Scientia vincere tenebras! Science should help us see in the darkness

Paul T. Sergeant*, Johanna Takkenberg* and John Pepper*

* Department of Cardiac Surgery, Gastroheusberg University Hospital, Leuven, Belgium
b Department of Cardio-Thoracic Surgery, Erasmus University Medical Center, Rotterdam, Netherlands
* NIHR Biological Research Unit, Royal Brompton Hospital, London, UK
Corresponding author. Department of Cardiac Surgery, Gastroheusberg University Hospital, Herestraat 49, 3000 Leuven, Belgium. Tel: +32-16-344219; e-mail: paul.sergeant@uzleuven.be (P. T. Sergeant).

Keywords: Risk • Quality of care • Ranking

Medical professionals receive the unique privilege from society to transgress the physical and mental integrity of the patient, conditional on proof of benefit. This was originally an issue for surgeons (the surgical professions), but over time most medical domains have become interventional with catheters or with powerful drugs, and these interventions seriously disturb several physiological systems, leading to morbidity and mortality.

Patients undergoing these procedures often navigate in the darkness of uncertainty, most certainly when this damage to their physical and mental integrity is associated with risk. It is, therefore, the duty of society to evaluate this unique privilege for all doctors. This can be done through the evaluation of the operational systems, the application of international therapeutic guidelines or by the evaluation of the outcomes. This manuscript focuses on the outcome analysis.

There are many challenges in the evaluation of medical interventions, and adult cardiac surgery provides us with a few interesting but difficult examples of these challenges, as outlined in this editorial. We will address the dilemma of ‘time zero’ and ‘time end’, the conceptual challenges in proving risk and benefit, data and model considerations including the uncertainty of any scientific observation and the ‘fata morgana’ of seeking one model that fits all.

THE OBSERVATION INTERVAL: TIME ZERO AND TIME END

The mission of the interventional cardiologist or of the interventional cardiac surgeon is not simply to bypass a few arteries or to replace/repair a valve. Our mission is not organ-specific. Our mission is holistic: to improve the patient’s quality and quantity of life.

The correct evaluation of quality should be the optimal balance between early risk and late result; ‘late’ should be defined differently for each pathology since every pathology, treated with interventions, has a different time impact on the quality and quantity of life of the patient. Risk is a possible surrogate for quality but with considerable limitations. In life, we sometimes accept larger costs (risks) for larger benefits (quality), and there may be a danger in avoiding early risks associated with a particular treatment, when in the long-term this treatment is associated with greater benefits compared with a treatment with a low early risk. It is optimal if risk is absent, but some would dare to state that there is no direct correlation between risk and quality. A single sternotomy would certainly carry less risk vs multiple valve repairs and extensive coronary reconstructions; this lesser risk is in most patients no surrogate for better quality of care.

Society must understand that the longer the period of observation, the greater will be the correlation between quality and benefit as distinct from the risk of the intervention. If society intends to focus on risk with all its limitations, then again we must clearly define this interval with a complete understanding of the clinical practice and the principles of outcome analysis. In the past ‘time zero’ was often ill defined. Possibilities were: the date of the intervention, the moment the patient entered the interventional area, the start of anaesthesia or maybe the start of the actual intervention. In the past ‘time end’ of the interval was most frequently the date of discharge from the primary hospital or the closure of the first 30 days after ‘time zero’.

Hospital stay is not an event but a biased medical decision influenced, also and correctly so, by social and possibly economic drivers, rarely only by medical drivers. The 30-day interval was created when intensive care was not as efficient as it is today and when most patients with serious events would decease within that interval. This interval, therefore, lacks all scientific validity. It is also irrelevant to the patient and society and they have the right to an exact evaluation of risk. The obvious limitation to starting the observation interval at the moment of the intervention is 3-fold: the exclusion of non-referred or refused patients, the exclusion of patients for whom an alternative interventional approach is proposed and the exclusion of morbidity and mortality on the waiting list. Even in 2013 and in Western European Countries, patients are sometimes waiting several years for cardiac interventions. But it needs to be stated that short waiting intervals do not exclude mortality or morbidity on the waiting list (Table 1).

Modern Medicine demands that therapeutic decisions be made by multidisciplinary teams [1–3] where all possible treatments are discussed, and therefore, the correct ‘time zero’ should be the moment of the multidisciplinary discussion. It should be assumed that all eligible patients are referred according to the latest guidelines to a multidisciplinary team; this is the basis for good clinical practice and should be independent of the hour of the day or the day of the week. As a consequence, all discussed patients should be included in the same dataset, irrespective of the therapeutic pathway chosen. Today, most datasets are procedure- or department-specific and fail to address the correct ‘time zero and end’.
The selection of an appropriate and scientifically correct ‘time end’ should be based on outcome science. Only a scientifically defined interval can have legal consequences. So if the project has the intention to study quality of care, then this interval should be as far as possible in follow-up, possibly 5 or 10 years. Similarly, the higher the early risk, the longer the observation interval should extend. If the project has the intention to only study risk of care, then this interval should be based on the mathematical hazard functions. Using this approach, we will come to an evidence-based disease management [4] driven by multidisciplinary teams.

The risk, at each moment in time, after a time zero, is usually a curve that starts high and then gradually diminishes until it reaches a plateau; later on, the risk of having the event will often rise again. The correct ‘time end’ is the moment the risk reaches the lowest plateau. For death after coronary artery bypass grafting (CABG), this interval is minimally 2–3 months [5, 6], for valves it might extend even longer, for transcatheter aortic valve implantation mostly 6 months, for some congenital cardiac pathologies it is 1 year, for stroke after CABG it is 10 days and for infants after CABG it is 5 days. The correct observation interval will vary by pathology, by comorbidity, by therapy and by studied event. If the observation interval is too short, then a considerable number of events will be missed for the analysis and possible inference building.

The question remains whether we are ever able to prove benefit, even if we can apply the proposed time zero and time end. It all depends on the robustness of our knowledge of the natural history of the disease, and this knowledge is limited and usually outdated for most disease entities. This applies in particular to those patients who are asymptomatic, and who usually have a good quality of life and therefore a lot to lose with an early intervention. But it is also challenging to obtain proof of benefit for the new and rapidly growing group of elderly patients for whom lesser invasive treatment innovations have become available in the past decade. This elderly patient group has a lot to gain with an intervention but at what price? Given their relatively short life expectancy, the weighing of costs and benefits of invasive treatment becomes increasingly important as the main goal of treatment is to provide a better quality and quantity of life. The main challenge in this patient domain is to find the right balance between overtreatment, under-treatment and cost containment, as society’s willingness and ability to pay becomes increasingly limited. Twenty-first century registries should be designed with a focus on determining this optimal balance in an era in which new technologies emerge continuously, and provide to patients both potential benefit and potential (still largely unknown) harm.

### DATA AND MODEL CONSIDERATIONS

**Correction for variability: variable selection and modelling approach**

Patients are not comparable with laboratory environments where research is done on animals with the same weight, gender, age and genetic origin. Luckily, the human species is rich in variability and if we want to stay within scientific rigour similar to laboratory environments, we need to correct ‘completely’ for this variability.

The prediction of rare events and thereby the possible correction for variability can be based on different methods [8], such as statistical forecasting, expert judgment, decomposed judgment, structured analogies, judgment adjustment of statistical forecasting, Delphi method, prediction markets and scenario planning.

The project leaders of the UK National Audit have chosen for this purpose the statistical forecasting method. This method is very dependent upon the reference database; a sparse or inappropriate database makes this method unreliable and in addition misplaced causality is often embedded. Therefore, the database will need to register and correct for the variability that possibly influences the studied outcome.

This variability is often corrected with the variables we have at our disposal, but rarely with the variables we should have available mathematically in order to create well-performing correction models. We all know the demographic variability in age and gender, but race, as an example, becomes very difficult to classify today and therefore difficult to include in the modelling. It is well known that Afro-Caribbeans have a tendency towards hypertension, but many patients have difficulty to classify their racial roots. Any variability in whatever domain that can possibly influence risk or quality of care needs to be collected and corrected for: whether that is quality of life, diurnal, circadian, frailty, educational, biological, neurological, hormonal, immunological, hepatic, renal, pulmonary, orthopaedic, muscular … and cardiac. It could very well be that we do not even know which variability is needed.

### PROVING BENEFIT

The project leaders of the UK National Audit have chosen for this purpose the statistical forecasting method. This method is very dependent upon the reference database; a sparse or inappropriate database makes this method unreliable and in addition misplaced causality is often embedded. Therefore, the database will need to register and correct for the variability that possibly influences the studied outcome.

This variability is often corrected with the variables we have at our disposal, but rarely with the variables we should have available mathematically in order to create well-performing correction models. We all know the demographic variability in age and gender, but race, as an example, becomes very difficult to classify today and therefore difficult to include in the modelling. It is well known that Afro-Caribbeans have a tendency towards hypertension, but many patients have difficulty to classify their racial roots. Any variability in whatever domain that can possibly influence risk or quality of care needs to be collected and corrected for: whether that is quality of life, diurnal, circadian, frailty, educational, biological, neurological, hormonal, immunological, hepatic, renal, pulmonary, orthopaedic, muscular … and cardiac. It could very well be that we do not even know which variability is needed.

---

**Table 1:** Mortality on the waiting list for isolated primary or repeat CABG at KU. Leuven, Belgium by year of surgery and with the added average waiting time in calendar days for the same list.

<table>
<thead>
<tr>
<th>Year of surgery</th>
<th>Average waiting time in calendar days</th>
<th>% Mortality on the waiting list</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000</td>
<td>18</td>
<td>0.6%</td>
</tr>
<tr>
<td>2001</td>
<td>25</td>
<td>0.3%</td>
</tr>
<tr>
<td>2002</td>
<td>21</td>
<td>0.3%</td>
</tr>
<tr>
<td>2003</td>
<td>29</td>
<td>0.6%</td>
</tr>
<tr>
<td>2004</td>
<td>19</td>
<td>0.9%</td>
</tr>
<tr>
<td>2005</td>
<td>17</td>
<td>0.2%</td>
</tr>
<tr>
<td>2006</td>
<td>16</td>
<td>0.4%</td>
</tr>
<tr>
<td>2007</td>
<td>13</td>
<td>0%</td>
</tr>
<tr>
<td>2008</td>
<td>18</td>
<td>0%</td>
</tr>
</tbody>
</table>
Most of this variability is not collected for practical and economic reasons and this means that we cannot correct for this variability. This is a serious issue. If indeed some specific variability would have an impact on risk and/or quality of care from a scientific perspective, then the multidisciplinary team needs to have that information and in the correct format before the patient is given appropriate differential therapy. If urgency does not give time to collect, then this major limitation needs to be included in the inference building around appropriate differential therapy.

There is currently a rise in the number of scoring systems using only administrative data, readily available within hospital structures and sometimes already needed for reimbursement of care. They usually have age, gender, urgency and some history of variables. In some countries, a limited number of crude clinical variables are available in administrative data. These systems go for pure pragmatism and reject completely scientific accuracy, even though some of them reach area under the curve (AUC) values in the 0.75 range. They are highly appreciated by Governments since they exclude the often less cooperative medical professional from the system.

Some variability is so rare that even if we could register it, it would still not be included in a correcting algorithm. Indeed, the laws of statistics demand sufficient density of risk in a certain domain of risk before that variable can be selected. A typical example is body mass index. This is rarely selected in scoring systems, for the simple reason that body mass index has only a minimal impact if it varies between 20 and 40 and also since medical professionals often exclude cachectic or very obese patients from multidisciplinary therapeutic discussions and from therapies. In the past, it was impossible to perform angiographic studies in very obese patients. One needs sufficient number of patients (density) with the extremes of variability to allow appropriate modelling coefficients. So we need sufficient number of patients with a body mass index of <20 or 18, and alternatively <40 or 50 body mass index before this variable can be included in our modelling. This obliges us either to include these patients in our reference classes or databases or to find solutions for persons who have exceptional variability in non-registered domains. Many frequently used scoring systems, in addition, ignore possible interaction between variables. Interaction is the situation in which the simultaneous influence of two variables on outcome is not additive. Examples could be the combination of insulin-treated diabetes and very high body mass index or gender and age. Ignorance of interaction between variables in a model may lead to valuable loss of model performance.

The project leaders of the UK National Audit have selected the EuroSCORE variable list as their variable list. This list has been discussed in hundreds of manuscripts and excludes correction for a considerable number of domains. This is a major limitation that decreases the scientific and thereby the legal accuracy. If the inference building is then made public, a massive liability rises through this lack of correction for variability. In the UK National Audit, three categories of patients are excluded but reported internally: emergency or salvage surgery, a patient requiring cardiopulmonary resuscitation (CPR) en route to operating theatre or those requiring CPR prior to induction of anaesthesia. This is a major step forward, but does not correct or solve the issue of patients with outliers in comorbidity. The UK National Audit excludes private hospitals because their case mix is known to be very different and because case attainment is incomplete. Both reasons are very difficult to understand and promote the idea that care is given at two different speeds, that patients with certain risk conditions prefer private hospitals and finally that the correcting systems are inappropriate to correct or solve for the case-mix differences between NHS and private care. How certain are we that they are able to correct for the variability between NHS hospitals. Transplant patients or patients with a ventricular assist device insertion or cardiac trauma are studied in a different register.

Another issue is the identification of the surgeon as a variable. This is outdated in modern therapeutic approaches. The UK National Audit considers the surgeon the person responsible for the clinical team providing the treatment. Medical professionals should work and are working in teams and should, therefore, be audited in team format. Anaesthesiologists, radiologists, intensive care specialists and other medical professionals are medical specialists of their own right and all form an essential element of a therapeutic pathway. Surgeons do not decide which other medical professional is going to be involved in the clinical pathway of his patient. Therefore, he or she cannot and should not have any responsibility in their functioning. In addition, experience has shown that negatively outlying performance (therefore not worse performance) has had career consequences for surgeons but rarely for the team. In the UK National Audit, surgeons are discriminated vs the other medical professionals. Surgeons with <30 cases are excluded from the analysis, which is not happening for anaesthesiologists or intensive care specialists with <30 cases.

The format of the variables and outcomes

Variability should be registered under its correct format. A history of renal, hepatic or pulmonary failure is a nominal variable, but science has progressed so much in registration of variability that most variability can be stored in a continuous format for many decades. This also includes frailty [9], quality of life [10] and other variability outside of the classical domains.

A patient with normal vital capacity % or forced expiratory volume % but on drug therapy for chronic obstructive pulmonary disease or similar condition will often be scored similar to a patient who is on oxygen-therapy with extremely low vital capacity or diffusion values. This avoidable oversimplification in registration and thereby in correction increases the uncertainty once more.

Mortality, stroke and infarcts are events that happen in time and for as long as patients live. They are time-related events and need to be studied under a time-related format. The same is valid for ventilation times, intensive care unit and hospital stay.

The Project leaders of the UK National Audit have selected the EuroSCORE variable list. Nearly, all of these variables oversimplify domain variability by using nominal (yes/no) registration, rarely categorical. This is indeed an additional avoidable system limitation. The UK National Audit has optimized the coefficients of the original EuroSCORE into three alternatives (refitted, refitted SCTS modified and recalibrated logistic). But coefficients are mathematical transformers adding weight to the registered value of a variable in relation to an event. The nominal simplification is not corrected by an optimization of the coefficient. The UK National Audit studies the postoperative time-related events in a non-time-related format. This adds additional structural imprecision.

Missing data

The assumption by the UK National Audit that the risk was not present or classified at its lowest possible value for variables with missing values is an appropriate decision. The assumption that the patient has died if the discharge data are missing is a fragile assumption but most likely highly motivational for participating centres, as they will push their limits to avoid missing discharge data.
The number of variables in the model for correction of variability

The Law of Parsimony states that if a model is not improved by adding a variable then the model with the fewest number of variables should be selected. This law, often known as ‘Occam’s Razor’, was mentioned in Posterior Analytics by Aristotle. A single variable such as age under a continuous format, with an appropriate transformation and an appropriate coefficient can, on its own, bring the discrimination power of a hospital mortality model after cardiac surgery up to the 0.65 and even higher AUC value, from the random 0.5 discrimination power value. Adding variables should improve the discriminatory and calibration power. An example is the ACEF [11] score with only three variables. This risk-correcting model performs as well as the traditional ones. The ACEF manuscript identified that the positive predictive value (the correct prediction of the patients having the event) was only 8% for the ACEF score (AUC = 0.80) and 7.7% for the Cleveland Clinic score (AUC = 0.81), 6.6% for the Log EuroSCORE (AUC = 0.79) and 6.7% for the Add EuroSCORE (AUC = 0.78)! So the conclusion is most certainly not that the 3-variable ACEF score is similarly good but probably that it is similarly underperforming. This very low performance should not astonish us in the presence of low prevalence of mortality and average AUC. Table 2 clearly identifies that, even with exceptional sensitivity and specificity but very low prevalence, only very low positive predictive values can be observed.

So from a statistical perspective, models should have as few as possible variables, whereas from a legal correction-for-variability perspective, we need to guarantee to the medical professional that we will correct for ‘all event-important’ variability. It is clear that there is a conflict of interest. Rare variables will hardly improve the models but possibly correct for variability. So if we decide to limit for practical and mathematical reasons the number of variables, we must dare to state that our correction is not complete and we must offer a solution for this incompleteness. Models should, therefore, not be used for ranking because they, by definition, will fail in correcting for variability. In addition, ranking does not cope with the general concept of the ‘uncertainty of an observation’, mandating reports to add uncertainty values to the observed and corrected-for values. Rankings [12] are an imprecise statistical method to report cardiac surgery mortality rates and prone to fluctuation. Hence, reshuffling of ranks can be expected solely due to chance. Models will, therefore, at their best, identify outliers and not lesser performance. Lesser performance demands a series of additional steps down to the individual records with the risk that the charts or our lack of knowledge do not identify the ‘uncorrected for’ variability.

Model performance

Mathematical models are often evaluated using the discriminatory power and the calibration. The discriminatory power is presented as the AUC where sensitivity is plotted vs 1-specificity. So the AUC gives a general evaluation and is not easily interpreted by the scholar unfamiliar with this type of analysis. When the model predicts at random, the AUC is 50% covered or the AUC is 0.5. The appreciation of a scoring system will depend upon the prevalence of the event. In cardiac surgery, most studied events are in the 1–5% range or so-called ‘rare event’ range. In the presence of rare events, we will need to understand in more detail how this AUC is obtained. It turns out that most scoring systems used for scoring rare events have an AUC in the 0.75–0.81 range but obtain these values through a correct prediction of the survivors and a very low correct prediction of the patients suffering the event (see previous paragraph on positive predictive value). So the interested reader should look for much more detailed information about the elements composing the obtained AUC, such as the positive predictive value, the negative predictive value, the misclassification rates. With positive prediction rates of an event at <10%, one can hardly speak about good prediction, most certainly not when the reputation of a medical institute or professional is at risk. An $R^2$ test could also add to the understanding of the completeness of the correction.

The Hosmer and Lemeshow [13] test divides the dataset in groups of equal size, ranked according to the risk prediction. It then uses simple comparisons between the group predicted and observed event rates. This gives the scholar information about the calibration of the scoring system. The problem with risk distribution in cardiac surgery is that these equal size groups do not have gradually rising predicted risk, but that most risk is located in the last group. This somewhat reduces the calibration value. The exclusion of some high-risk patients reduces this problem considerably. A positive Hosmer and Lemeshow test is a sign of ill-calibration, but a negative Hosmer and Lemeshow test is no proof of good calibration.

The UK National Audit obtains AUC in the range 0.78–0.79. No information is given about the misclassification rates and the negative or positive predictive values or ratios. The UK National Audit obtains evidence of ill-calibration in two of the three models used (not in the recalibrated EuroSCORE model).

**THE FATA MORGANA OF ONE MODEL FITS ALL**

It is an illusion to think that one model fits all. For the purposes of benchmarking, a general model like EuroSCORE seemed appropriate at the time of its inception in the 1990s. However, with time and with the emergence of several large cardiac surgery registries, our knowledge of cardiac surgery has deepened, and we have come to realize that cardiac surgery comprises an extremely wide variation of procedures with widely varying determinants and outcomes. In this light, and with the notion that mortality risk is steadily decreasing, even for the purposes of benchmarking it is highly questionable that we can maintain the static EuroSCORE model (or its newer variant EuroSCORE II) and its limited number of variables as the model that fits all. For the purposes of benchmarking, there is a need for dedicated predictors in the different large treatment domains (CABG, valve and CABG + valve), and dedicated models that adequately weigh these dedicated predictors in the context of the treatment domain of interest and dynamically adjust with the changing patient profile and patient outcome.

<table>
<thead>
<tr>
<th>Table 2: Effects of prevalence in the presence of models with a sensitivity and a specificity of 95%</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Prevalence (%)</strong></td>
</tr>
<tr>
<td>-------------------</td>
</tr>
<tr>
<td>0.1</td>
</tr>
<tr>
<td>1.0</td>
</tr>
<tr>
<td>2.0</td>
</tr>
<tr>
<td>5.0</td>
</tr>
<tr>
<td>50</td>
</tr>
</tbody>
</table>
In addition to benchmarking, there is increasing interest to use models for the purposes of patient information and risk prediction. In this setting, it is even more important to develop models that fit the disease and to move beyond the large treatment domains, to ensure valid estimates of outcome, although at the cost of reliability in less common subgroups.

The UK National Audit has decided to go for a single ‘all cardiac’ surgery equation. This reduces considerably the task but reduces also dramatically the possibilities for correction for variability. Valve patients and coronary patients have totally different anatomical and haemodynamic predictors. None of these is present in the EuroSCORE data list. By not registering these different predictors and by not allowing them to come into specific models, pragmatism has won over scientific and therefore legal certainty (or uncertainty).

The training and innovation

The UK National Audit annual report 2010–2011 gives very limited information about the technical elements of the cardiac procedures performed. Some of these technical components will have an early or late outcome effect and are most certainly possible surrogates for quality of care. For primary CABG a very slow increase in the use of the left internal mammary artery has been identified gradually reaching the 95% with no information about more extensive arterial grafting. For degenerative mitral valve surgery, it is still apparent that many centres do not even reach the 50% repair rate. The overall repair rate is again a very crude parameter, and more interesting information would be the components of repair: annular and/or leaflet and/or chordal repair. More extensive arterial revascularization and complex mitral valve repair were innovative techniques in the early 1980s. This mandates the scientific community to study the possible impact of public reporting on a shift towards early risk aversion and the absence of implementation of innovative techniques vs late benefit. How education and training fits into this perspective is unclear.

The public reporting

Public reporting of data about individual persons is a fragile balance between the rights of the individual and the rights of society. Fung et al. [14] have studied the evidence whether publishing patient care performance data improves quality of care. They stated that the evidence is scant and that it is possible that there is a beneficial effect at hospital level but that the evidence on effectiveness, safety and patient-centredness remains uncertain. This possible effect at hospital level can be induced through gaming and exclusion of high or higher risk patients. It is sometimes correct to redirect patient care to more conservative approaches, but inappropriate gaming has been documented in the presence of public reporting [15] of physician-specific outcome data even when physicians believe that the procedure might be beneficial.

The UK National Audit has made an enormous leap forward in the analysis of early outcome data at regional and supra-regional levels. It is of utmost interest to any medical professional or any health authority to have read the ‘technical review of the United Kingdom National Audit Cardiac Surgery Governance’ but also to understand the complexity of the problem. This National Audit project has chosen to go for public transparency, but pragmatic approaches and simplifications have been taken at each step in this project. Pragmatic approaches induce limitations and these need to be made as transparent as the reported medical data.

Hospitals can only hope that negative outliers are not immediately classified as negative underperformers but that appropriate care and attention is then given to these units after case-by-case revision of their unfortunate performance with appropriate attention to uncorrected variability and inappropriate observation interval. Those units where the events happen in secondary hospitals or before/after the analysed intervals and those units that fail to implement innovative techniques or fail to implement training will escape for the moment the critical review of society.

Let us rethink medical care, the analysis and the transparency of its performance according to the evolutions in mathematics, patient care, outcome science and in medicine. Let us come to a holistic evidence-based disease management across and not within therapeutic domains.

REFERENCES