Practice variations, chance and quality of care

James M. Brophy,*† MD; Lawrence Joseph,†‡ PhD

Editorial

Éditorial

From the *Department of Medicine, Centre hospitalier de l'Université de Montréal, Montreal, Que; the †Department of Epidemiology and Biostatistics, McGill University, Montreal, Que.; and the ‡Division of Clinical Epidemiology and Centre for the Analysis of Cost Effective Care, Department of Medicine, Montreal General Hospital, Montreal, Que.

CMAJ 1998;159:949-52

See related article page 926

In this issue Dr. William A. Ghali and colleagues publish their provocative study of Canadian (excluding Quebec) rates of in-hospital death after coronary artery bypass grafting (CABG) for the period 1992/93 to 1995/96 (page 926). For 50 357 CABG cases, they observed an overall rate of in-hospital death of 3.6%. After adjustment for sociodemographic factors, coexisting conditions and disease severity, the rate of in-hospital death ranged from 1.95% to 5.76%. The authors suggest that there may be clinically meaningful differences in the quality of care among the 23 institutions studied. However, partly because of the uncertainties in the methodology, the authors do not favour public disclosure of the findings for individual hospitals. Some salient points merit further discussion.

What is the risk of death after CABG?

This study provides a reliable overall estimate, based on a sample size of more than 50 000 interventions, of contemporary rates of in-hospital death after CABG. Although Quebec hospitals were excluded from this analysis (because they do not provide data to the Canadian Institute for Health Information [CIHI]), an examination of the equivalent Quebec administrative database revealed that 19 417 CABG procedures were performed over the same period in that province with a comparable unadjusted overall mortality rate of 4.3% (unpublished data [J.M.M.]). The general consistency of these results is reassuring.

However, despite continuing improvements in anesthesiology, cardiac surgery and postoperative care, outcomes research has shown that rates of in-hospital death have remained relatively constant since 1987.[1] This paradox is perhaps explained by changes in case-mix, whereby older and sicker populations now undergo cardiac surgery. To better assess individual risks it would be helpful if Ghali and colleagues were eventually to publish their results according to clinically germane subgroups.

Is there more noise than signal?

Before embarking on a sophisticated statistical analysis, an appreciation of the validity of the CIHI database is required. For example, is it reasonable that the 1995/96 adjusted mortality rate for hospital A is only a small fraction of that of the next best performer, or is there some other explanation? Did all hospitals equally record the secondary diagnoses that were used to build the logistic regression model? Were the recorded confounding variables present on admission, or were they complications of surgery? If the latter, adjustments to mortality rates would be inappropriate. How robust are the results to different models? Previous studies have shown that the rates of in-hospital death after CABG may be dependent on the choice of model.[2] How confident can we be in the results when the expected (adjusted) rates differ substantially from the observed ones (as indicated by the deviation from the 45° line in Fig. 1 on page 928), especially when the directions and magnitudes of the changes differ greatly from one hospital to another? Although it is reassuring that the necessary adjustments to account for varying severity have been carried out, it is worrisome that a coefficient that adequately adjusts the rates for one hospital may not work well for another, because of local practice variations. These questions highlight the degree of prudence we must bring to an interpretation of these results.

Nevertheless, Ghali and colleagues did observe a 3-fold difference between the highest and lowest overall mortality rates, and, while acknowledging that both chance and unmeasured confounding (severity of illness) may play a role, they believe that some of the differences are due to variable quality of care. This ratio of highest to lowest rate is referred to as the extremal quotient. Intuitively, an observed difference this large must surely indicate a process exceeding the play of chance. But is this necessarily true?

In this example the extremal quotient is 3.3 (5.76/1.76), and the probability that chance alone could produce this ratio is exceedingly small, less than 0.001. However, random variation may be an important consideration for less extreme data. For example, there is an approximately 50% probability that chance alone could produce an extremal quotient as great as 1.7. In general, the extremal quotient has several undesirable properties and becomes more unstable, with larger 95% confidence intervals, for low event rates, uneven population distributions or small populations or if there is a chance that one individual could be counted more than once in the numerator; all of these are potential problems in database analysis. Also, employing this ratio leads to ignoring and consequently wasting all of the information between the 2 extremes. Simulation studies have shown that large values of the extremal quotient may in fact occur by chance. For example, with very rare events (1 in 1000) an extremal quotient of 11 is likely to occur by chance alone.[3]

The problems of standard statistical methodology and the limitations of *p* values in interpreting the results of clinical trials are increasingly appreciated,[4] and similar issues arise in this type of small-area analysis. First of all, *p* values are the result of testing a null hypothesis that nobody seriously believes, for example, that there are absolutely no differences among the 23 hospitals in the study by Ghali and colleagues. As the number of hospitals increases it would be surprising if the null hypothesis was not eventually rejected, even if there are only trivial differences among institutions. In a study of 50 000 patients, a *p* value of 0.05 is actually rather weak evidence against the null hypothesis, because it implies a small effect size.[5] A *p* value calculated from a *t* statistic, for example, will be small if the *t* statistic is large in absolute value. This can occur either because the observed difference between the 2 samples being compared is large or because the standard error of the estimate is small. The latter automatically happens as the sample size increases, so that one cannot necessarily equate small *p* values with clinically meaningful differences in studies with large sample sizes. In such large studies, a *p* value of 0.05 is probably driven to a greater extent by the low standard error than by a meaningful difference in sample means. A further limitation of *p* values is that they are only useful for significance testing; they do not address the more interesting issue of estimation of the magnitude of between-hospital variation. How then can a true signal of poor-quality care be separated from background noise?

In their analysis Ghali and colleagues assume that all 23 hospitals are independent, yet they consider all patient data interchangeable and incorporate all the data into a single logistic regression model. This approach is problematic conceptually, because it ignores potential interaction between the regression variables and hospital practice. For example, local expertise could cause the prognostic influence of some confounders to vary from one hospital to another.

As an alternative, hierarchical modelling[6] assumes that participating hospitals are a random sample from a superpopulation of all hospitals where patients may undergo CABG and attempts to model hospital heterogeneity. Such heterogeneity can include both variations in regression coefficients, where hospital-specific coefficients can be estimated, and variations in rates. These estimates will be a compromise between the data for individual hospitals and the pooled information, when warranted, with the data dictating the degree of pooling. This approach of borrowing and sharing information from the different hospitals leads to superior global statistical properties, albeit occasionally at the cost of a more conservative estimate for an individual centre. In practical terms, hierarchical modelling results in a "shrinkage" of the extreme results toward the centre, which permits the estimation of individual hospital mortality rates and the probability of differences among hospitals, rather than uninformative, obscure *p* values. With this technique, a probability density function for mortality rates can be constructed for each hospital, which can then be used for drawing inferences. The results from a simple hierarchical analysis that ignores any possible errors associated with the adjustment model appear in Fig. 1.

Having recognized the limits of the *p* value as a criterion for measuring variation in practice patterns, it is necessary to decide what criterion is suitable to determine whether meaningful differences exist among hospitals. From our hierarchical model, we can directly calculate the probability of differences between 2 hospitals. For example, we are almost 100% certain that a difference in mortality rate exists between hospitals A and W, a difference that is visually represented by the absence of overlap between the 2 curves in Fig. 1. The improved prediction of the difference in mortality rate, as estimated by hierarchical modelling, is 2.8%, as compared with the originally reported difference of 3.81%. Hospitals B through S are also probably better performers than hospital W (data not plotted). Similar comparisons can be made between other pairs of hospitals. Although these comparisons are en-

lightening, it is inappropriate to use the best hospital, a posteriori (after the data have been collected), as the gold standard in judging quality of care. Even if there were no true differences among the hospitals, 2 of the hospitals will necessarily be at the extremes of the observations. An analysis that is stimulated uniquely by an examination of the data should be approached with suspicion, independent of whether the data originate from a clinical trial or an observational study.

Another method would be to consider the overall predictive probability density for the next randomly selected hospital as the gold standard; this value could then be compared with each hospital's distribution of mortality rates. However, because the spread of the predicted curve is determined by all of the data and is therefore influenced by the presence of the potentially poor performers, the power to detect substantial differences may be limited. A possible improvement would be the following: when evaluating a given hospital, i, the predictive distribution for the next hospital could be formed on the basis of the remaining $n - i$ hospitals (see Fig. 1). For example, when evaluating hospital W, the distribution that represents the expected range in morality rates for the other 22 hospitals should be formed. With this technique, there is an 81.7% chance that hospital W's mortality rate will exceed the upper value of the 95% confidence limit for the next randomly selected hospital. In other words, we are about 82% sure that hospital W is in the worst 2.5% of the population of hospitals represented by our sample.

This analysis includes only the between-hospital variation and ignores the uncertainty surrounding the adjustments to the observed data. The additional uncertainty will widen the probability curves, further diminishing our confidence in identifying possible poor performers and supporting the conservative approach to public dissemination advocated by Ghali and colleagues.

Where do we go from here?

Obviously all centres should receive these data and continue their individual quality-control programs but, given the generally favourable distribution of mortality

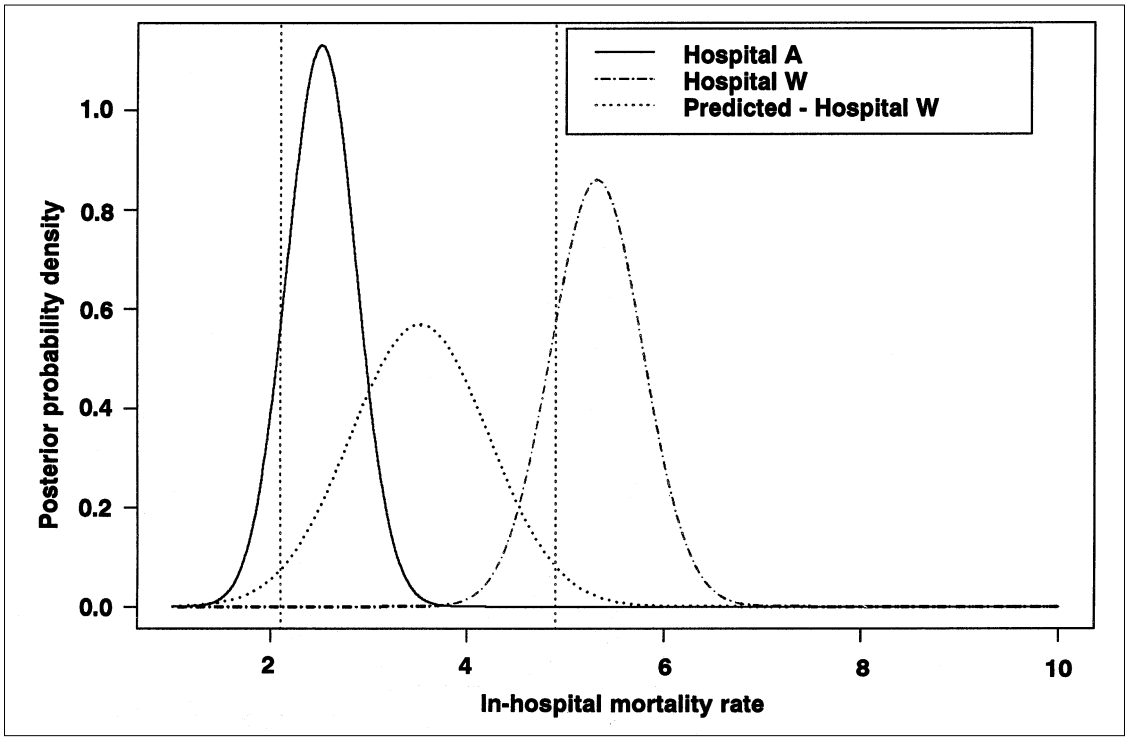

Fig. 1: Plots of the probability density for rates of in-hospital death for hospitals A and W, as well as for predicted rate of death for the next hospital, selected at random from the 22 hospitals other than hospital W. The absence of any overlap between the curves for hospitals A and W implies that the 2 rates are almost certainly different. The vertical lines represent the 95% confidence interval for this overall predicted rate of in-hospital death. The probability that the rate for hospital W falls within this range is the percentage of the area under hospital W's probability density curve that lies to the left of the upper 95% confidence limit (18.3%). Conversely, the probability that hospital W is a poor performer is defined as the area under its probability density curve to the right of the upper 95% confidence limit (81.7%). Similarly, the probability that hospital A is truly a better performer is 18.5%. Other hospitals would be situated between these extremes and have been excluded for visual clarity.

rates among the hospitals, it may be more useful to attempt to shift overall population mortality rates after surgery[7] than to concentrate solely on individual poor performers. It is important to ensure that those who need surgery receive it, but it is equally important to ascertain that those receiving it truly need it.

For example, 52% of this cohort underwent CABG after urgent admission. This aggressive, interventionalist approach emanates largely from the United States, where the number of patients who undergo CABG after myocardial infarction is 3 times greater than in Canada,[8] without an appreciable difference in mortality rates.[9,10] Some have suggested that anticipated improvements in quality of life justify the higher rates of revascularization procedures,[9] but this hypothesis is far from being firmly established or accepted.[11] Recent randomized trials of patients with Q-wave myocardial infarction[12–14] and non-Q-wave or unstable angina[15,16] have demonstrated no advantages to systematic early investigation and revascularization. A recent prospective registry of 8000 consecutive patients with acute coronary symptoms also failed to demonstrate an association between death or myocardial infarction and use of invasive procedures.[17]

When viewed in the abstract, the overall rate of in-hospital death reported by Ghali and colleagues seems small, but from a clinical perspective the implication is that 1 of every 28 patients will die after CABG. Clearly, this reality should not be forgotten in our interventional enthusiasm and is an important element in the assessment of risk–benefit ratios. Continued vigilance in selecting patients is required to ensure that the benefits of surgery will exceed the risk. For certain high-risk patients, for example those with left main coronary artery disease, the overall gain in life expectancy clearly favours surgery. However, when surgery is proposed for purported quality-of-life improvements, particularly in elderly patients, it is essential that patients be informed of the concrete risks, as outlined in the article by Ghali and colleagues, and of our uncertainty about future benefits.

In-hospital death is only one domain of quality of care. Now that the rate of in-hospital death associated with CABG in Canadian centres has been established, other facets of quality of care, including appropriate patient selection, accessibility, complications, long-term survival, quality of life and resource utilization, await investigation.

References

1. Grumbach K, Anderson GM, Luft HS, Roos LL, Brook R. Regionalization of cardiac surgery in the United States and Canada. *JAMA* 1995;274:1282-8.
2. Localio AR, Hamory BH, Fisher AC, TenHave TR. The public release of hospital and physician mortality data in Pennsylvania. A case study. *Med Care* 1997;35:272-86.
3. Diehr P, Cain K, Connell F, Volinn E. What is too much variation? The null hypothesis in small area analysis. *Health Serv Res* 1990;24:741-71.
4. Brophy JM, Joseph L. Placing trials in context using Bayesian analysis. GUSTO revisited by reverend Bayes. *JAMA* 1995;273:871-5.
5. Raftery AE. Bayesian model selection in social research (with discussion). In: Marsden PV, editor. *Sociological methodology*. Oxford: Blackwell; 1995. p. 111-96.
6. Carlin JB. Meta-analysis for 2 × 2 tables: a Bayesian approach. *Stat Med* 1992;11(2):141-58.
7. Rose G. The strategy of preventive medicine. Oxford: Oxford University Press; 1992.
8. Pilote L, Califf RM, Sapp S, Miller DP, Mark DB, Weaver WD, et al. Regional variation across the United States in the management of acute myocardial infarction. GUSTO-1 Investigators. Global utilization of streptokinase and tissue plasminogen activator for occluded coronary arteries. *N Engl J Med* 1995;333:565-72.
9. Mark DB, Naylor CD, Hlatky MA, Califf RM, Topol EJ, Granger CB, et al. Use of medical resources and quality of life after acute myocardial infarction in Canada and the United States. *N Engl J Med* 1994;331:1130-5.
10. Rouleau JL, Moye LA, Pfeffer MA, Arnold JM, Bernstein V, Cuddy TE, et al. A comparison of management patterns after acute myocardial infarction in Canada and the United States. The SAVE investigators. *N Engl J Med* 1993;328:779-84.
11. Brophy JM, Joseph L. Quality of life after myocardial infarction: Canada versus the United States. *N Engl J Med* 1995;332:469-70.
12. SWIFT (Should We Intervene Following Thrombolysis?) Trial Study Group. SWIFT trial of delayed elective intervention v conservative treatment after thrombolysis with anistreplase in acute myocardial infarction. *BMJ* 1991;302:555-60.
13. TIMI Study Group. Comparison of invasive and conservative strategies after treatment with intravenous tissue plasminogen activator in acute myocardial infarction. Results of the thrombolysis in myocardial infarction (TIMI) phase II trial. *New Engl J Med* 1989;320:618-27.
14. Madsen JK, Grande P, Saunamaki K, Thayssen P, Kassis E, Eriksen U, et al. The Danish multicenter randomized study of invasive vs. conservative treatment in patients with inducible ischemia after thrombolysis in acute myocardial infarctions. *Circulation* 1997;96:748-55.
15. TIMI III Investigators. Effects of tissue plasminogen activator and a comparison of early invasive and conservative strategies in unstable angina and non-Q wave myocardial infarction: results of the TIMI IIIB trial. *Circulation* 1994; 89:1545-56.
16. Boden WE, O'Rourke RA, Crawford MH, Blaustein AS, Deedwania PC, Zoble RG, et al, for the Veterans Affairs Non-Q-Wave Infarction Strategies in Hospital (VANQWISH) Trial Investigators. Outcomes in patients with acute non-Q-wave myocardial infarction randomly assigned to an invasive as compared with a conservative management strategy. *N Engl J Med* 1998;338: 1785-92.
17. Yusuf S, Flather M, Pogue J, Hunt D, Varigos J, Piegas L et al, for the OASIS Registry Investigators. Variations between countries in invasive cardiac procedures and outcomes in patients with suspected unstable angina or myocardial infarction without initial ST elevation. *Lancet* 1998;352:507-14.

Reprint requests to: *Dr. James Brophy, Service de Cardiologie, Centre hospitalier de l'Université de Montréal, Pavillon Notre-Dame, 1560, rue Sherbrooke Est, Montréal QC H2L 4M1; fax 514 896-4710; jbroph@po-box.mcgill.ca*