-forberedt-pa/f/5-4-905956. [Accessed: 22-Nov-2018].

[49] M. November Chie, L.,Weingart S.N., “Physician-Reported Adverse Events and Medical Errors in Obstetrics and Gynecology,” 2008.

[50] S. Andreasen, B. Backe, R. G. Jørstad, and P. Øian, “A nationwide descriptive study of obstetric claims for compensation in Norway,” *Acta Obstet. Gynecol. Scand.*, vol. 91, no. 10, pp. 1191–1195, Oct. 2012.

[51] D. C. Classen, “Adverse Drug Events in Hospitalized Patients<subtitle>Excess Length of Stay, Extra Costs, and Attributable Mortality</subtitle>,” *JAMA J. Am. Med. Assoc.*, vol. 277, no. 4, p. 301, 1997.

[52] S. Andreasen, B. Backe, S. Lydersen, K. Øvrebø, and P. Øian, “The consistency of experts’ evaluation of obstetric claims for compensation,” *BJOG An Int. J. Obstet. Gynaecol.*, vol. 122, no. 7, pp. 948–953, Jun. 2015.

[53] “California Medical Association. Report on the medical insurance feasibility study. San Francisco: Sutter, 1977.,” 1977.

[54] H. H. Hiatt *et al.*, “A Study of Medical Injury and Medical Malpractice,” *N. Engl. J. Med.*, vol. 321, no. 7, pp. 480–484, Aug. 1989.

[55] L. Leape and T. Brennan, “THE NATURE OF ADVERSE EVENTS IN HOSPITALIZED PATIENTS Results of the Harvard Medical Practice Study II,” *N. Engl. J. Med.*, vol. 324, pp. 377–384, 1991.

[56] W. C. Richardson *et al.*, *To Err is Human: Building A Safer Health System - Institute of Medicine*. 2000.

[57] G. R. Baker, "HARVARD MEDICAL PRACTICE STUDY," *Qual. Saf. Heal. Care*, vol. 13, no. 2, pp. 151–152, Apr. 2004.

[58] A. Gawande, E. Thomas, M. Zinner, and T. Brennan, "The incidence and nature of surgical adverse events in Colorado and Utah in 1992," *Surgery*, 1999.

[59] T. Schiøler, H. Lipczak, and B. Pedersen, "Incidence of adverse events in hospitals. A retrospective study of medical records," *Ugeskr.*, 2001.

[60] G. Baker, P. Norton, and V. Flintoft, "The Canadian Adverse Events Study: the incidence of adverse events among hospital patients in Canada," *Can. Med. ...*, 2004.

[61] C. Vincent, G. Neale, and M. Woloshynowych, "Adverse events in British hospitals: preliminary retrospective record review.," *BMJ*, vol. 322, no. 7285, pp. 517–9, Mar. 2001.

[62] P. Davis, R. Lay-Yee, R. Briant, and W. Ali, "Adverse events in New Zealand public hospitals I: occurrence and impact," 2002.

[63] J. James, "A new, evidence-based estimate of patient harms associated with hospital care," *J. Patient Saf.*, 2013.

[64] A. K. Jha, I. Larizgoitia, C. Audera-Lopez, N. Prasopa-Plaizier, H. Waters, and D. W. Bates, "The global burden of unsafe medical care: analytic modelling of observational studies," *BMJ Qual. Saf.*, vol. 22, no. 10, pp. 809–815, Oct. 2013.

[65] "Pasientskader i Norge 2015 målt med Global Trigger Tool - Google-søk," 2016.

[66] E. T. Deilkås *et al.*, "Exploring similarities and differences in hospital adverse event rates between Norway and Sweden using Global Trigger Tool," *BMJ Open*, vol. 7, no. 3, p. e012492, Mar. 2017.

[67] K. G. Shojania and E. J. Thomas, "Trends in adverse events over time: why are we not improving?," *BMJ Qual. Saf.*, vol. 22, no. 4, pp. 273–277, Mar. 2013.

[68] C. P. Landrigan, G. J. Parry, C. B. Bones, A. D. Hackbarth, D. A. Goldmann, and P. J.

Sharek, “Temporal trends in rates of patient harm resulting from medical care.,” *N. Engl. J. Med.*, vol. 363, no. 22, pp. 2124–2134, 2010.

[69] I. Christiaans-Dingelhoff, M. Smits, L. Zwaan, S. Lubberding, G. Van Der Wal, and C. Wagner, “To what extent are adverse events found in patient records reported by patients and healthcare professionals via complaints, claims and incident reports?,” *BMC Health Serv. Res.*, vol. 11, 2011.

[70] H. Jick, “Drugs — Remarkably Nontoxic,” *N. Engl. J. Med.*, vol. 291, no. 16, pp. 824–828, Oct. 1974.

[71] D. C. Classen, S. L. Pestotnik, R. S. Evans, and J. P. Burke, “Description of a computerized adverse drug event monitor using a hospital information system.,” *Hosp. Pharm.*, vol. 27, pp. 774, 776–779, 783, 1992.

[72] J. D. Rozich, C. R. Haraden, and R. K. Resar, “Adverse drug event trigger tool: a practical methodology for measuring medication related harm.,” *Qual. Saf. Health Care*, vol. 12, no. 3, pp. 194–200, 2003.

[73] G. S. Takata, W. Mason, C. Taketomo, T. Logsdon, and P. J. Sharek, “Development, testing, and findings of a pediatric-focused trigger tool to identify medication-related harm in US children’s hospitals.,” *Pediatrics*, vol. 121, pp. e927–e935, 2008.

[74] D. C. Classen, R. C. Lloyd, L. Provost, F. a. Griffin, and R. Resar, “Development and Evaluation of the Institute for Healthcare Improvement Global Trigger Tool,” *J. Patient Saf.*, vol. 4, no. 3, pp. 169–177, Sep. 2008.

[75] D. a Kennerly, M. Saldaña, R. Kudyakov, B. da Graca, D. Nicewander, and J. Compton, “Description and evaluation of adaptations to the global trigger tool to enhance value to adverse event reduction efforts.,” *J. Patient Saf.*, vol. 9, no. 2, pp. 87–95, Jun. 2013.

[76] C. von Plessen, A. M. Kodal, and J. Anhøj, “Experiences with global trigger tool reviews in five Danish hospitals: an implementation study.,” *BMJ Open*, vol. 2, no. 5, pp. 1–8, Jan. 2012.

[77] P. Davis and New Zealand. Ministry of Health., “Adverse events in New Zealand

public hospitals : principal findings from a national survey," 2001.

[78] E. Deilkås, G. Bukholm, J. Lindstrøm, and M. Haugen, "Monitoring adverse events in Norwegian hospitals from 2010 to 2013," *BMJ Open*, 2015.

[79] S. C. Hartwig, S. D. Denger, and P. J. Schneider, "Severity-indexed, incident report-based medication error-reporting program.," *Am. J. Hosp. Pharm.*, vol. 48, pp. 2611–2616, 1991.

[80] R. J. Baines, M. Langelaan, M. C. de Bruijne, and C. Wagner, "Is researching adverse events in hospital deaths a good way to describe patient safety in hospitals: a retrospective patient record review study.," *BMJ Open*, vol. 5, no. 7, p. e007380, Jan. 2015.

[81] J. C. Benneyan, R. C. Lloyd, and P. E. Plsek, "Statistical process control as a tool for research and healthcare improvement.," *Qual. Saf. Health Care*, vol. 12, no. 6, pp. 458–64, Dec. 2003.

[82] J. C. Benneyan, R. C. Lloyd, and P. E. Plsek, "Statistical process control as a tool for research and healthcare improvement.," *Qual. Saf. Health Care*, vol. 12, pp. 458–464, 2003.

[83] "COSMIN checklist." [Online]. Available: www.cosmin.nl.

[84] K. Schildmeijer, L. Nilsson, J. Perk, K. Arestedt, and G. Nilsson, "Strengths and weaknesses of working with the Global Trigger Tool method for retrospective record review: focus group interviews with team members.," *BMJ Open*, vol. 3, p. e003131, 2013.

[85] D. C. Classen *et al.*, "Measuring Patient Safety: The Medicare Patient Safety Monitoring System (Past, Present, and Future).," *J. Patient Saf.*, 2016.

[86] J. Garrett, R. Paul, and C. Sammer, "Developing and Implementing a Standardized Process for Global Trigger Tool Application Across a Large Health System," *... J. Qual. ...*, vol. 39, no. 7, 2013.

[87] J. M. Naessens, T. J. O'Byrne, M. G. Johnson, M. B. Vansuch, C. M. McGlone, and J. M. Huddleston, "Measuring hospital adverse events: assessing inter-rater reliability and

trigger performance of the Global Trigger Tool.," *Int. J. Qual. Health Care,* vol. 22, no. 4, pp. 266–74, Aug. 2010.

[88] K. Mevik, F. Griffin, T. Hansen, E. Deilkås, and B. Vonen, "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," *BMJ Open*, 2016.

[89] R. Zimmerman *et al.*, "Aiming for zero preventable deaths: using death review to improve care and reduce harm.," *Healthc. Q.*, vol. 13 Spec No, pp. 81–7, Jan. 2010.

[90] E. J. Thomas, S. R. Lipsitz, D. M. Studdert, and T. A. Brennan, "The reliability of medical record review for estimating adverse event rates," *Ann Intern Med*, vol. 136, no. 11, pp. 812–816, 2002.

[91] P. D. Hibbert *et al.*, "The application of the Global Trigger Tool: a systematic review," *Int. J. Qual. Heal. Care*, vol. 28, no. 6, pp. 640–649, Sep. 2016.

[92] A. J. Forster, J. Andrade, and C. Van Walraven, "Validation of a discharge summary term search method to detect adverse events," *Journal of the American Medical Informatics Association*, vol. 12, no. 2. pp. 200–206, 2005.

[93] P. M. Kilbridge *et al.*, "Computerized Surveillance for Adverse Drug Events in a Pediatric Hospital," *J. Am. Med. Informatics Assoc.*, vol. 16, no. 5, pp. 607–612, 2009.

[94] D. C. Stockwell *et al.*, "Development of an Electronic Pediatric All-Cause Harm Measurement Tool Using a Modified Delphi Method.," *J. Patient Saf.*, vol. 00, no. 00, pp. 1–10, Aug. 2014.

[95] H. Rognebakke, "Kvalitetssikring av rapport om GTT-gjennomgang i norske sykehus," Oslo, 2011.

[96] E. von Elm *et al.*, "The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: Guidelines for Reporting Observational Studies," *PLoS Med.*, vol. 4, no. 10, p. e296, Oct. 2007.

[97] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data.," *Biometrics*, vol. 33, no. 1, pp. 159–74, Mar. 1977.

[98] K. Walshe, "Adverse events in health care: issues in measurement," *JAMA*, vol. 265,

no. 15, pp. 1993–1994, Mar. 2000.

[99] M. Fiszman, W. W. Chapman, D. Aronsky, R. S. Evans, and P. J. Haug, “Automatic Detection of Acute Bacterial Pneumonia from Chest X-ray Reports,” *J. Am. Med. Informatics Assoc.*, vol. 7, no. 6, pp. 593–604, Nov. 2000.

[100] R. G. Newcombe, “Two-sided confidence intervals for the single proportion: Comparison of seven methods,” *Stat. Med.*, vol. 17, no. 8, pp. 857–872, 1998.

[101] D. W. Baker and S. D. Persell, “Criteria for Waiver of Informed Consent for Quality Improvement Research,” *JAMA Intern. Med.*, vol. 175, no. 1, p. 142, Jan. 2015.

[102] Council for International Organizations of Medical Sciences, “International ethical guidelines for biomedical research involving human subjects.,” 2002.

[103] World Health Organization, “Ethical issues in Patient Safety Research interpreting existing guidance,” 2013.

[104] L. Harrington, “Quality improvement, research, and the institutional review board,” *J. Healthc. Qual.*, 2007.

[105] R. A. Hayward, M. Heisler, J. Adams, R. A. Dudley, and T. P. Hofer, “Overestimating outcome rates: Statistical estimation when reliability is suboptimal,” *Health Serv. Res.*, vol. 42, pp. 1718–1738, 2007.

[106] P. Michel, J. L. Quenon, A. M. de Sarasqueta, and O. Scemama, “Comparison of three methods for estimating rates of adverse events and rates of preventable adverse events in acute care hospitals.,” *BMJ*, vol. 328, no. 7433, p. 199, Jan. 2004.

[107] C. Sammer *et al.*, “Developing and Evaluating an Automated All-Cause Harm Trigger System,” *Jt. Comm. J. Qual. Patient Saf.*, vol. 0, no. 0, Feb. 2017.

[108] H. Lau and K. C. Litman, “Saving lives by studying deaths: Using standardized mortality reviews to improve inpatient safety,” *Joint Commission Journal on Quality and Patient Safety*, vol. 37, no. 9. pp. 400–408, 2011.

[109] V. B. Haukland Ellinor, Mevik Kjersti, von Plessen Chrisitan, Nieder Carsten, “The contribution of adverse events to death in hospitalised patients,” *Submitted*.

[110] P. J. Sharek *et al.*, “Performance characteristics of a methodology to quantify adverse events over time in hospitalized patients.,” *Health Serv. Res.*, vol. 46, no. 2, pp. 654–78, Apr. 2011.

[111] V. Good and M. Saldana, “Large-scale deployment of the Global Trigger Tool across a large hospital system: refinements for the characterisation of adverse events to support patient safety,” *BMJ Qual. …*, 2011.

[112] C. R. Blyth, “On Simpson’s Paradox and the Sure-Thing Principle,” *J. Am. Stat. Assoc.*, vol. 67, pp. 364–366, 1972.

[113] M. Zegers, M. C. de Bruijne, C. Wagner, P. P. Groenewegen, G. van der Wal, and H. C. W. de Vet, “The inter-rater agreement of retrospective assessments of adverse events does not improve with two reviewers per patient record.,” *J. Clin. Epidemiol.*, vol. 63, no. 1, pp. 94–102, Jan. 2010.

[114] J. Faber and L. M. Fonseca, “How sample size influences research outcomes.,” *Dental Press J. Orthod.*, vol. 19, no. 4, pp. 27–9, 2014.

[115] K. J. O’Leary *et al.*, “Comparison of traditional trigger tool to data warehouse based screening for identifying hospital adverse events.,” *BMJ Qual. Saf.*, vol. 22, no. 2, pp. 130–8, Feb. 2013.

[116] J. Ferranti *et al.*, “A Multifaceted Approach to Safety,” *J. Patient Saf.*, vol. 4, no. 3, pp. 184–190, Sep. 2008.

[117] D. C. Classen, “Computerized surveillance of adverse drug events in hospital patients,” *Qual. Saf. Heal. Care*, vol. 14, no. 3, pp. 221–226, Jun. 2005.

[118] A. K. Jha *et al.*, “Identifying Adverse Drug Events,” *J. Am. Med. Informatics Assoc.*, vol. 5, no. 3, pp. 305–314, 1998.

[119] D. W. Bates, R. S. Evans, H. Murff, P. D. Stetson, L. Pizziferri, and G. Hripcsak, “Detecting adverse events using information technology.,” *J. Am. Med. Inform. Assoc.*, vol. 10, no. 2, pp. 115–128, 2003.

[120] D. C. Stockwell, E. Kirkendall, S. E. Muething, E. Kloppenborg, H. Vinodrao, and B. R. Jacobs, “Automated adverse event detection collaborative: electronic adverse event

identification, classification, and corrective actions across academic pediatric institutions," *J Patient Saf*, vol. 9, no. 4, pp. 203–210, 2013.

[121] V. Lemon and D. C. Stockwell, "Automated Detection of Adverse Events in Children," *Pediatric Clinics of North America*, vol. 59, no. 6. pp. 1269–1278, 2012.

[122] H. J. Murff, V. L. Patel, G. Hripcsak, and D. W. Bates, "Detecting adverse events for patient safety research: a review of current methodologies," *J. Biomed. Inform.*, vol. 36, no. 1–2, pp. 131–143, Feb. 2003.

[123] G. B. Melton and G. Hripcsak, "Automated detection of adverse events using natural language processing of discharge summaries.," *J. Am. Med. Inform. Assoc.*, vol. 12, no. 4, pp. 448–57, Jan. 2005.

[124] C. M. Rochefort *et al.*, "Accuracy and generalizability of using automated methods for identifying adverse events from electronic health record data: a validation study protocol," *BMC Health Serv. Res.*, vol. 17, no. 1, p. 147, Dec. 2017.

[125] M. Benson, A. Junger, A. Michel, and G. Sciuk, "Comparison of manual and automated documentation of adverse events with an Anesthesia Information Management System (AIMS).," *Stud. Heal.*, 1999.

[126] L. I. Iezzoni *et al.*, "Identifying complications of care using administrative data.," *Med. Care*, vol. 32, no. 7, pp. 700–15, Jul. 1994.

[127] H. J. Murff, A. J. Forster, J. F. Peterson, J. M. Fiskio, H. L. Heiman, and D. W. Bates, "Electronically screening discharge summaries for adverse medical events," *J. Am. Med. Informatics Assoc.*, vol. 10, pp. 339–350, 2003.

[128] J. F. E. Penz, A. B. Wilcox, and J. F. Hurdle, "Automated identification of adverse events related to central venous catheters.," *J. Biomed. Inform.*, vol. 40, no. 2, pp. 174–82, Apr. 2007.

[129] D. W. Bates, R. S. Evans, H. Murff, P. D. Stetson, L. Pizziferri, and G. Hripcsak, "Detecting Adverse Events Using Information Technology," *J. Am. Med. Informatics Assoc.*, vol. 10, no. 2, pp. 115–128, Mar. 2003.

[130] C. M. Rochefort, A. D. Verma, T. Eguale, T. C. Lee, and D. L. Buckeridge, "A novel

method of adverse event detection can accurately identify venous thromboembolisms (VTEs) from narrative electronic health record data," *J. Am. Med. Informatics Assoc.*, vol. 45, no. (43), pp. 992–8, Oct. 2014.

[131] B. Hazlehurst, H. R. Frost, D. F. Sittig, and V. J. Stevens, "MediClass: A system for detecting and classifying encounter-based clinical events in any electronic medical record.," *J. Am. Med. Inform. Assoc.*, vol. 12, no. 5, pp. 517–29, Jan. 2005.

[132] B. Hazlehurst, A. Naleway, and J. Mullooly, "Detecting possible vaccine adverse events in clinical notes of the electronic medical record," *Vaccine*, vol. 27, no. 14, pp. 2077–2083, 2009.

[133] C. J. McDonald, "The barriers to electronic medical record systems and how to overcome them," *J Am Med Inf. Assoc*, vol. 4, no. 3, pp. 213–221, 1997.

[134] B. Kaplan, "Reducing barriers to physician data entry for computer-based patient records," *Top. Health Inf. Manage.*, vol. 15, no. 1, pp. 24–34, 1994.

[135] G. B. Melton and G. Hripcsak, "Automated detection of adverse events using natural language processing of discharge summaries.," *J. Am. Med. Inform. Assoc.*, vol. 12, no. 4, pp. 448–57, Jan. 2005.

[136] D. W. Bates, A. C. O'Neil, L. A. Petersen, T. H. Lee, and T. A. Brennan, "Evaluation of screening criteria for adverse events in medical patients.," *Med. Care*, vol. 33, no. 5, pp. 452–62, May 1995.

[137] Z. Obermeyer and E. J. Emanuel, "Predicting the Future - Big Data, Machine Learning, and Clinical Medicine.," *N. Engl. J. Med.*, vol. 375, no. 13, pp. 1216–9, Sep. 2016.

[138] F. Cabitza, R. Rasoini, and G. F. Gensini, "Unintended Consequences of Machine Learning in Medicine," *JAMA*, vol. 318, no. 6, p. 517, Aug. 2017.

[139] A. K. Jha *et al.*, "Identifying adverse drug events: Development of a computer-based monitor and comparison with chart review and stimulated voluntary report," *J. Am. Med. Informatics Assoc.*, vol. 5, no. 3, pp. 305–314, May 1998.

[140] A. Donabedian, "Evaluating the quality of medical care. 1966.," *Milbank Q.*, vol. 83, no. 4, pp. 691–729, 2005.

APPENDICES

Norwegian GTT Trigger sheet

Norwegian GTT Categories of harm and severity

Approval from the Regional Committee for Medical and Health and Research Ethics

Letter from the Norwegian Social Science Data Service

Approval from the Data Protection Office at the Nordland Hospital trust

Trial invitation letter

Informed consent form

Paper I-III

Trigger	Care module Triggers		Medication Module Triggers
C1	Transfusion or use of blood products	M1	Clostridium difficile positive stool
C2	Code/arrest/rapid response team	M3	INR greater than 6
C3	Acute dialysis	M4	Glucose less than 2.8 mmol/l
C4	Positive blood culture	M5	Rising BUN or serum creatinine greater than 2 times baseline
C5	X-ray or Doppler studies for emboli or DVT	M6*	Vitamin K administration
C6	Decrease of greater than 25% in hemoglobin or hematocrit	M7*	Benadryl (Diphenhydramine) use
C7	Patient fall	M8*	Romazicon (Flumazenil) use
C8	Pressure ulcers	M9*	Naloxone (Narcan) use
C9	Readmission within 30 days	M10*	Anti-emetic use
C10*	Restraint use	M11*	Over-sedation/hypotension
C11	Healthcare-associated infection	M12*	Abrupt medication stop
C12	In-hospital stroke	M13*	Other
C13	Transfer to higher level of care		**Intensive Care Module Triggers**
C14	Any procedure complication	I1	Pneumonia onset
C15*	Other	I2	Readmission to intensive care
	Surgical Module Triggers	I3	In-unit procedure
S1	Return to surgery	I4	Intubation/reintubation
S2	Change in procedure		**Perinatal Module Triggers**
S3	Admission to intensive care post-op	P1*	Terbutaline use
S4	Intubation/reintubation/BiPap in PACU	P2	3rd- or 4th-degree lacerations
S5*	X-ray intra-op or in PACU	P3	Platelet count less than 50,000
S6	Intra-op or post-op death	P4	Estimated blood loss > 500 ml (vaginal) or > 1,000 ml (C-section)
S7	Mechanical ventilation greater than 24 hours post-op	P5	Specialty consult
S8*	Intra-op epinephrine, norepinephrine, naloxone, or romazicon	P6*	Oxytocic agents
S9	Post-op troponin level greater than 40 ng/l	P7	Instrumented delivery
S10	Injury, repair, or removal of organ because of accidental injury	P8	General anesthesia
S11	Change in anesthesia procedure	P9	Apcar score <7 after 5 minute
S12	Insertion of artery catheter or central venous catheter	P10	Induced labour
S13	Surgery more than 6 hours		**Emergency Department Module Triggers**
S14*	Any operative complication	E1	Readmission to ED within 48 hours
		E2	Time in ED greater than 6 hours

***Non-Automatic Triggers**

Severity of adverse event

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

Type of adverse event

Hospital acquired infections

- Urinary tract infection
- CVC infection
- Ventilator associated pneumonia
- Other infection
- Lower respiratory infection

Surgical complications

- Infection after surgery
- Respiratory complications after surgery
- Return to surgery
- Injury, repair or removal of organ
- Occurrence of any operative complication
- Switch in surgery

Bleeding/thrombosis

- Thrombosis/Embolism
- Bleeding
- Bleeding after surgery

Patient fall /fracture

- Patient fall
- Fracture

Other

- Other
- Allergy
- Medical technical harm
- Deterioration and chronic illness
- Medication harm
- Obstetric harm
- Pressure ulcer

REK
REGIONALE KOMITEER FOR MEDISINSK OG HELSEFAGLIG FORSKNINGSETIKK

Region:	Saksbehandler:	Telefon:	Vår dato:	Vår referanse:
REK vest	Arne Salbu	55978498	07.11.2012	2012/1691/REK vest
			Deres dato:	Deres referanse:
			25.09.2012	

Vår referanse må oppgis ved alle henvendelser

Barthold Vonen
Nordlandssykehuset HF
Prinsens gt 164
8000 Bodø

2012/1691 Validering av Global Trigger Tool som målemetode for kartlegging av pasientskader

Forskningsansvarlig: Nordlandssykehuset HF
Prosjektleder: Barthold Vonen

Vi viser til søknad om forhåndsgodkjenning av ovennevnte forskningsprosjekt. Søknaden ble behandlet av Regional komité for medisinsk og helsefaglig forskningsetikk (REK vest) i møtet 18.10.2012. Vurderingen er gjort med hjemmel i helseforskningsloven § 10, jf. forskningsetikklovens § 4.

Prosjektomtale
Global Trigger Tool (GTT) er en metode som alle helseforetak er pålagt å bruke for å kartlegge pasientskader i egen virksomhet. I dette prosjektet ønskes det å undersøke om GTT metoden er et robust og sensitivt verktøy brukt på dagens elektroniske pasientjournaler. Studien innebærer for det første to delstudier på selve GTT-metoden hvor det skal benyttes registerdata. I en tredje delstudie skal det gjøres journalgjennomganger fra minimum 240 tilfeldig utplukkede sykehusopphold. Her skal en bruke GTT metoden for å kartlegge evt pasientskader/uønskede hendelser. Det skal hentes inn samtykke for journalgjennomgangene. Det søkes om fritak fra samtykkekravet for bruken av registerdataene (utenom journalene).

Vurdering
Søknad/protokoll
Det hersker stor usikkerhet omkring spørsmålet om både omfang og alvorlighetsgrad av pasientskader ved norske sykehus. Både for pasientenes egen del og for samfunnets evne til å foreta nødvendige prioriteringer innenfor helsevesenet, er det svært viktig at det finnes gode og sammenlignbare oversikter over pasientskadene. Som politisk tema er dette også høyaktuelt. Derfor er dette en søknad REK Vest mener er svært viktig. Komiteen mener også at protokollen er egnet til å besvare de spørsmål en reiser.

Rekruttering/samtykke
Datamateriale hentes fra to kilder:

1.
Registerdata: Data fra GTT ved 7 enheter ved Norlandssykehuset i perioden 2010. Det gjennomgås årlig ca 1680 pasientopphold. Disse opplysningene skal anonymiseres og overføres til en database. Det søkes om fritak fra samtykkekravet for disse pasientene. Dette er begrunnet med at dataene skal konverteres til en

forskningsdatabase hvor koblingsnøkkel er fjernet. Videre opplyses det at arbeidet med å etablere forskningsdatabasen gjøres av personell utenfor selve forskningsprosjektet og i regi av forskningsansvarlig.

Adgang til bruk av helseopplysninger som er innsamlet i helsetjenesten til forskning er regulert i helseforskningslovens § 35. Vilkårene for å kunne tillate dette uten innhenting av samtykke, er at forskningen skal være av vesentlig interesse for samfunnet og at hensynet til deltakernes velferd og integritet er ivaretatt.

REK Vest mener at samfunnsnytten er godt dokumentert. Slik en har lagt opp anonymiseringsprosessen, mener komiteen at hensynet til deltakernes velferd og integritet også er godt ivaretatt. REK Vest vil godkjenne søknaden på dette punkt.

2.
Nye helseopplysninger: Journalgjennomgang av 240 tilfeldig utplukkede sykehusopphold ved Nordlandssykehuset. Denne delen er samtykkebasert.

Det vedlagte utkast til forespørsel er imidlertid av dårlig kvalitet. En må bestrebe seg på å benytte et mer allment tilgjengelig språk hvor det er på en enklere måte beskrives hva deltakelse innebærer. REK Vest ønsker å få det reviderte skrivet tilsendt, før endelig vedtak fattes.

Informasjonssikkerhet
Det opplyses at koblingsnøkkel oppbevares ved egen institusjon og at personidentifiserbare opplysninger oppbevares på institusjonens server. REK Vest forutsetter at koblingsnøkkel og personidentifiserbare opplysninger oppbevares separat.

Vedtak
Søken utsettes i påvente av tilbakemelding på ovennevnte merknad.

Vennligst benytt skjema for tilbakemelding som sendes inn via saksportalen til REK http://helseforskning.etikkom.no.

Med vennlig hilsen

Jon Lekven
komitéleder, dr.med.

Arne Salbu
rådgiver

Kopi til: *kso@nlsh.no*

REK
REGIONALE KOMITEER FOR MEDISINSK OG HELSEFAGLIG FORSKNINGSETIKK

Region:	Saksbehandler:	Telefon:	Vår dato:	Vår referanse:
REK vest	Øyvind Straume	55978497	11.12.2012	2012/1691/REK vest
			Deres dato:	**Deres referanse:**
			26.11.2012	

Vår referanse må oppgis ved alle henvendelser

Barthold Vonen

2012/1691 Validering av Global Trigger Tool som målemetode for kartlegging av pasientskader

Forskningsansvarlig: Nordlandssykehuset HF
Prosjektleder: Barthold Vonen

Vi viser til tilbakemelding om forhåndsgodkjenning av ovennevnte forskningsprosjekt. Tilbakemeldingen ble behandlet av leder av REK Vest på fullmakt. Vurderingen er gjort med hjemmel i helseforskningsloven § 10, jf. forskningsetikkloven § 4.

Vurdering:

Tilbakemelding
REK Vest krevde at informasjonsskrivet ble forfattet i et mer allment tilgjengelig språk. Et revidert skriv foreligger nå.

Ny vurdering i REK
REK Vest finner det nye informasjonsskrivet tilfredsstillende og har ingen ytterligere innvendinger til prosjektsøknad.

Vedtak:
REK Vest godkjenner prosjektet i samsvar med søknad og tilbakemelding.

Sluttmelding og søknad om prosjektendring
Prosjektleder skal sende sluttmelding til REK vest på eget skjema senest 30.06.2016, jf. hfl. §12.
Prosjektleder skal sende søknad om prosjektendring til REK vest dersom det skal gjøres vesentlige endringer i forhold til de opplysninger som er gitt i søknaden, jf. hfl. § 11.

Klageadgang
Du kan klage på komiteens vedtak, jf. forvaltningslovens § 28 flg. Klagen sendes til REK vest. Klagefristen er tre uker fra du mottar dette brevet. Dersom vedtaket opprettholdes av REK vest, sendes klagen videre til Den nasjonale forskningsetiske komité for medisin og helsefag for endelig vurdering.

Med vennlig hilsen

Jon Lekven
komiteleder

Øyvind Straume
seniorkonsulent

Kopi til: postmottak@nlsh.no

Besøksadresse:
Haukeland Universitetssykehus, Sentralblokken, 2. etg, Rom 4617

Telefon: 55975000
E-post: rek-vest@uib.no
Web: http://helseforskning.etikkom.no/

REK
REGIONALE KOMITEER FOR MEDISINSK OG HELSEFAGLIG FORSKNINGSETIKK

Region:	Saksbehandler:	Telefon:	Vår dato:	Vår referanse:
REK vest	Anna Stephansen	55978496	02.03.2018	2012/1691/REK vest
			Deres dato:	
			05.02.2018	

Vår referanse må oppgis ved alle henvendelser

Barthold Vonen
SKDE

2012/1691 Validering av Global Trigger Tool som målemetode for kartlegging av pasientskader

Forskningsansvarlig: Nordlandssykehuset HF, Nordlandssykehuset HF
Prosjektleder: Barthold Vonen

Vi viser til søknad om prosjektendring datert 05.02.2018 for ovennevnte forskningsprosjekt. Søknaden er behandlet av leder for REK vest på fullmakt, med hjemmel i helseforskningsloven § 11.

Vurdering
REK vest omfatter det slik at prosjektendringen innebærer ikke innsamling av nye data. Det er testing av validiteten til GTT som er formålet med prosjektendringen. Videre søker prosjektlederen om forlengelse av prosjektet til 05.07.2018.

Vurdering:
REK vest merker seg at prosjektet er gått ut på dato 31.08.2017. Vi gjør oppmerksom på at søknad om forlengelse av prosjektet skal sendes inn før prosjektsluttdato.

Vedtak
REK vest godkjenner prosjektendringen i samsvar med forelagt søknad.

Klageadgang
Du kan klage på komiteens vedtak, jf. helseforskningsloven § 10 og forvaltningsloven § 28 flg. Klagen sendes til REK vest. Klagefristen er tre uker fra du mottar dette brevet. Dersom vedtaket opprettholdes av REK vest, sendes klagen videre til Den nasjonale forskningsetiske komité for medisin og helsefag for endelig vurdering.

Med vennlig hilsen

Marit Grønning
dr.med.
Avdelingsdirektør, professor

Anna Stephansen
sekretariatsleder

Kopi til: *postmottak@nlsh.no; postmottak@nlsh.no*

Besøksadresse:
Armauer Hansens Hus (AHH), Tverrfløy Nord, 2 etasje. Rom 281. Haukelandsveien 28

Telefon: 55975000
E-post: post@helseforskning.etikkom.no
Web: http://helseforskning.etikkom.no/

Norsk samfunnsvitenskapelig datatjeneste AS
NORWEGIAN SOCIAL SCIENCE DATA SERVICES

Harald Hårfagres gate 29
N-5007 Bergen
Norway
Tel: +47-55 58 21 17
Fax: +47-55 58 96 50
nsd@nsd.uib.no
www.nsd.uib.no
Org.nr. 985 321 884

Kjersti Mevik
Nordlandssykehuset
Serviceboks
8375 LEKNES

Vår dato: 01.12.2014 Vår ref: 40442/3/IB/LR Deres dato: Deres ref:

AVSLUTTET SAKSBEHANDLING

Vi viser til meldeskjema mottatt 23.10.2014 for prosjektet:

40442 *Validering av Global trigger tool som metode for kartlegging av pasientskader*

Vi viser også til REK-godkjenningen som var vedlagt meldeskjema, og telefonsamtale 28.11.14 med bekreftelse på at REK-godkjenningen dekker hele prosjektet.

REK har vedtatt at prosjektet faller inn under helseforskningslovens bestemmelser. REK sin godkjenning er tilstrekkelig for behandling av personopplysninger i prosjektet.

Personvernombudet avslutter dermed saksbehandlingen av meldingen uten å realitetsbehandle denne. Vi avslutter også all videre oppfølgning av prosjektet.

Ta gjerne kontakt dersom noe er uklart.

Vennlig hilsen

Katrine Utaaker Segadal

Inga Brautaset

Kontaktperson: Inga Brautaset inga.brautaset@nsd.uib.no

Kopi: Institutt for klinisk medisin, UiT Norges arktiske universitet

NORDLANDSSYKEHUSET
NORDLÁNDA SKIHPPIJVIESSO

PERSONVERNOMBUD

Vår ref: 12.17. Saksbehandler: Alisa Larsen Dato: 26.06.17

ANBEFALING AV BEHANDLING AV PERSONOPPLYSNINGER

Viser til melding om behandling av personopplysninger, mottatt 21.06.

Tittel: Validering av GTT som målemetode for kartlegging av pasientskader

Formål med prosjektet: Å teste verktøyet GTT som brukes til kartlegge pasientskader. Metoden går ut på å screene pasientjournaler etter utvalgte triggere (lab verdi, fall, infeksjoner som kan oppstå i pasientforløpet) som kan indikere at en pasientskade har skjedd. Målet med studien er å finne den optimale utvalgsstørrelsen som trengs for å estimere antall skader, om utskiftning av de som screener påvirker resultatet og om ett automatisk verktøy kan erstatte den manuelle granskningen

Tidspunkt for prosjektet (til/fra): 01.01.2013 – 31.12.17.

Forskningsprosjektet krever forhåndsgodkjenning av REK. Personvernombudets (PVO) rolle er å ha oversikt over forskningsprosjekter samt se til at informasjonssikkerheten og personvernet blir ivaretatt.

Det forutsettes at prosjektet gjennomføres i tråd med de opplysningene som er gitt i selve meldingen samt i øvrig korrespondanse og samtaler. Videre forutsettes det at bestemmelsene i lov om behandling av personopplysninger og lov om helseregistre og behandling av helseopplysninger med forskrifter følges. Prosjektet må videre gjennomføres i henhold til annet relevant regelvert, herunder de alminnelige regler om taushetsplikt.

- Dersom registeret skal brukes til annet formål enn det som er nevnt i meldingen må det meldes særskilt i hvert enkelt tilfelle.
- Dersom prosjektet har varighet på mer enn tre år skal prosjektansvarlig hvert tredje år sende bekreftelse til personvernombud på at behandlingen skjer i overensstemmelse med søknaden og vilkårene som er nevnt i denne godkjennelsen.
- Det skal gis tilbakemelding til personvernombudet når registret er slettet.

Med hjemmel i personopplysningslovens forskrift § 7-12 godkjennes det at behandlingen av personopplysningene kan gjennomføres med de vilkårene som nevnt ovenfor.

Postadresse:
Nordlandssykehuset HF, Pb 1480,
8092 BODØ

Med hilsen
NORDLANDSSYKEHUSET HF

Alisa Larsen
Informasjonssikkerhetsrådgiver/Personvernombud

Vedlegg 1

Vedlegg – forskningsprosjekt

Helseforskningsloven

§ 10.Søknad om forhåndsgodkjenning
Søknad om forhåndsgodkjenning av et forskningsprosjekt skal sammen med forskningsprotokollen sendes til den regionale komiteen for medisinsk og helsefaglig forskningsetikk.

Den regionale komiteen for medisinsk og helsefaglig forskningsetikk skal foreta en alminnelig forskningsetisk vurdering av prosjektet, og vurdere om prosjektet oppfyller kravene stilt i denne loven eller i medhold av denne loven. Den regionale komiteen for medisinsk og helsefaglig forskningsetikk kan sette vilkår for godkjenning.

Vedtak vedrørende forhåndsgodkjenning kan påklages til Den nasjonale forskningsetiske komité for medisin og helsefag, jf. lov 30. juni 2006 nr. 56 om behandling av etikk og redelighet i forskning § 4.

Departementet kan gi forskrifter om krav til søknaden, om saksbehandlingsfrister for den regionale komiteen for medisinsk og helsefaglig forskningsetikk, og om de nærmere vilkårene for forhåndsgodkjenning

Forskrift om behandling av personopplysninger

§ 7-12.Personvernombud
Datatilsynet kan samtykke i at det gjøres unntak fra meldeplikt etter personopplysningsloven § 31 første ledd, dersom den behandlingsansvarlige utpeker et uavhengig personvernombud som har i oppgave å sikre at den behandlingsansvarlige følger personopplysningsloven med forskrift. Personvernombudet skal også føre en oversikt over opplysningene som nevnt i personopplysningsloven § 32.

Forespørsel om deltakelse i forskningsprosjektet:

Validering av Global Trigger Tool som målemetode for kartlegging av pasientskader

Bakgrunn og hensikt

Forskningsprosjektet skal undersøke om en ved bruk av GTT – Global Trigger tool - kan finne og dokumentere uønskede hendelser og skader på pasienter som følge av behandling i norske sykehus. Alle norske helseforetak er pålagt å bruke GTT-metoden for å kartlegge pasientskader i egen virksomhet. Antall, type og alvorlighet ved pasientskader rapporteres regelmessig til et sentralt register og offentliggjøres. I vår studie skal vi undersøke om GTT-metoden gir pålitelig data også når det brukes på dagens elektroniske pasientjournaler. For å kunne si noe om dette, ønsker vi å analysere data fra 240 pasientopphold ved Nordlandssykehuset med tanke på antall, type og alvorlighetsgrad av mulige behandlingsrelaterte skader og uønskede hendelser. Vi ber med dette om din tillatelse til å bruke journaldata fra opphold ved Nordlandssykehuset i dette arbeidet.

Hva innebærer studien?

Studien foregår ved at 240 sykehusopphold ved Nordlandssykehuset HF i perioden 01.01.2010 – 31.12.2010 trekkes tilfeldig av det totale antallet innleggelser ved sykehuset i samme periode. Ditt opphold ved Nordlandssykehuset i perioden (dato fylles inn) er trukket ut. Vi ber med dette om din tillatelse til at journaldata fra dette oppholdet kan gjennomgås av 1 lege og 1 sykepleier fra prosjektgruppa for å finne ut om det inntraff uønskede hendelser og om du ble påført skader. Er du pårørende ber vi om at du gir samtykke på vegne av pasienten.

Mulige fordeler og ulemper

For deg som pasient innebærer studien ingen ulemper eller direkte fordeler. Hvis du samtykker til denne undersøkelsen vil helsepersonell som deltar i dette forskningsprosjektet få innsyn i din pasientjournal. Finner vi at du har opplevd en alvorlig uønsket hendelse eller blitt skadet som følge av behandlingen, vil du bli kontaktet og informert om dette og du vil få tilbud om samtale med en av de som har gjennomgått journalen din. Vi vil ikke lete etter eventuelle nye lidelser/diagnoser, men kun vurdere om det forelå en pasientskade eller uønsket hendelse som følge av behandlingen du mottok i løpet av det aktuelle oppholdet.

Hva skjer med informasjonen om deg?

Informasjonen som registreres om deg skal kun brukes slik som beskrevet ovenfor. Når alle data fra sykehusoppholdet er gjennomgått, blir eventuelle skader eller uønskede hendelser registrert og lagret atskilt fra journalen din uten ditt navn, fødselsnummer eller andre direkte eller indirekte identifiserbare opplysninger (anonymisert). Det vil heller ikke være mulig å identifisere de enkelte deltagere i de publiserte resultatene av studien. Dersom vi senere ønsker å bruke de opplysningene vi har samlet inn til et annet forskningsprosjekt, vil du bli forespurt og videre bruk forutsetter at du samtykker også til det.

1.1 Personvern

Opplysninger som ønskes registrert om deg skal hentes fra Nordlandssykehuset elektroniske journalsystem. I vår studie skal dette ikke koples til andre lokale/nasjonale registre eller bli overlatt til andre forskere. Nordlandssykehuset ved administrerende direktør Paul Martin Strand er ansvarlig for håndtering og lagring av data.

1.2 Rett til innsyn og sletting av opplysninger om deg

Hvis du sier ja til å delta i studien, har du rett til å få innsyn i hvilke opplysninger som er registrert om deg. Du har videre rett til å få korrigert eventuelle feil i de opplysningene vi har registrert. Dersom du trekker deg fra studien, kan du kreve å få slettet alle innsamlede data, med mindre opplysningene allerede er benyttet i analyser eller i vitenskapelige publikasjoner. Studien er finansiert gjennom forskningsmidler fra Helse Nord og resultatene fra studien blir publisert i nasjonale og internasjonale fagtidsskrifter.

Frivillig deltakelse

Det er frivillig å delta i studien. Du kan når som helst og uten å oppgi noen grunn trekke ditt samtykke til deltagelse uten at dette vil noen konsekvenser for deg i din fremtidige kontakt med Nordlandssykehuset. Dersom du ønsker å delta, undertegner du samtykkeerklæringen på siste side snarest mulig og returnerer dette i vedlagt konvolutt. Dersom du senere ønsker å trekke deg eller har spørsmål til studien, kan du kontakte:

Tittel: LIS lege kir avdeling Kjersti Mevik, kjersti.mevik@nlsh.no

Telefon:75534000

Mvh

Kjersti Mevik

Stipendiat phD/LIS lege kir avd

Barthold Vonen

Prosjektleder/medisinsk direktør

Samtykke til deltakelse i studien *Validering av Global Trigger Tool som målemetode for kartlegging av pasientskader*

Jeg har lest prosjektbeskrivelsen ovenfor og gir samtykke til at journaldata fra min eller en annens person (i de tilfeller hvor samtykke gis av pårørende) brukes i studien:

(navn, dato, sted)

Bruk vedlagt ferdig frankert svarkonvolutt.

Paper I

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

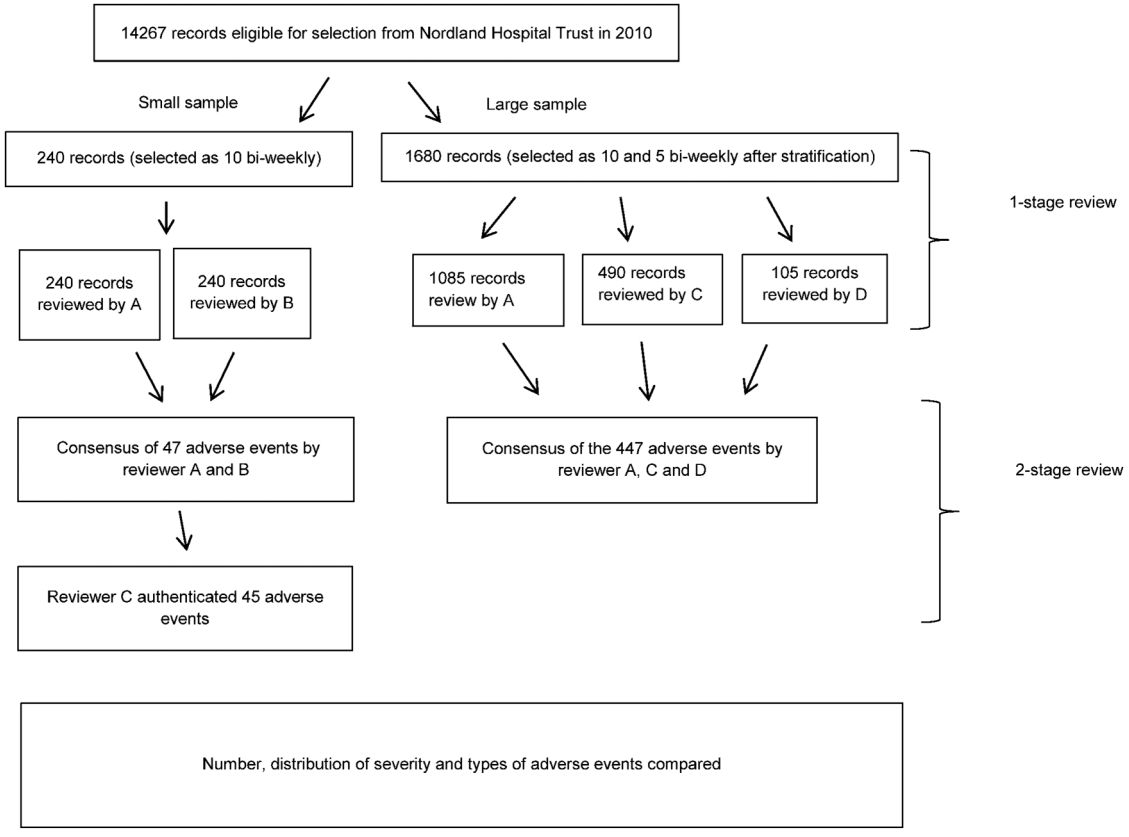


Figure 1 Overview of the study design. Non-physician reviewers; reviewer A and B, physician reviewers; reviewer C and D.

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		
n	Large sample 1680	Small sample 240	Overall population 14 267	Large vs small sample	Large vs overall population	Small vs overall population
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

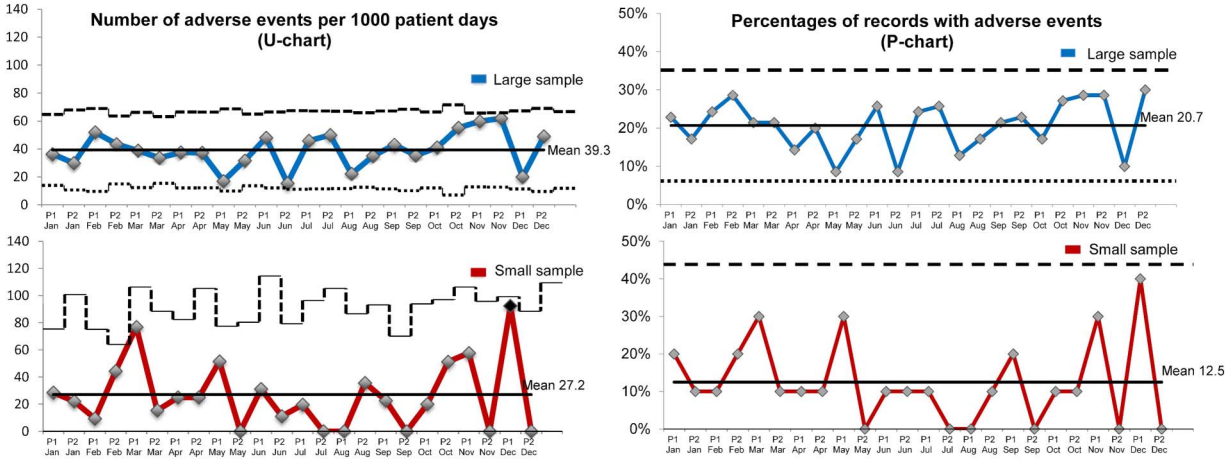


Figure 2 Comparison of statistical process control charts (U-chart) and (P-chart) between large and small samples. Dashed line = upper control limits; dotted line = lower control limits.

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.

Paper II

ISQua®

International Journal for Quality in Health Care, 2016, 28(4), 492–496
doi: 10.1093/intqhc/mzw054
Advance Access Publication Date: 9 June 2016
Article

OXFORD

Article

Is inter-rater reliability of Global Trigger Tool results altered when members of the review team are replaced?

KJERSTI MEVIK[1], FRANCES A. GRIFFIN[2], TONJE ELISABETH HANSEN[1], ELLEN DEILKÅS[3], and BARTHOLD VONEN[4]

[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, and [4]Nordland Hospital Trust, Post box 1480, 8092 Bodø, Norway and Institute for Community Medicine, The Artic University of Norway, Post box 6050, Langnes 9037 Tromsø, Norway

Address reprint requests to: Kjersti Mevik, Nordland Hospital Trust, Post box 1480, N-8092 Bodø, Norway.
Tel: +47 75534000; Fax: +47 75534111; E-mail: kjersti.mevik@nlsh.no

Accepted 15 April 2016

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: A total of 120 records were selected as biweekly samples of 10 from discharge lists between 1 July and 31 December 2010 for a 3-fold review.

Intervention: Replacement of review team members was tested to assess impact on inter-rater reliability and adverse events measurment.

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

Results: Substantial inter-rater reliability regarding the presence and severity level of adverse events was obtained between Teams I and II, while moderate inter-rater reliability was obtained between Teams I and 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.

Key words: 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 (GTT) 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 predefined 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 GTT 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 the presence and severity of identified adverse with the GTT.

METHODS

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 with 43 750 patient days in the period from July to December 2010. A total of 120 inpatient records were obtained by selecting 10 records randomly from the hospital discharge lists biweekly for the period of 1 July to 31 December 2010. Due to resources available, we found that 120 records selected from a 6-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 <24 h (to avoid any patients for observation) and <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 GTT [3] was applied with the adapted 57 triggers in the Norwegian translation (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 min per record. Examples of triggers included a given procedure, a laboratory 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 GTT 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 on a regular basis and internal to the trust. They were experienced with the GTT method, having previously used the GTT for at least 2 years. No additional training was done just before the 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.

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

Study design

The records were reviewed using the hospital's electronic patient journal system in sets of 10 records from each biweekly 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 the presence and severity level of adverse events identified by the GTT method.

Statistical analysis

To describe characteristics of the records, descriptive statistics were used presented as frequencies, mean values, medians and ranges. The level of agreement between Teams I and II and between Teams I and III in terms of inter-rater reliability was assessed using kappa statistic for nominal data (agreement on the 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 were 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 analyses were 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 (standard deviation (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). A total of 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 (Fig. 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 the findings of Team I are listed. There was disagreement in four records between Teams I and II and in seven records between Teams I and 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.

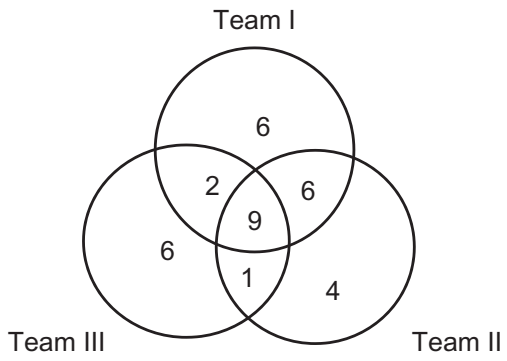


Figure 1 Venn diagram of number of adverse events identified by Teams I, II and III.

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.

Inter-rater reliability

Agreements were substantial on the presence of adverse events between Teams I and II and moderate between Teams I and III (Table 4). The agreement in terms of number of adverse events and severity levels was substantial between Teams I and II and moderate between Teams I and 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 the 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 during 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

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[a]	0.640 (0.434–0.846)	0.468 (0.232–0.703)
Number of adverse events[b]	0.661 (0.479–0.842)	0.468 (0.278–0.694)
Severity level[b]	0.652 (0.469–0.836)	0.442 (0.260–0.624)

[a]Unweighted kappa analysis.
[b]Weighted kappa analysis.
CI, confidence interval.

in Team II in neither of the biweekly 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 depends on a subjective 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 are generalizable to other users of the IHI GTT. 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.

Acknowledgements

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 T.E.H. Frank Federico and Carol Haraden for the development of the study design and Alexander Ringdal for help with organizing the data.

Funding

This work was supported by the North Norwegian Health Authority as a PhD grant to 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 (4 March 2016, date last accessed).
3. Griffin F, Resar R. IHI Global Trigger Tool for measuring adverse events. IHI Innovation Series White Paper 2007;1–44. http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:IHI+Global+Trigger+Tool+for+Measuring+Adverse+Events#0 (26 November 2014, date last accessed).
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 (20 July 2014, date last accessed).

Appendix

Trigger	Care module triggers
C1	Transfusion or use of blood products
C2	Code/arrest/rapid response team
C3	Acute dialysis
C4	Positive blood culture
C5	X-ray or Doppler studies for emboli or DVT
C6	Decrease of >25% in hemoglobin or hematocrit
C7	Patient fall
C8	Pressure ulcers
C9	Readmission within 30 days
C10	Restraint use
C11	Healthcare-associated infection
C12	In-hospital stroke
C13	Transfer to higher level of care
C14	Any procedure complication
C15	Other
	Surgical module triggers
S1	Return to surgery
S2	Change in procedure
S3	Admission to intensive care post-op
S4	Intubation/reintubation/BiPap in PACU
S5	X-ray intra-op or in PACU
S6	Intra-op or post-op death
S7	Mechanical ventilation >24 hours post-op
S8	Intra-op epinephrine, norepinephrine, naloxone, or romazicon
S9	Post-op troponin level >40 ng/l
S10	Injury, repair or removal of organ because of accidental injury
S11	Change in anesthesia procedure
S12	Insertion of artery catheter or central venous catheter
S13	Surgery more than 6 hours
S14	Any operative complication
	Medication module triggers
M1	*Clostridium difficile* positive stool
M3	INR > 6
M4	Glucose <50 mg/dl
M5	Rising BUN or serum creatinine >2 times baseline
M6	Vitamin K administration
M7	Benadryl (Diphenhydramine) use
M8	Romazicon (Flumazenil) use
M9	Naloxone (Narcan) use
M10	Antiemetic use
M11	Over-sedation/hypotension
M12	Abrupt medication stop
M13	Other
	Intensive care module triggers
I1	Pneumonia onset
I2	Readmission to intensive care
I3	In-unit procedure
I4	Intubation/reintubation
	Perinatal module triggers
P1	Terbutaline use
P2	Third- or fourth-degree lacerations
P3	Platelet count <50,000
P4	Estimated blood loss >500 ml (vaginal) or >1000 ml (C-section)
P5	Specialty consult
P6	Oxytocic agents
P7	Instrumented delivery
P8	General anesthesia
P9	Apcar score <7 after 5 minutes
P10	Induced labour
	Emergency Department Module Triggers
E1	Readmission to ED within 48 hours
E2	Time in ED >6 hours

Paper III

International Journal for Quality in Health Care, 2018, 1–6
doi: 10.1093/intqhc/mzy210
Article

Article

Is a modified Global Trigger Tool method using automatic trigger identification valid when measuring adverse events?

A comparison of review methods using automatic and manual trigger identification

KJERSTI MEVIK[1], TONJE E. HANSEN[2], ELLEN C. DEILKÅS[3], ALEXANDER M. RINGDAL[4] and BARTHOLD VONEN[5,6]

[1]Department of surgery, Nordland Hospital Trust, Post box 1480, N-8092 Bodø, Norway, [2]Nordland Hospital Trust, PO 1480, N-8092 Bodø, Norway, [3]Unit for Health Service Research, Akershus University Hospital, N-1478 Lørenskog, Norway, [4]Division of informatics, Nordland Hospital Trust, Post box 1480, N-8092 Bodø, Norway, [5]Center for Clinical documentation and Evaluation, North Norway Regional Health Trust, N-9038 Tromsø, Norway, and [6]Institute for community medicine, The Artic University of Norway, N-9037 Tromsø, Norway

Address reprint requests to: Kjersti Mevik, Nordland Hospital Trust, PO 1480, N-8092 Bodø, Norway. Tel: +4797123977; E-mail: Kjersti.mevik@nlsh.no

Editorial Decision 27 August 2018; Accepted 14 September 2018

Abstract

Objectives: To evaluate a modified Global Trigger Tool (GTT) method with manual review of automatic triggered records to measure adverse events.
Design: A cross-sectional study was performed using the original GTT method as gold standard compared to a modified GTT method.
Setting: Medium size hospital trust in Northern Norway.
Participants: One thousand two hundred thirty-three records selected between March and December 2013.
Main outcome measure: Records with triggers, adverse events and number of adverse events identified. Recall (sensitivity), precision (positive predictive value), specificity and Cohen's kappa with 95 % confidence interval were calculated.
Results: Both methods identified 35 adverse events per 1000 patient days. The modified GTT method with manual review of 658 automatic triggered records identified adverse events ($n = 214$) in 189 records and the original GTT method identified adverse events ($n = 216$) in 186 records. One hundred and ten identical records were identified with adverse events by both methods. Recall, precision, specificity and reliability for records identified with adverse events were respectively 0.59, 0.58, 0.92 and 0.51 for the modified GTT method. The total manual review time in the modified GTT method was 23 h while the manual review time using the original GTT method was 411 h.
Conclusions: The modified GTT method is as good as the original GTT method that complies with the GTTs aim monitoring the rate of adverse events. Resources saved by using the modified GTT method enable for increasing the sample size. The automatic trigger identification system may be developed to assess triggers in real-time to mitigate risk of adverse events.

Key words: Global Trigger Tool, automatic trigger identification, adverse events, record review, hospital care

Introduction

Identifying and measuring adverse events is important as they entail substantial burden to patients and health providers [1]. In addition, the economic burden of adverse events is considerable [2]. Adverse events have commonly been identified through voluntary incident reporting but this approach significantly underestimates the actual number of adverse events as it relies on healthcare providers willingness and opportunity to report [3]. Hence, trigger tools, first described by Jicks [4] and refined by Classen *et al.* [5], were developed to identify and measure adverse events. Patient records are screened for specific elements (triggers) in the records. Once a trigger is identified a more in-depth review is performed to determine if an adverse event may have occurred [6]. The trigger search is performed in randomly selected records, usually a limited number that is manageable [7]. The Institute for Healthcare Improvement (IHI) refined the trigger tools further and developed the Global Trigger Tool (GTT) which has successfully been advocated with the aim to monitor adverse events in adult inpatients [8]. The GTT is an easy two-step method of retrospective manual review of record samples: Two primary reviewers (nurses) individually review the records for specific triggers and determine if the triggers represent any adverse events, before reaching consensus (Step 1). A secondary reviewer (physician) authenticates their findings (Step 2) [8]. In Norway, all hospitals are instructed by the commissioning documents from the ministry of health to perform the GTT [9].

Many have considered the GTT as the best method to identify and measure adverse events. Results from the GTT demonstrates that one of five hospitalized patients experience at least one adverse event [10–12]. However, the practical disadvantages of the GTT, being resource-intensive due to time and personnel required, limits widespread use and adoption. Automatic identification of triggers in electronic health records (EHRs) provides a digital, standardized and cost-effective approach to measure adverse events [13]. Rather than a reviewer searches for triggers, algorithms are written to automatically identify triggers. The benefits are promising, once the algorithms are written, as manual review is only performed in the automatic triggered records [14–16]. However, the validity of automatic systems in comparison to other methods measuring adverse events varies [12–14].

We developed an automatic trigger identification system that identifies 42 of the GTT triggers. We included the system in a modified GTT method where manual review to identify adverse events was limited to only automatic triggered records, illustrated in Fig. 1. We considered the original GTT method with all manual review steps as the gold standard. This study aimed to evaluate the modified GTT methods ability to identify and measure adverse events using the original GTT method as a reference standard.

Methods

Study design

The study is an explorative cross-sectional study comparing a modified GTT method to the original GTT method to identify and measure adverse events.

Setting

The study was performed at a 524-bed trust with three hospitals in Nordland County, Northern Norway. The trust has approximately 14 000 discharges and 90 000 patients days per year in the somatic adult wards. EHRs (DIPS®, ASA) were implemented in the trust in 1992. The EHRs includes both free text (i.e. discharge summaries, operative reports, pathology reports, radiology results, transfer of service notes, admission notes, medical progress notes and notes from other healthcare professionals) and indexed variables (i.e. laboratory results, admissions and discharge data, diagnosis and procedure codes). In Norwegian hospitals medication administration, prescriber orders and recording of vital parameters are still hand-written and scanned into the EHRs but will be digitalized and indexed within the next two years in clinical information systems. The trust implemented the GTT in 2010 with bi-weekly review of 70 records randomly selected from the seven main units discharge lists [17].

Participants

The records included in the study were originally selected for the trusts GTT review in the period 1 March to 31 December 2013. Patient records were excluded if the patient was admitted for less than 24 h, discharged from psychiatric or rehabilitation units, and was aged 17 years or younger, as the triggers were not developed for these patients [8]. Approval for the study was obtained from the Data Protection Official in Nordland Hospital trust and by the Regional Committee for Medical and Health Research Ethics (ref 2012/1691). The committee approved a waiver for informed consent as the study fulfilled criteria described by Baker *et al.* [18].

Definition of triggers

The Norwegian translation of the GTT includes 57 triggers (supplementary file A) [19]. The triggers are events recorded in the clinical data such as; abnormal lab values, readmission within 30 days,

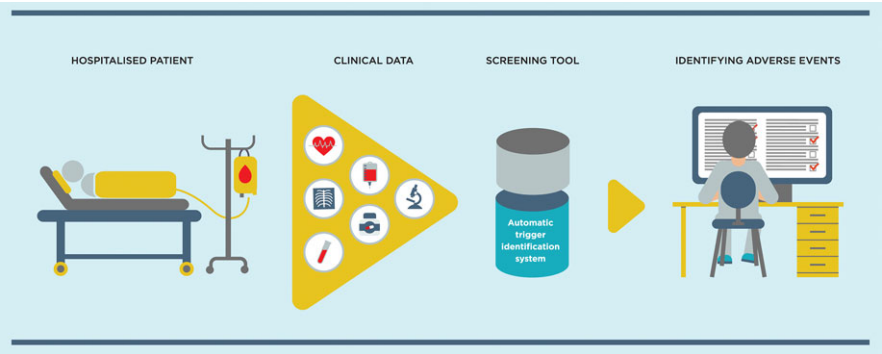


Figure 1 The modified GTT method.

return to surgery, blood transfusion or administration of drugs such as anti-dot or anti-emetic drugs [8]. Some of the triggers are adverse events by its nature, for example third- or fourth-degree perianal lacerations, pressure ulcer and injury, repair or removal of organ because of accidental injury. However, most of them are just indicators that an adverse event may have occurred. A more in-depth review is necessary to decide if the triggers are associated with any adverse events.

Definition of an adverse event

The definition of an adverse event adopted from the GTT was used by both methods when deciding if an adverse event was present when performing manual review of the triggered records [8]: 'Unintended physical injury resulting from or contributed to by medical care that requires additional monitoring, treatment or hospitalization, or that results in death'.

The adverse events were categorized according to severity with the adapted definitions from the National Coordinating Council for Medication Error Reporting and Prevention Index (NCC MERP) for categories E-I [20]:

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

Review methods

The original GTT method

Two primary reviewers (nurses) reviewed the records individually in a specific order to register any presence of triggers. Once a trigger was identified, a more in-depth review was performed to investigate if the trigger was associated with an adverse event according to the described definition. All performed within a 20-min time limit. A secondary reviewer (physician) authenticated the primary reviewers' findings. There was no time constraint for the secondary reviewer. Griffin *et al.* estimated that the secondary reviewer uses two hours per 20 records, confirming or deleting adverse events identified by the primary reviewers [8]. The secondary reviewer reviewed only the relevant parts of the records identified with adverse events.

The modified GTT method

The automatic trigger identification system can only identify triggers, not adverse events. The system identifies triggers based on algorithms. We have included examples used in such algorithms in Supplementary file B. The algorithms for indexed variables (e.g. INR >6, glucose <2.8 or diagnoses/procedures codes) are based on queries. Algorithms for free text (e.g. patient fall, specialty obstetric consult, induced labour) are based on information extractions and recognitions of text strings and patterns through text mining analysis. All conditions and words representing the actual trigger (e.g. patient fell out of bed, patient slipped in the bathroom) are extracted. In addition, the system omits the information if exclusion criteria are met (e.g. the anastomosis fell in place, the catheter fell out). The automatic trigger identification system included 42 triggers used in the Norwegian GTT (see Supplementary file A). Nine triggers were excluded as the information for these triggers are hand written and scanned into the EHR. The automatic trigger identification system cannot identify these triggers; use of anti-dot drugs, use of anti-emetic drugs, vitamin K administration, hypotension and abrupt medication stop. The three triggers labelled 'other' related to respectively medication, general and surgical care were not included in the system, as they do not correspond to a specific adverse event but used when reviewers identify an adverse event without finding a corresponding trigger. Finally, we opted to exclude three triggers rarely identified in our previous manual review of 6720 records from 2010 to 2013.

The records, both triggered and non-triggered, were presented in an interface along with information regarding triggers identified (e.g. type of trigger and which note/lab test/radiology or pathology report the triggers are detected in). One physician performed manual review of the triggered records to decide if the triggers were associated with any adverse events and if so, theirs severity and type. The described definition of an adverse event was applied. The manual review time used in each record was recorded. No time constraint was applied.

Statistics

One thousand four hundred records from the trusts GTT review of a 10-month period were selected. One hundred sixty-seven records were excluded as the data from the automatic trigger identification system was missing for these records, leaving a total of 1233 records included in the study.

The objective of this study was to evaluate the modified GTT method. Paired *t*-test was used to compare the number of triggered records, number of records with identified adverse events and number of identified adverse events between the methods. A *P*-value <0.05 was regarded significant. We calculated recall (sensitivity), precision (positive predictive value) and specificity with their respective 95 % confidence intervals (CI) to evaluate the validity of the modified GTT method using the original GTT method as gold standard:

$$\text{Recall} = \frac{\text{No. of correct positive records identified by the modified GTT method}}{\text{No. of positive records identified by gold standard}}$$

$$\text{Precision} = \frac{\text{No. of correct positive records identified by the modified GTT method}}{\text{Total no. of positive records identified by the modified GTT method}}$$

$$\text{Specificity} = \frac{\text{No. of correct negative records identified by the modified GTT method}}{\text{No. of negative records identified by gold standard}}$$

Recall represents the proportion of 'correctly' identified records with adverse events by the modified GTT method. Precision represents the proportion of records with adverse events identified by the modified GTT method that also were identified by the original GTT method. Specificity represents the proportion of 'correctly' identified records with no identified adverse events by the modified GTT method. For reliability, we used Cohen's Kappa to measure agreement of the results (inter-rater reliability) between the methods, taking into account the agreement occurring by chance. The following interpretations from Landis and Koch were used for the Cohen's 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) [21]. A 95 % CI was set. The CI for recall, precision and specificity was calculated using the Wilson score method [22]. CI for Cohen's kappa was $\kappa \pm 1.96$*SE. All analyses were performed using SPSS (version 22.0; SPSS Chicago, IL).

Results

Fifty eight percent (716) of the patients were women and average age was 58 years (range; 18–102, standard deviation (SD); 22). Mean length of stay was 5 days (range; 1–65, SD; 6).

The modified GTT method identified a total of 1216 triggers in 658 records while the original GTT method identified a total of 1267 triggers in 626 records. The number of the individually triggers identified by each method are included in Supplementary file C. In 110 identical records, both methods identified adverse events. In 79 records, the modified GTT method identified adverse events alone and vice versa in 76 records (Fig. 2). The recall, precision, specificity and Cohen's kappa with their respective 95 % CI of the modified GTT method are presented in Table 1. Figure 3 displays the types of adverse events identified by the two methods which differed between the methods. Number of records identified with adverse events and number of identified adverse events according to severity are presented in Table 2.

The modified GTT method identified 34.7 adverse events (n = 214) per 1000 patient days by manual review of 658 automatic triggered records for the 10-month period. Adverse events were identified in 28.7 % (n = 189 records) of the automatic triggered records (n = 658 record). Mean manual review time used per record was 2 min (range 0.2–21.5) and the total manual review time was 23 h.

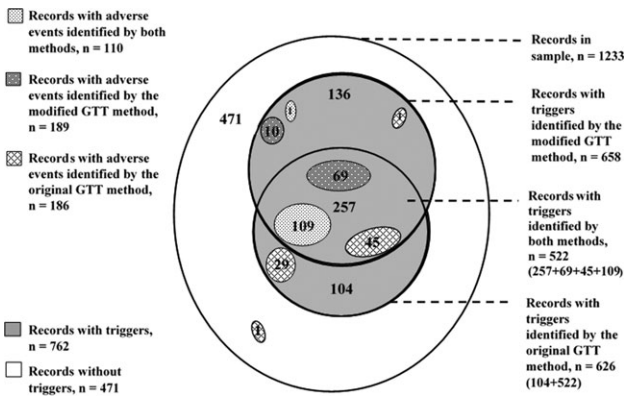


Figure 2 Records identified with triggers and adverse events by the modified GTT method and the original GTT method.

The original GTT method identified 35.0 adverse events (n = 216) per 1000 patient days of 626 manual triggered records in the same 10-month period. Adverse events were identified in 15.1 % (n = 186 records) of the records reviewed manually for triggers and adverse events (n = 1233 records). Total manual review time of 1233 records was 411 h.

Discussion

The aim of this study was to evaluate a modified GTT method with automatic trigger identification to identify and measure adverse events using the original GTT method as gold standard. We found that the modified GTT method is a valid, reliable and efficient method to monitor the rate of adverse events. The modified GTT method demonstrated major decrease in review time compared to the original GTT method. Both methods identified a rate of 35 adverse events per 1000 patient days. There was no significant difference between the methods regarding the severity of the identified adverse events. The modified GTT method comply with the GTTs aim to monitor the rate of adverse events over time consistently, but not completely.

The values of a 'new' measure are related to values from a reference measure performed at the same time and are defined as the

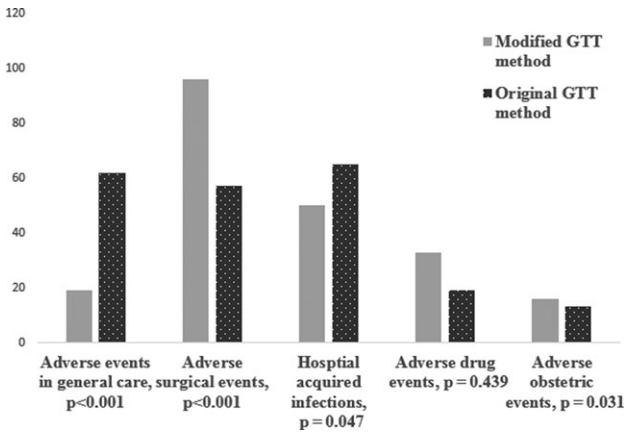


Figure 3 Adverse events identified according to types. Adverse events in general care: allergy, bleeding, patient fall, fracture, medical technical event, thrombosis/embolism, deterioration of chronic disease and other events. Adverse surgical events: infection after surgery, return to surgery, injury or removal of an organ by accident, bleeding after surgery, respiratory complication after surgery, switch in surgery and any other surgical complication. Hospital acquired infection: urinary tract infection, lower respiratory infection, ventilatorassociated infection, central vein catheter associated infection and other infections.

Table 1 Validity and reliability of the modified GTT method versus the original GTT method (gold standard)

Variable	Recall[a] (CI)[b]	Precision[c] (CI)[b]	Specificity[d] (CI)[b]	Cohen's Kappa[e] (CI)[b]
Triggered records	0.83 (0.80–0.86)	0.79 (0.76–0.82)	0.78 (0.74–0.81)	0.61 (0.56–0.66)
Records with adverse events	0.59 (0.52–0.66)	0.58 (0.51–0.65)	0.92 (0.91–0.94)	0.51 (0.44–0.57)
Records with adverse events within the common triggered records	0.71 (0.63–0.77)	0.61 (0.54–0.68)	0.81 (0.77–0.85)	0.50 (−0.31–1.30)

[a]Recall represent the proportion of 'correctly' records identified with triggers or adverse events by the modified GTT method.
[b]95 % confidence interval (CI).
[c]Precision represent the proportion of records with triggers or adverse events that were confirmed by the original GTT method.
[d]Specificity represents the proportion of 'correctly' records with no identified adverse events by the modified GTT method.
[e]Cohen's Kappa is the inter-rater reliability of the modified GTT method and the original GTT method evaluated by a 2 × 2 table.

Table 2 Number of adverse events and records with adverse events according to severity identified by the modified GTT method and the original GTT method

Severity category	Original GTT method		Modified GTT method		Records with adverse events: Modified GTT method vs. Original GTT method		Number of adverse events: Modified GTT method vs Original GTT method	
	Adverse events	Records with adverse events	Adverse events	Records with adverse events	*P**-value	CI 95%	*P**-value	CI 95%
E	120	109	95	90	0.08	−0.032–0.002	0.045	−0.04–0.00
F	87	80	97	91	0.29	−0.008–0.026	0.38	−0.01–0.03
G	5	5	12	12	0.09	−0.001–0.012	0.09	−0.001–0.01
H	1	1	1	1	1.00	−0.002–0.002	1.00	−0.002–0.002
I	3	3	9	9	0.03	0.000–0.009	0.03	0.00–0.01
Total	216	198***	214	203**	0.81	0.01–−0.02	0.90	−0.03–0.024

Notes: Severity level according to the NCC MERP index.

**P*-value of paired sample t-test.

***14 admissions with two more adverse events with different severity level.

****12 admissions with two or more adverse events with different severity level.

concurrent validity of the measure. Concurrent validity is evaluated by recall (sensitivity), precision (positive predictive value) and specificity, which we calculated for the modified GTT method. A review of current literature did not find any reference to evaluation of the validity of the GTT [23] but studies have demonstrated that the GTT identifies more adverse events than other methods [10, 11]. The purpose of the GTT is, with an easy method, to select those patients that may have experienced an adverse event by the use of triggers as screening criteria. We adopted this purpose when we evaluated the modified GTT method. We recorded therefore only the unique number of identified triggers in the triggered records and did not considered excessive testing of the individually triggers as this was beyond the scope of the study.

The modified GTT method demonstrated an efficient method to identify and measure adverse events with a total of 23 h to complete the manual review of 658 automatic triggered records compared to 411 h of review of 1233 records with the original GTT method. The modified GTT method reviewed only the triggered records thereby reducing the number of records to be manual reviewed by 50 %. This reduction enables for increasing the sample size without applying further resources. Critics have argued that the recommend sample size, 10 records bi-weekly, in the GTT is too low to estimate the rate of adverse events for an institution. Thus, sampling size should correspond to the hospital size [17]. Extrapolation, which is used when estimates are made on small samples, increases the random variability. Infrequent adverse events can also be missed when only samples of records are reviewed [24]. Increasing the sample size makes the results regarding the rate of adverse events more valid [17, 25].

The manual review processes differed somewhat between the two methods. Only one reviewer, a physician, performed the subsequent manual review of automatic triggered records in the modified GTT method. The original GTT method included two primary reviewers and a secondary reviewer authenticating their consensus findings. Reviewers in both methods were experienced reviewers of the GTT. The aim of the study was to assess if the rate of adverse events altered when we modified the GTT with manual review of only automatic triggered records. Hence, we do not consider the differences of the manual review processes as a bias.

Poor to moderate agreement between reviewers and between review teams have been demonstrated [26, 27].We believe the agreement can be improved by using an automatic trigger identification system. First, automatic identification of triggers in the EHR excludes the variability of manual identified triggers as triggers based on index information (i.e. blood transfusion and dialysis) have demonstrated higher agreement than triggers derived from free text (i.e. pressure ulcers, patient fall) [28]. Second, the manual trigger identification could suffer from the time constraints excluding possible triggers causing adverse events to be missed [27, 29]. Automatic trigger identification does not have a time constraint and all present triggers are identified. These issues make the identified adverse events based on automatic trigger identification more standardized and comparable than adverse events identified by manual trigger identification. Moreover, with further development, the automatic trigger identification system can provide a platform to identify patients at risk of adverse events in real-time. Such systems could be used to improve clinical outcome, optimize treatment, reduce the financial burden of patient harm and most importantly; reduce the suffering of the patients due to adverse events [24, 30]. However, the development of such methods requires both technical and economic inputs.

Strength and limitations

The main strength of the study is that we demonstrated a valid and efficient method to identify and measure adverse events.

Our study has some limitations. First, fifteen of the original 57 triggers were excluded in our automatic trigger identification system, but nine of them can be included when all patient data are digitalized and indexed in clinical information systems. Second, record reviews depend on that the necessary data are documented in the EHR. Records could be incomplete regarding documentation of adverse events causing adverse events to be missed. Third, this study has been performed in one hospital only. Modification of the automatic trigger identification system must be applied before adoption.

Conclusions

Our study demonstrated that the modified GTT method with automatic trigger identification is a valid, reliable and efficient method to identify and measure adverse events to comply the aim of the GTT

in respect to the original GTT method. We therefore recommend that the modified GTT method should be preferred as it offers an efficient alternative to the common costly and time-consuming approaches mainly used to identify and measure adverse events. The resources saved by using the modified GTT method are considerable and this enables for increase of the sample size.

Supplementary material

Supplementary material is available at *International Journal for Quality in Health Care* online.

Acknowledgements

The authors would like to thank the review teams at our trust who conducted the manual reviews. We also thank the SAS Institute® for contributing in developing our automatic trigger identification system implementing the algorithms. A special thanks to Christian Hardahl in SAS® reviewing the manuscript regarding logics, to Tom Wilsgaard for help with the statistics, to Fran Griffin reviewing the manuscript and to Laila Bjølgerud for making the graphical illustration.

Funding

This work was supported by The Northern Norway Regional Health Authority by a PhD grant to K.M.

References

1. Jha A, Pronovost P. Toward a safer health care system: the critical need to improve measurement. *JAMA* 2016;**315**:1831–2.
2. Agbabiaka TB, Lietz M, Mira JJ *et al*. A literature-based economic evaluation of healthcare preventable adverse events in Europe. *Int J Qual Health Care* 2017;**29**:9–18.
3. Macrae C. The problem with incident reporting. *BMJ Qual Saf* 2016;**25**: 71–5.
4. Jick H. Drugs—remarkably nontoxic. *N Engl J Med* 1974;**291**:824–8.
5. Classen DC, Pestotnik SL, Evans RS *et al*. Description of a computerized adverse drug event monitor using a hospital information system. *Hosp Pharm* 1992;**27**:774, 776–779, 783.
6. Resar RK, Rozich JD, Classen D. Methodology and rationale for the measurement of harm with trigger tools. *Qual Saf Health Care* 2003;**12**: ii39–i45.
7. Garrett J, Paul R, Sammer C. Developing and implementing a standardized process for Global Trigger Tool Application across a large health system. *J Qual* 2013;**39**:292–7.
8. Griffin F, Resar R. *IHI Global Trigger Tool for Measuring Adverse Events*, 2nd edn. IHI Innovation Series white paper. Cambridge, MA: Institute for Healthcare Improvement, 2009.
9. Deilkås E, Bukholm G, Lindstrøm J *et al*. Monitoring adverse events in Norwegian hospitals from 2010 to 2013. *BMJ Open* 2015;**5**:e008576.
10. Hibbert PD, Molloy CJ, Hooper TD *et al*. The application of the Global Trigger Tool: a systematic review. *Int J Qual Heal Care* 2016;**28**:640–9.
11. 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.
12. 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.
13. Musy SN, Ausserhofer D, Schwendimann R *et al*. Trigger tool-based automated adverse event detection in electronic health records: systematic review. *J Med Internet Res* 2018;**20**:e198.
14. Stockwell DC, Kirkendall E, Muething SE *et al*. Automated adverse event detection collaborative: electronic adverse event identification, classification, and corrective actions across academic pediatric institutions. *J Patient Saf* 2013;**9**:203–10.
15. Jha AK, Kuperman GJ, Teich JM *et al*. Identifying adverse drug events. *J Am Med Inform Assoc* 1998;**5**:305–14.
16. Lemon V, Stockwell DC. Automated detection of adverse events in children. *Pediatr Clin North Am* 2012;**59**:1269–78.
17. 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.
18. Baker DW, Persell SD. Criteria for waiver of informed consent for quality improvement research. *JAMA Intern Med* 2015;**175**:142.
19. Deilkås E, 'Gjennomføring av journalundersøkelse med Global Trigger Tool (GTT) i den norske pasientsikkerhetskampanjen,' 2011.
20. Hartwig SC, Denger SD, Schneider PJ. Severity-indexed, incident report-based medication error-reporting program. *Am J Hosp Pharm* 1991;**48**: 2611–6.
21. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;**33**:159–74.
22. Newcombe RG. 'Two-sided confidence intervals for the single proportion: comparison of seven methods,'. *Stat Med* 1998;**17**:857–72.
23. Hanskamp-Sebregts M, Zegers M, Vincent C *et al*. Measurement of patient safety: a systematic review of the reliability and validity of adverse event detection with record review. *BMJ Open* 2016;**6**:e011078.
24. Sammer C, Miller S, Jones C *et al*. Developing and evaluating an automated all-cause harm trigger system. *Jt Comm J Qual Patient Saf* 2017; **43**:155–65.
25. Good V, Saldana 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.
26. 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.
27. 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.
28. 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.
29. Unbeck M, Schildmeijer K. Is detection of adverse events affected by record review methodology? An evaluation of the 'Harvard Medical Practice Study' method and the' Global Trigger. *Patient Saf* 2013;**7**:10.
30. Jensen K, Soguero-Ruiz C, Oyvind Mikalsen K *et al*. Analysis of free text in electronic health records for identification of cancer patient trajectories. *Sci Rep* 2017;**7**:46226.