The European Thoracic Database project: composite performance score to measure quality of care after major lung resection

Alessandro Brunellia,*, Richard G. Berrisfordb, Gaetano Roccc, Gonzalo Vareld, on behalf of the European Society of Thoracic Surgeons Database Committee

a Umberto I Regional Hospital Ancona, Italy
b Royal Devon & Exeter NHS Foundation Trust, Exeter, United Kingdom
c National Cancer Institute, Pascale Foundation, Naples, Italy
d Salamanca University Hospital, Salamanca, Spain

Abstract

Background: Performance measurement is an essential element of quality improvement initiatives. The objective of this study was to develop a composite performance score (CPS) incorporating processes and outcomes measures available in the European Society of Thoracic Surgeons (ESTS) Database and apply it to stratify performance of participating units. Methods: A total of 1656 major lung resections for malignant primary neoplastic disease were collected in the ESTS database from 2001 through 2003 and were analyzed. For the purpose of this study only data collected from units contributing more than 50 consecutive cases were included. Three quality domains were selected: preoperative care, operative care, and postoperative outcome. According to best available evidence the following measures were selected for each domain: preoperative care (% of predicted postoperative carbon monoxide lung diffusion capacity (ppoDLCO) measurement in patients with predicted postoperative forced expiratory volume in one second (ppoFEV1) < 40%), operative care (% of systematic lymph node dissection), and outcomes (risk-adjusted cardiopulmonary morbidity and mortality rates). Morbidity and mortality risk models were developed by hierarchical logistic regression and validated by bootstrap analyses. Individual processes and outcomes scores were rescaled according to their standard deviations and summed to generate the CPS, which was used to rate units. Results: CPS ranged from 4.4 to 3.7. Individual scores were poorly correlated with each other. Two units were negative outliers and two positive outliers (outside 95% confidence limits). Compared to the rating obtained by using the risk-adjusted mortality rates, all units changed their positions when ranked by CPS. Conclusions: The composite performance score methodology may support future peer-based organizational quality benchmarking initiatives and may be used for regulatory and credentialing purposes.

Keywords: Lung resections; Quality of care; Audit; Composite performance score; Process of care; Outcome; Database

1. Introduction

Performance measurement is an essential element of quality improvement initiatives and health care reform. Until recently, quality improvement has focused on short-term outcomes (i.e. operative mortality and morbidity). However, it is recognized that such clinical outcomes are only one measure of overall health care quality [1,2]. More comprehensive quality indicators include intraoperative and perioperative processes of care that impact hospital outcomes, secondary prevention, and long-term health.

Like intelligence or musical ability, quality is an abstract construct that cannot be measured directly and it is characterized by one or more latent (unobserved) variables or traits [3]. To quantify abstract constructs we typically rely upon some combination of measurable surrogates that are thought to be associated with or contribute to that underlying trait.

The European Society of Thoracic Surgeons (ESTS) Database Committee and Audit and Clinical Excellence ad hoc Committee were appointed to devise and implement instruments of performance monitoring across European thoracic surgery units. A European Thoracic Database was used to collect thoracic procedures from European units to provide a platform for quality management initiatives. Although a risk model for in-hospital mortality after lung resection (ESOS) was developed, validated and applied to this purpose [2,4], mortality provided only limited information.
about individual providers’ performance and more comprehensive and reliable quality indicators were needed.

A recent document from the STS quality measurement task force elegantly explained the conceptual framework and the statistical consideration in the development of composite performance scores (CPSS) in cardiac surgery [3,5]. Based on that methodology, we have developed a composite performance score for lung surgery. The objective of this analysis was to generate a composite score from the data available in the European Thoracic Database to provide an instrument for performance monitoring for quality improvement, regulation and credentialing.

2. Patients and methods

Data was prospectively captured from the European thoracic surgery database (version 1) on thoracic surgical procedures from units across Europe. The database was designed to contain a minimum set of core variables and endpoints. Twenty-seven units from 14 European countries consistently submitted data from January 2001 through December 2003 without external local data audit. Data was scrutinized for possible inconsistencies and missing values to improve quality of data in the context of a voluntary database.

One thousand, six hundred and fifty six patients submitted to major lung resections (1281 lobectomies and 375 pneumonectomies) for malignant primary neoplastic disease in the 10 units contributing more than 50 consecutive cases to the European database (period 2001–2003) were analyzed. These cases represent approximately 70% of the total major lung resection in the database.

We selected three quality domains assessing preoperative care (process indicator), operative care (process indicator), postoperative care (outcome indicator). Based on the best available evidence (the integration of the best research evidence with clinical expertise and patient values [6]) the following indicators were selected.

2.1. Process indicators

Based on published guidelines,

1. The proportion of patients with predicted postoperative carbon monoxide lung diffusion capacity (ppoDLCO) estimated in the eligible group with a low predicted postoperative forced expiratory volume in one second (ppoFEV1) (defined as ppoFEV1 <40%) [7,8].
2. The proportion of patients with a systematic lymph nodes dissection according to ESTS published guidelines [9].

2.2. Outcome indicators

1. In-hospital mortality.
2. In-hospital cardiopulmonary morbidity (pneumonia, atelectasis requiring bronchoscopy, adult respiratory distress syndrome, mechanical ventilation longer than 24 h, pulmonary edema, pulmonary embolism, myocardial ischemia, cardiac failure, arrhythmia, stroke, acute renal insufficiency).

These two outcome indicators were risk-modeled as follows. An initial screening of preoperative and operative variables (age, gender, predicted postoperative forced expiratory volume in one second in percent of predicted normal values—ppoFEV1, type of operation—(lobectomy or pneumonectomy), presence of extended resection (lung resection associated with resection of chest wall, diaphragm, pericardium, or other mediastinal structures)) was performed by univariate analyses using unpaired Student’s t-test or Mann–Whitney test for numerical variables (with or without normal distribution, respectively) and the chi-square test or Fisher’s exact test whenever appropriate for categorical variables. Significant variables (p < 0.1 at univariate analysis) were then used as independent predictors in hierarchical regression models. Hierarchical or multilevel random-effect multivariate regression analysis takes into account the between-provider and within-provider variations. Provider-specific estimates are shrunken towards the mean for all providers and this methodology has been shown to provide a more reliable estimate of a single provider performance [10,11] by better reflecting the true mortality rate of low volume providers and accounting for clustering (non-random allocation) of observations within providers [12,13]. The regression models were further validated by bootstrap analyses with 1000 samples.

Finally, risk-adjusted morbidity and mortality rates were calculated for each unit by dividing the observed to the predicted outcome multiplied by the mean observed outcome rate in the total population. Risk-adjusted outcome rates (morbidity or mortality) are regarded as the outcome rates a unit would have if its case-mix would be similar to the average case-mix in the entire population.

2.2.1. Determination of final composite score

Final composite score combined the two process measures (percent of ppoDLCO in patients with ppoFEV1 <40%, and percent of systematic lymph nodes dissection in patients with primary lung cancer) with the two risk-adjusted outcome measures (mortality and cardiopulmonary complications rates) into a single comprehensive quality score.

To assure consistent directionality (increasingly positive values reflecting better performance), mortality rates were converted to survival rates (risk-standardized survival rate = 100 – risk-standardized mortality rate), and morbidity rates were converted to ‘freedom from morbidity’ rates (risk-standardized freedom from morbidity rate = 100 – risk-standardized morbidity rate).

Account was taken of the differing measurement scales of the domain-specific scores. Theoretically, each measure has
the same scale, 0—100%, in reality, measurement scales differ dramatically. Process adherence rates are widely dispersed and range from close to 0% to almost 100% (a range of 100%). In contrast, risk-standardized survival rates are tightly clustered in a narrow interval ranging from about 95% to 99% (a range of 4%). To account for these differences, the scales of measurement were standardized by the reciprocal of their standard deviations before the individual domain scales of measurement were standardized by the reciprocal to 99% (a range of 4%). To account for these differences, the rescaling method was applied to all the outcome and process indicators before summarizing the final composite score of each unit.

All data were at least 95% complete. All patients with postoperative events had complete data. Missing data were imputed by averaging the non-missing values for numerical variables and by taking the most frequent category for categorical variables. All the statistical tests were two-tailed and a significance level of 0.05 was accepted. The analysis was performed by using the STATA 8.2 (Stata Corp., College Station, TX) statistical software.

3. Results

The characteristics of patients operated on in the 10 units analyzed in this study are shown in Table 1 along with the individual morbidity and mortality rates. Table 2 shows the results of the multilevel regression analyses for in-hospital morbidity and mortality. Factors reliably associated with cardiopulmonary complications were age \((p < 0.0001)\), ppoFEV1 \((p < 0.0001)\) and the presence of an extended resection \((p < 0.0001)\). The resulting model for morbidity was:

\[
\ln \left( \frac{R}{R-1} \right) = 0.13 + 0.0036 \times \text{age} - 0.002 \times \text{ppoFEV1} + 0.14 \times \text{extended resection}.
\]

Factors reliably associated with mortality were age \((p = 0.02)\) and the presence of an extended resection \((p < 0.0001)\). The resulting model for mortality was:

\[
\ln \left( \frac{R}{R-1} \right) = -0.04 + 0.0009 \times \text{age} + 0.05 \times \text{extended resection}.
\]

Table 3 summarizes the process and outcome rates in each unit in the study. The average rates of DLCO measurement (in patients with ppoFEV1 <40%), systematic lymph nodes dissection, probability of non-complication and probability

<table>
<thead>
<tr>
<th>Variables</th>
<th>Coefficients</th>
<th>SE</th>
<th>p values</th>
<th>Bootstrap%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Morbidity</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>0.13</td>
<td>0.07</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>0.0036</td>
<td>0.0009</td>
<td>&lt;0.0001</td>
<td>92%</td>
</tr>
<tr>
<td>ppoFEV1</td>
<td>-0.002</td>
<td>0.0005</td>
<td>0.0001</td>
<td>87%</td>
</tr>
<tr>
<td>Extended resections</td>
<td>0.14</td>
<td>0.03</td>
<td>&lt;0.0001</td>
<td>99.5%</td>
</tr>
<tr>
<td>Mortality</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>-0.04</td>
<td>0.02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>0.0009</td>
<td>0.0003</td>
<td>0.02</td>
<td>65%</td>
</tr>
<tr>
<td>Extended resection</td>
<td>0.05</td>
<td>0.01</td>
<td>&lt;0.0001</td>
<td>98%</td>
</tr>
</tbody>
</table>

Morbidity model: c-index 0.65, Hosmer–Lemeshow goodness of fit 2.3 \((p = 0.9)\); mortality model: c-index 0.71, Hosmer–Lemeshow goodness of fit 9.8 \((p = 0.3)\).

Table 3 Process and risk-adjusted outcome measures in each unit under analysis.

<table>
<thead>
<tr>
<th>Unit</th>
<th>%DLCO measured</th>
<th>%Systematic lymphadenectomy</th>
<th>Risk-adjusted probability of non-complications</th>
<th>Risk-adjusted probability of survival</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>86.6</td>
<td>94.5</td>
<td>97.4</td>
</tr>
<tr>
<td>B</td>
<td>93</td>
<td>97.7</td>
<td>67.5</td>
<td>94.4</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>97.7</td>
<td>87.4</td>
<td>96.6</td>
</tr>
<tr>
<td>D</td>
<td>0</td>
<td>96.7</td>
<td>74.4</td>
<td>95.3</td>
</tr>
<tr>
<td>E</td>
<td>0</td>
<td>96.4</td>
<td>86.7</td>
<td>98.2</td>
</tr>
<tr>
<td>F</td>
<td>53</td>
<td>92.4</td>
<td>92.1</td>
<td>99.4</td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td>95</td>
<td>85.6</td>
<td>97</td>
</tr>
<tr>
<td>H</td>
<td>13</td>
<td>97</td>
<td>87.2</td>
<td>98.5</td>
</tr>
<tr>
<td>I</td>
<td>4</td>
<td>97.5</td>
<td>91.1</td>
<td>96.8</td>
</tr>
<tr>
<td>L</td>
<td>0</td>
<td>97</td>
<td>86.4</td>
<td>100</td>
</tr>
</tbody>
</table>

Results are expressed as percentages.
of survival in the entire population were 16.3% (SD 31.6), 56.7% (SD 37.8), 80.6% (SD 10.9), 97.5% (SD 2.1), respectively. From a quality management perspective, the different indicators are associated with each other in such a ratio that a 1% decrease in mortality corresponds to 5% decrease in complications, 19% increase in systematic nodal dissection and 16% increase in DLCO measured in pts with marginal ppoFEV1.

Composite scores were generated by adding the rescaled individual process and outcome scores of each unit, and they ranged from individual process and outcome scores of each unit, and they ranged from 4.4 to 3.7 (Table 4). The correlation coefficients between the individual four rescaled scores (processes and outcomes) were all below 0.5, with the exception of non-complication score and survival score (processes and outcomes) were all below 0.5, with the exception of non-complication score and survival score (processes and outcomes) were all below 0.5, with the exception of non-complication score and survival score (processes and outcomes) were all below 0.5, with the exception of non-complication score and survival score (0.68).

Table 4
Summary of the individual rescaled process and outcome scores in each unit.

<table>
<thead>
<tr>
<th>Unit</th>
<th>DLCO score (preop process)</th>
<th>Lymphadenectomy score (intraop process)</th>
<th>Mortality score</th>
<th>Mortality score</th>
<th>CPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>−0.516</td>
<td>0.78</td>
<td>1.122</td>
<td>0</td>
<td>1.386</td>
</tr>
<tr>
<td>B</td>
<td>2.427</td>
<td>1.03</td>
<td>−2.171</td>
<td>−1.765</td>
<td>−0.478</td>
</tr>
<tr>
<td>C</td>
<td>−0.516</td>
<td>−1.3</td>
<td>0.256</td>
<td>−0.471</td>
<td>−2.03</td>
</tr>
<tr>
<td>D</td>
<td>−0.516</td>
<td>−1.5</td>
<td>−1.329</td>
<td>−1.235</td>
<td>−4.58</td>
</tr>
<tr>
<td>E</td>
<td>−0.516</td>
<td>−1.3</td>
<td>0.171</td>
<td>0.471</td>
<td>−1.175</td>
</tr>
<tr>
<td>F</td>
<td>1.161</td>
<td>0.85</td>
<td>0.829</td>
<td>1.176</td>
<td>4.017</td>
</tr>
<tr>
<td>G</td>
<td>−0.516</td>
<td>0.07</td>
<td>0.037</td>
<td>−0.235</td>
<td>−0.645</td>
</tr>
<tr>
<td>H</td>
<td>−0.104</td>
<td>0.97</td>
<td>0.232</td>
<td>0.647</td>
<td>1.744</td>
</tr>
<tr>
<td>I</td>
<td>−0.389</td>
<td>0.166</td>
<td>0.707</td>
<td>−0.353</td>
<td>0.131</td>
</tr>
<tr>
<td>L</td>
<td>−0.516</td>
<td>0.58</td>
<td>0.134</td>
<td>1.529</td>
<td>1.728</td>
</tr>
</tbody>
</table>

CPS: composite performance score obtained by summing all the individual rescaled scores.

Table 5
Rating of units by their composite score and, by reference, their risk-adjusted mortality rates.

<table>
<thead>
<tr>
<th>Units</th>
<th>Rank position by risk-adjusted mortality</th>
<th>Rank position by CPS</th>
<th>Positions moved</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>5</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td>10</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>C</td>
<td>8</td>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>D</td>
<td>9</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>E</td>
<td>4</td>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>F</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>G</td>
<td>6</td>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>H</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>I</td>
<td>7</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>L</td>
<td>1</td>
<td>3</td>
<td>2</td>
</tr>
</tbody>
</table>

4. Discussion

The current climate of public accountability, financial restraints, pay-for-performance and ranking culture demands that performance of health care providers be monitored through reliable instruments. In our specialty outcome endpoints are still the most widely used quality indicators [2,4,17—21], however these may be quite insensitive to quality, require long-term evaluation to aggregate the sufficient number of events for a reliable statistics and fail to provide immediately actionable information to providers [1].

More comprehensive quality indicators are needed that may capture most aspects of the entire process of care and for which providers may be reliably made accountable.

The ESTS thoracic database was designed to provide a tool for a fair evaluation of performance across different European thoracic surgery units. Several reports have already been published on the development and use of a risk-adjusted mortality model derived from data collected in the database [2,4]. More recently the ESTS database committee in cooperation with the ESTS Audit and Clinical Excellence Committee and in response to the growing demand for a more reliable instrument of performance rating that may be used as a tool for credentialing and regulatory purposes decided to build a comprehensive performance score from data available in the version 1 of the European database.

Several principles underpinned the development of a comprehensive quality rating system for lung surgery:

1. Quality assessment should be at the level of the program or hospital rather than the individual surgeon.
2. Quality scores should consider structure, process, and outcomes [22].
3. Quality scores should assess three temporal domains: preoperative, operative, and postoperative [3].
4. Quality scores should be interpretable and actionable by providers.
5. Initial quality reports should focus on lung resection as the most representative operation in our specialty.
6. All quality measures should be available as data elements within the European thoracic surgery database.

The use of a composite score for measuring quality of care is appealing in light of the abstract nature of the quality
itself, which cannot be directly measured. Although quality improvement requires attention to each individual aspect of quality, there are many settings in which the users of quality measures are most interested in the bottom line. Thus, composite indicators seem particularly useful for summarizing and comparing the quality of care delivered by healthcare providers. This is particularly true in many surgical areas where the small sample size and the low rate of adverse events diminish the statistical utility of outcomes comparisons [23]. Thus, composite indicators may provide a quantitative basis for clinicians, organizations, and planners aiming to achieve improvement in care and the process by which patient care is provided. They may serve many purposes: document the quality of care, make comparisons (benchmarking) over time between providers or within the same provider, make judgements and set priorities (i.e., choosing a hospital or surgery; organizing medical care), support accountability, regulation, and accreditation, support quality improvement, and support patient choice of providers.

It was noteworthy that results obtained applying the CPS differed substantially from those obtained by applying a more traditional endpoint such as risk-adjusted mortality. In fact, all units changed their position in the ranking and 40% moved two or more positions. This may have important administrative and managerial implications.

This score is far from being perfect and future refinements in data collection and the inclusion of new outcome or process indicators are needed to improve its reliability and precision. A mechanism for independent validation of the completeness and accuracy of each center’s data was not present and should be implemented. It would be desirable to include multiple processes within each domain in order to assess the quality of care in a more comprehensive manner. Unfortunately, we were limited to those process indicators that were present in the first version of the European database. A second version of the database has been recently launched that could allow the inclusion of multiple evidence-based process indicators (https://www.thoracicdata.org).

For a composite score to be valid, there must be a strong link between processes and outcomes that can come from previously published evidence. This prior evidence supporting the link between process and outcomes may be weak or non-existent for many processes even when they are truly linked to outcomes.

We did not include structural indicators in the CPS. However, a structural indicator is very often regarded as a surrogate in the absence of valid risk-adjusted outcome endpoints. Furthermore all units under analysis participated to the database, which was considered per se a structural indicator common to all centers. Each individual indicator reflects a different aspect of quality. Therefore, some information is invariably lost when only a single summary score is reported, particularly when measures are poorly correlated and non-redundant.

We recognize that some important areas of performance may not be addressed or may be relatively undervalued, and aggregation may also obscure individual areas of strength or weakness. The ability to decompose the composite into its individual components is critical. This allows providers to analyze their performance in specific areas and to formulate improvement strategies [5]. The CPS as developed in this analysis can be easily decomposed and the units may be evaluated according to their individual scores related to preoperative, operative and postoperative domains.

Determining a set of valid exclusion criteria for process measures has proved challenging in many areas of medicine, and information on contraindications to various process measures is absent in the European database. As a result, some patients who are appropriately denied a practice owing to a contraindication will be misclassified as representing a process failure. The true overall percentage compliance with process measures will therefore be underestimated because some ineligible patients are included. Limitations also exist with respect to various outcome measures. Because certain outcomes (morbidity) are difficult to define precisely, it is possible that variation in coding practices could account for some of the observed differences between providers.

The encouraging results of this study lead us to consider future implementations in agreement with similar conclusions reached by the STS task force for quality measurement in cardiac surgery [5]:

- construction of separate composite scores for processes and outcomes, assessing their degree of congruence, and studying each separately to determine their ability to discriminate overall performance,
- consideration of new or revised measures that provide important incremental information about overall care,
- ongoing assessment of the relationship between process measures and short- or long-term outcomes, with potential elimination of process measures that demonstrate limited clinical effectiveness,
- monitoring provider-specific scores longitudinally to assess their consistency, and to study which domains and individual measures appear most predictive of subsequent provider performance.

If composite scores are to be used for a rating system, choices must be made regarding the use of an internal or external benchmark, the number of rating tiers, the cut points used to define each tier, and the statistical criteria used to assign providers to tiers.

- research is needed to identify the optimum time window for estimating and reporting performance, and the methods for dealing with underperformance.

The availability of a reliable rating system may make possible to implement peer-based, provider-initiated, organizational quality initiatives [24,25], such as the creation of European thoracic quality benchmark program. Those units participating in the database will be subject to scoring and rating on a voluntary, non-coercive, non-punitive, cooperative way. Their performance will be reviewed annually during a benchmark meeting lead by ESTS database representatives. Rating will be confidentially revealed and an in-depth analysis of individual critical pathways of care reviewed in the outlier’s centers. Possible visits to individual centers of high performance may be scheduled and made available to representatives of units in the lower performance tiers in order to gain insights for improving their activity. The success of such an initiative may be evaluated by longitudinal review of performance on a yearly basis. The
The main objective of this program would be to globally improve the quality of the entire European thoracic surgery community and to apply clinical governance mechanisms that may protect our profession from the increasing public and forensic scrutiny.

It must be emphasized that the aim of the present project was to provide an example of methodology on how to construct and use a composite performance score in our specialty and in the context of multicenter database and not to provide a ready-made instrument for quality assessment.

In conclusion, based on data and variables from the European Thoracic Database, a composite performance score, incorporating process and outcome measures of quality, has been developed, with the ultimate intent to provide a methodological template for implementing instruments for a reliable provider profiling and performance monitoring. Future refinement is needed in order to apply this tool for international quality initiatives and regulatory and credentialing purposes.

References