Interactive CardioVascular and Thoracic Surgery 20 (2015) 691–693
doi:10.1093/icvts/ivv060 Advance Access publication 18 March 2015

Cite this article as: Sergeant P, Takkenberg JJM, Noyez L. The sequelae of misinterpretating surgical outcome data. Interact CardioVasc Thorac Surg 2015;20:691–3.

The sequelae of misinterpretating surgical outcome data

Paul Sergeant[a,*], Johanna J.M. Takkenberg[b] and Luc Noyez[c]

[a] Serrey Consulting, Sint Joris Winge, Belgium
[b] Department of Cardio-thoracic Surgery, Erasmus MC, Rotterdam, Netherlands
[c] Department of Cardio-thoracic Surgery, Radboud University Nijmegen Medical Center, Nijmegen, Netherlands

* Corresponding author. Reigersweide 16, 3390 Sint Joris Winge, Belgium. Tel: +32-475-742771; e-mail: paulsergeant133@gmail.com (P. Sergeant).

Keywords: Audit • Risk control • Quality control • Variable life-adjusted display curves • Scoring systems

On a normal working day, a plot is presented to You that will change your life forever. The plot (Fig. 1A) represents your variable life-adjusted display (VLAD) curves depicted versus VLAD curves from your 'competing' colleagues. This VLAD curve, a real case, a real plot (the surgeon is not an author of this letter), presents a plot of your cumulative sum of the difference in expected and observed outcome; the unit of measurement is 'lives'. The expected outcome is based on an adjustment or scoring system. The VLAD curve, a valuable tool in the monitoring of a process, is based on the statistical principle of process variability. In your case, the lines separated and a difference in quality of care seemed to be identified. In similar instances, as in your case, careers have been damaged.

The first thing any informed observer picks up is the absence of the uncertainty of the scientific observation, but in addition 22% of your case volume was missing (without agreement) excluding operations supervised by you, not because of a specific higher risk. The same has happened with the plots of the 'competing' surgeons, even though there might be a difference in their concept of patients eligible for supervised procedures. Later on, it was identified that there were considerable events in their excluded cases and none in your own excluded ones. Figure 1B presents 100% of your case volume with the 99% confidence limits, clearly indicating that your VLAD curve falls within the scientific uncertainty of an observation. But the issue is not closed, on the contrary!

Outcome analysis is a science and there are fundamental laws of science and nature that need to be respected and were not in this case. Some examples are the uncertainty of a scientific observation, the continuity and the variability of nature. In addition, outcome analysis is the analysis of rare events, with all its specificities. It is the intention of this editorial to address these limitations and propose a checklist to follow if similar events should repeat itself.

The evaluation of a rare event can be approached using different pathways [1], as there are statistical forecasting, expert judgement, decomposed judgement, structured analogies, adjusted judgement, Delphi, prediction markets and scenario planning. Your plot used a VLAD format of depiction, but the underlying process used was *statistical forecasting* with well-known limitations. If the reference database (the original database from which an adjustment equation is created) or the mathematical model is sparse or inappropriate, then the adjustment process becomes unreliable. Misplaced causality is possibly embedded and the issue of frame blindness is not addressed.

'Sparsity' identifies the richness of a database: the number and format of the variables, the completeness versus all possible variability and the density of the outlying values. If a clinical database does not have a sufficient number of patients with extreme obesity, then this variable can never be corrected for. In addition, not just the surgeon needs to be identified and corrected for, but all major players in the complete hospital care process, from admission to discharge. Even if one has a perfect coefficient for the separate rare comorbidities, then a correct estimation of the even more rare combination of these rare comorbidities creates statistical issues that need to be corrected for (or the patients deleted from analysis). It is obvious that sparsity refers also to the quality of data input and checking (audit and consistency). 'Inappropriate' identifies how the case mix of the reference(s) database relates to the index database (yours), to the observation interval and also to the quality of the adjustment system. An observation of the events of the hospital stay or even first month after a therapy has no relevance to the quality of care of that therapy, unless the expected survival before the therapy would be expressed in minutes or hours. 'Causality' is the direct or indirect relation between a set of factors and a phenomenon or event. 'Frame blindness' is solving the wrong problem because one has created a mental framework for one's decision, overlooking the best options or losing sight of important objectives.

In your plot, the issue of the reference database is very complex since there are several databases. The reference database of the adjustment system is the core dataset used for the adjustment scoring system. But in this process, you are not just compared with the scoring system but also with the performance of other competing surgeons, without any additional correction; it needs to be proven that no correction was needed. In that comparison, the reference databases are the datasets where the VLAD curves of the competing surgeons originate. The index database is the dataset from which your VLAD curve originates.

The reference database of the adjustment system (Table 1) was the EuroSCORE 1 dataset [2, 3]. The case mix of this database was never compared, but is very different from the index database.

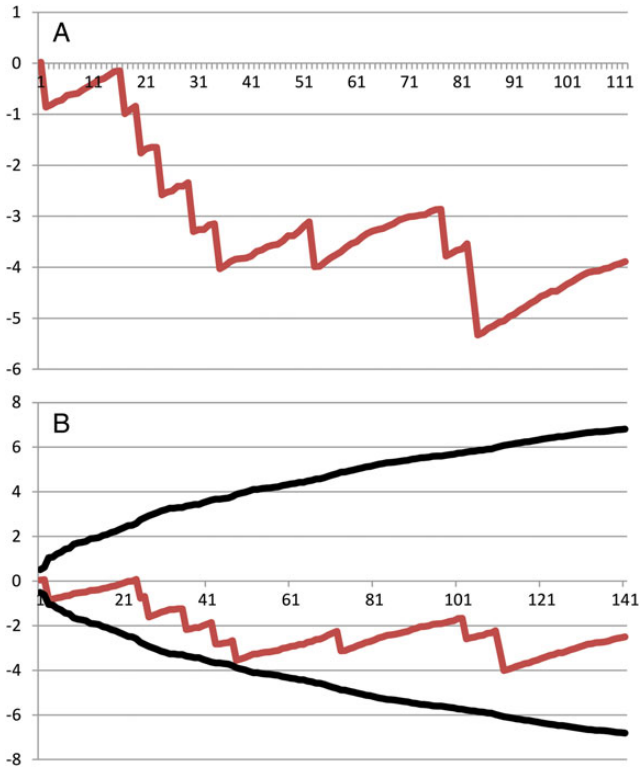

Figure 1: (**A**) The VLAD curve of the involved surgeon (with 22% of the actual case volume deleted). The *x*-axis is the number of procedures over 1 year. (**B**) The VLAD curve that could have been plotted, including all the patients treated by the involved surgeon that year and including the 99% confidence limits. VLAD: variable life-adjusted display.

The outcome event of the adjustment database was primary hospital mortality, very different from the early risk and even more from the 'quality of care' of the procedure. This outcome event is a real event, not a medical decision. The outcome event interval is biased since it was not based on hazard function analysis. The time frame of the adjustment database is very different. The variables do not cover all possible variability. Most variability was only documented in a single expression. There was absolutely no information about the density of the extremes of risk. The variables were not stored in their original format. There was no proof of any consistency in the creation of the different datasets composing the adjustment database. The completeness and the accuracy of the data or of the follow-up were never audited. Some quality criteria of the adjustment system are known, most are not. A borderline reasonable receiving operating characteristics/area under the curve (close to 0.8) was reached in the internal validation and in some external databases, mostly due to the appropriate prediction of the survivors but a very poor one (<10% correct) of the deceased patients (positive predictive value). On the 14 possible quality criteria, the adjustment database scored 13 possible or certain elements of bias. Recalibrating the coefficients would not repair these biases.

The index and competing databases need to undergo the same critical analysis. Even though your own index database or the databases of the competing surgeons could be physically registered into the same software, their data process, completeness and flow could be different. The power of the index versus the competing database sizes has not been calculated. The confidence limits have not been calculated; what is seen on Figure 1B is our own addition. The case mix differences have not been calculated, in addition there was no common definition about the supervision

Table 1: Checklist of quality criteria of the reference database of the adjustment system and of the index and competing databases

1. How good is the reference database
   - Is the adjustment database appropriate in case mix?
   - Is the adjustment based on a generic equation or on a domain-specific equation?
   - Is the adjustment database appropriate in outcome event for the desired assessment?
   - Is this outcome event a real event, or is this a medical decision or intervention?
   - Is the adjustment database appropriate in outcome interval for the desired assessment?
   - Is the adjustment database appropriate in timeframe?
   - Do the variables of the adjustment database cover all variability?
   - Do the variables of the adjustment database exist in different expressions?
   - Is there enough density of the extremes of risk in the adjustment database?
   - Are the variables of the adjustment database stored and analysed in their original format?
   - Is there consistency in the creation of the adjustment database?
   - Has there been an internal/external quality audit of the adjustment database?
   - Are the data and the follow-up of the adjustment database complete?
   - How good are the quality criteria of the adjustment system (scoring system)? Are they known and do they reach acceptable values?
2. How good and comparable are the index and competing databases?
   - Is there sufficient power to compare the competing with the index databases?
   - If there is sufficient power, have the confidence limits been calculated and presented?
   - If there is sufficient power, are the competing and index databases similar in case mix?
   - Do the adjustment databases, the index and competing databases include the same outcome event?
   - Do the index, the competing and the adjustment databases carry the same variables and with the same definitions?
   - Is the consistency in creation of the index and the competing databases similar?
   - Has there been a similar internal/external quality audit of the index and the competing databases?
   - Are the data and the follow-up of the index and the competing databases complete?
   - Do the index and competing databases carry all the procedural and institutional variables and have they been corrected for?
   - How good has the adjustment system scored on the dataset, combining the index and the competing datasets in ROC/AUC, PPV NPV…?

ROC: receiving operating characteristics; AUC: area under the curve; PPV: positive predictive value; NPV: negative predictive value.

cases. The outcome event of the adjustment, the index and the competing databases are ill-defined. The variables and definitions were neither defined nor analysed versus the adjustment database. The consistency, apparently different, was not analysed. There has been no internal/external audit, no outlier or no completeness analysis of the index and competing databases. No additional variability has been corrected for.

On that same, apparently not so normal, working day, your life changes and your surgical career is over. A personal tragedy, but what will happen to the system you work in? Your colleagues take over the patients you used to operate on and will probably experience a downward shift in their VLAD curve, and/or your colleagues will take a more conservative approach in accepting patients for surgery. It is the patient who will be denied appropriate care. A surgeon works in a complex system; the correct approach in dealing with negative outliers is to first analyse the system before, if at all, pointing to an individual in the system.

The VLAD and similar methods originate from the industry, where quality control can easily be translated as 'on target with minimal variation'. In the medical field, however, there are much more inherent variabilities, case mix, risk variables, etc. As statistical forecasting is highly dependable on the quality of the data (collection process), the appropriateness of the applied model and the methodological approach used, we may not assume that a negative outlier is by definition an underperformer. Therefore, if a VLAD shows a run of bad events, suggesting that the process is out of control, this can only be interpreted as a warning signal, and nothing else. It must be the start of further investigation of the process, beginning with the data registration itself and including both content and organization. Additionally, it requires predefined analytical methods that adequately correct for case mix to ascertain a systematic approach and to avoid 'cherry picking' as illustrated in the case above. Usually, several years of practice are needed to form the basis of such a plot. Sometimes, it will demand exclusion of patients from the analysis because their variability cannot be corrected for, even though clinical judgement considers these patients high risk. Finally, we need to implement in clinical practice periodical evaluations of quality of cardiac surgery care including—but surely not solely depending on—VLADs, and committing all medical specialists in the system who are involved in the care for cardiac surgery patients, in order to achieve continuous *improvement* in the healthcare process.

What happened to You on that particular working day, reflects an undesirable attitude among a surgical specialty that has always been excelling by standing out and taking risks for patients. The misinterpretation of surgical outcome data as illustrated above will lead to risk avoidance, mediocracy and suboptimal care for patients requiring cardiac surgery.

REFERENCES

[1] Goodwin P, Wright G. The limits of forecasting methods in anticipating rare events. Technol Forecast Soc Change 2010;77:355–68.
[2] Roques F, Nashef S, Michel P, Gauducheau E, de Vincentiis C, Baudet E *et al*. Risk factors and outcome in European cardiac surgery analysis of the EuroSCORE multinational database of 19030 patients. Eur J Cardiothorac Surg 1999;15:816–23.
[3] Michel P, Roques F, Nashef S. The EuroSCORE project Logistic or additive EuroSCORE for high risk patients. Eur J Cardiothorac Surg 2003;23:684–7.