Multicentric analysis of performance after major lung resections by using the European Society Objective Score (ESOS)

Alessandro Brunelli a,*, Gonzalo Varela b, Paul Van Schil c, Michele Salati a, Nuria Novoa b, Jeroen M. Hendriks c, Marcelo F. Jimenez b, Patrick Lauwers c, on behalf of the ESTS Audit and Clinical Excellence Committee

a Division of Thoracic Surgery, Umberto I Regional Hospital, Via Santa Margherita 23, Ancona, Italy
b Thoracic Surgery Service, Salamanca University Hospital, Spain
c Department of Thoracic and Vascular Surgery, University Hospital Antwerp, Belgium

Received 28 May 2007; received in revised form 28 September 2007; accepted 3 October 2007; Available online 21 December 2007

Abstract

Objective: Outcome endpoints are still the most widely used indicators of performance. However, they need to be risk-adjusted in order to be reliable instruments of audit. Recently, the European Society Objective Score (ESOS) was developed from the online European Thoracic Surgery Database as an audit tool. In this study, we applied for the first time the ESOS.01 to assess the performance of three European thoracic surgery units during three successive years of activity.

Methods: This study is a retrospective analysis performed on prospective databases. We analysed 695 patients submitted to pneumonectomy (117) or lobectomy (578) for lung neoplasm at three European dedicated thoracic surgery units (unit A 264 patients, unit B 262, unit C 169) from January 2004 through December 2006. Qualified thoracic surgeons performed all the operations. No patients in this series were in the original ESOS development set. ESOS.01 was used to estimate the risk of in-hospital mortality in all patients. Observed and predicted mortality rates were then compared within each unit by the z-test.

Results: Cumulative observed mortality rates in units A, B and C were 2.3% (six cases), 2.7% (seven cases) and 4.1% (seven cases), respectively. We were not able to find statistically significant differences between observed and ESOS-predicted mortality rates. The comparison of risk-adjusted mortality rates between units did not show significant differences (unit A 3.9%, unit B 3.3%, unit C 5.6%).

Conclusions: The use of ESOS.01 revealed that the performances of all units were in line with the predicted ones during each period under analysis and did not differ between each other. The results of our study warrant future efforts to refine the ESOS model and to develop other risk-adjusted outcome indicators with the aim to establish European benchmarks of performance.

© 2007 European Association for Cardio-Thoracic Surgery. Published by Elsevier B.V. All rights reserved.

Keywords: Lung resection; Lung cancer; Mortality; Quality of care; Comparative audit; Risk model

1. Introduction

In the present era of managed care system, cost-containment, public accountability, pay-for-performance and ranking culture, an increasing pressure is exerted on our profession to deliver high quality of surgical care.

Although Quality is an abstract construct that cannot be directly measured [1] several indicators have been used as surrogates thought to be associated with it. Measuring the quality of clinical practice requires focusing on its structure, process and outcomes [2].

In our specialty, outcome endpoints are still the most widely used indicators of quality. However, whenever used to evaluate the providers’ performance, outcomes need to be adjusted for the different case-mixes that may characterize different institutions or, even, different periods of activity. The use of crude outcome rates may, in fact, lead to misleading information with inherent important clinical and administrative consequences for the unit, and to unethical risk-averse behaviours.

With the intent to provide an instrument for a fair comparative audit, the ESTS Thoracic Surgery Database project recently published the European Society Objective Score (ESOS.01) for in-hospital mortality [3].

The objective of the present multicentric study was to apply ESOS.01 to assess the performance of three different European thoracic surgery units after major lung resection in patients with lung tumours, with the intent to provide a methodological and practical framework for institutional comparative audit.
2. Patients and methods

All 695 patients submitted to major lung resections (578 lobectomies, 117 pneumonectomies) for lung neoplasms at three European dedicated general thoracic surgery units (unit A 264 patients, unit B 262 patients, unit C 169 patients) during a 3-year period (January 2004—December 2006) were included in the analysis. This is an observational analysis performed on prospectively collected data. Patients’ information was prospectively collected at each individual centre in electronic, periodically audited datasets which were approved by the local hospital institutional review boards. Informed consent was obtained from all patients to collect and use their data for quality management and analysis purposes. Data were collected at each unit by a trained physician and periodically audited by a designated Clinical Audit Lead who was responsible for their accuracy and completeness. The local in-house databases are currently used as continuous quality improvement instruments at the three participating units.

For the purpose of this study, the variables and endpoints of interest were centralized in a merged dataset, after being anonymised for both patients and surgeons. The resulting merged dataset was thoroughly evaluated to detect and possibly correct incomplete or inconsistent data by the principal investigators of the study (AB, GV, PVS).

In all the units, surgery was contraindicated in those patients with a predicted postoperative forced expiratory volume in 1 s (ppoFEV1) and predicted postoperative carbon monoxide lung diffusion capacity (ppoDLCO) less than 30% of predicted, in association with a poor exercise capacity (height at maximal stair climbing test lower than 12 m or maximum oxygen consumption at cycle-ergospirometry less than 10 ml/kg min) or in the presence of hemodynamic instability despite optimization of treatment. As a rule, lung resections were performed through a muscle-sparing thoracotomy by certified thoracic surgeons. In the postoperative period, patients of unit A and B were admitted to a dedicated general thoracic surgery ward immediately after operation, resorting to ICU only in case of complications needing invasive assisted ventilation or invasive continuous monitoring. In unit C, patients were admitted to ICU for a period of 24—48 h and then transferred back to the thoracic ward. Postoperative treatment was standardized in all units and focused on early mobilization, chest physiotherapy and physical rehabilitation, thoracotomy pain control, antibiotic and anti-thrombotic prophylaxis. Postoperative chest pain was controlled by means of epidural or continuous intravenous analgesia, which were titrated to keep the pain visual score below 4 (in a scale ranging from 0 to 10) during the first postoperative 48—72 h.

3. Statistical analysis

All variables of interest in the dataset were complete. The ESOS.01 [3] logistic equation \( \ln(R/1-R) = -5.8858 + 0.0501 \text{Xage} - 0.0218 \times \text{ppoFEV1} \) was applied to all patients to estimate their risk of in-hospital mortality.

Predicted postoperative forced expiratory volume in 1 s (ppoFEV1) was expressed as percentage of predicted normal for age, gender and height according to the standard equation [4] and calculated in all units according to the number of obstructed/non-obstructed segments removed during operation [5].

The characteristics of the patients were compared between units by means of the Mann–Whitney test (numerical variables) and Chi-square test (categorical variables).

Predicted and observed in-hospital mortality rates in each unit were then compared by Fisher’s exact test. Risk-adjusted mortality rates were also estimated for each unit by multiplying the observed/predicted mortality ratio of each unit by the cumulative observed mortality rate in the entire population [6]. Risk-adjusted mortality rates represent the mortality rate a unit would have if its case-mix were similar to the average case-mix in the study and were compared by the Fisher’s exact test.

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.

4. Results

The characteristics of the patients in the three units are shown in Table 1. The three units had a different case-mix of patients. In fact, compared to the other units, elderly patients were more frequently operated on in unit B, ppoFEV1 was higher in unit A and a greater number of pneumonectomies were performed in unit C.

Cumulative observed mortality rates in unit A, B and C were 2.3% (six cases), 2.6% (seven cases) and 4.1% (seven cases), respectively.

Fig. 1 shows the distribution of quartiles of expected mortality risk in each unit. Unit A had a significantly lower proportion of patients with higher expected risk of mortality compared to the other two units \( (p < 0.0001) \).

The comparison of observed versus predicted mortality rates within each unit did not show any statistically significant differences (Table 2).

Risk-adjusted mortality rates in unit A, B and C were 3.9%, 3.3% and 5.6%, respectively, without any significant differ-

---

Table 1  
Characteristics of the patients in the three units under analysis

<table>
<thead>
<tr>
<th>Variables</th>
<th>Unit A (264 patients)</th>
<th>Unit B (262 patients)</th>
<th>Unit C (169 patients)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years)</td>
<td>64.5 (10.7)</td>
<td>66.4 (10.7)</td>
<td>64.6 (9.9)</td>
</tr>
<tr>
<td>Elderly (&gt;70 years)</td>
<td>89 (34%)</td>
<td>115 (44%)*</td>
<td>55 (33%)</td>
</tr>
<tr>
<td>PpoFEV1%</td>
<td>75.7 (26.2)%</td>
<td>67.4 (16.3)%*</td>
<td>63 (17.3)%</td>
</tr>
<tr>
<td>PpoDLCO%</td>
<td>n/a</td>
<td>63.1 (16.1)%</td>
<td>59.1 (18.5)%*</td>
</tr>
<tr>
<td>Pneumonectomy (n, %)</td>
<td>40 (15%)</td>
<td>35 (13%)</td>
<td>42 (25%)*</td>
</tr>
</tbody>
</table>

Results are expressed as means ± SD unless otherwise indicated. (*) \( p < 0.05 \) compared to units A and C. (§) \( p < 0.05 \) compared to unit B and C. (\( ) \ p < 0.05 \) compared to units A and B (ppoDLCO not assessed in unit A).
ences between the units (unit A vs unit B, \(p = 1\), unit A vs unit C, \(p = 0.6\), unit B vs unit C, \(p = 0.5\)).

Table 3 shows the comparison of observed versus predicted mortality rates within each unit subdivided by year of operation. Even in this case, each unit appeared to work in line with the expected mortality rates within each year of activity.

We then restricted the analysis to the patients with the highest risk of postoperative death (>95th percentile, corresponding to 4% of mortality risk). In this high-risk group of patients (8 patients in unit A, 12 in unit B and 11 in unit C), no mortality was observed in each unit.

5. Discussion

The increasing pressure exerted by administrators, financing institutions and the public to monitor and improve the quality of surgical care requires an accurate evaluation of the performance through the use of reliable instruments.

Although imperfect in many circumstances [7], outcome indicators are still the most widely used endpoints to evaluate the performance of surgical units. Mortality is certainly one of the most used outcome indicators, since it is an unambiguous endpoint and it is regarded particularly attractive by both the public and the payers. Patients and insurances are in fact likely to prefer units with the lowest mortality rates.

Nevertheless, if outcomes are used for comparative audit, they need to take into account the fact that patients are not randomized between different units, and different selection processes and referral patterns may determine considerable differences in the case-mix. As outcomes are known to reflect not only the quality of the process delivered to the patients but also (and in a variable degree according to the type of outcome) the severity of operation and the physiological conditions of patients at the time of surgery, they need to be necessarily adjusted for the patients’ risk factors.

With the aim to develop an instrument for a fair evaluation of performance, the European Society of Thoracic Surgeons implemented an online European Thoracic Surgery Database to prospectively capture data on thoracic surgical procedures from units across Europe. The database was designed to contain a minimum set of core variables and endpoints. A total of 27 units from 14 European Countries consistently contributed to submit data from January 2001 through December 2003 in a voluntary fashion and without external local data audit. An ESTS Database Committee was formed and was deemed responsible for scrutinizing data for possible inconsistencies, missing values, or errors in the effort to guarantee the highest quality of data as possible in the context of a voluntary database. The first product of this project was the development of the European Society Objective Score (ESOS), a risk-adjusting model for in-hospital mortality following lung resection for lung tumours [3]. ESOS was derived from approximately 1700 patients and validated in an external sample of 1128 patients randomly drawn from the dataset. The model contains only two predictors (age and ppoFEV1) and was deemed to have a good face and content validity, and a satisfactory precision in the validation set.

The objective of the present multicentric investigation was to apply ESOS.01 as an instrument to evaluate the performance of three dedicated European thoracic surgery
units with the ultimate intent to provide a methodological and practical framework for institutional comparative audit in Europe.

We limited the analysis to patients submitted to major lung resections for lung tumours. Although the original model was developed from all types of lung resections we decided not to include minor resections as the mortality in this sample in all three units was virtually absent. Lung resection for benign diseases was also not included since the original ESOS was developed from lung tumours patients only and its accuracy in patients with benign disease is still to be verified.

In-hospital mortality was selected as the outcome indicator since this was the endpoint originally modelled by ESOS [3]. Although other mortality time frames may be more indicated and may perhaps reflect different aspects of quality, in our case we did not observe any additional 30-day mortality in those patients discharged from the hospital.

We found that the distribution of the predicted risk of death between the three units was different. Unit A had a significantly lower proportion of patients with a higher expected risk of mortality, underlining the importance to use a risk-stratification tool to adjust for different case-mixes at different institutions.

The comparison of predicted and observed mortality rates within each unit did not show any statistically significant differences.

In order to compare the mortality rates between the units, we then calculated the risk-adjusted mortality rates for each unit. These rates reflect the mortality a unit would have if its case-mix were similar to the average case-mix in the study. No differences in risk-adjusted mortality rates were noted between the three units.

Risk-adjusting models allow the analysis of success cases. When the analysis was in fact restricted to those patients with the highest risk of death (>95th percentile of expected mortality), all units showed an excellent performance with no observed mortality. This may trigger an in-depth evaluation of the processes of care leading to the favourable outcome in these high-risk patients, and, if feasible, lead to their generalization to other patients or other centres.

The results of this study should be interpreted with caution due to the inherent limitation of the mortality endpoint when used as quality monitoring instrument. In spite of the fact that 3 years of activity were used in this study, this period of time may not have been enough to aggregate the sufficient number of patients and detect statistically significant differences between predicted and observed outcomes, as also indicated by the wide confidence intervals of the differences. Sample size may raise serious problems if mortality is used as the sole indicator for judging the performance. Even the most stable model generated from the largest multicentric population has, in fact, to be applied for audit purposes to individual units or, worse, to individual surgeons [8].

On the other hand, ideal indicators should provide rapidly actionable information in order to implement corrective measures to improve the quality of care without undesirable wasting of time and perhaps preventable life-losses [9].

One way to obviate this problem may be the use of multiple risk-adjusted outcomes for a more comprehensive and reliable provider profiling [10]. In fact, each different outcome may reflect different aspects of quality [11].

The results generated in this study show that a risk model developed from a multi-institutional international dataset may be easily applied for comparative audit at local sites and that future efforts are warranted to refine the model and develop other indicators of quality.

In our view, these future efforts should focus on the refinement of data collection in the most cost-saving and quality-controlled way as possible, the development of process-based indicators of quality that may obviate problems of sample size and provide more actionable information on the performance, the use of breakthrough reliability statistics for cross-validation of risk-models (bootstrap) [12—14], the identification of mechanisms for dealing with underperformance, and the research of the optimal time-frame for evaluating and reporting the performance.

In conclusion, in this era of performance ratings, which are increasingly used not only for quality improvement, but also for public reporting and for pay-for-performance, we need to be evaluated in the most objective way as possible. As physicians, we must assume complete responsibility at the local, national, and international level to influence health policy-makers in order to use the most rigorous and accurate performance methodologies [15].

Organizational quality initiatives, such as ESOS, should always be interpreted in a dynamic fashion and subject to refinements and improvements. In this regard, international cooperative efforts (such as the ESTS Thoracic Surgery Database) will be the ideal platform for implementing future quality monitoring and improvement activities and for setting European benchmarks of performance.

References

Appendix A. Conference discussion

Dr D. Wood (Seattle, WA): I am impressed at the consistency between the three units and I think that shows when you have a modelling system that can risk-adjust how good units can have very small differences in outcomes. I think that is a strength of a thoracic surgery database and the quality improvement that is inherent in database participation.

A comment and a suggestion rather than a question would be that the STS General Thoracic Surgery Database has now developed a risk model for pulmonary resection as well, a model developed completely independently from the ESTS. I think it would be useful for the two databases to cross-test each other’s modelling and see how well they fit into another database. I think that is an opportunity for collaboration between our two societies and two databases and could potentially result in an even stronger risk adjustment model.

Dr S. Cassivi (Rochester, MN): Thank you for the wonderful presentation. It is very important that we talk about quality. It is the issue of the day and as surgeons we should be at the forefront. Your ability to have a risk-adjusted model which seems to be reproducible among three good surgical units is already a huge start. You have found a relatively simple observation that implicates two factors for risk adjustment: age and FEV1.

In your discussion, you mentioned that currently there is no good way of using the current outcome data to determine changes that could be made to improve on a particular centre’s results. Clearly, outcome data do not translate easily into definable changes that can be made because they don’t tell us why one surgical unit achieves better results than another. A different strategy may be to look at process measures as indicators of performance. These measures can more easily translate into significant changes in practice. I know you agree that these too are important measures of a surgical unit’s performance. Can you explain how you propose to convert this laudable project into a process measure project?

Dr Brunelli: I think the next step will be to switch from outcome indicators to indicators of performance which are much more informative on where and how to improve our practice.

Dr T. Rice (Cleveland, OH): Is mortality really an end-point that should be looked at in the analyses of performance?

Dr Brunelli: Mortality is probably the most widely used outcome indicator of performance especially in our specialty. The problems with mortality are multiple. The most important one is that fortunately mortality after lung resection is rare only about 1–3%. This means that to aggregate the sufficient number of events to perform a reliable and meaningful statistic takes time and time is not what we need when we want to assess performance. You need rapid feedback; you don’t want to waste time and potentially preventable life losses in this process. Maybe we can improve our evaluation tools by adding new (intermediate) outcomes such as morbidity, ICU admission, residual functional state, etc and then ultimately by developing reliable process indicators.

In a similar fashion as our colleagues in STS cardiac surgery database, who were able to develop a composite score for quality assessment including outcome and process indicators in a single score, I hope in the future we will be able to produce a similar instrument and I hope in this regard, I look forward into a cooperation between North America and Europe to have more members and a more powerful database.