Intensive care unit benchmarking
Verburg, I.W.M.

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
8 | General Discussion
8.1 Introduction

This thesis contributes to the process of benchmarking intensive care unit (ICU) performance. Benchmarking helps to identify areas of improvement and thus initiate quality improvement activities. We focused on models for adjusting observed length of stay for patient characteristics and on methods of presenting values of quality indicators to stakeholders. Although several papers have been written about prognostic models for ICU length of stay, these models are less often used in practice than models for in-hospital mortality of ICU patients. Several ICU quality registries, including the Dutch National Intensive Care Evaluation foundation (NICE) registry, report crude outcomes for ICU length of stay as quality indicators for the efficiency of ICU care. Case-mix adjusted quality indicators for ICU length of stay have not been available. However, benchmarking ICUs is less meaningful if case-mix correction is omitted. Therefore prognostic models to properly adjust ICU length of stay for case-mix are important. This thesis addressed the development and assessment of prognostic models for ICU length of stay, mainly for benchmarking purposes but also for bed capacity planning and staff resourcing and the identification of patients with unexpectedly long ICU length of stay.

Benchmark reports need to be reliable and understandable for the general public and care providers, who need to take action depending on the results presented. Different quality indicators may capture different aspects of the quality of ICU care. These differences may influence the choice of quality indicators registries present. This thesis addressed whether ICU quality indicators, based on in-hospital mortality, ICU readmission and ICU length of stay, are independent of each other. As there are many ways of presenting quality indicators, this thesis addressed the utility of league tables and funnel plots in ICU benchmarking.

This chapter provides an overall discussion of the research presented in this thesis in the context of the research questions defined in chapter 1:

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?

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

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?

For each research question, the main findings are summarized and discussed.
8.2 Prognostic models for intensive care unit length of stay

The first research question 'Is it feasible to predict ICU length of stay accurately using regression methods and patient characteristics only?' was addressed in chapter 2 and chapter 3. In chapter 2, we systematically reviewed the suitability of existing models for predicting ICU length of stay for three reasons for predicting ICU length of stay: 1) benchmarking; 2) planning capacity in terms of number of beds and staff; and 3) identifying individual patients or groups of patients with unexpectedly long ICU length of stay to drive direct quality improvement. In chapter 3, we examined the utility of different regression methods in predicting individual patient ICU length of stay following an unplanned ICU admission.

In the systematic review, we found eleven studies describing the development and validation of 31 prediction models. These included the eight models we developed in chapter 3 [56], and three studies [26, 56, 67] externally validating the APACHE IV model [28]. None of these models fulfilled all of our requirements for patient level predictions of ICU length of stay. In particular, the accuracy of patient-level predictions was insufficient. To investigate the use of these models for benchmarking, we also evaluated model performance at ICU level. Two models fulfilled our requirements for accuracy at ICU level [26, 28]. However, they did not fulfill the definition of moderate calibration [62].

The predictive performance of the models we developed was also disappointing. The percentage of variance in the observed values of ICU length of stay explained by the models was less than 20% at patient level and they had prediction errors of multiple days. The differences in predictive performance between the models were generally small.

ICU discharge decisions often do not only depend on a patient’s recovery, but on organizational circumstances such as the availability of beds on general nursing wards and the need to free up ICU beds for other patients. These organizational circumstances depend on structural factors related to the ICU and hospital. In chapter 2 and chapter 3, we deliberately chose not to include ICU and hospital level covariates in our models, because we wished to investigate the feasibility of predicting ICU length of stay for correcting for case-mix differences when comparing ICUs. In chapter 4, we hypothesized that information on ICU organizational characteristics would be required to model ICU length of stay accurately. This was addressed in the second research question 'What is the role of ICU organizational characteristics in predicting ICU length of stay?'. We examined the association between ICU organizational characteristics and ICU length of stay after correcting for patient case-mix. In addition, we investigated whether ICU length of stay predictions based only on patient characteristics could be improved by adding ICU organizational characteristics to the model.

We found that six ICU organizational characteristics were associated with ICU
length of stay. These were: number of hospital beds; number of ICU beds; fellows in training for intensivist available; number of ICU nurses; nurses to patient ratio; and discharged in a shift with 100% bed occupancy. Including these ICU organizational characteristics in a multivariate model significantly improved model fit, but not predictive accuracy. We concluded that predicting ICU length of stay using a model based on patient and ICU organizational characteristics is not better than using one based on patient characteristics only.

The results of the chapters 2 and 4 have to be put into the context of the application of a prognostic model. For planning required capacity and identifying patients with unexpected long ICU length of stay, predictions needs to be reliable for individual patients. The results of chapter 2 and chapter 3 showed that model accuracy at patient level was insufficient and that including ICU organizational characteristics did not improve these predictions.

Our work has sparked debate in the scientific literature after it was published. Straney and co-workers [105] have argued that poor model performance at patient level may not be indicative for poor utility of a model for benchmarking purposes. Similarly, Kramer [187] has claimed that some of the models included in our review described in chapter 2 (specifically, the models presented in [28] and [55]) could be used for benchmarking since these have an accuracy based on $R^2$ of 50 to 70% across ICUs.

We agree that, for benchmarking purposes, a prognostic model needs to predict average ICU length of stay at ICU level accurately. However if a model fails to also predict patient-level outcomes accurately, we cannot exclude the possibility that there exists significant residual variation in case mix. Therefore we remain cautious with recommending such a model for benchmarking in practice. However, mindful of the argument put forward, we recently examined model performance at ICU level for the model presented in chapter 5. We found an percentage of explained variance of 64%, which indicates that the accuracy ICU level predictions would be sufficient for the use of benchmarking. The calibration plot of mean predicted ICU length of stay against mean observed ICU length of stay based on 2% percentiles of predicted ICU length of stay was found to be satisfactory, figure 5.6. The regression line through the curve showed $\alpha=0.06$ (-0.01 to 0.14) and $\beta=0.99$ (0.97 to 1.01), table 5.7 through table and figure . We do recommend that further research is performed on accuracy for subgroups of patients and the calibration of models for the NICE registry.

Future research should address the limitations in our work. We did not have information regarding the availability of step down units or beds on other wards in hospitals. This information could improve patient level ICU length of stay. It may also be useful to explore the utility of Sequential Organ Failure Assessment (SOFA) scores in predicting ICU length of stay for planning purposes, as previous studies have [68, 188].

In this thesis, we addressed ICU length of stay as a continuous outcome. However, for benchmarking the binary outcome prolonged ICU length of stay may be useful.
This binary outcome has been considered in for patients admitted to the ICU following cardiac surgery, as these patients often have a short, fixed ICU length of stay [189, 190]. Other studies have addressed the binary outcomes prolonged ICU or hospital length of stay [67, 188, 191] based on a continuous prediction of length of stay. Other modeling approaches, such as machine learning techniques based on SOFA scores have been used to predict ICU length of stay and prolonged ICU length of stay [188]. Furthermore, Markov models predicting the state of the patient for each following day [65, 68] may have a role in predicting ICU length of stay for planning capacity and staff resourcing.

### 8.3 Presentation of values of quality indicators

#### 8.3.1 Association between quality indicators

The third research question 'Are case-mix adjusted ICU outcomes independent measures of ICU quality?' was addressed in chapter 5. In this chapter, we examined whether there is an association between case-mix adjusted quality indicators for: in-hospital mortality; readmission to the ICU within 48 hours after ICU discharge; and ICU length of stay. We examined these associations for the general Dutch ICU population and subgroups of ICU admissions presented on the publicly available part of the NICE website [19].

Pearson's correlation coefficients showed no significant association between the quality indicators at ICU population level and for most of the subgroups. These results were in line with other studies [39, 142]. However, we found mild associations between the performance of ICUs on the quality indicators in-hospital mortality and ICU length of stay for patients with low and high probabilities of in-hospital mortality. However, the correlation coefficients had opposite directions of association. For patients with a low probability of in-hospital mortality, there was a positive association between performance on in-hospital mortality and performance on ICU length of stay. In general, ICUs with better performance on case-mix adjusted in-hospital mortality also had better performance on case-mix adjusted ICU length of stay. For patients with a high probability of in-hospital mortality, there was a negative association between the observed values of the quality indicators. In general, ICUs with better performance on case-mix adjusted in-hospital mortality had worse performance on case-mix adjusted ICU length of stay. These results are intuitive in the sense that if very severely ill patients survive, long and intensive treatment is required.

Based on these results, we recommend that users of quality indicator values consider multiple indicators when judging or monitoring ICU quality of care. This is because different quality indicators capture different aspects of ICU performance. Furthermore, we suggest that users of quality information also have access to the values of quality indicators for subgroups of patients, especially those with low and high risks of in-hospital mortality.
In this thesis, we focused on three quality indicators related to the outcomes of ICU patients. Within the ICU domain many other quality indicators are also used. For example, within the NICE registry new actionable quality indicators for pain management, blood use, antibiotics and mechanical ventilation have been recently developed [192, 193]. Their aim is to provide tools to directly improve specific care processes. This is in contrast to the three quality indicators examined in this thesis, which provide a more general view on overall quality of ICU care. Future research should focus on the additional value of these actionable quality indicators of ICU care over and above quality indicators currently used by the NICE registry.

8.3.2 League tables

The fourth research question 'What is the rankability of league tables for in-hospital mortality of ICU patients, and how can it be improved?' was addressed in chapter 6. In this chapter, we assessed the rankability of league tables for case-mix adjusted in-hospital mortality and examined whether the reliability of these league tables can be improved if the assessment period is increased or if ICUs are grouped in clusters with similar performance.

The overall rankability of the league table developed in chapter 6 was 73%, meaning that when ranking Dutch ICUs based on case-mix adjusted in-hospital mortality 27% of all variation was caused by random variation within ICUs not explained by the model. Hence, presenting in-hospital mortality over a period of one year is not an accurate presentation of performance of Dutch ICUs, based on a 95% norm. However, other researchers have stated that a rankability of 75% is sufficient [153]. When compared to this norm, the rankability of Dutch ICUs can be viewed as approaching acceptability. Considering the potential impact of comparisons between healthcare institutions [47, 48], we believe that a high level of certainty is required and the 95% norm is required. In addition, this norm is well-accepted for decision making, it is generally accepted to make a decision even if there is up to 5% uncertainty whether that choice is really optimal. Judgements on the level of rankability that is acceptable may also depend on the potential societal impact of a particular league table.

To assess the impact of implemented improvement strategies, ICU performance is usually assessed each three, six or 12 months. The duration of the assessment period influences the number of admissions available within each ICU and, hence, the within-hospital variation. The rankability of our league table increased to 89% when data from a three year period was used. However, extending the assessment period may not always be desirable. This is because the types of patients admitted to an ICU, care processes and staff resources may change substantially during a longer assessment period.

As an alternative, we grouped ICUs in clusters with similar performance. The reliability improved to over 95% if we used fewer than nineteen clusters and was
optimal for seven clusters. These results form a clear starting point for further research. Commentators noticed that, if the table was reduced to seven clusters, the 95% confidence intervals of consecutive clusters overlapped but there were statistically significant differences between clusters spaced further apart. The use of four or five clusters may be better [194]. They also suggested that other clustering methods, such as K-means [195] and self-organizing maps [196] should be investigated.

League tables are used for decision making. Therefore, it is important that users are aware of their limitations and become acquainted with the concept of rankability. We believe our clustering approach could be useful for registries, such as NICE, when identifying underperforming and exemplary healthcare institutions. This approach may form a starting point for staff and directors from the lower clusters to improve clinical practice using information from the best performing cluster. In chapter 6, we addressed league tables for case-mix adjusted in-hospital mortality. Since the first part of this thesis addressed prognostic models for ICU length of stay, a logical next step would be to create a league table for ICU length of stay and to investigate its rankability.

8.3.3 Funnel plots

League tables are used to identify the worst and best performing ICUs and enable the worst to learn for the best. However, funnel plots can be used to compare individual ICUs to a national average and each other. The fifth research question 'Which steps and choices need to be made to construct reliable funnel plots?' was addressed in chapter 7. In this chapter, we provided guidance for data analysts on constructing funnel plots for quality assessment, focusing on binary quality indicators presented as proportions, risk adjusted rates, or standardized ratios. As an internal validation of the guidance, we used data from the NICE registry. In this chapter, we brought together literature on many aspects of the development of funnel plots. We identified six conceptual steps in constructing funnel plots [145]. These were: 1) policy (board of directors) level input; 2) checking the quality of prediction models used for case-mix correction; 3) ensuring that the number of observations per hospital is sufficient; 4) examining overdispersion of quality indicators; 5) examining associations between the values of quality indicators and hospital characteristics; and 6) funnel plot construction. With these steps we provided a framework in which standard operating procedures involving the construction of funnel plots are described. The content of the different steps is described in scientific papers. However, for some aspects, no absolute truth exists. This is the case for the values of performance measures used that indicate whether the quality of a prediction model is sufficient; the minimum number of events required to present a quality indicator; and how to correct for overdispersion [174, 177]. In this chapter, we recommended data analysts take each of these steps into account and prepare a statistical analysis plan before performing the analysis. In
this plan, they should specify which statistical tests they will use in each step and for which results of these tests they will decide that presenting a quality indicator using a funnel plot is acceptable.

To assess the usability of our guidelines, we performed an internal validation. We did not perform external evaluation. In the future, researchers should use data from another registry to conduct external usability and feasibility tests on our guidelines. Our results showed that it was appropriate to develop funnel plots for case-mix adjusted in-hospital mortality for all ICU admissions, but not for subgroups based on admission type. For these subgroups the number of admissions per ICU was too small or severity of illness expressed in the expected probability of mortality was associated with case-mix adjusted in hospital mortality. We recommend NICE and other quality registries to evaluate the process used for funnel plot construction, especially when presenting quality indicator values for subgroups of patients.

In conclusion, we expect that our guidelines will be useful for data analysts and registry employees striving for consistency in funnel plot construction over projects, employees and time. This is particularly true if these people and organizations wish to use standard operating procedures when constructing funnel plots. This may be required to comply with the demands of certification, such as International Organization for Standardization (ISO) 9001. In this chapter, we only considered funnel plots for binary quality indicators. We recommend registries pay attention to all methodological aspects of funnel plots when presenting quality indicators. Future research should focus on other types of data, including as normally and non-normally distributed continuous quality indicators, such as ICU length of stay.

At this end of this thesis, we conclude that benchmarking is a powerful method of assessing the quality ICU care. However, benchmarking is only meaningful if: observed differences between ICUs are adjusted for other sources of variance, such as differences in case-mix; random variation within ICUs is taken into account; registration bias is minimized; and quality indicators are presented in reliable and understandable ways.