Influences of definition ambiguity on hospital performance indicator scores: examples from The Netherlands

Helen A. Anema 1, Sabine N. van der Veer 2, Job Kievit 3, Elly Krol-Warmerdam 3, Claudia Fischer 1,4, Ewout Steyerberg 4, Dave A. Dongelmans 5, Auke C. Reidinga 6, Niek S. Klazinga 1, Nicolet F. de Keizer 2

1 Department of Public Health, Academic Medical Centre, University of Amsterdam, Amsterdam, The Netherlands
2 Department of Medical Informatics, Academic Medical Centre, University of Amsterdam, Amsterdam, The Netherlands
3 Department of Medical Decision Making, Leiden University Medical Centre, Leiden, The Netherlands
4 Department of Public Health, Erasmus Medical Centre, Rotterdam, The Netherlands
5 Department of Intensive Care, Academic Medical Centre, University of Amsterdam, Amsterdam, The Netherlands
6 Department of Intensive Care Medicine, Martini Hospital, Groningen, The Netherlands

Correspondence: Helen A. Anema, Department of Public Health, Academic Medical Center, University of Amsterdam, Room J2-211, PO Box 22660, NL-1100 DD Amsterdam, The Netherlands, tel: +31 20 56 64684, fax: +31 20 69 72316, e-mail: helen.anema@gmail.com

Research objective: Reliable and unambiguously defined performance indicators are fundamental to objective and comparable measurements of hospitals' quality of care. In two separate case studies (intensive care and breast cancer care), we investigated if differences in definition interpretation of performance indicators affected the indicator scores.

Design: Information about possible definition interpretations was obtained by a short telephone survey and a Web survey. We quantified the interpretation differences using a patient-level dataset from a national clinical registry (Case I) and a hospital's local database (Case II). In Case II, there was additional textual information available about the patients' status, which was reviewed to get more insight into the origin of the differences.

Participants: For Case I, we investigated 15,596 admissions of 33 intensive care units in 2009. Case II consisted of 144 admitted patients with a breast tumour surgically treated in one hospital in 2009.

Results: In both cases, hospitals reported different interpretations of the indicators, which lead to significant differences in the indicator values. Case II revealed that these differences could be explained by patient-related factors such as severe comorbidity and patients' individual preference in surgery date.

Conclusions: With this article, we hope to increase the awareness on pitfalls regarding the indicator definitions and the quality of the underlying data. To enable objective and comparable measurements of hospitals' quality of care, organizations that request performance information should formalize the indicators they use, including standardization of all data elements of which the indicator is composed (procedures, diagnoses).

Introduction

Performance indicators (PIs) are used to monitor and improve quality and safety of health care and to stimulate accountability and market processes in countries worldwide (e.g. USA (www.ahrq.gov), UK (www.hqip.org.uk) and Denmark (www.ikas.dk)). To accurately assess hospitals' performance, PIs need to be reliable measurement instruments, particularly when hospital performance rankings are published in lay press, or when indicator results are linked to reimbursement.

When indicators are self-reported, the reliability of PI scores is not self-evident, which can be partly explained by the complex process involved in indicator computation. For example, the computation of an indicator score often requires following steps of composite logic based on free-text descriptions in a manual (figure 1). Different hospitals can interpret these descriptions differently. Moreover, the available data are not always informative enough to decide whether a particular data element needs to be abstracted.

The complexity of the computation process is tightly connected to the probability of errors and consequently to the reliability of indicator scores. To keep the error rate to a minimum, the underlying data need to be complete, accurate, consistent and reproducible. Besides the indicator itself, all the underlying data elements need to be unambiguously defined. This is, however, not always as straightforward as it may seem. Yet, research regarding the extent of the effect of definition ambiguity on PI scores remains scarce, especially in the area of PI programmes aimed at health-care transparency or pay-for-performance programmes.

Here we investigated possible effects of definition ambiguity on PI scores within the area of intensive care (Case I) and breast cancer care (Case II). These PIs were chosen on basis of their national and international history in data-driven quality monitoring and quality improvement (see, for instance, UK, Spain and several other European countries). Moreover, as several abstraction steps are required to be taken in the calculation process, each with their own definitions and instructions, these PIs provide good examples of indicators of which interpretation uniformity is at stake. In Case I, we quantified possible effects of definition ambiguity using data from a national registry for a large number of hospitals. In Case II, we zoomed in on one hospital, and aimed to obtain more insight into possible mechanisms underlying the interpretation differences.

Methods

Intensive care

PI ‘Duration of Mechanical Ventilation (DMV)’

Mechanical ventilation is a frequently applied treatment at the intensive care unit (ICU). Long ventilation duration creates...
substantial discomfort for a patient\textsuperscript{15} and increases the risk of respiratory infections.\textsuperscript{16} It is therefore important to minimize mechanical ventilation duration.

Since 2007, the Dutch Health Care Inspectorate (IGZ)\textsuperscript{21} [Each year the IGZ publishes the updated definitions in instruction manuals on their public Website (www.igz.nl)] uses ‘Duration of mechanical ventilation’ as a process indicator to monitor national patient safety. Dutch hospitals are obliged to submit self-calculated indicator scores to the Inspectorate, which are subsequently published at www.ziekenhuistransparant.nl.

In 2009, the numerator of this indicator was defined as the total number of invasive or non-invasive ventilation days in a given period. The denominator is the total number of patients in that same period. A ventilation day is defined as a day at the ICU during which a patient was ventilated invasively or non-invasively, at any time.

**Data and participants**

To investigate the effect of definition ambiguity on the DMV indicator score, we used raw data from the National Intensive Care Evaluation (NICE) registry. This registry contains data from patients admitted to Dutch ICUs, and provides insight into the effectiveness and efficiency of ICU care in The Netherlands.\textsuperscript{22} Currently, 81 ICUs covering 90\% of all Dutch ICUs voluntarily participate in the NICE registry. Most ICUs participating in NICE...
The Dutch Collective Cancer Centers (CCC) \[17,18\] results are known \[NABON nota 2008; in Dutch; published by it is important not to delay the surgery >4 weeks after laboratory for the patient to decide on the specific treatment options. However, removal of the tumour'. The time between the final laboratory PI: Percentage of patients operated within 4 weeks

Breast cancer

When means are described, standard deviations (SD) are added. Significantly different results. Indicator scores are denoted in days, and variance—including a Bonferroni post-hoc test—to investigate if definitions and calculation instructions and determined the overall scores (percentages) for each definition separately. Differences between the separate indicator scores were tested by means of chi-square analyses. Finally, by checking the available textual information about the status of the patient, we explored what type of patients would or would not be included in the various numerator and/or denominator populations.

Results

Intensive care

Interpretations under investigation

The telephone survey with the hospitals resulted in three ways to interpret the definition of the DMV indicator (table 2). Interpretation A: for a specified period, calculating the numerator by adding the

## Table 1 Characteristics of included ICUs

<table>
<thead>
<tr>
<th>Number included</th>
<th>33</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type of hospital</td>
<td></td>
</tr>
<tr>
<td>Academic</td>
<td>5 (15)</td>
</tr>
<tr>
<td>Teaching</td>
<td>15 (45)</td>
</tr>
<tr>
<td>Non-teaching</td>
<td>12 (36)</td>
</tr>
<tr>
<td>Number of ICU beds</td>
<td></td>
</tr>
<tr>
<td>&lt;10</td>
<td>5 (15)</td>
</tr>
<tr>
<td>10–20</td>
<td>19 (58)</td>
</tr>
<tr>
<td>&gt;20</td>
<td>9 (27)</td>
</tr>
<tr>
<td>Median (IQR) number of admissions*</td>
<td>852 (641–1204)</td>
</tr>
</tbody>
</table>

IQR, interquartile range.
a: As submitted to the NICE registry in 2009.
Values are numbers (percentages), unless indicated otherwise.

The denominator is defined as the number of patients who had primary surgery of a primary (malign) breast tumour in 2009, including patients with a direct breast reconstruction. Patients who are primarily treated with neoadjuvant therapy are excluded from the denominator. The numerator is defined as the number of patients who were operated within 4 weeks after the laboratory results were known.

Data and participants

The data used for this case were obtained from a large academic hospital (>600 beds) encompassing female patients who were admitted in 2009 with a primary breast tumour \(N=144\). Patient-level data (e.g. patient ID, date of surgery) were routinely registered in a local data spreadsheet by a mamma care nurse practitioner, as part of clinical practice and internal quality monitoring. Also, a free-text field was used to provide for background information important for clinical decision-making.

Identifying different interpretations of the indicator definition

The possible interpretations of the indicator definition were derived from ongoing research by means of telephone interviews with three hospitals. ICUs were asked whether they experienced difficulties while collecting and computing the indicator score and what interpretation of the indicator they used (according to which algorithm) to calculate the indicator score.

Analysis

We calculated the ventilation duration on a patient level for each obtained interpretation of the DMV indicator. Next, we determined the denominators and numerators per ICU based on the different definition interpretations and calculated accordingly different versions of the DMV indicator score for each of the 33 included ICUs. Finally, we performed a univariate weighted analysis of two sets of indicators. One questionnaire item specifically targeted the difficulties that a hospital experienced while collecting and computing the indicator score.

We invited all 100 hospitals in The Netherlands, of which 24 were small (<320 beds), 48 intermediate (320–627 beds) and 28 large (>627 beds). Regarding hospital type, eight were university hospitals, 28 teaching and 64 non-teaching hospitals. In total, 41 hospitals (response rate, 41%) gave informed consent to participate in the study and returned the questionnaire. The responders included 5 small, 27 intermediate and 9 large hospitals; 2 university, 13 teaching and 26 non-teaching hospitals. The participating institutions did not differ from non-responders with regard to hospital size \(X^2(2) = 4.12, P = 0.128\) or type of hospital \(X^2(2) = 0.462, P = 0.794\).

Analyses

First we calculated the number of patients for the denominator and numerator based on the provided definitions and calculation instructions and determined the overall scores (percentages) for each definition separately. Differences between the separate indicator scores were tested by means of chi-square analyses. Finally, by checking the available textual information about the status of the patient, we explored what type of patients would or would not be included in the various numerator and/or denominator populations.

Breast cancer

PI: Percentage of patients operated within 4 weeks after laboratory results

‘In general, the first treatment against breast cancer is surgical removal of the tumour’. The time between the final laboratory result and the date of surgery is used to communicate the result to the patient, to discuss the case in a multi-disciplinary team and for the patient to decide on the specific treatment options. However, it is important not to delay the surgery >4 weeks after laboratory results are known \[NABON nota 2008; in Dutch; published by the Dutch Collective Cancer Centers (CCC)].\[17,18\]
duration of all non-invasive or invasive mechanical ventilation episodes at the ICU, and determining the duration of one episode by taking the difference between start and end time of that episode, using fractional days (i.e. number of hours divided by 24). The denominator for this interpretation was defined as all ventilated (non-invasive and invasive) ICU admissions during that same period. Interpretation B: for a specified period, calculating the numerator as defined for interpretation A, but including all ICU patients (ventilated and non-ventilated patients) admitted in that period to determine the denominator. Interpretation C: for a specified period, calculating the numerator as defined for interpretation A, but using calendar days instead of fractional days. The denominator was the same as for interpretation A. The term ‘fractional days’ used in interpretation A and B refers to the number of hours divided by 24. For example, when a patient is mechanically ventilated from 10 PM on one day until 5 AM the next (7 h), the value used for the numerator in interpretation A and B is 0.3 fractional days, whereas in interpretation C, this would be 2 calendar days.

Effect on indicator scores

Table 2 shows the different indicator scores observed for each interpretation. Calculating the DMV score based on interpretation B, including all ICU patients, yielded the lowest score [mean = 0.1 day (SD = 0.13)], whereas calculation on basis of calendar days (definition C) led to the highest ventilation score and the largest variance [mean = 3.7 days (SD = 0.7)]. The differences between interpretation A and B can be explained by the inflated denominator caused by including all ICU patients in interpretation B instead of all ventilated ICU patients in interpretation A [mean denominator A = 1373 (±959); B = 1634 (±1100)]. The difference between interpretation B and C originated in the numerator [mean numerator B = 273 (±137); C = 473 (±204)]. Moreover, these differences were statistically significant, as was shown by the univariate weighted analysis of variance [F(1,2) = 331.4, P < 0.01], and the separate post-hoc pairwise comparisons for interpretation A versus interpretation B [t(2) = 3.41, P < 0.01], and for interpretation B versus interpretation C [t(2) = −4.1343, P < 0.01].

Breast cancer

Interpretations under investigation

Five of 41 hospitals added detailed information in the questionnaire regarding the interpretation difficulties they observed. In all, differences interpretations were observed of the date that was used to start calculating the number of weeks before the date that surgery was performed. Three possible interpretations of the indicator could be formulated: (i) using the date of biopsy as an approximate, (ii) using the date of biopsy + 7 days as an approximate (definition used by CCC) and (iii) using the date that a patient was scheduled for surgery (placed on the waiting list).

Effect on indicator scores

Exclusion of the patients with neoadjuvant therapy (n = 10) and patients for whom the date of biopsy was missing (n = 6) resulted in a denominator population of 128 patients. Calculating the numerators for the three separate definitions (see table 3) revealed between-definition differences from 8 to 30%. These differences were significant for definition 1 and 2 and 1 and 3 [Diff = 22%, χ²(1) = 12.0, P < 0.01, Diff = 30%, χ²(1) = 24.7, P < 0.01].

Table 2 Differences in numerator, denominator and indicator score values between interpretations of the definition of duration of mechanical ventilation

<table>
<thead>
<tr>
<th>Interpretation</th>
<th>Value (SD)</th>
<th>Mean indicator score (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Numerator Calculation of the duration of one ventilation episode (difference between start and time of mechanical ventilation) using fractional days</td>
<td>237 (137)</td>
</tr>
<tr>
<td></td>
<td>Denominator Including only ventilated ICU admissions</td>
<td>1373 (959)</td>
</tr>
<tr>
<td>B</td>
<td>Numerator Calculation of the duration of one ventilation episode (difference between start and time of mechanical ventilation) using fractional days</td>
<td>237 (137)</td>
</tr>
<tr>
<td></td>
<td>Denominator Including all ICU admissions</td>
<td>1634 (1100)</td>
</tr>
<tr>
<td>C</td>
<td>Numerator Calculation of the duration of one ventilation (difference between start and end time of mechanical ventilation) episode using calendar days</td>
<td>473 (204)</td>
</tr>
<tr>
<td></td>
<td>Denominator Including only ventilated ICU admissions</td>
<td>1373 (959)</td>
</tr>
</tbody>
</table>

Values are means (SD) as calculated for all 33 ICUs.

Table 3 Numerator, denominator and indicator score values per interpretation of the definition of % patients who are operated within 4 weeks

<table>
<thead>
<tr>
<th>Interpretation</th>
<th>Value</th>
<th>Mean indicator score</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Numerator Number of patients with a primary breast tumour primarily treated with surgery &lt;4 weeks after biopsy date</td>
<td>78</td>
</tr>
<tr>
<td></td>
<td>Denominator All patients with a primary breast tumour primarily treated with surgery</td>
<td>128</td>
</tr>
<tr>
<td>B</td>
<td>Numerator Number of patients with a primary breast tumour primarily treated with surgery &lt;4 weeks after biopsy date + 7 days</td>
<td>105</td>
</tr>
<tr>
<td></td>
<td>Denominator All patients with a primary breast tumour primarily treated with surgery</td>
<td>128</td>
</tr>
<tr>
<td>C</td>
<td>Numerator Number of patients with a primary breast tumour primarily treated with surgery &lt;4 weeks after date on surgery waiting list</td>
<td>116</td>
</tr>
<tr>
<td></td>
<td>Denominator All patients with a primary breast tumour primarily treated with surgery</td>
<td>128</td>
</tr>
</tbody>
</table>
Of the 128 patients treated with surgery, nine were delayed according to interpretation A and C. Forty-one surgical interventions were delayed according to interpretation A but not C. Based on the additional medical information that was available in free text, we observed several reasons why patients were not timely available for surgery. For instance, nine patients decided to postpone their surgery. Other delays could be explained by severe comorbidity (n = 3), complications (n = 2) or immediate reconstructions (n = 3).

**General Discussion**

Our study showed that different interpretations of indicator definitions significantly altered the indicator score, implying the need to disambiguate indicator definitions before they can be used for healthcare transparency or pay-for-performance programmes.

The investigated interpretations of the ICU indicator DMV are real examples given by hospitals that use the IGZ construction manuals when collecting the data, compute and report the indicator scores. Some hospitals use computerized search algorithms to abstract the data from the various information systems. Owing to ambiguity in the definition, these search algorithms led to the inclusion of the non-ventilated patients in the denominator. In turn, the inflated denominator led to low ventilation scores. Recently, the unit of the IGZ definition changed from calendar days to calendar hours (= fractional days). As hospitals are not obliged to submit patient-level data, they can accidently submit indicator scores that are incorrectly composed based on calendar days instead of hours, as was observed. Nevertheless, even when the indicator score is composed on basis of fractional days using computerized data abstraction, the definition of start and end data lacks specificity to accurately write the syntax. For instance, to test a patient’s ability to breathe independently, the ventilation is deliberately stopped, after which the ventilator is reconnected. Does this create a new start or is it a continuation of breathing time, as the breathing tube itself is not removed? Should the patient be admitted or ventilated in the reporting year? Also, medical concepts underlying indicator definitions, such as diagnostic categories or medical procedures, are not always well-defined. In our ICU case study, for example, the indicator definition does not specify if ventilatory support through a nasal mask should also be considered as non-invasive ventilation.

Similar results were observed for the breast cancer indicator Percentage of patients operated within 4 weeks after laboratory results are known. The significant difference between definition interpretations can be explained by the lack of a uniform definition of ‘date of lab result in the computer’. Hospitals solved this problem by using ‘date of biopsy’, ‘date of biopsy + 7 days’ or ‘date of patient on surgery waiting list’ as an approximate. Also, several patients chose to delay the surgery or they were not fit enough to be operated on short notice. Whether these patients were included or excluded depended on the chosen indicator definition. As an indicator is expected to measure quality of care provided by a care taker, such variation in patient characteristics (e.g. casemix) threaten the validity of the indicator definitions.1

To support hospitals with correctly calculating indicator scores and to improve data quality, various steps can be made. Arts et al.12 proposed a generic framework that can be used to improve data quality in medical registries. Several suggested elements can be relevant for the area of health-care transparency and benchmarking (e.g. define a data dictionary, a data collection protocol, compose data checks or set up a quality assurance plan). Good-quality data start with good definitions, definitions that are uniformly defined, preferably within an international context. Moreover, a formalization process during the indicator development phase seems vital, as every hospital collects its own data and has its own local data infrastructure, and the organizing institutions (IGZ, DHTP) have no insight into the underlying data. To create unambiguous definitions, the language of all individual concepts needs to be formalized; a process that is well-known from the area of computer science.24,25

In Sweden, for example, National Quality Registries (www.kvalitetsregister.se) have been developed, and are currently managed by representatives of the professional groups that use them. The data in these registries are regarded as reliable and valid (e.g. the Swedish National Total Hip Arthroplasty registry26), and internationally these data registries are recognized as providing a unique opportunity to monitor and continuously improve the health services. Importantly, these initiatives are strongly supported by the Swedish government.

Besides tightening the indicator definitions, a possible solution to eliminate between-hospital differences caused by ambiguous definition might be provided by centrally coordinating the data abstraction (NICE registry, NKR). This would at least suggest that all indicator scores are based on comparable underlying data and calculated in a similar manner. Alternatively, organizations that develop PIs should define them in a formalized way and should ideally publish the PIs as queries based on standard information models and terminologies,25 which can be run by individual hospitals. In Germany, health-care quality is housed in a national independent institute for quality and patient security (BQS) (http://www.bqs-institut.de/). It centrally organizes data abstraction and data handling, suggesting that the indicator scores are based on one interpretation of the definition.

In summary, with this article we hope to increase the awareness on pitfalls regarding the indicator definitions and the quality of the underlying data, even when indicators already have been used for several years. Indicator developers, users and the scientific field thus need to engage more in research, using both qualitative and quantitative methods, targeting the development of reliable PIs.

**Acknowledgements**

This research was carried out under the auspices of the Consortium Quality of Care of the Dutch Federation of University Medical Centres (NFU), a collaboration of the eight university medical centres that focuses on research, education and policy in the domain of quality of care.

**Funding**

Disclosure of funding received: The Dutch Federation of University Medical Centres (NFU) has received a grant of the Dutch ministry of Health, Welfare and Sport to carry out this research.

**Conflicts of interests:** None declared.

**Key points**

- Although developed within a multi-disciplinary team, PI definitions are not always unambiguously defined, as is shown by the different interpretations of the definitions that are given by Dutch hospitals.
- Different interpretations of indicator definitions significantly altered the indicator score, rendering them useless for health-care transparency purposes or pay-for-performance programmes.
- This research demonstrates the need to centrally coordinate the data abstraction, as it would at least suggest that all indicator scores are based on comparable underlying data and calculated in a similar manner.
- Moreover, organizations that define PIs should do this in a formalized way and should ideally publish the PIs as queries based on standard information models and terminologies.
References