A technical review of the United Kingdom National Adult Cardiac Surgery Governance Analysis 2008–11

Graeme L. Hickeyab, Rebecca Cosgriffb, Stuart W. Grantbc, Graham Cooperd, John Deanfielde, James Roxburghf and Ben Bridgewaterf,gh

a University of Manchester, Manchester Academic Health Science Centre, The Northwest Institute for Bio-Health Informatics, Manchester, UK
b University College London, National Institute for Cardiovascular Outcomes Research (NICOR), The Institute of Cardiovascular Science, London, UK
c Department of Cardiothoracic Surgery, University of Manchester, Manchester Academic Health Science Centre, University Hospital of South Manchester, Manchester, UK
d Department of Cardiothoracic Surgery, Northern General Hospital, Sheffield, UK
e Department of Cardiothoracic Surgery, St Thomas’ Hospital, London, UK
f Department of Cardiothoracic Surgery, Northern General Hospital, Sheffield, UK
gh Corresponding author. University of Manchester, Manchester Academic Health Science Centre, University Hospital of South Manchester, Southmoor Road, Manchester M23 9LT, UK. Tel: +44-161-2912511; e-mail: ben.bridgewater@uhs.nhs.uk (B. Bridgewater).

Received 24 May 2013; received in revised form 5 August 2013; accepted 21 August 2013

Summary

The Society for Cardiothoracic Surgery in Great Britain and Ireland (SCTS) has published named mortality data since 2001. The importance of accurate and robust clinical outcome reporting has been emphasized by a number of high-profile cases in England. In this article, we give a technical review of the United Kingdom National Adult Cardiac Surgery Governance Analysis 2008–11. The statistical and analytical assumptions and methods are discussed in order to add an additional layer of transparency to the clinical governance process and precipitate scrutiny with the aim of optimizing future analyses.

Keywords: Risk-adjustment • Registry data • Performance indicators • Healthcare regulation • Outliers

INTRODUCTION

The Society for Cardiothoracic Surgery in Great Britain and Ireland (SCTS) has published named mortality data since 2001 and by consultant surgeon since 2005 [1–3]. This transparency in reporting clinical outcomes was prompted by the public inquiry into paediatric cardiac surgery at the Bristol Royal Infirmary Inquiry [4]. Publication of clinical outcome data has recently become a mandatory requirement for nine surgical specialties and interventional cardiology in England and Wales from 2013 [5].

The importance of accurate and robust clinical outcome reporting has recently been emphasized by the controversy surrounding the provisional publication of the paediatric cardiac surgery governance analysis for England 2009–12. This publication led the Medical Director of the National Health Service (NHS) to order the immediate closure of the Leeds unit with the underpinning analysis concluded as being a contributing factor to this decision [6]. There have been numerous other high-profile cases such as these in the field of cardiac surgery; notable cases include Winnipeg, Canada and Oxford, UK [7, 8].

The benefits of analysis and public reporting of outcomes data are thought to have improved overall quality and the ability of patients to choose providers if they wish [9]. It has also been suggested that transparency can facilitate the early identification of units with suboptimal quality and that the process of ‘openness’ may help to maintain public trust in the profession and healthcare providers [10]. Potential downsides are the possibility that higher risk patients may be denied surgery because of governance processes, the risk of incorrectly identifying providers with high mortality rates and the associated adverse consequences. A full consideration of all of these issues is beyond the scope of this review; however, they are covered elsewhere [3, 11].

In this article, we give a technical review of the United Kingdom National Adult Cardiac Surgery Governance Analysis 2008–11, which was published online (http://www.scts.org/patients/default.aspx) in March 2013. The objectives of this review are 3-fold: (i) to add an additional layer of transparency to the clinical governance process; (ii) precipitate scrutiny with the aim of optimizing future analyses and (iii) disseminate the methodology for potential reuse by other professional societies and regulators.

MATERIALS AND METHODS

Data extraction and cleaning

A complete extract of prospectively collected data by the SCTS in Great Britain and Ireland for all adult cardiac surgery procedures recorded in the UK was obtained from the National Institute of Cardiovascular Outcomes Research (NICOR) central cardiac database on 20 March 2012. Definitions of database variables used for the study are available at http://www.ucl.ac.uk/nicor/audits/Adultcardiacsurgery/datasets.
The flow of the data from surgeon-input to analysis has been described elsewhere [12]. Briefly, data entered locally by surgeons are validated at the unit-level by database managers prior to upload via a web-portal to NICOR. At this stage, further validation is performed according to logical rules and missing data reports are generated for primary variables (e.g. EuroSCORE risk factors, patient identifiers and outcome data). The data are then forwarded to an academic healthcare informatics department for data cleaning.

The complete data cleaning process has been previously described [12]. Briefly, duplicate records are removed, transcriptional discrepancies recoded and clinical and temporal conflicts resolved. The data cleaning is performed by the analyst responsible for the governance analysis in collaboration with surgeons and the audit manager. All cleaning is made reproducible by programming a series of scripts, which are updated following each new data extract. At this stage, and prior to analysis, data for the last 3 years are returned to each contributing hospital for local validation, and units update their records in the central registry repository where necessary. The final extract used for this clinical governance analysis was made from the National Adult Cardiac Surgery Audit (NACSA) database on 12 October 2012.

Exclusion criteria

The clinical governance analysis was for the period 1 April 2008 to 31 March 2011; all other records were excluded. Records identified as being within-admission reoperations were removed, leaving only the original procedure for each admission spell. This is to ensure that only the primary procedure for each admission spell is included and prevent outcomes being counted more than once for each patient. Should a patient be reoperated on by a different consultant surgeon in the same admission, the outcome is still associated with the consultant responsible for the initial procedure. Transplants, primary ventricular assist device insertions and trauma cases were excluded from the analysis as they are analysed within separate national audit programmes.

Records corresponding to emergency or salvage surgery, defined as any unscheduled patient with on-going refractory cardiac compromise with no delay in surgery irrespective of the time of day (SCTS defined emergency surgery) or any patient requiring cardiopulmonary resuscitation en route to the operating theatre or prior to the induction of anaesthesia (SCTS defined salvage surgery), were excluded but reported internally. This is because available risk models have been found to perform inadequately for clinical governance purposes in this cohort of patients [13]. Private hospitals, i.e. those that operate outside of the NHS, were also excluded from the primary analysis since most data are not collected from these units. We have conducted a separate analysis of the available private hospital data against the NHS hospital framework, but the data from that analysis are not included in the results of this manuscript. It should be noted that in the UK, there are a mixture of NHS hospitals, which are the majority, and some private hospitals. Some NHS hospitals also provide surgery to private patients, but usually in relatively small numbers. For the purposes of the governance analysis, we have primarily compared the outcomes across the NHS hospitals (including both NHS and private patients).

Some consultants contributed a small number of records over the study period for reasons including, but not limited to: retirement; newly appointed consultant or they predominantly perform thoracic or congenital surgery. We have made the decision that it is unfair to report on these consultants; therefore, all consultants who performed <30 procedures (equivalent to an average of 10 per annum) were excluded from the analysis. However, their data are retained for risk-adjustment and standard-setting purposes.

Missing data

Most missing data are resolved during the validation stages of the data transfer. During this process, some registry-specific nuances become apparent. For example, pulmonary artery pressure should be recorded in the registry as a continuous value. However, many units instead record a nominal value considered to be normal (e.g. 0 or 1).

The SCTS has an established policy for the handling of missing data. First, missing and conflicting data for in-hospital mortality status are backfilled and validated via record linkage to the Office for National Statistics (ONS) census database, which records details of all deaths in England and Wales. After all reasonable attempts to backfill these data using record linkage or notifying the base, any remaining missing discharge status data are mapped to in-hospital death. ONS tracking data are not available for patients in Scotland and Northern Ireland, which does not enable us to backfill missing data for these patients, but the incidence of missing data for mortality was very low from these countries. For the final analysis dataset after backfilling discharge status data, in Scotland there were 0 (0.00% of Scottish records) missing discharge statuses; in Wales and Northern Ireland, there were 3 missing discharge statuses each (0.06 and 0.11% of Welsh and Northern Irish records, respectively) and for England, there were 23 missing discharge statuses (0.02% of English records).

For records where data required to calculate a EuroSCORE variable were missing, it was assumed that the risk factor was not present, i.e. equivalent to the reference level. In the case of patient age at the time of surgery, an age of <60 is imputed, which attracts the smallest adjustment in the logistic EuroSCORE model [14].

Healthcare provider assignment

The SCTS reports outcomes for two levels of healthcare providers: (i) the base hospital and (ii) the responsible consultant surgeon. Although the consultant may not actually perform the cardiac surgery, they are considered responsible for the clinical team providing the treatment. We think this is a reasonable policy for several reasons. First, patients in the UK usually see a named consultant prior to surgery, and it is that surgeon who explains the risk and benefits of the available treatment options, obtains informed consent, and then carries the overall responsibility for the conduct of surgery and the aftercare. It is fully accepted that the decision-making is supported by the multidisciplinary ‘heart team’ and care is delivered by the various medical and non-medical specialties within the hospital. Secondly, consultant surgeons, and particularly groups of surgeons working together, are important drivers for change in organization care when that is required. In the UK, it has been decided that it is reasonable to report at individual surgeon level to drive quality, increase accountability and support governance and patient choice.

Each record in the cardiac surgery registry contains a hospital identifier code and a consultant General Medical Council (GMC) number. Records where data other than a GMC number have been inputted, for example, a name or initials, are mapped to a valid number by accessing the public GMC register (http://www.gmc-uk.org/doctors/register/LRMP.asp) and liaising with the base
Outcomes

The outcome for this study was risk-adjusted in-hospital mortality. Ancillary to this is the in-hospital mortality rate, where in-hospital mortality was defined as death due to any cause during the base hospital admission for cardiac surgery. We follow the indirect risk-adjustment methodology as described by Hannan et al. [15] and Racz and Sedransk [16]. Briefly, for patient i in hospital j, we define \( y_{ij} \) as 1 if the patient died in-hospital, and 0 otherwise. Also, we define \( \hat{p}_{ij} \) as the predicted mortality (a probability value between 0 and 1) estimated using appropriate methodology (see below). Letting \( n_j \) denote the number of patients treated in healthcare provider \( j \), the expected mortality rate (EMR\(_j\)) and observed mortality rate (OMR\(_j\)) for hospital \( j \) are respectively calculated as:

\[
\text{EMR}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} \hat{p}_{ij}
\]

\[
\text{OMR}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} y_{ij}
\]

With these, we define the risk-adjusted mortality rate (RAMR\(_j\)) for healthcare provider \( j \) as:

\[
\text{RAMR}_j = \frac{\text{OMR}_j}{\text{EMR}_j} \times \theta
\]

where \( \theta \) is the mean population mortality rate for patients undergoing cardiac surgery during the study period.

Risk-adjustment and model validation

The logistic EuroSCORE is known to greatly over-predict the mortality rate in the UK population [17]. We continue to verify this status through model assessment (see below). A number of different risk-adjustment methods were considered for the 2008-11 analysis, including:

(i) Refitted logistic EuroSCORE: a multiple logistic regression model between the outcome variable and the EuroSCORE variables as determined for the 2008-11 data.

(ii) Refitted SCTS modified EuroSCORE: the modified EuroSCORE was originally developed by the SCTS for the 2004-07 governance analysis by fitting a multiple logistic regression model between the outcome variable and four operation variables (coronary artery bypass graft (CABG), tricuspid valve, mitral valve and other cardiothoracic surgery), offset by the precalculated logistic EuroSCORE [18, 19]. The approach was replicated with the 2008-11 data.

(iii) Recalibrated logistic EuroSCORE: a multiple logistic regression model between the outcome variable and the log-odds (‘logit’) transformed EuroSCORE [20].

For approach 3, the linearity assumption was assessed by first categorizing the logit-transformed EuroSCORE into 10 equally sized groups (‘deciles’). A logistic regression model is then fitted between the outcome variable and the categorical variable. The mean value of the continuous dependent variable within each category is plotted against the estimated model coefficient. Departures from a straight-line indicate the need to transform the dependent variable [21]. An iterative process was implemented until an adequate transformation was reached.

For any given model, a ‘fitted’ value, \( \hat{p} \), was calculated by plugging-in the appropriate corresponding covariates into the model with the estimated model coefficients. In all cases, this is equivalent to calculating the predicted value for a logistic regression model.

Model validation was assessed by evaluating the goodness-of-fit using the Hosmer-Lemeshow test and calibration plots, and discrimination by calculating the area under the receiver operating characteristic curve (AUROC).

The Hosmer-Lemeshow test enables detection of departures from calibration within different groups. Contributing \( \chi^2 \) statistics from each decile group were summed. To test the hypothesis of whether the model predictions match the observed probabilities within the subgroups, the overall test statistic was evaluated with a \( \chi^2 \) distribution. For each approach (which is an internal model validation), the null distribution is assumed to have 8 degrees of freedom; for analysing the logistic EuroSCORE (which is an external model validation), the null distribution is assumed to have 10 degrees of freedom [21].

Calibration plots are a type of graphical representation of the Hosmer–Lemeshow test. The calibration plots presented show for each model the mean predicted probability of outcomes against the observed proportion of outcomes for 10 approximate equally sized groups based on the ranked predicted risks calculated according to the models. Overlaid is the line of equality, which would indicate a model with perfect calibration, and a non-parametric locally weighted scatter-plot smoothing (LOESS) curve for each model showing the general relationship [22]. Approximate 95% confidence intervals (CI) for the observed mortality proportions are shown as error bars, estimated by the Agresti-Coull approximation [23].

The AUROC will typically range from 0.5, indicating that the model performs no better than chance, to 1.0. The threshold for acceptable model performance is generally considered to be >0.7 and >0.8 for excellent discrimination [21]. Standard error estimates for the AUROC according to Hanley and McNeil [24] are used to calculate approximate 95% CI. A graphical representation of the function that maps the logistic EuroSCORE to the recalibrated value is presented and AUROC-invariance property investigated.

All risk-adjustment approaches and validations were performed using the R (version 2.15.2) statistical computing software (http://www.R-project.org/) [25]. The AUROC and standard error was calculated using the Hmisc package [26]. A P-value of <0.05 was considered statistically significant.

Defining divergence

The SCTS outlined its thresholds for defining divergence of performance in Maintaining patients’ trust: modern medical professionalism [27]. Briefly, the target standard to which each healthcare provider’s average in-hospital rate is the mean population mortality rate, \( \hat{p} \). It is expected that there will be variation in this standard due to within-unit sampling variability. The underpinning statistical model is a binomial distribution with \( n \) trials and standard proportion, \( \theta \). With such a model, the standard graphical descriptor is the funnel plot [28]. Other graphical descriptors such as the
caterpillar plot have been criticized due to the implicit league-
table structure.

For hospital-level reporting, two threshold limits (‘funnels’) were
constructed to account for varying degrees of sampling variation:
(i) a two-sided 95% control limits and (ii) two-sided 99% control
limit. For surgeon-level reporting a further threshold was included:
(iii) a one-sided 95% control limit corrected for multiple compar-
isons. This allows simultaneous inferences about the healthcare
providers to be made, whereas thresholds 1 and 2 are limited to
individual hypothesis tests.

In all cases, a normal approximation to the exact binomial limits
was used. That is, for a given sample size \( n \), the 100 \((1 - \alpha)\)%
control limit is given by \( \theta \pm Z_{\alpha/2} \sqrt{\theta(1 - \theta)/n} \), where \( Z_{\alpha/2} \) is the
100\((1 - \alpha/2)\)-th percentile of the normal distribution. For thresh-
olds 1 and 2, \( \alpha \) equals 0.05 and 0.01, respectively. To correct for
multiple comparisons, the Bonferroni correction is made. Without
applying the Bonferroni correction, a one-sided 95% control limit
at sample size \( n \) would be \( \theta + Z_{0.05} \sqrt{\theta(1 - \theta)/n} \). Applying the ad-
justment to the \( m \) consultants included in the analysis simply
yields \( \theta + Z_{0.05/m} \sqrt{\theta(1 - \theta)/n} \).

It is not unusual for analysis of performance indicators to
show larger variability in outcomes than would be expected by
chance alone. This phenomenon, known as over-dispersion,
can lead to an excess of units being classi-

There were sufficient data to calculate 96 083 (91.11%) logistic
EuroSCORE values. Missing data were <1% for all variables except
neurological dysfunction (1.46%); creatinine >200 \( \mu \text{mol/l} \) (1.80%);
left ventricular function (2.39%); active infective endocarditis
(2.12%) and recent myocardial infarction (1.28%). For 9374 (8.89%)
records with at least 1 missing EuroSCORE data, we imputed the
missing values to the baseline reference. Despite the EuroSCORE
being missing in \(~9\%\) of records overall, most records only had 1
(5.87%) or 2 (2.51%) missing EuroSCORE variables, and only 0.51%
of records had \( \geq 3 \) missing EuroSCORE variables.

Prior to recalibration of the logistic EuroSCORE, assessment of
the linearity between the logit-transformed EuroSCORE and
outcome variable provided evidence of a polynomial relationship

\[
Z_\text{if} = \frac{\text{RAMR}_j - \theta}{\sqrt{\theta(1 - \theta)/n}} \quad \text{for each unit } j
\]

Results

A total of 111 132 records were extracted from the NACSA data-
base from the period 1 April 2008 to 31 March 2011. Of these,
641 were excluded as they corresponded to within admission reo-
perations. A further 154 records were excluded as corresponding
to transplantations, trauma surgery or primary ventricular assist
device insertion. Excluding data from four contributing private
hospitals reduced the number of records by 1538. Of the remain-
ing 108 799 records, the distribution of operative urgency was:
missing \( \text{(n} = 237 \text{retained for analysis)} \); elective \( \text{(n} = 75 241 \text{)} \); 
urgent \( \text{(n} = 29 979 \text{)} \); emergency \( \text{(n} = 3036 \text{)} \) and salvage \( \text{(n} = 306 \text{)} \).
The 3342 emergency and salvage surgery records were excluded
from the analysis, leaving 105 457 records for governance pur-
poses. A total of 37 NHS hospitals and 294 consultant surgeons
were included. There were 20 records with missing surgeon id-
ifiers spanning 10 hospitals. These were retained for model fitting
and standard setting.

*Results*

Figure 1: Assessment of the linearity between the logit-transformed outcome
and logit-transformed logistic EuroSCORE. Each point corresponds to the mean
logit EuroSCORE with a particular decile and the estimated regression coefficient
for that category. The curvature in the points implies a polynomial relationship.

Figure 2: Figure showing the relationship between the logistic EuroSCORE and
the recalibrated logistic EuroSCORE (red line). The black line indicates equality.
A cubic polynomial was found to satisfy the assumption and provide a good model fit. The final recalibrated model is

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = -1.4916 + 0.5580q - 0.0269q^2 + 0.0164q^3$$

where $q$ is the logit-transformed EuroSCORE and $\hat{p}$ is the recalibrated predicted mortality. One can apply the inverse logit-function to extract $\hat{p}$. Figure 2 describes the relationship between $q$ and $\hat{p}$.

Figure 3 shows calibration plots for: (i) the logistic EuroSCORE; (ii) refitted logistic EuroSCORE; (iii) refitted modified EuroSCORE and (iv) recalibrated EuroSCORE. As expected, the logistic EuroSCORE greatly over-predicted the observed mortality rate ($\chi^2 = 2795.7; P < 0.001$). The refitted logistic EuroSCORE and refitted modified EuroSCORE were a considerable improvement ($\chi^2 = 40.8; P < 0.001$ and $\chi^2 = 91.1; P<0.001$, respectively), but there was evidence of miscalibration in some deciles.

The recalibrated EuroSCORE (as defined above) was well-calibrated ($\chi^2 = 6.9; P = 0.545$).

The models showed a similar level of discrimination. The refitted modified EuroSCORE had the highest AUROC (0.791; 95% CI: 0.783–0.799), followed by the refitted logistic EuroSCORE (0.789; 95% CI: 0.781–0.797) and then the logistic EuroSCORE and recalibrated EuroSCORE (both 0.783; 95% CI: 0.775–0.791). The invariance between the AUROC for the logistic EuroSCORE and recalibrated EuroSCORE is a mathematical property of the recalibration formula. Simply, for any two logistic EuroSCORE values, $q_1 > q_2$, as the recalibration formula is monotonically increasing (which can be informally gauged from Fig. 2), it means the predicted values, $\hat{p}_1$ and $\hat{p}_2$, will also satisfy $\hat{p}_1 > \hat{p}_2$. This fact preserves the AUROC and can be formally proved by calculus (not shown here).

The mean in-hospital mortality was 2.74% and was used to set the standard, $\theta$, for measuring divergence from in the analysis. For the 37 hospitals analysed, the over-dispersion factor was $\hat{\varphi} = 4.02$, which was greater than the critical value of 1.47. For the 275 consultants with a sufficient number of cases for inclusion, the
over-dispersion factor was 1.614, which was greater than the critical value of 1.17. Figures 4 and 5 show the published funnel plots for the 2008–11 governance analysis.

DISCUSSION

The SCTS has published the results of its governance analysis for all elective and urgent adult cardiac surgery procedures between April 2008 and March 2011. These can be accessed through a web-portal (http://www.scts.org/patients/default.aspx). From the inception of this recent governance programme to the publication of the results are a series of analytical decisions that needed to be taken. Statisticians and healthcare analysts might consider some decisions to have limitations. However, there are pros and cons to each decision made, which are discussed here.

Outcomes

To monitor outcomes, we have used risk-adjusted in-hospital mortality as our measure of choice as it is undoubtedly an important outcome from the patients’ perspective. From a statistical viewpoint, it occurs with sufficient frequency in our specialty to be useful to discriminate between providers of care [30]. Other measures of mortality are potentially available such as risk-adjusted 30- or 90-day mortality (which is thought to be a better measure overall as it picks up the majority of surgical related deaths), but these are more difficult to measure requiring post-discharge patient tracking. Owing to the purpose of this analysis being to compare performance of healthcare providers, we think that these subtleties of methodology are not vital in identifying potential problems, as long as the chosen measure is applied consistently across providers [31, 32]. For relatively lower-risk procedures, patients are becoming more interested in other outcomes such as length-of-stay and postoperative complications that lead to morbidity. The counterpart to the SCTS in the USA, The Society of Thoracic Surgeons (STS), now reports composite outcome measures based on the domains mortality, morbidity and evidence-based process measures, which are combined into a ‘star rating’ [33, 34]. Although these measures are very attractive, as a society whose responsibility is to its members, we have tried to avoid summarizing complex and multidimensional data into a single metric.

In cardiac surgery, we have a series of potentially definable subgroups, such as CABG, aortic valve replacement (AVR) surgery, combined CABG and valve surgery. Arguably, patients will be more interested in the performance of healthcare providers for the specific operation type they are to receive. Despite this, we have decided to use ‘all’ cardiac surgery (with the exception of emergency and salvage surgery) as the group for governance analyses due to simplicity. Internally, however, we do monitor two subgroups: isolated first-time CABG and isolated first-time AVR.

Units

We have decided to monitor hospitals and consultants separately. However, not all consultants have worked consistently in one single hospital due to relocation or multiple units coming under the control of a single NHS trust. During the study period, of the 294 consultants contributing data, 18 of them (6.12%) have worked in >1 NHS hospital. If we treated the units to be measured as consultants within hospitals, then we would need to either make some assumptions or adjust for the correlation between some units. In fact, we have already made a rather strong assumption in the risk-adjustment: that each patient is independent. Hierarchical modeling has been advocated as a more appropriate method of developing a risk-adjustment since patients can be nested within healthcare providers (consultants or hospitals). More sophisticated hierarchical models could account for crossed-effects, namely patients treated by consultants within hospitals, and even account for consultants practicing in >1 hospital [35].

It should be recognized that the consultant-level analyses are a proxy for a healthcare team. It is not always the case that a consultant performs the cardiac operation on every patient; a trainee of associate-grade surgeon might perform the surgery, with or without the supervision of the responsible consultant. Incorporating this structure into the analysis would only add unwanted complexity.

Missing data

As with all large-scale registries, missing data are common [36]. We work with units to ensure that data are complete to a high degree, particularly for data used in historical risk-adjustment.
models. In addition, we use record linkage to minimize missing data on outcome variables. Despite our best efforts, it is not possible to completely avoid missing data. Removing records with missing data would be statistically flawed and contrary to the aims of the governance analysis.

Our approach is somewhat contentious: missing independent data are mapped to a value that provides the smallest possible (if any) adjustment for risk and outcome data are mapped to the worst-case scenario. However, our approach serves the purpose of minimizing the number of missing data by actively encouraging engagement in the governance programme by individual units. For many variables, the approach is not unreasonable based on expert understanding of the data collection process from procedure to upload, discussion with database managers and regular validation exercises. However, on the positive side, we also publish lists of data completeness by unit through our online national report (http://bluebook.scts.org/), which allows units who perform well to take opportunities to celebrate their success. We would also add here that any analysis that leads to identification of high mortality rates should automatically be followed by an examination of local factors, starting with data quality; it is very possible that apparently high risk-adjusted mortality rates can be due to missing risk factor data, and we give units the opportunity to discover and resolve these issues during our validation processes.

There is a vast methodological literature on proposals for handling missing data. The standard advocated approach for this type of missing data problem is multiple imputation [36]. Multiple imputation uses iterated sequential conditional modelling to generate multiple complete datasets. The analyses are combined using a sophisticated methodological framework. In a general context, the superiority of this method to our default approach has been demonstrated [37].

As noted, we have focused our efforts on variables that go into historical risk-adjustment models. This allows us to focus limited resources on variables that have been important to previous governance analyses. Moving towards a new risk-adjustment model that incorporates previously unanalysed variables will likely be met with an increase in the number of missing data. This would lead to delays in analyses (and subsequent dissemination of results) while individual units are notified and allowed time to update their data. Additionally, there is a substantial effort required to augment and amend the structure of clinical registries to collect variables that are not already collected. For this reason, model design must be planned in advance, utilizing the most contemporary data on variables of importance in cardiac surgery prognostic modelling, and we know from recent studies that there is a need to reconsider some of the variables included in future [38].

Similarly, we know that some variables, such as creatinine as a measure of renal failure, are included as binary variables in the EuroSCORE, but models should be more discriminating by inclusion of continuous data, which are unfortunately not currently available to us for these analyses [39].

**Risk-adjustment**

The risk-adjustment methodology, used to adjust for the diverse case-mix, needs to be sophisticated to represent contemporary outcomes or false reassurance may be derived [17]. Monitoring against a continually moving standard (because quality is improving) has made our methods more complex requiring a contemporary recalibration of the original EuroSCORE model [14]. Moreover, inadequate risk-adjustment can lead to false assurances being made, triggering of false alerts, or provoke risk-averse behaviour [40, 41]. For this reason, risk-adjustment is considered the most important aspect of the governance analysis from both statistical and political viewpoints. We have already made the decision to exclude emergency surgery from the governance analyses on the grounds of risk-adjustment being ineffective in this cohort [13]. It should be emphasized that the development of the model here was for adjustment purposes, i.e. to allow fair comparison between healthcare providers, and not as a prediction model for patient-level use.

We opted to use a logistic regression model based on a transformation of the ubiquitous logistic EuroSCORE. This is effectively an extension of a model calibration assessment method [20]. The advantage of this is that a simple formula, easily interpretable by individual units, is derived for a variable that inherently adjusts for preoperative risk factors, thus having ‘face value’, while also having a low rate of missingness. Importantly, however, the model was well calibrated and the *a priori* discrimination preserved. A disadvantage is that this approach does not account for the varying contemporaneous adjustments each risk factor might have. Other methods, including complete model refitting were examined and did not have any great effect on inferences.

There are a number of considerations we did not focus on in this analysis. First, EuroSCORE II has since replaced the logistic EuroSCORE [42]. EuroSCORE II is more contemporary than its predecessor; however, it was developed for prospective data collected in 2010–2 years after the study window began. Furthermore, the SCTS does not currently collect all of the EuroSCORE II variables, and others were only collected after 2010 [43]. Secondly, hierarchical (also known as ‘random effects’) modelling is a popular approach for risk-adjustment. Notwithstanding the added complexity, Hannan et al. [15] demonstrated in the context of a large state-wide governance analysis that performance was similar. Thirdly, it has been demonstrated that cardiac surgery prediction models have been subject to systematic calibration drift with time. This naturally calls for the application of dynamic models [44]. The SCTS plans to review all of these approaches for future governance analyses.

The recent publication of the paediatric cardiac surgery governance analysis in the UK recalibrated the associated risk-adjustment model by multiplying the predicted mortality by the national OMREMR [45]. This follows the same methodology applied by the STS (cf. the ‘calibration factor’) [46]. Notwithstanding the widespread use of this recalibration methodology, Jin et al. [46] notes a fundamental technical problem with the method. Namely that predicted mortality values of >100% can be obtained and different inferences yielded for the complement of the outcome, namely discharged alive.

**Definitions of divergence**

The SCTS previously established criteria for defining divergence [27]. We translated the SCTS definitions of divergence using a normal approximation based on the observed mortality proportion. There are two issues with this.

First, the control limits overlaid for the funnel plots of risk-adjusted mortality rates against case volume treated the risk-adjusted mortality as an observed proportion, as has been done elsewhere [18, 40, 47–49]. Strictly speaking, this is incorrect as the variance will depend not only on θ and n, but also on the EMR. In principle, to overcome this, case volume can be replaced by ‘adjusted case volume’ (that depends on θ, n and EMR for each unit); however, this approach has
not been studied so far, and would produce funnel plots that would be complicated to interpret. Alternatively, one could replace case volume with expected number of outcomes, or use separate funnel limits for each healthcare provider. As it stands the funnel control limits applied in Figs 4 and 5 should be viewed as being mapped from the analysis on crude mortality, where they still remain correct.

Secondly, the normal approximation to the binomial distribution is asymptotic; for very small sample sizes, it will be inappropriate. Over the 3-year study period, each hospital has an appreciably large sample size, and so the approximation is not considered to be an issue. Similarly, for individual consultants, the fact that we excluded those with very small numbers of procedures means that the assumption should not be unreasonable. Of the 275 consultants included in the final analysis, 85% had performed ≥182 operations—the minimum sample size required to satisfy the ‘rule of thumb’ that $n \times \text{OMR} \geq 5$; the median sample size was $n = 394$. In future analyses, we intend to consider using exact limits; however, this is complicated by the issue of over-dispersion. Alternatively, one might consider using a different normal-based approximation, for example based on log-transformed Z-scores [50].

Over-dispersion was also accounted for by inflating the standard deviations of the control limits to avoid inappropriate classification of units. It is important that attempts to understand over-dispersion are made. First, the choice of indicator may be partially accountable due to decreasing numbers of outcomes; a 30- or 90-day perioperative mortality rate might bring the system ‘in control’ as it will avoid the dependence on separate hospital discharge policies. For example, a hospital may appear to perform better based on in-hospital mortality, but this may be an artefact of a systematic policy to discharge patients after a fixed number of days. Secondly, comparing all cardiac procedures is equivalent to ‘comparing apples and oranges’, which necessitates further consideration to separate analyses and risk-adjustment within major procedure groups. Thirdly, the risk-adjustment model may require improvement. Our approach was based on the EuroSCORE, which is being a preferred method for handling over-dispersion, and this is being evaluated for future governance analyses [52]. We implemented a means of evaluating multiple hypotheses of consultants having unusually high RAMRs by including a one-sided control limit adjusted for multiple comparisons. The Bonferroni adjustment is known to be conservative; however, it is probably the simplest approach available. Other statistical approaches for controlling the family-wise error rate (the probability of making ≥1 type I errors) have been proposed, and will be considered for future studies [53].

Other limitations

In addition to the various technical limitations described above, we would also acknowledge that, by the very nature of a hospital-based operative mortality analysis we have looked at a relatively narrow time window of patient care, i.e. in-hospital care, rather than the wider process of care for that particular disease or procedure, including preoperative and post-discharge management. We have also looked purely at results from operations actually undertaken, due to a lack of data on those patients referred for, but not ultimately undergoing surgery.

CONCLUSIONS

In the light of the alternatives, our governance programme is a delicate balance of pragmatism and statistical robustness, which we believe should always remain as simple as analytically tolerable to ensure that it is understood by clinicians and the public. The decisions have been made; the results have been published. The purpose of this article is to disseminate our analysis blueprints to the wider cardiac surgery community, allowing for open critiquing in order to drive improvements for future governance analyses. The implications of a national cardiac surgery analysis should not be understated. The results determine the revalidation of surgeons in the UK and play a key role in the decision-making process when a unit is faced with closure. Nonetheless, it should be emphasized that classifying healthcare providers as unusual, or ‘outliers’ as described by some, does not imply poor practice. Results may also be artefacts of data quality, unusual case-mix or other aberrancies. However, it allows medical institutes, regulators and professional societies, who have limited resources, to focus their attention to a small number of units who might benefit from closer inspection and it is only after local consideration of all issues that any judgement should be made about ‘quality’. Governance analysis has to remain accurate and transparent in order to avoid unintended negative consequences [41].

ACKNOWLEDGEMENTS

The authors thank all members of the Society for Cardiothoracic Surgery in Great Britain and Ireland who contribute data to the SCTS database. The National Institute for Cardiovascular Outcomes Research, UCL London, provided the data for this study. The National Adult Cardiac Surgery Audit is commissioned and funded by the Healthcare Quality Improvement Partnership (HQIP).

Funding

We are grateful to Heart Research UK [Heart Research UK grant RG2583] for funding the analytical part of the governance analysis.

Conflict of interest: none declared.

REFERENCES