Analysis of unintended events in hospitals: inter-rater reliability of constructing causal trees and classifying root causes

MARLEEN SMITS¹, JASPER JANSSEN¹, RIEKIE DE VET², LAURA ZWAAN², DANIELLE TIMMERMANS², PETER GROENEWEGEN¹,3,4 AND CORDULA WAGNER¹,2

¹NIVEL, Netherlands Institute for Health Services Research, Utrecht, The Netherlands, ²Department of Public and Occupational Health, EMGO Institute, VU University Medical Center, Amsterdam, The Netherlands, ³Department of Sociology, Utrecht University, Utrecht, The Netherlands, and ⁴Department of Human Geography, Utrecht University, Utrecht, The Netherlands

Abstract

Background. Root cause analysis is a method to examine causes of unintended events. PRISMA (Prevention and Recovery Information System for Monitoring and Analysis) is a root cause analysis tool. With PRISMA, events are described in causal trees and root causes are subsequently classified with the Eindhoven Classification Model (ECM). It is important that root cause analysis tools are reliable, because they form the basis for patient safety interventions.

Objectives. Determining the inter-rater reliability of descriptions, number and classifications of root causes.

Design. Totally, 300 unintended event reports were sampled from a database of 2028 events in 30 hospital units. The reports were previously analysed using PRISMA by experienced analysts and were re-analysed to compare descriptions and number of root causes (n = 150) and to determine the inter-rater reliability of classifications (n = 150).

Main outcome measures. Percentage agreement and Cohen’s kappa (κ).

Results. Agreement between descriptions of root causes was satisfactory: 54% agreement, 17% partial agreement and 29% no agreement. Inter-rater reliability of number of root causes was moderate (κ = 0.46). Inter-rater reliability of classifying root causes with the ECM was substantial from highest category level (κ = 0.71) to lowest subcategory level (κ = 0.63). Most discrepancies occurred in classifying external causes.

Conclusions. Results indicate that causal tree analysis with PRISMA is reliable. Analysts formulated similar root causes and agreed considerably on classifications, but showed variation in number of root causes. More training on disclosure of all relevant root causes is recommended as well as adjustment of the model by combining all external causes into one category.

Keywords: adverse events, patient safety, incident reporting and analysis, medical errors, human factors

Introduction

Since the early nineties, many studies have been performed to examine the incidence of adverse events, with reported incidence rates ranging from 3 to 17% of all hospital admissions. A quarter to half of the adverse events was considered preventable [1–9]. These studies have increased the sense of urgency to take effective countermeasures in order to improve patient safety in hospitals. Improvements can only be achieved if the interventions tackle the dominant underlying causes. When examining causal factors, not only adverse events are informative, but also other unintended events. Unintended events are a broader group of events, including near misses, which do not necessarily result in patient harm. They are considered as a useful source of information on causal factors as well, because they occur more frequently than mere adverse events and—at least near misses—are believed to have the same underlying failure factors as accidents that do reach the patient [10, 11].

In order to identify the causes of unintended events, there are a number of analysis tools, brought together under the umbrella term Root Cause Analysis. One common tool is causal tree analysis. Causal trees are the basis for the PRISMA method (see below), which is becoming the
standard in Dutch hospitals to analyse root causes of unintended events and has also been used in studies abroad [12, 13]. The corresponding taxonomy to classify root causes distinguishes technical, organizational, human and patient-related factors. It was used as a foundational component for the conceptual framework for the World Health Organization World Alliance for Patient Safety’s International Classification for Patient Safety [14, 15].

**PRISMA root cause analysis**

PRISMA, an acronym for Prevention and Recovery Information System for Monitoring and Analysis, was originally designed for the chemical industry, but was expanded for application in health care (PRISMA-medical) in 1997 [16]. Unintended events are analysed in three main steps.

In step one, an event is described by means of a causal tree. At the top of the tree a short description of the event is placed, as the starting point for the analysis. Below the top event, all direct causes that can be identified are mentioned. These direct causes often have their own causes. By continuing to ask ‘why’ for each event or action beginning with the top event, the majority of causes are revealed. This way a structure of causes arises, until the root causes are identified at the bottom of the tree [16–18]. PRISMA is suitable for events varying in complexity, ranging from one to more than 10 root causes. Figure 1 shows an example of a relatively simple structure of a causal tree. The example is derived from our database.

In step two, the identified root causes are classified with the Eindhoven Classification Model (ECM) of PRISMA [16–18]. The ECM taxonomy distinguishes five main categories of causes: technical, organizational, human, patient-related and other factors. These main categories can be subdivided into 20 subcategories (see Table 1). The technical and organizational categories are latent conditions and the human categories are active failures, consistent with Reason's system approach [19, 20]. The human categories are based on the widely established framework for classifying human errors by the skill-, rule- and knowledge-based behaviour model of Rasmussen [21].

In the final step, all classifications of a group of unintended events are added up to make a so-called PRISMA profile, a graphical representation (for example, a bar plot) of the relative contributions of the causal factor categories of the ECM. Prevention strategies can be directed at the most frequently occurring (combinations of) root causes [16–18].

**Inter-rater reliability of PRISMA**

In order to be valuable as input for patient safety interventions, there has to be a certain degree of consistency in which different individuals make causal trees and classify root causes (inter-rater reliability). Without a satisfying inter-rater reliability, the recommendations based on the results of root cause analysis are questionable. Inter-rater reliability scores of causal trees and classifications can be calculated when different individuals independently analyse the same unintended event in one hospital, but they can serve as a proxy for the more common situation of different individuals analysing similar unintended events in different hospitals. A high inter-rater reliability score is supporting for the confidence that these analysts perform the analyses in the same way. It is important for researchers and hospital staff that are involved in the analysis of unintended events to have a reliable root cause analysis tool.

Previously, two small studies have evaluated the inter-rater reliability of the second step of the PRISMA method: classifying root causes with the ECM. Wijers et al. [22] examined the reliability with only 19 root causes. The inter-rater reliability at highest main category level was substantial ($k = 0.71$), but it was only fair for the lowest subcategory level ($k = 0.39$). The study of Dabekausen et al. [23] included a double review of only 10 events with PRISMA rail (PRISMA for the railway industry). The inter-rater reliability at main category level was substantial ($k = 0.69$). No results were reported for subcategories.

**Study objectives**

The present study is the first study in any industry that examines the inter-rater reliability of both formulating causal trees and classifying root causes by means of a large database of analyses with PRISMA. Our objectives were to determine the inter-rater reliability of (i) the descriptions of root causes in the causal tree, (ii) the number of root causes used in the causal tree and (iii) the classification of root causes with the ECM.

**Methods**

**Study design and setting**

From October 2006 to February 2008, a large cross-sectional study was conducted to examine the root causes of
<table>
<thead>
<tr>
<th>Level 0 (system)</th>
<th>Level 1 (main category)</th>
<th>Level 2 (SRK not subdivided)</th>
<th>Level 3 (complete ECM)</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Latent conditions</td>
<td>Technical</td>
<td>External</td>
<td>External</td>
<td>Technical failures beyond the control and responsibility of the investigating organization.</td>
</tr>
<tr>
<td>Design</td>
<td>Design</td>
<td>Design</td>
<td></td>
<td>Failures due to poor design of equipment, software, labels or forms.</td>
</tr>
<tr>
<td>Construction</td>
<td>Construction</td>
<td>Construction</td>
<td></td>
<td>Correct design, which was not constructed properly or was set up in inaccessible areas.</td>
</tr>
<tr>
<td>Materials</td>
<td>Materials</td>
<td>Materials</td>
<td></td>
<td>Material defects not classified under Design or Construction.</td>
</tr>
<tr>
<td>Organizational</td>
<td></td>
<td></td>
<td></td>
<td>Failures at an organizational level beyond the control and responsibility of the investigating organization, such as in another department or area (address by collaborative systems).</td>
</tr>
<tr>
<td>Transfer of knowledge</td>
<td>Transfer of knowledge</td>
<td></td>
<td></td>
<td>Failures resulting from inadequate measures taken to ensure that situational or domain-specific knowledge or information is transferred to all new or inexperienced staff.</td>
</tr>
<tr>
<td>Protocols</td>
<td>Protocols</td>
<td></td>
<td></td>
<td>Failures relating to the quality and availability of the protocols within the department (too complicated, inaccurate, unrealistic, absent or poorly presented).</td>
</tr>
<tr>
<td>Management priorities</td>
<td>Management priorities</td>
<td></td>
<td></td>
<td>Internal management decisions in which safety is relegated to an inferior position when faced with conflicting demands or objectives. This is a conflict between production needs and safety. Example: decisions that are made about staffing levels.</td>
</tr>
<tr>
<td>Culture</td>
<td>Culture</td>
<td></td>
<td></td>
<td>Failures resulting from collective approach and its attendant modes of behaviour to risks in the investigating organization.</td>
</tr>
<tr>
<td>Active errors</td>
<td>Human</td>
<td>External</td>
<td>External</td>
<td>Human failures originating beyond the control and responsibility of the investigating organization. This could apply to individuals in another department.</td>
</tr>
<tr>
<td>Knowledge-based behaviour</td>
<td>Knowledge-based behaviour</td>
<td></td>
<td></td>
<td>The inability of an individual to apply their existing knowledge to a novel situation. Example: a trained blood bank technologist who is unable to solve a complex antibody identification problem.</td>
</tr>
<tr>
<td>Rule-based behaviour</td>
<td>Qualifications</td>
<td>The incorrect fit between an individual’s training or education and a particular task. Example: expecting a technician to solve the same type of difficult problems as a technologist.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------------------</td>
<td>----------------</td>
<td>---------------------------------------------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coordination</td>
<td></td>
<td>A lack of task coordination within a healthcare team in an organization. Example: an essential task not being performed because everyone thought that someone else had completed the task.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Verification</td>
<td></td>
<td>The correct and complete assessment of a situation including related conditions of the patient and materials to be used before starting the intervention. Example: failure to correctly identify a patient by checking the wristband.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intervention</td>
<td></td>
<td>Failures that result from faulty task planning and execution. Example: washing red cells by the same protocol as platelets.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Monitoring</td>
<td></td>
<td>Monitoring a process or patient status. Example: a trained technologist operating an automated instrument and not realizing that a pipette dispensing reagents is clogged.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Skill-based behaviour</td>
<td>Slips</td>
<td>Failures in performance of highly developed skills. Example: a technologist adding drops of reagents to a row of test tubes and then missing the tube or a computer entry error.</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Tripping</td>
<td>Failures in whole body movements. These errors are often referred to as ‘slipping, tripping or falling’. Examples: a blood bag slipping out of one’s hands and breaking or tripping over a loose tile on the floor.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Other factors

<table>
<thead>
<tr>
<th>Patient related</th>
<th>Patient related</th>
<th>Patient related</th>
</tr>
</thead>
<tbody>
<tr>
<td>Other</td>
<td>Other</td>
<td>Other</td>
</tr>
</tbody>
</table>

*Note:* Level 0 is the system factors with 3 categories; Level 1 is the main category with 5 categories; Level 2 is the subcategory level with no subdivision within the skill- and rule-based factors (15 categories); Level 3 is the complete ECM with 20 categories.
unintended events at 30 hospital units in the Netherlands: 10 emergency units, 10 surgery units and 10 internal medicine units. These units were located in 21 hospitals: 2 university hospitals, 8 tertiary teaching hospitals and 11 general hospitals. During the study, 2028 written unintended event reports were gathered from healthcare providers in the participating units. Unintended events were broadly defined as all events, no matter how seemingly trivial or commonplace, that were unintended and could have harmed or did harm a patient [24]. Reports were analysed using PRISMA by an analyst from a research team of five PRISMA analysts. They all had followed a one-day PRISMA training session. Each analyst collected report forms from healthcare providers, interviewed them with some additional questions in order to enhance understanding of the event and wrote down the questions and responses. No interviews were held with staff in other hospital departments than the department under investigation. Using this information, the analyst constructed a PRISMA causal tree for each event. This resulted in 2028 causal trees and 3015 (classified) root causes. Only 2.4% of the root causes were classified as ‘Other’, an indication for the comprehensiveness of the ECM. The research team got together twice a month and discussed any difficulties they encountered in performing the analyses. A Frequently Asked Questions list was composed to record the problems and solutions (see online supplementary material).

In order to perform the present inter-rater reliability study, a random sample of 10 unintended events was selected from each unit, resulting in a total sample of 300 unintended events. The study was conducted over an 18-week period from January till May 2008. Half of the sample (150 events; 7.4% of the complete data set) was used to determine the inter-rater reliability of the descriptions (objective a) and number of root causes (objective b) and the other half to determine the inter-rater reliability of classifying root causes with the ECM (objective c). These 300 unintended events were assigned to a second analyst of the research team who was unfamiliar with the event.

Assessment of agreement in formulating root causes (objectives a and b)

For each of the 150 unintended events used for objectives a and b, a second analyst of the research team drafted a causal tree based on the same information about the event as had been available for the initial analyst of the team, that is the written reports and the additional information from the interviews. Each second causal tree was compared with the original causal tree for agreement in the content of the root cause descriptions by an independent third researcher.

The causal tree with the largest number of root causes was the basis for the comparison. The number of root causes in the largest causal tree did not exceed four. The 150 causal trees consisted of 257 root causes to be compared. A set of two root causes was evaluated as agreeing (A), partly agreeing (P) or not agreeing (N). If the root causes of the two analysts agreed in content, irrespective of the precise use of words, the set was coded as ‘A’. If the root causes had some overlap in content, but not completely, the set was coded as ‘P’. Sets were coded as ‘N’ in two cases: (i) if the root causes involved two different subjects and (ii) if the number of root causes in one causal tree differed between both analysts. In the latter case, the additional root causes of one of the analysts were evaluated as ‘N’, because when a root cause is not present in the causal tree of one of the analyst, there is no agreement for this cause.

Using the rules for comparison of causal trees that were formulated by the authors (Table 2), 150 pairs of causal trees were given a code ranging from one to four characters, depending on the number of root causes in the largest causal tree of the two. The pairs were assigned an overall causal tree score from 0 to 3.

Besides this evaluation of descriptions of root causes (objective a), the agreement in the number of root causes that analysts used in the causal trees was also determined (objective b), irrespective of the content of these root causes.

Assessment of agreement in classifications with the ECM taxonomy (objective c)

For objective c, the other group of 150 unintended events was used. A second analyst classified the root causes of 150 original causal trees that were drafted by the first analyst. Next to the original causal tree (excluding the classifications of root causes), all information in the event report was at the disposal of the second analyst. There were 212 root cause classifications to be compared. Classification codes of both analysts were compared on the four levels of the ECM (see Table 1). Level 0 is the system factors level, which makes a

<table>
<thead>
<tr>
<th>Number of root causes in (largest) causal tree</th>
<th>Overall score for causal tree</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 points</td>
<td>1 point</td>
</tr>
<tr>
<td>1</td>
<td>N</td>
</tr>
<tr>
<td>2</td>
<td>NN</td>
</tr>
<tr>
<td>3</td>
<td>NNN</td>
</tr>
<tr>
<td>NNP</td>
<td>ANP</td>
</tr>
<tr>
<td>4</td>
<td>NNNN</td>
</tr>
<tr>
<td>NNNP</td>
<td>APPN</td>
</tr>
<tr>
<td>NNNNA</td>
<td>APPN</td>
</tr>
<tr>
<td>NNP</td>
<td>PPPP</td>
</tr>
<tr>
<td>PPPN</td>
<td></td>
</tr>
</tbody>
</table>

A, agreement between first and second analyst in one pair of root causes; P, partial agreement; N, no agreement.

Note: In case of a one letter code, the causal tree to be compared has one root cause; in case of a two letter code, the causal tree to be compared has two root causes, et cetera.
distinction between latent factors, active failures and other factors. Level 1 is the main category level. At this level, the taxonomy has five main categories of causes: technical, organizational, human, patient-related and other. Level 2 involves 15 subcategories with no further subdivision of human skill-based, rule-based and knowledge-based categories. Level 3 is the complete ECM taxonomy with 20 subcategories. We distinguished these four levels, because researchers might want to report results about causal factors on more aggregated levels than the complete ECM. If doing so, it would be useful to know the inter-rater reliability of classifying the root causes at these levels.

**Statistical analysis**

The inter-rater reliability of the descriptions of root causes was expressed using a mean score (range 0–3), according to the rules for comparison of descriptions of root causes, as presented in Table 2. The inter-rater reliability of the number of root causes between two analysts was determined using Cohen's weighted kappa (κ) with 95% confidence intervals (CIs). Quadrate weighting was applied, giving different weights to disagreements according to the magnitude of the discrepancies [25]. Disagreement between two adjacent numbers (for example, analyst 1 having two root causes and analyst 2 having three) contributed less to the weighted kappa coefficient as more extreme disagreements (for example, analyst 1 having one root cause and analyst 2 having four). The agreement in number of root causes was also expressed as the percentage of unintended events for which there was agreement between analysts.

The inter-rater reliability scores of the classification of root causes on the four levels of the ECM taxonomy were expressed as Cohen's kappa with 95% CIs and as the percentage of root causes for which there was agreement. Because the classifications either were identical or not (and nothing in between), no weighting was applied.

A κ-value between 0.00 and 0.20 was classified as ‘slight’; between 0.21 and 0.40 as ‘fair’; between 0.41 and 0.60 as ‘moderate’; between 0.61 and 0.80 as ‘substantial’ and between 0.81 and 1.00 as ‘(almost) perfect’ [26]. Data were analysed using SPSS 14.0 and Microsoft Excel 2003 for Windows.

### Results

**Inter-rater reliability of formulating root causes (objectives a and b)**

There was agreement (A) in the descriptions of 138 of the 257 root causes in the causal tree (54%), partial agreement (P) for 44 root causes (17%) and no agreement/no equivalent (N) for 75 root causes (29%). Table 3 shows some examples of the assessment of agreement in the descriptions of root causes. The mean score for overall causal tree agreement, according to the rules for comparison of causal trees (Table 2), was 2.0 (SD = 1.04) on a scale from 0 to 3.

The inter-rater reliability for the number of root causes used in the causal tree was moderate (κ = 0.46, 95% CI: 0.18–0.75), with a percentage agreement of 69%.

**Inter-rater reliability of classifying root causes (objective c)**

The overall inter-rater reliability was substantial for all four levels of the ECM: Level 0 (κ = 0.71, agreement 87%),
Level 1 (κ = 0.70, agreement 85%), Level 2 (κ = 0.67, agreement 75%) and Level 3 (κ = 0.63, agreement 68%). Within each main type of causes with subcauses (human, technical and organizational), the reliability was also substantial. The human categories had a lower reliability than the other two main categories, although not significant (Table 4).

Most inconsistencies between analysts occurred in classifying external root causes; mainly in the choice between organizational external and human external (see Table 1 for explanation of causal factors). For example, the root cause ‘Bladder catheter set was packed incorrectly at medical supplies company’ was classified as human external by one analyst and as Organizational External by the other analyst.

When all external factors (human external, technical external and organizational external) are placed in one category and the kappa values are recalculated, they improve at all levels of the ECM: Level 0 (κ = 0.76, agreement 85%), Level 1 (κ = 0.76, agreement 84%), Level 2 (κ = 0.73, agreement 81%) and Level 3 (κ = 0.68, agreement 74%).

Moreover, many inconsistencies occurred when the first analyst classified a root cause as external, whereas the second analyst chose for one of the internal root cause codes. For example ‘A cardiologist on duty for emergency department did not load the battery of his telephone’ was classified as human external by the first analyst and as human rule-based intervention by the second analyst. This discrepancy occurred only in one direction: in each case the first analyst, probably having more knowledge about the boundaries of the unit under investigation, chose an external factor and the second analyst, who performed the double rating for the reliability study, chose an internal factor.

### Discussion

#### General findings and interpretation

The mean score for agreement in content of causal trees was 2.0 on a scale from 0 to 3. More than half (54%) of the descriptions of root causes agreed, 17% agreed partly and for 29% there was no agreement or no equivalent. We consider these results as satisfactory, as most root causes showed complete agreement and since we believe that even partly agreeing root causes have the potential to end in the same ECM classification code. We know from comparing the causal trees that in many of the partly agreeing root causes, the basic information needed was present in the causal tree, but the root cause descriptions differed too much to judge them as totally agreeing.

The inter-rater reliability for the number of root causes used in the causal tree was only moderate (κ = 0.46). Beside the possibility that the analysts disagreed on the number of root causes, there are two alternative reasons why the number of root causes in the causal trees of two analysts differed. The first is that one of the analysts might have missed some relevant information in the unintended event description and therefore did not make enough ‘roots’ in the causal tree. The other possibility is that two root causes might have been combined into one. For example, one

### Table 4

<table>
<thead>
<tr>
<th></th>
<th>Level 0 (system)</th>
<th>Level 1 (main category)</th>
<th>Level 2 (SRK not subdivided)</th>
<th>Level 3 (complete ECM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agreement (%)</td>
<td>67.9</td>
<td>65.6 (0.58–0.72)</td>
<td>75.0</td>
<td>73.9</td>
</tr>
<tr>
<td>Kappa (95% CI)</td>
<td>-</td>
<td>0.70 (0.58–0.82)</td>
<td>0.67 (0.59–0.76)</td>
<td>0.63 (0.58–0.68)</td>
</tr>
<tr>
<td>Overall (κ = 212)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Humana (κ = 135)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Technicala (κ = 19)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Organizationala (κ = 20)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

* Root causes in which both analysts agreed on this main category level. For technical and organizational factors, calculations are only relevant for Level 3. For human causes, calculations are relevant for Levels 3 and 2 (SRK categories are human cause categories).
analyst formulated a root cause as: ‘Unannounced maintenance of digital patient record system on busy hours’, whereas the other formulated two root causes: ‘Maintenance of digital patient record system was not announced in advance’ and ‘Maintenance planned during hectic period of the day’.

The inter-rater reliability of classifying root causes according to the ECM was substantial at all levels of the ECM, indicating that analysts were adequately able to determine the same classification for a given root cause ($0.63 < \kappa < 0.71$). The higher reliability scores in our study for the subcategory levels may be due to the fact that we used a research team of experienced analysts who got together twice a month during the data collection period of 1.5 years to discuss any difficulties encountered in performing the analyses. These difficulties mainly concerned the choice of one or another subcategory.

**Comparison with previous studies**

As indicated in the introduction, the inter-rater reliability of classifying root causes with the ECM has previously been examined in very small studies [22, 23]. In our study of the inter-rater reliability of classifications (objective c), a large data set of 150 events with 212 root causes was used, resulting in more trustworthy kappa values. Our study confirms the findings of both studies with regard to the inter-rater reliability of the classifications at main category level. The higher reliability scores in our study for the subcategory levels may be due to the fact that we used a research team of experienced analysts who got together twice a month during the data collection period of 1.5 years to discuss any difficulties encountered in performing the analyses. These difficulties mainly concerned the choice of one or another subcategory.

**Limitations**

Our study has some limitations. The process of questioning nurses and physicians who were involved in the unintended event, in order to make the event description complete, was not included in our reliability study. The second analyst used the additional questions and answers that the first analyst had gathered from the healthcare providers beforehand. The first analyst wrote down this information on paper and made it available for the second analyst; the interview was not repeated. As a result, the inter-rater reliability of formulating a causal tree may be somewhat overestimated, because the second analyst probably would have asked different questions, possibly leading to different conclusions. It was, however, not possible to carry out a second independent interview process, since a healthcare provider cannot report an unintended event twice without being biased by questions raised by the first analyst.

A second limitation is that it was rather subjective to determine if two root causes agreed, partly agreed or not at all. We were, however, quite stringent in comparing descriptions of causal trees. We based our comparison on the tree which had the largest number of root causes of both trees, instead of the smallest number. When we had chosen for the latter, with 205 root causes to compare instead of 257, the agreement scores would have been higher: 67% agreement (A), 21% partial agreement (P) and 12% no agreement (N). The mean score for agreement according to the rules for comparison of descriptions (Table 2) would have been higher as well: 2.3 (SD = 1.02). The reason for these differences is that the 52 root causes identified by only one of both analyst were all valued as $\pi$.

Finally, the causal trees in our study were relatively small with on an average 1.5 root causes per event. There are two possible explanations for this. Firstly, we examined not just adverse events (with harm for patients) but also minor unintended events. Adverse events will probably result in larger causal trees because of the complexity of these events: often many causal factors are involved simultaneously. The unintended events we examined included small deviations from standard practice (with often only a few root causes). A second explanation for the relatively small causal trees is that we asked only questions to staff in the participating units and not in other units within the hospitals. Not only did this make the execution of our study more practical; but also we wanted to give our advice at unit level because this would increase the recognizability for the reporters and therefore the likelihood that the advice would be acted on. When the PRISMA analysis revealed causes present in other units, we classified them as external. It is possible that an external factor in fact had more underlying root causes. These remained hidden because of our choice to examine the participating units only: we had not enough information to make a further subdivision. We do not think that the size of the causal tree influences the inter-rater reliability of the classifications. However, it may have influenced the inter-rater reliability of the number of root causes. The likelihood that two analysts find two root causes is larger than the likelihood they both find eight, for example. Thus, the kappa value for number of root causes might have been smaller in a sample with larger causal trees.

**Implications for research and practice**

The results of this study indicate that causal tree analysis with PRISMA is a reliable tool for research into causes of unintended events at hospital units. As most discrepancies in ECM classifications were found in external causes, we suggest to leave out the distinction between organizational external, human external and technical external, and simply combine the external categories into a new ECM category ‘External’, for all failures beyond the control and responsibility of the investigating organization/department. In many cases, analysts do not have enough information about other organizations/departments to determine whether either an organizational-, or human- or technical external factor contributed to the event. This is because a PRISMA investigation ends when the system boundary is reached, that is when the accompanying measures are outside the range of influence of the organization/department [17, 18].

Moreover, a general recommendation for users of PRISMA is to get training in making causal trees and classifying root causes, as training of analysts increases reliability [27]. While the inter-rater reliability of the number of root causes was only moderate, we recommend to pay specific attention to revealing all root causes of an unintended event by tracing any additional causal factors in an event description and by dismantling all causes in the causal tree into separate root causes. It is important not to miss any relevant causal
factors, because the PRISMA profile gives direction to the development of preventive strategies and it is constructed by aggregating all root causes of a group of unintended events. With a training session addressing this issue, the inter-rater reliability for number of root causes might increase.

Supplementary data

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

Acknowledgements

We thank Sanne Lubberding who was involved in the initial design and coordination of the reliability study. We thank Hanneke Merten, Inge van Wagtendonk and Sanne Lubberding for their time and effort in performing the double analyses of the reported events.

Funding

The Dutch Patient Safety Research Program has been initiated by the Dutch Society of Medical Specialists (in Dutch: Orde van Medisch Specialists) with financial support from the Ministry of Health, Welfare and Sport. The Program is carried out by EMGO Institute/VUmc and NIVEL.

References


Accepted for publication 28 May 2009