Eur J Vasc Endovasc Surg **24**, 505–510 (2002)
doi:10.1053/ejvs.2002.1763, available online at http://www.idealibrary.com on IDEAL®

A Bayesian Hierarchical Approach to Comparative Audit for Carotid Surgery

G. Kuhan[1], E. C. Marshall[2], A. F. Abidia[1], I. C. Chetter[1] and P. T. McCollum*[1]

[1] *Academic Vascular Unit, Hull Royal Infirmary, Anlaby Road, Hull, HU3 2JZ, U.K.*
[2] *Department of Epidemiology and Public Health, Imperial College School of Medicine,*
St Mary's Campus, Norfolk Place, London, W2 1PG, U.K.

Objectives: *the aim of this study was to illustrate how a Bayesian hierarchical modelling approach can aid the reliable comparison of outcome rates between surgeons.*
Design: *retrospective analysis of prospective and retrospective data.*
Materials: *binary outcome data (death/stroke within 30 days), together with information on 15 possible risk factors specific for CEA were available on 836 CEAs performed by four vascular surgeons from 1992–99. The median patient age was 68 (range 38–86) years and 60% were men.*
Methods: *the model was developed using the WinBUGS software. After adjusting for patient-level risk factors, a cross-validatory approach was adopted to identify "divergent" performance. A ranking exercise was also carried out.*
Results: *the overall observed 30-day stroke/death rate was 3.9% (33/836). The model found diabetes, stroke and heart disease to be significant risk factors. There was no significant difference between the predicted and observed outcome rates for any surgeon (Bayesian* p*-value > 0.05). Each surgeon had a median rank of 3 with associated 95% CI 1.0–5.0, despite the variability of observed stroke/death rate from 2.9–4.4%. After risk adjustment, there was very little residual between-surgeon variability in outcome rate.*
Conclusions: *Bayesian hierarchical models can help to accurately quantify the uncertainty associated with surgeons' performance and rank.*

Key Words: Carotid endarterectomy; Comparative audit; Hierarchical models; Bayesian analysis; WinBUGS.

Introduction

Performance measures for surgeons and units are increasingly used to introduce professional accountability and to set clinical standards.[1] Traditionally, crude outcome rates have been used to compare surgical or hospital performances. However there is little doubt that true comparisons can only be achieved after adjustment for case-mix.[2–5] Even after such adjustment, classical "fixed-effects" methods employed to quantify a surgeon's effect on outcome of procedures often fail if missing or "noisy" data are encountered. Furthermore, comparing outcome of surgeons with small caseloads can be a difficult task.[6]

Several authors have highlighted such methodological concerns with the traditional approach to performance assessment,[7,8] advocating instead the use of hierarchical or multi-level models in which it is assumed that the (latent) surgeon-specific effects are drawn from some common distribution. The theoretical advantages of these models have been known for a long time[9] and they have been widely used in geographical epidemiology,[10,11] educational research,[12,13] and more recently, in health performance assessment.[7,14–18] Such models in general, and the Bayesian hierarchical approach[19] adopted here, in particular, allow the appropriate pooling of information across surgeons to overcome problems of small sample sizes.[16] This pre-empts the need for an arbitrary decision to be made as to whether there is enough information (i.e. a surgeon has performed enough operations) to allow reliable inference to be drawn. Bayesian hierarchical models also provide a coherent inference framework that permits the incorporation of multiple sources of variability – including that arising from missing covariate or outcome data.

Early identification of "divergent" surgeons[14] – surgeons whose outcome cannot be assumed to be drawn from the same distribution as that of their

* Please address all correspondence to: P. T. McCollum, Hull Royal Infirmary, Anlaby Road, Hull, HU3 2JZ, U.K.

1078–5884/02/060505 + 06 $35.00/0

peers – will hopefully prevent incidences similar to the Bristol heart scandal.[20] There is increasing pressure from the media and consumer groups to rank and produce "league tables" of performance of surgeons. Identifying surgeons who are not divergent yet are significantly above or below average could provide the opportunity to inform procedural practice with a view to improving outcome rates across the board. A surgeon's rank, however, has an associated uncertainty that must be quantified accurately before inferences regarding relative performances can be made.[21] Estimates of this uncertainty can be obtained very easily as a by-product of the current analysis.

It should be stressed that the methods outlined in this paper are only appropriate for retrospective analysis. There is an emerging literature proposing alternative methodologies for continuous time medical surveillance.[22–26] The aims of the current study were to illustrate the use of a Bayesian two level hierarchical model to identify "divergent" surgeons and to carry out a ranking exercise, which will enable reliable comparison of surgical performance.

Methods and Analysis

A series of 836 CEAs performed by four vascular surgeons from two units from 1992 to 1999 were available for analysis. Data on 67 risk factors were collected using pre-printed sheets and later entered into a database based on Access 97 (Microsoft, Redmond, U.S.A). The number of risk factors was reduced to 15 based on previous studies in the literature and on univariate analysis[27] (Table 1). The outcome endpoints were the occurrence of major stroke or death within 30 days of the procedure. Major stroke was defined by any neurological deficit lasting more than 7 days. One unit had prospectively collected data (41%) while the other unit had mixed prospective and retrospective data (59%). The outcome data were obtained by outpatient follow up by surgeons and case note review. Data on deaths occurring outside the hospital were obtained from the local registry office for deaths. Median patient age was 68 (range 38–86) years and 60% were men.

Table 1. Selected risk factors for the model.

Age
Sex
Hypertension
Heart disease
Diabetes
Stroke
Renal failure
Contralateral internal
carotid artery occlusion
Respiratory disease
Side of operation
Shunt
Patch
ASA grade
Surgeon
Vascular unit

A logistic regression model was fitted at the first level of our model including the risk factors and a surgeon-specific parameter. These surgeon-specific parameters reflect the effect of all unmeasured covariates on an individual's risk of death/stroke following a CEA performed by that particular surgeon. At the second level of the hierarchy, they were assumed to be drawn from a common Normal population distribution. The estimated variance of this distribution was used to quantify the residual variability in outcome between surgeons (on a logit scale) after adjustment for differences in patient case-mix. Since we adopt a Bayesian approach to inference, the model is completed through specification of prior distributions for all the model unknowns. In the absence of strong prior information, we assume uninformative priors for all regression coefficients, and the mean and variance of the Normal random effects distribution.

The regression coefficients and associated 95% Bayesian Credible Intervals (95% BCI) were computed via the Gibbs sampler, a Markov chain Monte Carlo (MCMC) technique,[28] which was implemented using WinBUGS software.[29] The exponential of these coefficients was taken to obtain Odds Ratio (OR) estimates and their 95% BCI's. Missing covariate data were imputed at each iteration of the Gibbs sampler and so the estimates of all parameters were fully adjusted for this additional associated uncertainty. The 30-day stroke or death risk for various combination of risk factors were obtained to identify a high-risk group.

The model was used to ascertain whether any surgeons could be considered "divergent". Following Spiegelhalter *et al.*[20] we excluded each surgeon in turn, fitted the model to the data from the remaining surgeons and then, given the observed case-mix of his/her patients, predicted the expected death/stroke rate for the excluded surgeon. The latter was labelled as divergent if his observed death/stroke rate was significantly different from that predicted. This comparison is summarised by way of the Bayesian probability, or p-value[30] defined here as $p =$ probability that the predicted rate is less than that observed. This quantity is computed via MCMC by introducing a dummy indicator variable for the excluded surgeon which takes the value 1 at a given iteration if, at that iteration, the value of his/her predicted mortality rate is less than or equal to the observed rate, and 0 otherwise. The mean of the values of this indicator variable over all iterations yields the required p-value. A probability close to 1 (say > 0.95) could be cause for concern since it indicates that prediction is nearly always lower than that observed, thus casting doubt on the model or, equivalently, suggesting that the mortality rate for that surgeon is significantly different from his peers.

The built in rank function in WinBUGS was used to identify surgeons whose effect is in the extremes of this distribution. Results presented here were based on multiple runs of length 10 000 following a burn-in of 1000 iterations to achieve convergence.[31]

Results

Observed stroke or death rate was 3.9% (33/836). The prevalence of missing data for each variable is given in Table 2. Intra operative shunt was used in 60% (501/839) and 61% (513/839) were repaired using a patch. Diabetes (OR = 2.65, 95% BCI (1.1, 6.2)), heart disease (OR = 2.23, 95% BCI (1.04, 4.73)) and previous stroke (OR = 3.13, 95% BCI (1.54, 7.12)) were the significant risk factors identified by the model. This agrees with the results from earlier modelling using backward elimination on SPSS at a 5% significance level.[27] The median risk of 30-day stroke or death for the 3 risk factors and the various combinations are shown in Table 3. An individual in the highest risk group (presence of all 3 risk factors) had a median stroke or death risk of 12.7% (3.2, 36.6) compared to the 0.6% (0.14, 1.60) of an individual presenting with none of the three major risk factors.

The observed 30-day stroke or death risk varied from 2.9–4.4% for the four vascular surgeons. The caseloads were also variable from 102 to 383. There was no significant difference between the observed and the predicted 30-day stroke or death risk for any of the surgeons (Table 4). No surgeon was labelled divergent and an assumption of a common distribution of surgeon effects seemed appropriate. Focussing now on this distribution, there was little variability in outcome after adjustment for significant risk factors between the four surgeons studied (Fig. 1). This variability can be quantified on a more interpretable scale by considering a patient's risk of death/stroke if operated on by a "high-risk" surgeon compared to a "low-risk" surgeon (OR = 1.3 95% BCI (1.05, 4.93)). A high-risk surgeon is defined as one whose surgeon-specific effect lies in the upper tail (mean +1 s.d.) of the random effects distribution, whereas a low-risk surgeon's effect lies in the lower tail (mean −1 s.d.).

The median rank for all surgeons was 3.0 with associated 95% BCI of 1.0–5.0.

Table 2. Prevalence of missing data for each risk factor.

Risk factor	(%) missing	Risk factor	(%) missing
Age	0 (0/836)	Contralateral ICA	9.5 (80/836)
Sex	0 (0/836)	Respiratory disease	0.2 (2/836)
Hypertension	0.8 (8/836)	Side of operation	0.2 (2/836)
Heart disease	0.9 (8/836)	Shunt	1.6 (14/836)
Diabetes	0.2 (2/836)	Patch	1.3 (11/836)
Stroke	0.2 (2/836)	Renal failure	0.8 (7/836)

Table 3. Thirty day stroke or death rates for each risk factors.

Risk factor	Probability of 30 day stroke/ death (mean)	95% credible interval
None	0.60	(0.14, 1.60)
Stroke	1.11	(0.27, 3.76)
Diabetes	1.31	(0.28, 5.23)
Heart disease	1.58	(0.47, 4.54)
Diabetes and stroke	2.87	(0.64, 10.91)
Stroke and heart disease	3.42	(0.94, 10.6)
Diabetes and heart disease	5.38	(1.01, 15.10)
Diabetes, stroke and heart disease	12.71	(3.2, 36.60)

Discussion

Since the introduction of clinical governance there has been greater emphasis to introduce performance measures for surgeons and units. Case-mix adjusted outcome rates are needed for accurate comparison of surgical performance. Comparison of median predicted rates after risk adjustment with the observed outcome rates indicate surgeons 2 and 4 to have higher mortality rate than that expected. However the 95% credible intervals and the Bayesian *p*-values show that none of the predicted outcome was significantly different from the observed (Table 4). That is, there were no divergent performers amongst the population of surgeons in this study. Due to the high level of variability associated with the estimates of surgical performance, the estimated rank for each surgeon had huge associated credible intervals covering the whole range (Fig. 1). This indicates how league tables, which do not consider risk adjustment for patient-level factors and do not account for random variation, should be interpreted with caution. In a previous study a simpler fixed-effects logistic regression model was implemented to compare the surgeons performance and the difference in the outcomes after risk adjustment was not found to be statistically significant.[27] Comparison of this model to the more complex Bayesian two level hierarchical model is made throughout the course of the discussion.

The current model identified diabetes, heart disease and previous stroke to be significant risk factors for 30-day stroke or death following CEA. Diabetes is a well-known risk factor for stroke and myocardial

Table 4. Observed and predicted 30 day stroke/ death risk for each surgeon.

Surgeon	No of procedures	Observed 30 day stroke/ death risk (%)	Median predicted 30 day stroke/ death risk (%)	95% credible interval	Probability of predicted risk to be less than observed
1	237	4.2	4.7	(0.4, 12.6)	0.5612
2	114	4.4	4.1	(0.0, 9.6)	0.7011
3	102	2.9	4.0	(0.0, 9.8)	0.4712
4	383	3.9	2.3	(0.2, 5.7)	0.9180

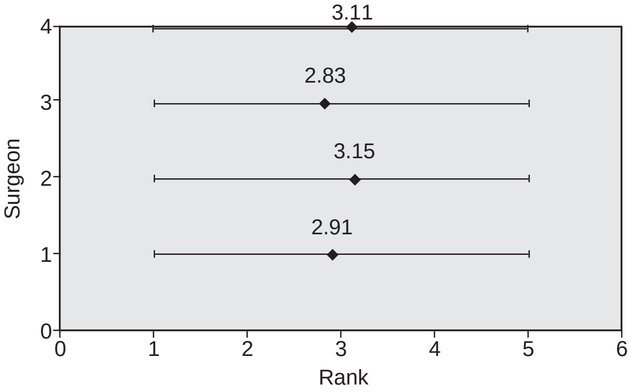

Fig. 1. Mean rank and 95% intervals for each surgeon.

infarction.[32] In the North American Symptomatic Carotid Endarterectomy Trial diabetes was associated with a two-fold increase in the risk of peri-operative stroke or death.[33] Similarly pre existing heart disease and stroke are associated with an increased operative risk of stroke or death.[34,35] As expected, the same three risk factors emerged as significant in the classical approach.[27] As with the latter, the current model allowed estimation of risk of stroke or death for the various combinations of risk factors, which can be used to counsel patients before the operation (Table 3).

In the traditional approach the surgeon specific variables were introduced as fixed effects into the developed risk model. Odds ratios for each surgeon, relative to a designated reference, were then calculated based only on the data specific to that surgeon. Adopting this approach, a surgeon who performed few operations is more likely to have an extreme odds ratio due to chance alone. In contrast, the hierarchical model pools data across all surgeons to calculate the odds ratios and confidence limits thus making comparative audit more robust, and in our opinion, more reliable. In general, estimates based on large patient populations are preserved yet those based on sparse data are shrunk towards the population average. There is the danger of "over-shrinking" and potentially masking true, but low-volume, outliers. However, given the potential consequences of declaring a surgeon to be an outlier, we advocate the current, more conservative approach.

One of the aims of the current study was to illustrate how a hierarchical modelling approach allows the reliable estimation of the uncertainty associated with a surgeon's effect on outcome. The observed outcome rates varied from 2.9 to 4.4% for the four surgeons (Table 4). After risk adjustment, however, there was very little residual variability in the risk of death between surgeons. What variability there was could be attributed to differences in surgical performance but could equally be reflecting systematic yet unmeasured differences in patient case-mix. There will always be variability in outcome and one could argue that an aim of comparative audit should be to try and identify the cause of this variability through sensitive investigation.

Missing data are a major source of bias. Missing data however are inevitable in routinely collected administrative data and reflect the quality of data collection. In the current study the missing data was 1.4% for the selected variables. Contra lateral internal carotid disease had the highest percentage of missing data (9.5%). In the traditional analysis[27] 98 procedures were excluded due to missing data items. Many critics argue that 98 CEAs reflects approximately one to two years of workload in most vascular units and measures should be taken to incorporate that data. In the current study none of the missing data were excluded. Missing data was imputed at each iteration of the Gibbs sampler the additional source of uncertainty acknowledged in all estimates. Although we assume in the current paper, that those data were missing at random, the flexibility of the Bayesian approach is such that informative missing data mechanisms may easily be modelled.[36]

There were shortcomings as well as strengths in the methods employed in the current paper. The current model was developed on routinely collected data which may not be perfect.[14] After risk adjustment it was assumed the residual variability could be assigned to differences in surgical performance although like many performance assessment methods, ours is open to the criticism of insufficient case-mix adjustment. The same risk factors of 30-day mortality were identified under both the Classical and Bayesian approaches and both analyses led to a conclusion of no difference in performance between the surgeons.[27] The advantage of the approach taken in this study is that outcome data from all surgeons can be incorporated in one coherent inference framework, including those with missing covariate information or low caseload. The Bayesian approach avoids the need for arbitrary decision making *a priori* regarding the sufficiency (or not) of the information relating to a particular surgeon. We are also able to quantify both the variability in the surgeons' ranks and that in outcome between surgeons, after adjusting for major patient-specific risk factors. Advances in computing have meant that the MCMC methods are now increasingly used to tackle wide variety of problems in statistics. The current model could easily be extended to a third level to compare performances between vascular units after adjustment for patient and surgeon characteristics, or used for the comparative audit of other index procedures in vascular surgery.

Implications of identifying performance divergence will raise great concern for the patients and the surgeon involved. However it should be emphasised that risk models are tools for comparative audit and should not be applied blindly. Surgical competency should also be judged at a clinical level and not purely on outcome measures. Furthermore, identifying surgeon-effects to be in the tails of their distribution is not necessarily cause for alarm – as Poloneicki[37] points out, half of all surgeons will be below average. Nonetheless, this information could provide valuable feedback to improve practice.

Acknowledgement

This study was funded by a fellowship grant from the Royal College of Surgeons of England. There is no conflict of interests.

References

1 NHS management executive and Department of Health. Clinical governance: Quality in the new NHS 1999.

2 Bernstein AD, Parsonnet V. Bedside estimation of risk as an aid for decision-making in cardiac surgery. *Ann Thorac Surg* 2000; **69**: 823–828.

3 Champion HR, Sacco WJ, Copes WS *et al*. A revision of the Trauma Score. *J Trauma* 1989; **29**: 623–629.

4 Copeland GP, Jones D, Walters M. POSSUM: a scoring system for surgical audit. *Br J Surg* 1991; **78**: 355–360.

5 Edwards FH, Albus RA, Zajtchuk R *et al*. Use of a Bayesian statistical model for risk assessment in coronary artery surgery. *Ann Thorac Surg* 1988; **45**: 437–440.

6 Irvine CD, Grayson D, Lusby RJ. Clinical governance and the vascular surgeon. *Br J Surg* 2000; **87**: 766–770.

7 Normand S-L, Glickman ME, Gatsonis CA. Statistical methods for profiling providers of medical care: issues and applications. *J Am Stat Assoc* 1997; **92**: 803–814.

8 Thomas N, Longford NT, Rolph JE. Empirical Bayes methods for estimating hospital-specific mortality rates. *Stat Med* 1994; **13**: 889–903.

9 Efron B, Morris C. Steins paradox in statistics. *Sci Am* 1977; **236**: 119–227.

10 Bernardinelli L, Montomoli C. Empirical Bayes versus fully Bayesian analysis of geographical variation in disease risk. *Stat Med* 1992; **11**: 983–1007.

11 Clayton D, Kaldor J. Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics* 1987; **43**: 671–681.

12 Bryk AS, Raudenbush AS. Hierarchical linear models. Newbury Park CA: Sage; 1992.

13 Goldstein H, Rasbash J, Yang M, Woodhouse G, Pan H, Nuttal D. A multilevel analysis of school examination results. *Oxford Rev Ed* 1993; **19**: 425–433.

14 Aylin P, Alves B, Best N *et al*. Comparison of UK paediatric cardiac surgical performance by analysis of routinely collected data 1984–96: was Bristol an outlier? *Lancet* 2001; **358**: 181–187.

15 Goldstein H, Spiegelhalter DJ. League tables and their limitations: statistical issues in comparisons of institutional performance. *J Royal Stat Soc* 1996; **159**: 385–443.

16 Morris CN, Christiansen CL. Hierarchical models for ranking and for identifying extremes, with applications. In *Bayesian Statistics V* (eds J. O. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), pp., 278–295. Oxford, Oxford University Press.

17 Rice N, Leyland A. Multilevel models: applications to health data. *J Health Ser Res Policy* 1996; **1**: 154–164.

18 Spiegelhalter DJ. Mortality and volume of cases in paediatric cardiac surgery: retrospective study based on routinely collected data. *BMJ* 2002; **324**: 261–263.

19 Gelman A, Carlin JB, Stern HS, Rubin DB. Bayesian data analysis. London: Chapman and Hall; 1995.

20 Spiegelhalter DJ, Aylin P, Evans SJW, Murray GD, Best NG. Commissioned analysis of surgical performance using routine data: lessons from the Bristol inquiry. *J Royal Stat Soc* 2002; **165**: 1–31.

21 Marshall EC, Spiegelhalter DJ. Reliability of league tables of *in vitro* fertilisation clinics: retrospective analysis of live birth rates. *BMJ* 1998; **316**: 1701–1704.

22 Clark DE, Cushing BM, Bredenberg CE. Monitoring hospital trauma mortality using statistical process control methods. *J Am Coll Surg* 1998; **186**: 630–635.

23 Mohammed MA, Cheng KK, Rouse A, Marshall T. Bristol, shipman, and clinical governance: Shewhart's forgotten lessons. *Lancet* 2001; **357**: 463–467.

24 Poloniecki J, Valencia O, Littlejohns P. Cumulative risk adjusted mortality chart for detecting changes in death rate: observational study of heart surgery. *BMJ* 1998; **316**: 1697–1700.

25 Rossi G, Lampugnani L, Marchi M. An approximate CUSUM procedure for surveillance of health events. *Stat Med* 1999; **18**: 2111–2122.

26 Williamson GD, Weatherby HG. A monitoring system for detecting aberrations in public health surveillance reports. *Stat Med* 1999; **18**: 3283–3298.

27 Kuhan G, Gardiner ED, Abidia AF *et al*. Risk modelling study for carotid endarterectomy. *Br J Surg* 2001; **88**: 1590–1594.

28 Gilks WR, Richardson S, Spiegelhalter DJ. Markov chain montecarlo methods in practice. New York: Chapman & Hall; 1995.

29 Spiegelhalter DJ, Thomas A, Best NG, Gilks WR. BUGS: Bayesian inference using Gibbs sampling. Version 0.5. Cambridge: MRC Biostatistics Unit. 1995.

30 Gelman A, Meng X-L, Stern H. Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica* 1996; **6**: 733–807.

31 Cowles MK, Carlin BP. Markov chain Monte Carlo convergence diagnostics – a comparative review. *J Am Stat Assoc* 1996; **91**: 883–904.

32 Wolf PA, D'Agostino RB, Belanger AJ, Kannel WB. Probability of stroke: a risk profile from the Framingham Study. *Stroke* 1991; **22**: 312–318.

33 Barnett HJ, Taylor DW, Eliasziw M *et al*. Benefit of carotid endarterectomy in patients with symptomatic moderate or severe stenosis. North American Symptomatic Carotid Endarterectomy Trial Collaborators. *N Engl J Med* 1998; **339**: 1415–1425.

34 McCrory DC, Goldstein LB, Samsa GP *et al*. Predicting complications of carotid endarterectomy. *Stroke* 1993; **24**: 1285–1291.

35 Riles TS, Imparato AM. Preoperative risk factors for carotid endarterectomy. *Stroke* 1994; **25**: 2096–2097.

36 Best NG, Spiegelhalter DJ, Thomas A, Brayne CEG. Bayesian analysis of realistically complex models. *J Royal Stat Soc* 1996; **159**: 323–342.

37 Poloniecki J. Half of all doctors are below average. *BMJ* 1998; **316**: 1734–1736.

Accepted 13 August 2002