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):

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: https://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.
Individual and clustered rankability of ICUs according to case-mix adjusted mortality

Ilona W.M. Verburg, Nicolette F. de Keizer, Rebecca Holman, Dave A. Dongelmans, Evert de Jonge and Niels Peek

Critical Care Medicine 2016; 44(5):901-909
Abstract

**Background:** The performance of intensive care units (ICUs) can be compared by ranking them into a league table according to their risk-adjusted mortality rate. The statistical quality of a league table can be expressed as its rankability, the percentage of variation between ICUs due to unexplained differences. We examine whether we can improve the rankability of our league table by using data from a longer period or by grouping ICUs with similar performance constructing a league table of clusters, rather than individual ICUs.

**Methods:** We used data from intensive care units participating in the Dutch National Intensive Care Evaluation registry between 2011 and 2013. We developed a league table for risk-adjusted mortality rate with its rankability. The effect of assessment period was determined using a resampling procedure. Hierarchical clustering was used to obtain clusters of similar ICUs.

**Results:** We constructed league tables using 157,394 admissions from 78 ICUs with risk adjusted mortality rate (RAMR) between 5.9% and 13.9% per ICU over the inclusion period. The rankability was 73% for 2013 and 89% for the whole period 2011 to 2013. Rankability over the year 2013 increased till 98% when clustering ICUs, reaching an optimum at a league table of seven clusters.

**Conclusions:** We conclude that the rankability of a league table of Dutch ICUs based on risk-adjusted mortality rate was unacceptably low, when using data from a single year. We could improve the rankability of this league table by increasing the period of data collection or by grouping similar ICUs into clusters and constructing a league table of clusters of ICUs, rather than individual ICUs. Ranking clusters of ICUs could be useful for identifying possible differences in performance between clusters of ICUs.
6.1 Introduction

A range of audiences, including hospital staff and directors, insurance companies, politicians and patients, are interested in the quality of care in hospitals as expressed by outcome indicators [41, 47]. The quality of care in a hospital is often assessed in terms of case-mix or risk adjusted in-hospital mortality [16, 75, 87, 88]. Often, there exists no external benchmark value to which individual hospitals can be compared objectively, and instead they are compared amongst each other.

This type of comparison can be performed graphically, for instance, using a funnel plot [51]. In a funnel plot, an outcome measure, such as the risk adjusted mortality rate (RAMR) of a hospital for a given year, is plotted against a measure of its precision, such as the number of admissions to the hospital in that year. The value of the outcome measure for each hospital is compared to a set value, such as the RAMR of all hospitals together, using control limits that reflect the precision for each hospital. Funnel plots are useful tools for judging whether individual hospitals performed better or worse than the group as a whole, but cannot be used to examine whether particular hospitals performed better or worse than each other.

Another way of comparing hospital performance is to create a league table of hospitals ranked according to their performance on the quality indicator. Confidence intervals could be added to the ranks or RAMR of a specific intensive care unit (ICU) in this league table [41, 45, 46]. This method can discriminate between the hospitals with best and worst performance. As a league table identifies underperforming hospitals with a low rank in the league table, it enables their staff and directors to recognize the need for improvement and to identify the best performing hospitals with a high rank in the league table, from which they can use implemented methodology to improve clinical practice.

However, statisticians have raised concerns about the reliability of league tables [45–48]. League tables focus on observed differences between hospitals and ignore uncertainties in performance due to limited sample size. 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 hospital 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]. The rankability of a league table expresses the percentage of variation between hospitals that is due to unexplained differences. Higher values of rankability correspond to more reliable league tables [49, 50]. Ideally unexplained differences only represent true differences in the quality of care between hospitals. However, it is possible that the unexplained variation in RAMR between ICUs is due to residual confounding, such as patient characteristics which were not included in the risk adjustment model. Furthermore, unexplained variation could be caused by registration bias, rather than genuine differences in quality.
The rankability of hospital league tables varies greatly [50, 147, 148]. Researchers have found that the rankability of league tables for nine quality indicators for hospitalized patients in the Netherlands ranged from 37% to 71% [50]; was 8% for hospitalized patients with surgical site infections [147]; and was 38% for postoperative mortality in colorectal cancer surgery [148].

ICU clinicians and directors are the important users of benchmark information and clearly one of the main stakeholders in benchmark processes. Therefore, it is valuable if they are aware of the limitations of league tables and become acquainted with tools to mitigate these limitations. The concept of rankability has not previously been applied to ICU league tables, although risk adjusted mortality is often used to compare performance in critical care [1, 16, 29].

The objective of this study is 1. to assess the relationship between rankability and length of assessment period and 2. to develop a method to increase rankability bases on clustering methods, for ICUs based on RAMR, using data from a large registry in the Netherlands. As uncertainty is primarily influenced by sample size, we hypothesized that we could increase the rankability of our league tables by considering longer assessment periods or by ranking clusters of similar, instead of individual, ICUs.

### 6.2 Methods

#### 6.2.1 Data

The dataset used for this study consist of a cohort of admissions from 78 ICUs participating in the Dutch National Intensive Care Evaluation foundation (NICE) registry for the entire period from January 1st 2011 until December 31st 2013. Patients with ICU admission date between January 1st 2013 and December 31st 2013 were included in all analyses of this study. Furthermore, patients with ICU admission date between January 1st 2011 and December 31st 2013 were included in the analyses on increasing the assessment period. The NICE registry collects data on demographics, physiological and clinical data from the first 24 hours that patients are admitted to the ICU including among others all variables used in the Acute Physiology and Chronic Health Evaluation (APACHE) IV model [16]. Patients are followed until death before hospital discharge or hospital discharge. Anonymized data on ICU and hospital mortality are recorded. The NICE registry has been active since 1996 and in 2015 84 of 90 Dutch ICUs participate. To improve the quality of the data, the data are subject to quality controls, onsite data quality audits take place [15, 20] and data collectors participate in training sessions. The use of anonymized data does not require informed consent in the Netherlands. The data were officially registered in accordance with the Dutch Personal Data Protection Act. We included all admissions fulfilling the APACHE IV inclusion criteria [16] and excluded ICU admissions following cardiac surgery. We did not examine admissions following cardiac surgery, because cardiac surgery
is only performed in a small number of hospitals in the Netherlands and contains a large proportion of all ICU admissions in these hospitals having a low risk of ICU readmission or post-ICU in-hospital mortality [16, 142].

6.2.2 League table and rankability

We ranked the ICUs by RAMR using data from the year 2013. The RAMR is defined as the number of deaths actually observed in an ICU divided by the number of deaths predicted by the recalibrated APACHE IV model [146] standardized in-hospital mortality ratio (SMR), multiplied by the average hospital mortality rate in the entire study population. First level recalibration of the APACHE IV model was performed by fitting a logistic regression model with in-hospital mortality as dependent variable and the logit-transformed original APACHE IV mortality probability as the sole independent variable. We obtained 95% confidence intervals for the hospital ranks using a resampling procedure described in the Appendix [71, 149, 150].

The rankability of the RAMR league table is the percentage of variation between ICUs attributed to genuine differences in quality of care [49, 50]. The rankability as calculated consists of two components: the within-hospital variance (uncertainty); and the between-hospital variance (heterogeneity), interpreted as the differences in quality of care [49, 50]. To calculate the uncertainty and the heterogeneity regression methods were used. We computed rankability ($\rho$) using the estimated heterogeneity ($\tau^2$) in measured performance between institutions that cannot be explained by case-mix differences and the estimated uncertainty ($\sigma^2$) in measured performance due to sampling error. The rankability is defined as $100 \cdot \frac{\tau^2}{\sigma^2 + \tau^2}$. We provide details on how we calculated the rankability, heterogeneity and uncertainty in appendix 2.A.

Rankability ranges from 0% to 100%, with higher values corresponding to more reliable league tables. A rankability of 100% means that there is no uncertainty in the league table. Although some authors have proposed a lower acceptable limit of 75% for rankability [47], considering the potentially large social impact of outcome-based comparison of hospitals, we suggest a threshold of 95% for rankability. This is motivated by the field of statistics where a confidence interval of 95% is a broadly accepted norm.

We computed the rankability for all ICUs together. However, comparing institutions based on case-mix adjusted hospital mortality has found to be insufficient to correct for differences in patient case-mix between ICUs [146]. Therefore the rankability will be calculated for several subgroups of patients for a one-year period, 2013. In addition, rankability was computed separately for general, teaching, and university-affiliated hospitals as there may be unmodelled case-mix differences between these different types of hospital. Furthermore rankability was computed dividing ICUs in four volume groups based on the quartiles of the number of admissions in 2013 (430 or fewer; 431 to 610; 611 to 797; more than 797), as
within-hospital variance is influenced by sample size, and therefore by yearly patient volume. In order to compare rankability for more homogeneous subgroups of patients, rankability was computed separately for medical, elective surgical, and emergency surgical admissions. Patients undergoing unplanned ICU-admission differ considerably from patients undergoing elective surgery. Mortality is generally higher in unplanned ICU admissions than in planned ICU admissions.

6.2.3 Improving rankability

As clinical performance may change over time, ICU performance is usually assessed each three or six months, or every year or two years. Because the duration of the assessment period influences the number of admissions available to produce a league table, and hence within-hospital variance, we evaluated the effect of increasing the length of the assessment period on rankability, using a simulation method based on subsampling. More details are provided in the appendix 2.A. Extending the assessment period may not always be desirable as it slows down the speed of audit cycles. Furthermore, changes in quality of care are discounted in the results of a longer assessment period. We hypothesized that it is also possible to increase the reliability of a league table by ignoring performance differences between ICUs with similar RAMR values and treating these groups of ICUs as a single entity in the league table. In this case groups of optimally performing ICUs could potentially be separated from groups of underperforming ICUs. We expect that clustering ICUs will increase the heterogeneity between groups of ICUs and lower the uncertainty in measured performance, thus improving rankability. To evaluate the effect of grouping ICUs into clusters on rankability, we used agglomerative hierarchical clustering [151]. We analyzed the stability of the clusters using a resampling procedure which repeated the clustering procedure 1,000 times, after which it was assessed whether the resulting clusters were the same or different as the clustering obtained on the original dataset. The sampling method is described in more detail in the appendix 2.A. All statistical analyses were performed using the R statistical software, version 2.15.1 [97].

6.3 Result

6.3.1 Data

A total of 78 ICUs participated in the NICE registry for the whole period between 1st January 2011 and 31st December 2013, resulting in 229,233 ICU admissions, of which 195,592 (85.3%) fulfilled the APACHE IV inclusion criteria. Of these, 157,394 (80.5%) were not related to cardiac surgery. Of the 78 ICUs, six (8%) were in university hospitals, 26 (33%) were in teaching hospitals and 46 (59%) were in non-teaching hospitals. The median (interquartile range [IQR]) number
of admissions per ICU was 1,759 (1,236-2,306) in the whole period and 610 (IQR 430-797) in the year 2013 (Table 6.1).

Crude hospital mortality rate varied from 4.2% to 23.7% for the years 2011 to 2013, and varied from 3.5% to 21.4% for the year 2013. RAMR varied from 6.6% to 17.4% for the years 2011 to 2013 and from 6.9% to 17.7% for the year 2013.

Table 6.1: Rankability for case-mix corrected hospital mortality of ICU patients.

<table>
<thead>
<tr>
<th>Subgroup</th>
<th>Number of ICUs</th>
<th>Median (IQR) number of ICU admissions</th>
<th>Median (IQR) RAMR (%)</th>
<th>Rankability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Included admissions</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2013</td>
<td>78</td>
<td>610 (430-797)</td>
<td>9.8 (8.6-10.8)</td>
<td>73%</td>
</tr>
<tr>
<td>2011 to 2013</td>
<td>78</td>
<td>1,759 (1,236-2,306)</td>
<td>10.5 (9.4-11.7)</td>
<td>89%</td>
</tr>
<tr>
<td>Admission type (2013)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medical</td>
<td>78</td>
<td>344 (235-474)</td>
<td>13.8 (12.1-14.9)</td>
<td>64%</td>
</tr>
<tr>
<td>Urgent surgery</td>
<td>78</td>
<td>80 (53-112)</td>
<td>11.6 (9.1-13.2)</td>
<td>33%</td>
</tr>
<tr>
<td>Elective surgery</td>
<td>78</td>
<td>160 (97-259)</td>
<td>1.9 (1.1-2.7)</td>
<td>42%</td>
</tr>
<tr>
<td>Hospital type (2013)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>General</td>
<td>46</td>
<td>481 (337-608)</td>
<td>8.8 (7.6-10.1)</td>
<td>69%</td>
</tr>
<tr>
<td>Teaching</td>
<td>26</td>
<td>752 (669-915)</td>
<td>9.7 (8.7-10.7)</td>
<td>81%</td>
</tr>
<tr>
<td>University</td>
<td>6</td>
<td>1,287 (1,274-1,330)</td>
<td>13.0 (12.4-13.2)</td>
<td>55%</td>
</tr>
<tr>
<td>Admissions yearly (2013)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Up to 430</td>
<td>20</td>
<td>311 (280-362)</td>
<td>7.9 (7.2-9.6)</td>
<td>67%</td>
</tr>
<tr>
<td>431 to 610</td>
<td>19</td>
<td>519 (485-558)</td>
<td>9.5 (8.1-10.4)</td>
<td>66%</td>
</tr>
<tr>
<td>611 to 797</td>
<td>19</td>
<td>668 (632-705)</td>
<td>9.5 (8.5-10.8)</td>
<td>83%</td>
</tr>
<tr>
<td>797 or more</td>
<td>20</td>
<td>1,098 (911-1,372)</td>
<td>10.5 (9.9-11.3)</td>
<td>74%</td>
</tr>
</tbody>
</table>

IQR=interquartile range; RAMR=risk adjusted mortality rate

6.3.2 League table and rankability

We present our results on sample uncertainty, heterogeneity between ICUs and the rankability of case-mix corrected hospital mortality in Table 6.1. When using the data from a single year (2013) the rankability of case-mix corrected hospital mortality was 73%. The rankability of case-mix corrected hospital mortality was higher for medical ICU (64%) admissions than for admissions following emergency (33%) and elective surgery (42%). We obtained a higher rankability for teaching hospitals (81%) than for general hospitals (69%) and university affiliated hospitals (5%). We also obtained a higher rankability for the hospitals with 611 to 797 admissions (83%) than for the hospitals with less than or equal to 430 admissions (67%), the hospitals with 430 to 610 admissions (66%) or the hospitals with more than 797 admissions (74%). Figure 6.1 displays risk adjusted mortality rates and corresponding median ranks with 95% confidence intervals for each of the participating ICUs, based on data from the year 2013. The confidence intervals
around the RAMR overlap each other, which suggests that using a league table as presented in this figure does not make it possible to assess whether the ordering of the ICUs is reliable.

6.3.3 Improving rankability

We obtained a rankability of 89% for the whole period 2011 to 2013. In Figure 6.2, we show the (median [interquartile range]) rankability to the mean number of patients per ICU for the period 2011 to 2013, basing the league table on an increasing percentage of admissions from 5% to 100%. Figure 6.3 shows these results for the heterogeneity and uncertainty respectively. The heterogeneity between ICUs remained stable, due to stable between-hospital variance, at approximately 0.06. The uncertainty, within-hospital variance, decreased asymptotically because of an increase in sample size, to below 0.01. This resulted in an increase in median rankability from 28.7% to 90.4%, as the effect of the uncertainty in the denominator became smaller.

Figure 6.4 and 6.5 show the effects of hierarchical clustering of ICUs on rankability, heterogeneity and uncertainty, using data from the year 2013 only. Rankability increased from 73% when each of the 78 ICUs was considered as a separate cluster to 100% (trivially) when all ICUs were merged into a single cluster. The heterogeneity between clusters steadily increased as the number of clusters shrunk to seven, after which heterogeneity declined sharply, meaning that clusters with large between cluster variance were merged together. Uncertainty steadily decreased during the entire clustering procedure. We interpret this as an indication that seven is the optimal number of clusters to rank, for the ICUs in our study.

We define the cluster in which a specific ICU falls, when ICUs were grouped into seven clusters, as the reference cluster for this specific ICU. For the reference clusters the range in RAMR for the best performing cluster was 0.081 to 0.087 (2 ICUs) and the range in RAMR for the worst performing cluster was 0.170 to 0.203 (10 ICUs). We analyzed the stability of the clustering procedure using a resampling procedure which grouped the 78 ICUs into seven different clusters 1,000 times. Using the clustering obtained on the original data set as a reference, we found that ICUs were in the same cluster or a directly neighboring cluster in 89% of the comparisons. This means that the clustering procedure identifies performance differences in a robust way. Table 6.2 presents a summary of the median number of ICUs, number of admissions and RAMR in each of the seven clusters following the resampling procedure. The best and worst clusters contained the fewest ICUs. Figure 6.6 displays the median and (2.5%-97.5%) confidence interval of the RAMR for each of the seven clusters. We describe further results of this analysis in appendix 2.B.
Figure 6.1: Estimated median (2.5%-97.5%) rank and risk adjusted mortality rate (RAMR) for each ICU. The number of admissions is shown at the right side of the figure.
Figure 6.2: Relationship between the rankability (median [2.5%-97.5%]) and the mean number of patients per ICU for the period 2011 to 2013.

Figure 6.3: Relationship between the heterogeneity and uncertainty (median [2.5%-97.5%]) and the mean number of patients per ICU for the period 2011 to 2013.
Table 6.2: Estimated number of ICUs, number of admissions and risk adjusted mortality rate (RAMR) for each of the seven clusters over all sampling iterations (median [2.5%-97.5%]).

<table>
<thead>
<tr>
<th>Cluster number</th>
<th>Number of ICUs</th>
<th>Number of admissions</th>
<th>RAMR (2.5%-97.5%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4 (1-12)</td>
<td>1,979 (506-6,292)</td>
<td>0.082 (0.060-0.097)</td>
</tr>
<tr>
<td>2</td>
<td>10 (4-20)</td>
<td>6,199 (2,283-12,399)</td>
<td>0.105 (0.092-0.118)</td>
</tr>
<tr>
<td>3</td>
<td>14 (7-25)</td>
<td>9,877 (4,730-17,761)</td>
<td>0.122 (0.110-0.135)</td>
</tr>
<tr>
<td>4</td>
<td>16 (8-27)</td>
<td>12,062 (6,211-19,922)</td>
<td>0.137 (0.125-0.149)</td>
</tr>
<tr>
<td>5</td>
<td>15 (7-25)</td>
<td>11,166 (5,335-19,419)</td>
<td>0.151 (0.139-0.167)</td>
</tr>
<tr>
<td>6</td>
<td>10 (4-19)</td>
<td>7,512 (2,222-14,496)</td>
<td>0.169 (0.154-0.190)</td>
</tr>
<tr>
<td>7</td>
<td>4 (1-13)</td>
<td>2,098 (302-8,305)</td>
<td>0.197 (0.175-0.261)</td>
</tr>
</tbody>
</table>

Figure 6.4: Rankability when iteratively merging ICUs into clusters of ICUs.

Figure 6.5: Heterogeneity and uncertainty when iteratively merging ICUs into clusters of ICUs.
6.4 Discussion

In this study we developed a league table of 78 Dutch ICUs based on risk-adjusted mortality rates and estimated its rankability as a measure of reliability. Furthermore, we assessed the effect of grouping similar ICUs into clusters and increasing the assessment period on the rankability. The RAMR for the 78 ICUs ranged from 4.8% to 12.8% and the overall rankability for 2013 was 73%. The rankability varied among admission types, and was poor for admissions following both emergency (33%) and elective surgery (42%). Rankability also varied by hospital type and ICU volume, reaching the best rankability for teaching hospitals (81%) and moderately large ICUs (610-797 admissions per year, 83%). When data from the period 2011 to 2013 were used, rankability increased to 89%. The rankability of a league table based on data from 2013 increased to over 95% when ICUs were grouped into fewer than 19 homogeneous clusters. An optimal balance between rankability and heterogeneity was reached at seven clusters.

From a statistical point of view there was little difference between clustering ICUs into 19 cluster rather than 7 clusters. But grouping ICUs in a lower number of clusters (such as 7) forms a clearer starting point for further investigation. After performing the cluster analysis, further investigation should point out whether there are systematic differences in care practice between the clusters that have been identified. If this is indeed the case, it can enable staff and directors from hospitals in low-ranked clusters to improve their clinical practice by learning from ICUs in the higher-ranked clusters.

Rankability reflects signal-to-noise ratio for an entire league table. As such, it will
Rankability of ICUs according to mortality

increase when it is calculated using more, and decrease when it is calculated using less data. This probably explains our poor rankability for surgical admissions, but not the relatively high rankability of teaching hospitals and moderately large hospitals, as these categories of hospitals did not have more data than other categories.

ICU clinicians and directors are the important users of benchmark information and clearly one of the main stakeholders in benchmark processes. Therefore, it is valuable if they are aware of the limitations of league tables and become acquainted with tools to mitigate these limitations. As far as we know the concept of rankability has not been applied in the ICU domain before. Bakhshi-Raiez et al. [149] previously reported a high variability in the ranks of ICUs in league tables based on standardized mortality rate. Although finding a relatively low rankability of a league table or a large variability in the ranks for ICUs within a league table there could be specific ICUs within the league table for which the rank has a high certainty. This is the case whether the 95%-confidence intervals of the rank and RAMR are small for these specific ICUs. We did find a limited number of studies applying the concept of rankability in other care domains. Van Dishoeck et al. [50] found a poor rankability of 97 Dutch hospitals by Dutch Healthcare Inspectorate quality indicators. Yet in their study the rankability by hospital death after acute myocardial infarction was 58% (88 hospitals; median N 85.5 (4-720)), which is approximately reached in our study when 18% of the data or a three years period was used, figure 2 (mean number of admission per ICU 363; rankability 59%). In another study, the same research group calculated rankability for hospitalized patients with surgical site infections [147]. They found a rankability of only 8%, though the rankability was 80% for the subgroup of patients undergoing colonic resection. The low rankability could be caused by large differences between patients undergoing different types of surgery. Finally, another study assessed rankability for postoperative mortality rate, for 25,592 patients in 92 hospitals who underwent colorectal cancer surgery (15). Rankability after adjustment for case-mix was found to be 38%. We found this value for the rankability when the mean number of admissions was 162 over 78 ICUs (8% of the patient admissions of a three years period), figure 2.

In the literature there is limited evidence on the acceptable levels of rankability in practice [148, 152, 153]. Lingsma et al. [153] suggested that any ranking is meaningless if its rankability is less than 50%, of moderate quality if its rankability is between 50% and 75% and acceptable if its rankability is greater than 75%. However, considering the potentially large social impact of outcome based comparisons of care providers [47, 48], we believe that a high level of certainty is required. Therefore we have adopted the 95% norm because this is a well-accepted criterion for decision making, i.e. it is generally accepted to make a decision even though there is up to 5% uncertainty whether that is choice is really the best one. If we interpret league tables as tools for decision making about institutional quality of care, then we should be equally certain about the
information provided. As expected, in our study rankability increased to 89% when we extended the period of data collection from one to three years. Arguably, an assessment period of three years is too long to be useful in practice. Furthermore, it is conceivable that the clinical performance of individual ICUs will vary over this time period. For this reason we did not perform subgroup analyses over the three-year period. An advantage of grouping ICUs into clusters is that underperforming ICUs can choose from a larger group of better performing ICUs from which to learn. As long as the difference in performance between ICUs in a cluster is small, this procedure retains or even increases heterogeneity, while reducing uncertainty of the league table. In our case, heterogeneity gradually increased until the number of clusters was seven, after which it dropped quickly. This suggests that to judge and rank performance of the 78 Dutch ICUs, it is sufficient to distinguish between seven levels of performance and neglect finer-grained differences.

An alternative for comparing ICUs using a league table are graphical methods, such as funnel plots [51]. In a funnel plot, the performance of each individual ICU is plotted against a measure of its precision and compared to an external goal or a summary of performance across all ICUs considered. The control limits are based on the external goal or a summary of performance [93]. When the outcome variable for an ICU falls outside the 95% confidence interval of the funnel plot one can conclude that this ICU has a different performance compared to the group as a whole. But this method cannot be used to compare two ICUs with each other. By this reason funnel plots are useful tools for judging whether individual ICUs performed differently from the group as a whole, but cannot be used to examine whether particular ICUs performed better or worse than each other.

Our study has several limitations. Increasing the assessment period as we applied by using a resampling technique to express rankability for a smaller period (such as one year) using patients ICU admission data of a three years period could add additional variation, since changes in for example population, quality in care and treatment methods could influence the performance of ICUs. Ideally unexplained differences only represent true differences in the quality of care between hospitals. However, it is possible that the unexplained variation in RAMR between ICUs was due to residual confounding, such as patient characteristics which were not included in the APACHE IV model. Furthermore, unexplained variation could be caused by registration bias, rather than genuine differences in quality. However, to help minimize registration bias, data collected in the NICE registry is subject to thorough quality checks including onsite data quality audits [15, 20] and local data collectors participate in training sessions. In addition, our data came from ICUs, which participate voluntarily in the NICE registry. However, we believe that there is no substantial volunteer bias in our results, since over 90% of all ICUs in the Netherlands participate in NICE.
6.5 Conclusion

Based on a 95% norm, there was too much uncertainty to reliably rank Dutch ICUs on case-mix adjusted mortality when using data from a single year. A relatively high rankability was found for teaching hospitals and moderately large hospitals. In these groups relatively a large amount of heterogeneity between ICUs was found. Rankability increased with longer assessment periods and when ranking clusters of similar, rather than individual ICUs. This method does not influence the ordering of ICUs based on risk adjusted mortality. This clustering approach could be an useful alternative to identify underperforming ICUs and better performing ICU.

Acknowledgements

We acknowledge all participating ICUs in the NICE for their participation and hard work to collect and improve their processes based on these data.
Appendix 6.A: Methods

6.A.1 Simulation methods

League table and rankability

In order to obtain 95% confidence intervals for the rank and RAMR of ICUs, we calculated the uncertainty associated with the ranks and the RAMR using a resampling procedure. We drew 1,000 stratified samples of admissions with replacement. The samples were stratified by ICU and the size of the strata was equal to the total number of admissions to the ICU in 2013 [71]. We calculated the rank and RAMR of each ICU for each of the 1,000 samples and hence obtained the median and 2.5% and 97.5% percentiles (95% confidence interval) of the rank and RAMR for each ICU [149, 150]. We used the same set of 1,000 samples to determine the stability of the clusters of ICUs, see the section on clustering similar ICUs together in this appendix.

Increasing the period of assessment

We also investigated the effect of increasing the length of the assessment period on rankability using a resampling simulation method. We drew random samples, with replacement, stratified by ICU and varying in size between 5% and 100% of the total number of admissions to the ICU in the period 2011 to 2013, with increments of 1%. For each increment in sample size we generated 100 stratified samples and calculated the median and 95% confidence interval of the rankability using the results from the 100 samples.

6.A.2 Calculation of the rankability

We computed rankability ($\rho$) using the estimated heterogeneity in measured performance between institutions that cannot be explained by case-mix differences (between-hospital variance, $\tau^2$) and the estimated uncertainty in measured performance due to sampling error (within-hospital variance, $\sigma^2$). The rankability is defined as $100 \cdot \tau^2 / (\sigma^2 + \tau^2)$.

In order to estimate $\tau^2$ we used a random effect logistic regression model with ICU as random variable. We defined $\tau^2$ as the variance of the random effects. A fixed effect logistic regression model with ICU as a categorical variable was used to estimate the uncertainty parameter $\sigma^2$. The parameter was estimated as the median standard error of the estimated coefficients associated with individual ICUs in this model. We included the logit transformed recalibrated APACHE IV mortality risk as fixed covariate in the random effect model and as offset variable in the fixed effect regression model to adjust for patient case-mix. For further details on estimating $\tau^2$ and $\sigma^2$, we refer to the paper by Van Dishoeck et al. [50].
6.A.3 Clustering similar ICUs

Method

We grouped ICUs into clusters of ICUs with similar RAMR using the agglomerative hierarchical clustering procedure. During this procedure each ICU was initially considered as a separate cluster. We ordered the ICUs by RAMR and iteratively merged them into clusters. During each iteration, we merged the two neighboring clusters with the lowest value of the $\chi^2$-statistic testing the difference in RAMR. Thus, in each iteration the number of clusters was reduced by one. After each iteration we estimated uncertainty, heterogeneity and rankability of the league table using the remaining clusters. When performing the clustering procedure on the original dataset we found an optimum in heterogeneity and rankability when the ICUs were grouped into seven separate clusters (Figures 6.3 and 6.4). We refer to these as reference clusters. In order to determine whether an ICU really belongs to its reference cluster, we analyzed the stability of these clusters, by repeating the clustering procedure for the 1,000 samples of the original sample used to estimate uncertainty around the RAMR and rank of the league table. For each of the 1,000 samples, we performed the clustering procedure until we had exactly seven clusters and recorded to which cluster each ICU belonged. Cluster one consists of ICUs with the best performance according to the RAMR and cluster seven consists of ICUs with the worst performance according to the RAMR. Furthermore, we calculated the RAMR, number of admissions and number of ICUs in each of the seven clusters. We present the variability in cluster number compared to the number of the reference cluster for each ICU.

Appendix 6.B: Results

We analyzed the stability of the clustering procedure using a resampling procedure, which grouped the 78 ICUs into seven different clusters 1,000 times. Figure 6.7 displays a histogram of the clusters in which each ICU was classified and Figure 6.8 shows a histogram of the absolute difference (upwards or downwards) in cluster number compared to the reference cluster. In total 68,289 (87.6%) of the 78,000 differences in cluster number across all 78 ICUs are zero (32,340; 41.4%) or one (35,949; 46.1%) different from the number of the reference cluster. We found no relationship between the size of an ICU and the number of changes in cluster number. This indicates that our clustering procedure produces a stable result.
Figure 6.7: Distribution of the cluster numbers ICUs belong to when clustering ICUs in seven clusters for 1,000 samples of the original dataset.
Figure 6.8: Absolute difference in cluster number for each ICU compared to the reference cluster an ICU belongs to, for 1,000 samples of the complete dataset. The reference cluster is derived by applying the clustering procedure on the complete dataset until ICUs were clustered in seven groups.