Perspective

Improving cardiac surgical practice through outcomes analysis

Ben Bridgewater¹,∗ and Bruce Keogh²

INTRODUCTION

Over recent years there has been an increasing focus on the measurement of clinical outcomes in cardiac surgery. Data collection at local level is as old as cardiac surgery itself but there has been a gradual introduction of standardised datasets and national/international data collection systems which have led to IT based feedback systems and major publications of trends and outcomes [1–3]. Systems that have actively embraced data collection, analysis and feedback seem to have achieved excellent results, and possibly better outcomes than systems that have not [4].

A world supported and influenced by IT and social networking is changing, with major implications for all walks of life including the retail sector, sociology, politics and, of course, healthcare. One important aspect of the underlying cultural change is increasing transparency within society which has manifested itself in the UK with major scandals around the expenses of Members of Parliament, privacy laws and illegal phone message interception. A major event which led to changes in healthcare delivery and its regulation in the UK happened a little earlier with the events in Paediatric Cardiac Surgery in Bristol [5], but it was ‘patient power’ and a desire for transparency that led to the Public Inquiry into the happenings and the subsequent recommendations, which include the collection and publication of clinical outcomes. It seems unlikely that the developments in the use of, and external desire for, clinical outcome information will recede and with escalating costs of healthcare internationally, clear justification of clinical effectiveness of procedures will be increasingly necessary for commissioners/payers of healthcare. Cost constraint will by high on the list of objectives for providers of the services.

WHAT OUTCOMES CAN YOU MEASURE?

Cardiac surgery in the UK started to measure clinical outcomes in a structured national format as early as 1977 [6]. The initial program simply collected operation type and crude mortality by unit, to permit pooled national benchmarks to be disseminated and allow individual units to compare their outcomes to that mortality rate [7]. The data also allowed trends in activity and mortality to be tracked over time, giving useful information for clinicians, commissioners, politicians and planners of healthcare services.

This system gave no ability to adjust for case mix, and we have come to learn that operation type, age, gender, left ventricular function and other risk factors are important in determining operative outcomes, and these factors vary significantly between centres. More recently national groups have defined specific datasets for collection [1,2], and numerous models have been published which enable clinical outcomes to be adjusted for risk [8–10]. As the delivery of healthcare improves, the mortality associated with cardiac surgery has fallen, with the overall mortality for isolated coronary artery surgery in the UK now under 2%, which is a low incidence for the purposes of quality control and quality improvement [ Fig. 1]. This has driven interest in the collection of other clinical outcomes such as re-operations for bleeding, major sternal wound infection, prolonged post-operative length of stay and new post-operative renal intervention. It is intuitive that measuring and reducing an outcome like excessive blood loss or re-exploration for bleeding will be associated with better outcomes for patients [ Fig. 2]. An outcome such as new renal intervention is much less clear cut –
some surgeons have suggested this to be a good quality marker, but recent analysis in the UK has shown a marked increase in new renal intervention after cardiac surgery in recent years, with much improved outcomes for those patients [7]. This is likely to be due to changes in clinical philosophy with respect to earlier filtration for these patients, and such this particular outcome has limited place as a quality marker, either for benchmarking or improvement at this stage. Quality markers must be chosen with great care.

It is clear that risk-adjusted mortality has fallen in systems that routinely measure outcomes, but it is probable, but not definitive, that collecting, analysing and feeding back clinical outcomes data drives quality improvement, but the exact mechanisms by which it does so are not entirely understood. At the simplest level it is probably just the ‘Hawthorn effect’ in action [11], in that focussing consideration on an outcome improves that outcome. More sophisticated approaches using methodology to understand natural variance and minimise ‘special cause’ variance to optimise processes may well further improve outcomes [12]. These issues are complex and fall outside the scope of this manuscript.

Another approach to quality improvement and control is to define best practice process measures that are thought to be associated with optimal outcomes, and to measure and drive better compliance with these; this could typically include the use of the left internal mammary artery for coronary surgery and the prescription of appropriate medication at discharge for CABG surgery. This has been at the heart of payment for quality schemes such as those introduced by Premier in the US [14], and ‘Advancing Quality’ [15] in the UK. These groups have introduced ‘payment for performance’ to incentivise improvement from providers. The original work suggested that better compliance with best practice processes was associated with improved outcomes, but we are not aware that these findings have been reproduced in systems where there is greater existing focus on clinical outcomes as a primary marker of quality. However, using a balanced package of outcomes and process measures is probably optimal for improving quality, and there is now increasing interest in combining these measures of clinical quality with those of patient safety and patient experience [16], to define an overall standard of care.

Patient experience is an interesting quality marker in its own right. Recent adverse events at Mid Staffs Hospitals in the UK came to light as much from issues about poor patient experience as unacceptable clinical outcomes [13], and it seems likely that these two domains are linked, at least to

![Crude survival and mortality rates by procedure (n=269,610)](image)

Figure 1. The graph shows improvements in mortality in the UK for all the main operative groups over time. The degree of improvement is marked; between 2001 and 2008 the rates changed from 2.3% to 1.5% for isolated CABG ($\chi^2$-test $p < 0.001$), 2.6% to 1.7% ($\chi^2$-test $p < 0.001$) for all CABG, 5.2% to 3.5% for isolated valves ($\chi^2$-test $p < 0.001$) and 8.3% to 6.1% ($\chi^2$-test $p < 0.001$) for combined valve & CABG. These changes were seen despite increasing risk profiles. Acknowledgement Bridgewater B, Kinsman R, Walton P and Keogh B. Demonstrating quality: The sixth National Adult Cardiac Surgery database report: Publisher Dendrite Clinical Systems Ltd. ISBN 1-903968-23-2. 2009. Ref. [7].
some extent. Important distinction must be made between patient satisfaction, and patient experience measurement, with the latter allowing more granular information to allow generic improvements in service quality across the board and more specific interventions where required. Data on patient experience can come from analysis of complaints data, but it is better in our view to instigate a structured approach to collecting this information on a routine basis, with regular feedback to clinical teams and increasingly sophisticated tools and methods are being develop for this purpose [16].

**HOW DOES MEASURING OUTCOMES DRIVE QUALITY?**

It is becoming increasingly accepted that clinical benchmarking drives quality improvement [17], but the mechanisms by which it does so are far from understood. In the UK we have taken the approach that all hospitals and surgeons have an obligation to know how the outcomes of the care they deliver compares to their peers and that all comparisons must utilise risk adjustment. This is similar to the approach also used, for example, in New York state, by the STS, and in the Northern New England Cardiovascular disease study group [2,18,19]. This last collaboration has been particularly effective in focussing on measurement and improvement of quality, and we have reproduced this methodology on a regional network in northwest England. We have found geographically-based national, regional or sub-regional managed clinical networks are relatively easy (and inexpensive) to establish and they provide a good substrate for quality improvement, both through the use of data, stimulation of reflective practice and dissemination of best practice. The model delegates different responsibilities to the various players — at national (or regional) level the data needs to be collated, risk adjusted, analysed and packaged in a way that will stimulate improvements overall, and more significant improvements where necessary for outliers. At local level there is responsibility for ongoing benchmarking of outcomes. Detection of suboptimal results would typically lead to a process involving data validation to check the signal is valid, diagnostic work to understand the cause of the poor results, and an action plan to resolve the problem. The action plan may involve defining best practice for that particular issue and possibly use of external clinicians or specific experts to help refine care.

In the UK we run a system of national feedback and surveillance whereby results can be reviewed continuously using an online software package, and intermittent governance analyses are undertaken.

![Figure 2](image-url)
to look for abnormal outcomes [16]. Recent iterations have used risk adjustment with a contemporary recalibration of the logistic EuroSCORE. The use of an appropriate risk model is of utmost importance; the history of these algorithms is that they suffer ‘calibration drift’ as overall quality improves. Failure to accept this can create false reassurance and mitigate against quality improvement as results are compared to a historical, and falsely high, risk-adjusted mortality prediction. Our model involves significant adaption of the logit coefficients of the originally published EuroSCORE to render it fit for contemporary benchmarking.

We have defined variance from expected mortality at various levels, categorised ‘yellow, amber and red.’ These utilise statistical adjustment because of multiple comparisons issues (we compare around 45 hospitals and 300 surgeons simultaneously with a peer group mean), but the alerts effectively equate to 1 in 20, 1 in 100 and 1 in a 1000 probabilities that any abnormal finding are due to chance alone [Fig. 3]. This is highly controversial area and is considered in more depth in the recent report from the SCTS [16]. We also further categorise variance depending on whether it is a ‘one off’, persistent or recurrent event. On the most recent analysis of surgeons’ data, 4%, 2.5% and 0.3% of surgeons triggered at yellow, amber and red levels respectively. Developing methods of analysis which are both fair to surgeons and protective of patients remains a significant challenge and there is more work to be done in this area.

There remains a vigorous debate about the rights and wrongs of this programme, particularly amongst UK cardiac surgeons. The enthusiasts believe it stimulates overall improvements and protects the public from surgeons and hospitals with results that are not as good as expected [18,20]. Those who oppose the programme believe that places undue stress on the surgical community and stimulates risk adverse behaviours whereby some patient may be denied treatment altogether because surgeons are concerned about the implication of mortality on ‘their figures’ or, more subtly, they opt for a surgical strategy associated with the best short-term, outcomes, rather than the one which would give the lowest overall lifetime risk [21–27]. There is strong opinion and weak evidence on these issues but what seems very clear in the UK is that patients and politicians remain keen to see this type of data in the public domain, so the challenge is on for the profession to refine methodologies and conduct analyses which optimise the benefits whilst minimising any adverse negative consequences [28].

**WHAT ABOUT IMPROVEMENTS FOR SPECIFIC OPERATIVE GROUPS?**

An ideal quality improvement/governance surveillance programme would look at overall clinical outcomes (in the UK we run our analyses on risk adjusted ‘all’ adult cardiac surgical outcomes, with a number of surgeons also looking at isolated CABG.

---

**Figure. 3** Funnel plot individual surgeons mortality for first time isolated CABG in the UK.Acknowledgement Bridgewater B, Kinsman R, Walton P and Keogh B. Demonstrating quality: The sixth National Adult Cardiac Surgery database report: Publisher Dendrite Clinical Systems Ltd. ISBN 1-903968-23-2. 2009. Ref. [7].
excluding transplantation, primary ventricular assist devices and trauma), but would also support structured improvements for specific diseases or procedures. In the UK to date we have used the EuroSCORE which a generic risk adjustment model designed for all cardiac surgery. The STS have gone down a different route by developing procedure specific models for the various operative groups. The generic model has advantages of simplicity and the ability to pool all cases for governance and quality improvement purposes. The specific model gives better prediction for the procedures under consideration. We find the adage that ‘all models are wrong, but some are useful’ an important one in this context. The best predictive models have only moderate ability to discriminate between patients of different observed risk as evidenced by an area under the ROC curve of around 0.8 (an area of 1.0 would represent a perfect predictor), and key to successful use of any model requires a deep understanding of its strengths and weakness, and the algorithm should only be used for a purpose to which it is well suited.

One example of a specific area that has seen some considerable focus in recent years in that of mitral valve surgery for degenerative mitral disease. It is generally accepted that mitral repair is a better option for the majority of these patients than replacement [29] but it has also been demonstrated that many surgeons (and units) perform repair procedures in low volumes [3], the overall rate of repair for degenerative mitral valve disease is low [3] [Fig. 4] and lack of specific repair expertise is quoted as a reason for replacing the valve [30], and units with the largest volumes have lower risk adjusted mortality [31]. We have previously published practice standards for mitral repair surgery [28], and some networks are now actively collecting and benchmarking mitral valve activity data, including mortality rates and procedure type, to improve quality.

For this data to be optimally meaningful requires comprehensive data, and more extensive data than that commonly collected in standard national datasets; for example left ventricular dimensions in systole and diastole, sophisticated indices of left ventricular function, detailed information on intraoperative techniques, TOE findings and short and longer-term analyses of symptom status and degree of mitral regurgitation. Maximising the information that can be gleaned from this data will possibly require even more sophisticated procedure-specific risk models [32]. It may be that careful scrutiny of this type of information will lead to restructuring of current service provision, possibly with concentration of cardiological and surgical specific expertise. The arguments presented here are for mitral valve repair surgery, but similar issues apply to service provision for major aortic, congenital heart disease and other forms of sub-specialist surgery.

TRACKING THE INTRODUCTION OF NEW PROCEDURES

Defining clinical datasets for operative registries is a significant task — there needs to be clarity of purpose for the use of that registry at the outset, clinical fields must be clearly and unambiguously defined, ideally all definitions should be compatible with other registers and the dataset should include all important fields, but be free of redundancy or repetition to help maximise data quality and the utility of analyses. Revising and updating a dataset is a major task and experience has taught us that some stability with a dataset pays dividends with improving quality and completeness of data. Not changing a dataset is at odds with a register being useful in collecting newly introduced procedures and so a number of groups have been active at introducing specific novel procedure registers to collect activity and outcomes for emerging procedures to provide early experience data for commissioners and clinical benchmarking data for providers. For example a number of national and industry sponsored registers were established with the introduction of transcatheter aortic valve treatments and clinical data from these are now being published [33–36]. Data from these registers will support clinical trials to help customise appropriate treatments for patients as well as providing benchmarking data to help improve quality.

Three challenges inherent in developing databases to track new procedures are defining the necessary procedures to track at the earliest stage, ensuring the database base is implemented at the right time (considering learning curve issues and a need to effectively evaluate efficacy before novel procedures become ubiquitously adopted) and collecting the ‘right’ fields to determine risk adjusted outcomes for an emerging technology, before the important fields are determined through scientific study.

DISEASE SPECIFIC OR GENERALISED DATASETS

A further consideration about the use of clinical outcomes for driving quality, and providing governance oversight, is the relative merit between using highly specialised disease or procedure specific datasets against more generally available, but clinically lower level, administrative datasets. These latter data exist on both sides of the Atlantic and in the UK coding is based on the internationally recognised classification system ICD for diagnostic coding [37] and the National Office for Population and Censuses Survey (OPCS) Coding system for procedures [38]. The advantages of the administrative data are that they are already routinely collected and exist for all hospital admissions, allowing relative hospital performance to be measured using methods such as standardised mortality ratios, with the ability to drill down deeper into specialities or departments as required. Measurement of hospital level mortality using these methods has detected organisations with unacceptable performance in the UK [13]. A major advantage of these data is that, as they are already routinely collected, additional functionality requires only the additional resource required to provide routine or bespoke analysis function, and not the more expensive processes for data collection and collation. The obvious disadvantage of these systems is that they are higher level, and as such they do not allow the more granular collection of important risk factors or operative characteristics which are potentially of great importance in understanding divergent outcomes or tracking detailed clinically important trends.

Necessary inertia to change of the administrative dataset also renders them problematic for monitoring when there are rapid changes in clinical practice (such as the introduction of Troponin measurement to diagnose acute coronary syndromes and widespread introduction of primary angioplasty as a first line treatment for myocardial infarction). That said, a number of different groups claim to have produced risk prediction models from administrative data using limited variables such as age, sex, procedure type, urgency, social deprivation and previous admission history, which have as good a predictive ability as models derived from more complete clinical datasets. We suspect that many will push for more use of these administrative data for clinical purposes, at the expense of specialised datasets, and professional groups will continue to make the case for professionally-driven specific registries, but in reality optimal benefits will come from defining clarity of purpose with respect to functionality combined with appropriate linkage across different datasets.

RESEARCH SUBSTRATE

Rigorous collection of clinical outcomes data also provides a great opportunity for research. The SCTS adult cardiac surgical database contains overall 400,000 records, the majority of which can be linked to long-term mortality data. The STS database contains many more. These registries are in effect rich,
real world epidemiology laboratories in that they describe exactly what has happened for all patients. They are limited in terms of some of the data quality (missing data, lack of robust data validation, and the absence of possible important specific clinical fields) but they do allow analysis for hypothesis testing on huge numbers of patients. With low mortality rates for many cardiac procedures it is becoming increasingly difficult to demonstrate survival differences between different procedures (e.g. off-pump/on-pump surgery or bilateral/single internal mammary arteries) using randomised clinical trials, but it is possible to provide more power by analysing the large registries using appropriate statistical techniques. Unlike RCTs, which usually only enrol a small proportion of eligible patients, the databases reflect real world practice and also include greater coverage, rather than outcomes in specified clinical institutions with an interest in being involved in trials. Registries are however also subject to ‘confounding by indication’ whereby it is possible that factors not collected in a dataset are important in determining outcomes, and this may influence the overall results.

Furthermore, registries can actually be used to facilitate the development of clinical trials, by allowing better assessment of possible treatment effects to allow more accurate power calculations, or possibly for targeted recruitment to trials. For example a protocol is currently under development in the UK to test the need for antibiotic prophylaxis for patients with prosthetic valves undergoing dental or other treatments. This is using analysis of the SCTS database during planning, and we are exploring the potential to recruit existing appropriate patients for possible randomisation from the register.

With increasingly large, comprehensive and sophisticated databases, scientific advancement will come from a balance of clinical studies set against research from registers. Linkage of procedure specific, diagnostic and late mortality datasets will also allow a better understanding of outcomes along a patient’s journey through a disease, and numerous groups around the world are working on overcoming the technical, ethical, political, and legal obstacles to this type of outcomes research [7,39,40]. Technical obstacles include methods for anonymisation (or pseudonymisation) for patient records to ensure patient privacy and confidentiality, whilst still allowing linkage to be made. Ethical issues involve the possibilities of using patient data for research studies for which they may not have been specifically consented (and there are on going debates about the use of anonymised patient level data in this context). Political challenges include those of intellectual property attribution and who ‘owns’ the various aspects of dataset definitions, the data itself and the methods to analyses the data. Legal obstacles vary in different countries but breaches of data protection laws or confidentiality principles are rightly taken increasingly seriously. It will be challenging to introduce robust systems to maximise benefits from linked datasets, whilst robustly protecting patients’ rights.

References

[1] www.scts.org