**Intensive care unit benchmarking**

*Prognostic models for length of stay and presentation of quality indicator values*

Verburg, I.W.M.

---

**Creative Commons License** (see https://creativecommons.org/use-remix/cc-licenses):

**Other**

**Citation for published version (APA):**

1 | General Introduction
Due to a drive for continuous quality improvement, a pressure on accountability and budgetary constraints, the awareness for quality of care has grown among various stakeholders including care providers, healthcare managers, insurance companies, governmental bodies and patients. Healthcare institutions search for quality indicators to identify room for quality of care improvement [1–4]. This has led to the formulation of numerous quality indicators across all fields of clinical medicine. In some cases, it is obvious what the target (preferred) value should be. However, for many indicators, no clear target values are available to define good care. For example, mortality rates in hospitals should be as low as possible, but it would be unrealistic to demand that no hospitalized patients die. In the absence of clear external target values, healthcare institutions often compare themselves to their own historical values or with their peers in a process called benchmarking.

Ideally, the quality indicator values represent the true quality of care provided by an institution and differences in the quality indicator values between institutions would indicate that institutions with worse values could improve their quality of care. Figure 1.1 presents an adaption of a previously published schematic representation of causes of differences in quality indicator values between healthcare institutions [5]. Observed differences in quality indicators will ideally be the result of genuine differences in quality, but can also arise from noise caused by registration biases, differences in patient characteristics, residual confounding, and random variation. This noise may influence observed changes over time or differences between institutions and could lead to incorrect judgments, being made about institutions.

Errors may occur in the registration of patient data and may result in registration bias and inadequate data quality meaning that a quality indicator is less reliable [6]. Examples of causes of registration bias are differences in: how data are collected, such as the use of multiple electronic patient record systems; definitions; and interpretation. Quality registries should aim to minimize registration bias by standardization and monitoring data quality [7, 8].

In addition to random variation, differences in patient characteristics (case-mix) such as age, sex or severity of illness may influence care outcomes and thus quality indicator values. For example, higher in-hospital mortality rates or longer in-hospital length of stay may result from a more severely ill patient population. Fair and meaningful benchmarking requires correction for differences in patient case-mix. Prognostic models can partially correct for differences in patient case-mix. However, even after case-mix correction quality indicator values may still be influenced by patient characteristics, for which adjustment was not or inadequately performed. This is known as residual confounding. Theoretically, differences in the values of correctly constructed quality indicators should reflect true differences in the quality of care.
Defining meaningful quality indicators to measure the quality of care is difficult since different quality indicators may reflect different aspects of performance. No single quality indicator reflects the whole spectrum of healthcare performance. Therefore, a set of quality indicators reflecting structure, process, and outcomes of health care are often presented to identify room for quality of care improvement and to support policy decisions [9].

There are many possible ways of presenting quality indicator values to stakeholders to support them in making decisions about the quality of care. Methods of identifying institutions with outlying performance include: simple descriptive statistics; league tables; Bayesian ranking; the probability of being in the worst-ranked group of institutions; preset limits for acceptable performance, such as 95% confidence intervals; statistical process control (SPC) charts; funnel plots; variable life adjusted display (VLAD) curves; and risk-adjusted exponentially weighted moving average (RA EWMA) plots [10]. In this thesis, we focus on league tables and funnel plots.

All studies included in this thesis have been performed in the context of the Dutch National Intensive Care Evaluation foundation (NICE) registry, a quality registry for Dutch intensive care units (ICUs). The remainder of this chapter introduces the domain of intensive care and the NICE registry. In addition, prognostic models to adjust for case-mix differences and other unexplained differences, such as organizational aspects, are introduced. Furthermore, league tables and funnel
Chapter 1

plots and their ability to take random variation into account, are introduced as methods of presenting quality indicator values. The chapter concludes with the objective and an outline of the thesis.

1.2 Intensive care units and National Intensive Care Evaluation foundation

Intensive care is defined as 'a service for severely ill patients with potentially recoverable conditions, who can benefit from more detailed observation and more intensive treatment than can safely be provided in general wards or high dependency areas' [11]. This definition originates from the end of the 20th century. Since then intensive care has expanded and evolved. Nowadays, ICU care is very complex and delivered in a highly technical and labor-intensive environment. As these developments have occurred, the survival chances of critically ill patients has drastically improved. But the cost of intensive care has also increased substantially, resulting in a high proportion of the health care budget being spent on ICUs [12]. This all makes ICUs a particularly interesting part of the hospital to assess and improve performance.

For this reason, the NICE foundation [13, 14] was established in 1996 by a group of intensivists. Its purpose is to facilitate quality monitoring and quality improvement initiatives in Dutch ICUs. At the start of the registry, a small proportion of all Dutch ICUs voluntarily participated. Over the years the NICE registry has expanded and, currently, all Dutch ICUs participate.

Figure 1.2 gives an overview of data registration, analyses, and benchmarking by the NICE registry. All ICUs register a core dataset including demographic, diagnostic, and physiological data from the first 24 hours after ICU admission. Furthermore, they register outcome data, such as: ICU and in-hospital mortality; ICU readmission; and ICU and hospital length of stay. In addition, most ICUs participate in an additional quality indicator registry of NICE. This consists of structure, process, and outcome indicators chosen by the Dutch Society of Intensive Care (NVIC). Examples are: the number of ICU and hospital beds; staff resources; nurse-to-patient ratio; glucose regulation; and duration of mechanical ventilation. Furthermore, the NICE registry has several other optional registration modules. These are: complications; sepsis; sequential organ failure assessment (SOFA); and nursing workload. The analyses presented in this thesis use data from the core dataset. In addition, chapter 4 uses data from the NVIC quality indicators.

To improve the quality of the data collected, the NICE registry uses strict definitions and specifications of the data collected, described in a data dictionary. It provides participants with e-learning based training, provides participants with a mandatory training before starting the registration and performs data quality
General Introduction

Stakeholders
- Benchmark reports
- Web-based dashboard
- Public website
- Prepared datasets
- Analyses (e.g. case-mix correction)

Hospital

ICU

Internal hospital database

Figure 1.2: Overview of data registration, analyses and benchmarking by the NICE registry.

ICU patients input:
- Demographics
- Severity of illness
- Diagnosis

Patient outcome:
- ICU and in-hospital mortality
- Readmission to the ICU
- ICU and in-hospital length of stay

Data dictionary and training

Data extraction of registry modules

Data security and encryption

Prepared datasets

Upload data to NICE

Automatic data quality checks

NICE database

Analyses (e.g. case-mix correction)

Data validation report

Data reminder

Data quality audits

Audit report

Public website

Benchmark reports Web-based dashboard

Stakeholders
controls, such as automated checks on data entry and onsite data quality audits [6, 15].

ICU patients are very heterogeneous. They may have a high severity of illness or have undergone major surgery. Post-surgical patients often have low mortality rates and short ICU length of stay. Hence, it is meaningless to compare the outcome of different ICUs without proper case-mix correction. The NICE registry corrects the quality indicator in-hospital mortality for case-mix using the Acute Physiology and Chronic Health Evaluation (APACHE) IV model [16]. However, the NICE registry does not correct the quality indicator length of stay for case-mix. This thesis focusses on prognostic models used for case-mix adjusted ICU length of stay.

The NICE registry performs analyses on the registered data and supplies feedback on the values of a set of quality indicators to the participating ICUs through reports issued every 6 months and a web-based dashboard application. This enables ICUs to benchmark their performance to national values and to groups of ICUs of comparable size which can be used to identify critical points in the care process as starting point for improving quality of care. Like other ICU quality registries [17, 18], the NICE registry has offered ICUs the opportunity to make the outcomes of some quality indicators publically available on a website [19] since 2013. This makes performance transparent to all stakeholders.

Several methods are used within the NICE registry to present performance on quality indicators such as SPC charts [20], VLAD curves [21], RA EWMA charts [22], funnel plots, and caterpillar diagrams. Results are presented for the entire cohort and for subgroups of ICU patients [19].

Currently, data of more than 1,000,000 ICU admissions have been included in the NICE database and about 85,000 new admissions are registered annually. In 2016 the overall in-hospital mortality of ICU patients in the Netherlands was around 13%. The mean ICU length of stay was 2.9 days (median 1.0 day) for ICU survivors and 5.1 days (median 2.1 days) for ICU non-survivors. The overall percentage of patients readmitted to the ICU was around 6% and the overall percentage of patients readmitted to the ICU within 48 hours after ICU discharge around 2%.

1.3 Prognostic models for intensive care unit length of stay

The first part of this thesis addresses prognostic models for ICU length of stay. Since costs are strongly related to ICU length of stay [23, 24], ICU length of stay can play an important role in examining the efficiency of care. As discussed earlier in this section, ICU patients form a heterogeneous population with patients with a wide range of complex health issues, each of which may have a different association with ICU length of stay [25]. For example, severe trauma patients
may have a very long ICU length of stay while most post-surgical patients will be discharged from the ICU very quickly. Quality indicators for ICU length of stay are meaningless if they are not properly corrected for patient case-mix. For in-hospital mortality, prognostic models have been proposed and are widely implemented to adjust in-hospital mortality for patient case-mix. Prognostic models for ICU length of stay are less frequently used as little consensus exists on the best method for predicting ICU length of stay and the predictive performance of existing models is modest [26–29]. To date, the NICE registry presents crude mean and median ICU length of stay and does not correct reported values of ICU length of stay for differences in case-mix between ICUs.

The accurate prediction of ICU length of stay is challenging for three main reasons. Firstly, the distribution of ICU length of stay is typically strongly skewed to the right with both a long tail and an inflated number of values close to zero. Secondly, the association between severity of illness and ICU length of stay differs for ICU survivors and ICU non-survivors [30]. Thirdly, the characteristics of individual ICUs can be associated with patient level ICU length of stay. Examples are discharge policies and the availability of spare beds on general wards [31–34]. The focus of this thesis and of quality registries, such as the NICE, is primarily on benchmarking. However, in chapter 2 and chapter 4 of this thesis, we extent our focus to address three reasons for predicting ICU length of stay. These are: 1) benchmarking; 2) planning the number of beds and members of staff required to fulfill demand for ICU care within a given hospital or geographical area; and 3) identifying individual patients or groups of patients with unexpectedly long ICU length of stay to drive direct quality improvement [35, 36].

The requirements for an ICU length of stay prediction model differ between these situations. A model for benchmarking purposes needs to predict ICU length of stay reliably at the ICU level. A model for planning or identifying patients with unexpected long ICU length of stay needs to predict ICU length of stay reliably for individual patients. Including ICU organizational characteristics in a prognostic model for ICU length of stay will limit the usefulness of a prediction model for benchmarking purposes. This is because this type of model will adjust for a part of the variation in the quality indicator values that can be attributed to quality of care. However, for planning the number of beds and members of staff required, including ICU organizational characteristics might be valuable. In addition, a model to identify individuals or groups with unexpectedly long ICU length of stay needs to predict ICU length of stay reliably for individual patients. It might be valuable to include ICU organizational characteristics in this type of prognostic model for ICU length of stay.
Chapter 1

1.4 Presentation of quality indicator values

The main focus of the second part of this thesis is on the presentation of quality indicator values. As we previously described, registries often report a set of quality indicator values to a range of stakeholders. Besides ICU length of stay, in-hospital mortality and readmissions to the ICU are often used as quality indicators for ICU care. Several studies found that patient level outcomes of ICU care are interrelated and influence each other [37–44]. The second part of this thesis addresses whether it is sufficient to report a single quality indicator. This would be the case if ICUs that perform well on one quality indicator also perform well on other quality indicators. This would not be the case if different ICU quality indicators reflect different aspects of performance. This thesis also addresses league tables and funnel plots, which are visual methods of presenting quality indicators. We describe league tables and funnel plots below.

1.4.1 League tables

League tables are frequently used to present comparative performance results [10]. Figure 1.3 presents a league table for benchmarking ICUs, for the general ICU population reported to the NICE registry. In the league table, ICUs are ranked according to their values of the in-hospital standardized mortality ratio (SMR) over the year 2016. We present 95% confidence intervals around the values of SMR. The SMR is defined as the number of deaths actually observed in an ICU divided by the number of deaths predicted by the APACHE IV [16] model. The purpose of league tables is to discriminate between the best and worst performing institutions. It enables staff in underperforming institutions to recognize the need for improvement and to identify the best performing institutions with a high rank in the league table, from which they can learn and develop improvement strategies.

Although it is possible to add confidence intervals to the ranks of a league table [41, 45, 46], statisticians have raised concerns about the reliability of league tables to discriminate between institutions [45–48]. League tables focus on observed differences between hospitals. However, users of league tables may ignore uncertainties in performance due to limited sample size even though confidence intervals have been added. If the sampling error is substantial, and hence the signal-to-noise ratio is low [49], league tables may be unreliable. This is because the rank of a particular institution may be largely determined by chance rather than the underlying quality of care it provides. The reliability of a league table can be expressed in terms of its rankability [50]. Rankability expresses the percentage of variation between ICUs' observed quality indicator values that is due to unexplained differences rather than random variation. Unexplained differences hypothetically reflect differences in quality of care, see figure 1.1. Higher values of rankability correspond to more reliable league tables [49, 50]. The concept of reliability in terms of rankability has not previously been applied.
to ICU league tables, although risk adjusted mortality has been used in league tables to compare the performance of ICUs [41].

Figure 1.3: League table of Dutch ICUs based on case-mix adjusted in-hospital mortality.
1.4.2 Funnel plots

Funnel plots form an alternative for league tables. Funnel plots are graphical decision-making tools to assess and compare the clinical performance of a group of institutions on quality indicators against a pre-defined benchmark [51]. A funnel plot is an example of a Shewart control chart, which were originally meant for quality control at the Western Electric company in the 1920s. Although Shewart control charts were used in industry, where quality control of manufacturing and other processes and reporting business performance were essential for remaining competitive, they have only been applied in healthcare since around 1990 [10].

Figure 1.4 presents a funnel plot of ICU performance, based on the same data as used for figure 1.3. In a funnel plot, the value of a quality indicator for each institution is plotted against a measure of its variation, often the number of patients or cases used to calculate the quality indicator. Control limits indicate a range, in which the values of the quality indicator would, statistically speaking, be expected. The control limits form a funnel shape around the benchmark, which is presented as a horizontal line. If an institution falls outside the control limits, it is seen as performing differently than expected, given the value of the benchmark [51–53].

It is important that we can assume that institutions falling outside the control limits perform significantly different than expected, given the benchmark. It is also important that there is no reason to suspect that institutions falling inside the control limits are performing differently from the benchmark. Incorrect
judgements may have severe consequences, such as loss of trust among patients, insurance companies refusing to pay for care delivered or award new contracts, and demotivated health care staff. Hence, the methods used to construct funnel plots need to have a solid justification in statistical theory. Incorrectly constructed funnel plots could lead to highly consequential incorrect judgements about hospital performance.

Prior papers comparing hospital performance by funnel plots are mostly based on a single paper on funnel plot methodology [51]. In this paper, Spiegelhalter describes multiple methods of constructing control limits. We found no publications describing a guideline for producing funnel plots, which describe all steps required when producing a funnel plot.

1.5 Objectives of this thesis

This thesis aims to contribute to knowledge on improving the reliability and accuracy of ICU benchmarking and is divided into two themes: 1) prognostic models for ICU length of stay; and 2) the presentation of quality indicator values. The two research questions on prognostic models for ICU length of stay in this thesis are:

1. Is it feasible to predict ICU length of stay accurately using regression methods and patient characteristics only?

2. What is the role of ICU organizational characteristics in predicting ICU length of stay?

The three research questions on the presentation of quality indicator values are:

3. Are case-mix adjusted ICU outcomes mutually independent measures for ICU quality of care?

4. What is the rankability of league tables for in-hospital mortality of ICU patients, and how can it be improved?

5. Which steps and choices need to be made to construct reliable funnel plots?
1.6 Outline of this thesis

This thesis contains eight chapters. To address the research questions on prognostic models for ICU length of stay, three studies were performed. In chapter 2, we systematically reviewed the reporting and methodological quality of models predicting ICU length of stay. In chapter 3, we compared ordinary least square (OLS) regression, generalized linear models (GLM)s, and Cox proportional hazards (CPH) regression to predict individual patient ICU length of stay. In chapter 4, we added ICU organizational characteristics to a regression model correcting for patient case-mix and assessed the influence of these characteristics on ICU length of stay and the change in model performance.

To address the research questions on presentation of quality indicator values three studies were performed. In chapter 5, we examined the associations between outcome-based quality indicators for in-hospital mortality; readmission to the ICU within 48 hours of ICU discharge; ICU length of stay. In chapter 6, we evaluated the rankability of a league table of ICUs based on case-mix adjusted in-hospital mortality. In chapter 7, we conducted a literature search to identify the steps in the process of funnel plot development. We applied the steps identified to an example for crude proportion of mortality and SMR in the NICE registry.

This thesis concludes with chapter 8, which provides an overall discussion of the principal findings of the work.