Does the Aristotle Score predict outcome in congenital heart surgery?

Nicholas Kang a, Victor T. Tsang a,*, Martin J. Elliott a, Marc R. de Leval a, Timothy J. Cole b

a Cardiothoracic Unit, Great Ormond Street Hospital for Children NHS Trust, London WC1N 3JH, UK
b Centre for Paediatric Epidemiology and Biostatistics, Institute of Child Health, London, UK

Received 17 October 2005; received in revised form 16 January 2006; accepted 30 January 2006; Available online 4 May 2006

Abstract

Objective: The Aristotle Score has recently been proposed as a measure of ‘complexity’ in congenital heart surgery, and a tool for comparing performance amongst different centres. To date, however, it remains unvalidated. We examined whether the Basic Aristotle Score was a useful predictor of mortality following open-heart surgery, and compared it to the Risk Adjustment in Congenital Heart Surgery (RACHS-1) system. We also examined the ability of the Aristotle Score to measure performance. Methods: The Basic Aristotle Score and RACHS-1 risk categories were assigned retrospectively to 1085 operations involving cardiopulmonary bypass in children less than 18 years of age. Multiple logistic regression analysis was used to determine the significance of the Aristotle Score and RACHS-1 category as independent predictors of in-hospital mortality. Operative performance was calculated using the Aristotle equation: performance = complexity/C2 survival.

Results: Multiple logistic regression identified RACHS-1 category to be a powerful predictor of mortality (Wald 17.7, $p < 0.0001$), whereas Aristotle Score was only weakly associated with mortality (Wald 4.8, $p = 0.03$). Age at operation and bypass time were also highly significant predictors of postoperative death (Wald 13.7 and 33.8, respectively, $p < 0.0001$ for both). Operative performance was measured at 7.52 units. Conclusions: The Basic Aristotle Score was only weakly associated with postoperative mortality in this series. Operative performance appeared to be inflated by the fact that the overall complexity of cases was relatively high in this series. An alternative equation (performance = complexity/mortality) is proposed as a fairer and more logical method of risk-adjustment.

# 2006 Elsevier B.V. All rights reserved.

Keywords: Cardiac surgery; Risk; Surgical outcomes

1. Introduction

The Aristotle Score has recently been proposed as a complexity scoring system in congenital heart surgery. The score is based on three components: mortality, morbidity and technical difficulty. Estimates of these three elements were subjectively rated by a consensus panel, and a composite score was defined for 145 different procedures [1].

We sought to test whether the Aristotle Score predicted mortality at our institution. We also compared its predictive ability to that of the Risk Adjustment for Congenital Heart Surgery (RACHS-1) method [2]. In this study, only the Basic Aristotle Score was examined, rather than the comprehensive score, which requires up to 248 variables to be assessed. We also measured our institutional performance using the equation proposed by the Aristotle Committee.

Based on our experience, we discuss the relative merits and shortcomings of the Aristotle scoring system as an instrument for surgical outcome analysis. The method of calculating operative performance is also addressed, and an alternative equation proposed.

2. Material and methods

A 3-year dataset (between April 2000 and March 2003) of open-heart surgery at our institution, which has previously been validated and analysed as detailed in an earlier report [3], was examined in order to assess the Aristotle Score and its relationship to postoperative in-hospital mortality. Scores were assigned retrospectively using the scoring system published by the Aristotle Committee in June 2004.

Multiple logistic regression modelling [4] was used to examine the strength of the relationship between the Aristotle Score and in-hospital mortality, before and after controlling for other previously identified independent risk factors. These other factors were (1) age at operation, (2) bypass time and (3) RACHS-1 risk category [3]. Regression
models were constructed with (1) data from all patients able to be scored using the Aristotle system, and (2) data restricted to patients scored only by both the Aristotle and RACHS-1 systems.

Performance was calculated using the equation proposed by the Aristotle Committee: performance = complexity × survival. Mean complexity and mean survival were calculated from patients in whom an Aristotle Score was able to be assigned.

3. Results

Of 1085 cases in the dataset, 1068 (98.4%) were able to be scored using the Basic Aristotle Score. The remaining 17 cases were not defined by any of the 145 procedures listed in the scoring system, and were excluded from further analysis. The mean Aristotle Score was 7.9 and the median score was 8.0. Table 1 summarises the distribution of cases according to Aristotle Score complexity level and postoperative mortality.

In contrast, only 998 cases (92%) could be classified using the RACHS-1 method. The median RACHS-1 risk category was 3 (mean 2.4). Nine hundred ninety-two cases (91%) were able to be coded using both Aristotle and RACHS-1 systems. Restricting the risk model to these 992 cases, multiple logistic regression identified RACHS-1 category to be a powerful predictor of mortality (Wald 17.7, \( p < 0.0001 \)), whereas Aristotle Score was only weakly associated with mortality (Wald 4.8, \( p = 0.03 \)). Age at operation and bypass time were also highly significant predictors of postoperative death (Wald 13.7 and 33.8, respectively, \( p < 0.0001 \) for both). Age was modelled as \( 1/\sqrt{\text{age in days}} + 1 \). Table 2 summarises the results of the analysis for the Aristotle and RACHS-1 systems.

Using the same 992 cases, but excluding RACHS-1 category from the model, the Aristotle Score became slightly worse in its predictive value (Wald 4.0, \( p = 0.02 \)). Repeating the analysis with all patients who were able to be scored by the Aristotle system (n = 1068), the Aristotle Score became slightly more significant (Wald 5.9, \( p = 0.02 \)).

Table 1

<table>
<thead>
<tr>
<th>Complexity level</th>
<th>Aristotle Score</th>
<th>Cases (%) (n)</th>
<th>Mortality (%) (n)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.5—5.9</td>
<td>10 (109)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>2</td>
<td>6.0—7.9</td>
<td>35 (374)</td>
<td>3.2 (12)</td>
</tr>
<tr>
<td>3</td>
<td>8.0—9.9</td>
<td>37 (390)</td>
<td>5.1 (20)</td>
</tr>
<tr>
<td>4</td>
<td>10.0—15.0</td>
<td>18 (195)</td>
<td>8.7 (17)</td>
</tr>
<tr>
<td>Total</td>
<td>1.5—15.0</td>
<td>100 (1068)</td>
<td>4.6 (49)</td>
</tr>
</tbody>
</table>

Table 2

Comparison of Aristotle and RACHS-1 systems and results of multiple logistic regression analysis examining association with postoperative mortality

<table>
<thead>
<tr>
<th></th>
<th>Aristotle</th>
<th>RACHS-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cases coded (%)</td>
<td>98</td>
<td>92</td>
</tr>
<tr>
<td>Mean score</td>
<td>7.9</td>
<td>2.4</td>
</tr>
<tr>
<td>Median score</td>
<td>8.0</td>
<td>3</td>
</tr>
<tr>
<td>( P )-value</td>
<td>0.03</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Wald test</td>
<td>4.8</td>
<td>17.7</td>
</tr>
</tbody>
</table>

Fig. 1 displays the observed mortality for both Aristotle and RACHS-1 systems. Mortality is seen to rise monotonically for the RACHS-1 system, whereas the Aristotle Score is much less clearly related to mortality.

Our institutional performance was calculated as 7.88 (mean complexity) × 0.954 (mean survival) = 7.52 units.

4. Discussion

This study identified only a weak association between the Basic Aristotle Score and postoperative mortality in our series. This contrasts with the RACHS-1 method, which we found to be strongly predictive of mortality using the same dataset.

Al-Radi et al. [5] performed a similar comparative validation study of the Aristotle and RACHS-1 systems. They also found that the RACHS-1 method more consistently represented the probability of hospital death in a much larger series of over 13,000 operations spanning 22 years. However, they felt that neither system in isolation was adequate for risk adjustment to compare institutions.

We acknowledge the fact that the Aristotle scoring system takes into account factors other than expected mortality alone. By incorporating anticipated morbidity and technical difficulty as well, the score provides a subjective rating system for a wide range of procedures. The combination of these three elements has been termed ‘complexity’.

Our data confirm that some procedures, which we subjectively agree to have greater overall complexity, such as the arterial switch operation, in fact have a relatively low mortality. Conversely, some operations with subjectively less technical difficulty and overall complexity, such as bidirectional cavopulmonary shunt/Norwood stage II, had a higher mortality.

Therefore, the Aristotle scoring system may still have merit as a subjective estimate of the overall difficulty of a given procedure. This measure of difficulty, or complexity, however, has not been examined scientifically and validated statistically as a reliable tool for comparison of institutional outcomes. It may be that it is not possible to scientifically validate such a system, which is founded on subjective probability. This in fact was the original premise of the
Aristotle method, which argued that 'where no scientific answer is available, the opinion perceived and admitted by the majority shall have the value of truth'.

Nevertheless, there may be a danger when applying such a scoring system to identify poor performance. The Aristotle Committee have defined operative performance using the equation:

\[
\text{Operative performance} = \text{complexity} \times \text{survival}
\]

Comparing our performance to the data from the 26 European centres that participated in the EACTS congenital database for the purposes of the Aristotle Committee report [1], our results appear to be quite favourable (performance 7.5 vs 6.3 ± 0.4).

In 2005, the Society for Thoracic Surgeons published results from 16 U.S. centres for the period 1998–2001 [6]. In that report, mean complexity was 7.2 units, mean mortality 4.2% and mean performance 6.90 units (range 5.49–8.60). Again our own institution’s results would appear favourable to these, with only one U.S. centre having superior performance.

Are these comparisons fair?

It is important to recognise that this definition of performance is not the same as risk-adjusted survival. Consider the following hypothetical example: three surgeons are to have their performance measured. Surgeon A has an average complexity score of 5, and has a mortality of 5%. Surgeon B has an average complexity score of 10, and a mortality of 10%. Surgeon C has an average score complexity of 20, and a mortality of 20%.

Their ‘performance’ is now calculated:

- **Surgeon A**
  \[
  \text{Performance} = \text{complexity (5)} \times \text{survival (0.95)} = 4.75
  \]

- **Surgeon B**
  \[
  \text{Performance} = \text{complexity (10)} \times \text{survival (0.90)} = 9.00
  \]

- **Surgeon C**
  \[
  \text{Performance} = \text{complexity (20)} \times \text{survival (0.80)} = 16.00
  \]

Even though all three surgeons have mortality rates which are exactly proportional to the level of complexity of their cases, their ‘performance’ is not the same. The performance rating is greatly in favour of the surgeon undertaking the cases with the highest complexity score. Using this system benefits centres which receive referrals for more complex surgery, whilst penalising other centres which undertake more routine surgery, even if the results achieved are equivalent by standard risk-adjustment methods.

Our own institutional performance may therefore appear inflated since our average complexity was significantly higher than the other 26 European centres (7.9 vs 6.7) as well as the 16 U.S. centres (7.9 vs 7.2). It should be noted also that closed (non-bypass) cases have been excluded from our series, which may have affected these results.

An alternative method for calculating performance might be:

\[
\text{Performance} = \frac{\text{complexity}}{\text{mortality}}
\]

Using this equation, the three hypothetical surgeons in the abovementioned example would each have the same performance. In this way, the definition of performance becomes ‘complexity-adjusted survival’. This is a more intuitive method of calculating performance and does not favour providers with an overall higher complexity casemix.

5. Conclusion

Despite the comparisons made of the Aristotle and RACHS-1 systems, the present study is not intended to endorse one system over the other. Rather, we observe that the two systems are quite different in their design and application. We therefore believe that caution should be exercised when interpreting the results of a risk-adjusted analysis using either method alone.

We believe that the appropriate application of any risk stratification system is to audit the results of large series, identify overall trends in outcomes, and allow fair comparison of results amongst institutions. When divergent performance is recognised using such a method, it should serve only as a trigger for closer examination. More specific, detailed and comprehensive analysis using predetermined hypothesis testing is then required to draw valid conclusions.

Acknowledgements

Research at the Institute of Child Health and Great Ormond Street Hospital for Children NHS Trust benefits from Research and Development funding received from the NHS Executive.

References


