Author's response to reviews

Title: Use of hierarchical models to evaluate performance of cardiac surgery centres in the Italian CABG outcome study

Authors:

Paola D'Errigo (paola.derrigo@iss.it)
Maria E Tosti (mariaelena.tosti@iss.it)
Danilo Fusco (fusco@asplazio.it)
Carlo A Perucci (dir@asplazio.it)
Fulvia Seccareccia (fulvia.seccareccia@iss.it)

Version: 3 Date: 13 April 2007

Author's response to reviews: see over
Dear Referees,

We appreciate your comments. The following is a response to all of those comments. Also, your suggestions have been incorporated into the manuscript, as noted below. We resubmitted a new version per reviewers’ comments, wherein we yellow highlighted all the corrections and we supplemented the list of references as required.

Response to David Spiegelhalter comments:

Page 3: We agree that the choice of a random-effect rather than a fixed effect model deserves more emphasis. The 3rd paragraph of the background section now presents a more detailed argumentation, considers the references you suggested, and explains as follows: “Comparative data, especially if adjusted using a risk function empirically derived from the observed population, serve many purposes and have the potential to provide insight and improve the quality of care. Unfortunately, the existing standard single-level models, usually adopted in outcome studies, treat all patients as independent observations and ignore that they are grouped within hospitals. Patients undergoing a surgical intervention within the same hospital may be correlated, violating one of the basic assumptions of traditional regression analysis. Hierarchical (or multilevel) models consider the hospitals involved in the study as a random sample from a population of hospitals and partition the random variability of the data into two parts: that between different patients and that between different hospitals. The hospital-specific random error component is interpreted as representing differences in hospital quality. Consequently, hierarchical modelling is strongly advocated as a more appropriate statistical method for dealing with outcomes data in which individual patients are clustered within hospitals. [8-12] Moreover, hierarchical models account for regression-to-the mean by providing estimates of standardized mortality rates that are appropriately less extreme than the observed. Hospitals with small sample sizes are more likely to have extreme observed mortality rates because of chance variation, their true rates usually being less extreme than their observed. Estimates from hierarchical models provide more accurate assessments, with the most improvement for smaller hospitals because they experience greater regression-to-the mean. [11,13-15].”

Similar considerations have also been reported in the discussion section (PAGE 3). The same applies to your comment on page 10 as well.

Page 7: We have modified the last paragraph of the Statistical Methods section (PAGE 7) to clarify the question you have arisen. Actually, the focal point was not to assess whether centres were above or below the mean, but to rank centres performance and to identify low and high outliers.

The paragraph now explains as follows: “For each clinical centre, the group-level residual represents the distance between its estimated mortality (hospital’s mean effect) and the overall mortality estimated by the hierarchical model. These residuals can be used to rank centres
performance and to identify low and high outliers. Clinical centres performances were compared with the overall mean by graphically presenting group-level residuals with their 95% confidence intervals\textsuperscript{17}. The residuals were ordered from the smallest to the largest.\textsuperscript{20-22} Centres with residuals significantly less than zero (CI not overlapping the zero row) performed better than the overall population whereas centres with residuals significantly higher than zero performed worse.\textsuperscript{20-22} Spearman correlation coefficient of ranks and a scatter-plot were used to compare results from standard logistic regression with those obtained by the hierarchical model.”

Page 8: Concerning your observation on the use of center-level covariates together with a factor for clinical centre, we need to clarify that our approach was based on the knowledge that in “mixed effect” models some patient descriptors (i.e. gender) and hospital characteristics (i.e. teaching/non-teaching), being limited in the number of possible values, can be considered fixed. Therefore, in our model the “random effect” is only due to the variable “clinical centre”. On the other hand, a preliminary data analysis based on a classical single level approach revealed that the predictive power of the most relevant hospital-level covariates was negligible. In the first paragraph of (PAGE 9) the sentence “Group-level covariates were tested with a classical single-level logistic model, but not significant contribution was found” has been added.

However, our approach was based on suggestions reported in literature (DeLong E et al. Am Heart J 2003, Austin PC et al. Am Heart J 2003). In particular, we considered the observations of DeLong: “When assessing provider performance regarding surgical outcomes, an appropriate mixed model might include fixed effects for patient-level case-mix variables and hospital-level teaching status, and would include the individual providers as random effects”.

Page 9: We agree that more detailed results and comments on the risk-adjusted mortality rates are necessary. The last paragraph of (PAGE 9) now explains as follows: “Second-level residuals, obtained by the hierarchical model, were used to evaluate the effect of each clinical centre and are presented in figure 1 with their confidence intervals. In particular, 3 of the 64 centres analysed (4.7%) had a residual significantly lower than zero (RAMR ranging from 0.28 to 0.73) and 11 centres (17.2%) significantly higher (RAMR ranging from 5.25 to 10.44); 50 centres (78.1%) showed a residual not significantly different from zero. Using the classical approach 8 centres showed estimated mortality rates lower than the mean (RAMR ranging from 0.26 to 1.32) and 7 centres showed estimated mortality rates significantly higher (RAMR ranging from 4.37 to 8.76). The 3 centres identified by the hierarchical model as low outliers were found to perform better than the mean also using the classical approach. On the contrary, among the high outliers identified by the hierarchical model 4 centres performed not differently from the mean using the classical approach.”

Page 10: Your considerations have been dealt previously (see Page 3 comments).

Figure 1: The figure and its legend are now more clear and legible. Being inspired by the work of Merlo J et al. (J Epidemiol Community Health 2005) the vertical scale has been relabeled and is now more informative.

Response to Patrick S Romano comments:
Page 5, Par. 2: We agree that this section of Statistical methods needs more details. The second paragraph of (PAGE 5) now explains as follows: “Univariate analyses were used to compute crude odds ratios for all potential confounding factors. We identified the best risk adjustment model using a single-level multiple logistic regression to account for joint confounding. The potential predictive variables were selected using a conventional stepwise method with a cross validation procedure. First, all possible confounding variables were included in the model. Second, a backward stepwise method was used in order to identify independent associations with the outcome. Patients were randomly split into two equal-size samples: sample I was used to build the predictive model (n = 17,231); sample II was used as an independent database for model validation (n = 17,079). The entire data set was finally used to estimate the definitive coefficients and calculate their p-values, in order to provide more precise parameter estimates. A set of biologically plausible interaction hypotheses defined “a priori” was also tested (gender and age with the variable identifying clinical centres) [7].”

Page 5, Par. 3: We agree that the method used to identify outliers was not well explained. For this reason, the sentence “To identify hospitals significantly different from the mean the exact method was used [16]” has been added.

Page 6, end: In order to estimate 95% CI for the group level residuals, procedure suggested by Goldsetin and Healy (J R Statist Soc, A 1995, prev. ref. 16 now ref. 20) was followed. Differently from their approach, we used the $z_p$ value of 1.96 instead of 1.396, being not interested in the comparison of each pair of centres, but only in the comparison between each centre and the population mean mortality.

Page 8, Par. 2: We agree that this paragraph seems to be incomplete. It has been modified and now it explains as follows: “The single-level logistic model confirmed the following factors to be the most strongly associated with 30 day mortality after CABG surgery: emergency (OR = 3.89, CI = 3.12 - 4.85), shock (OR = 3.44, CI = 2.48 - 4.78), dialysis (OR = 6.66, CI = 4.91 - 9.04), pulmonary hypertension (OR = 6.29, CI = 3.69 - 10.71), and ejection fraction < 30 (OR = 3.14, CI = 2.35 - 4.20). A strong association was also found for previous CABG intervention (OR = 2.86, CI = 2.10 - 3.89).” However, in Table 2 all factors associated with mortality were reported.

Page 9, Par. 3: As already explained to the other reviewer, we agree that more detailed results and comments on the risk-adjusted mortality rates are necessary. The last paragraph of (PAGE 9) now explains as follows: “Second-level residuals, obtained by the hierarchical model, were used to evaluate the effect of each clinical centre and are presented in figure 1 with their confidence intervals. In particular, 3 of the 64 centres analysed (4.7%) had a residual significantly lower than zero (RAMR ranging from 0.28 to 0.73) and 11 centres (17.2%) significantly higher (RAMR ranging from 5.25 to 10.44); 50 centres (78.1%) showed a residual not significantly different from zero. Using the classical approach 8 centres showed estimated mortality rates lower than the mean (RAMR ranging from 0.26 to 1.32) and 7 centres showed estimated mortality rates significantly higher (RAMR ranging from 4.37 to 8.76). The 3 centres identified by the hierarchical model as low outliers were found to perform better than the mean also using the classical approach. On the contrary, among the high outliers identified by the hierarchical model 4 centres performed not differently from the mean using the classical approach.”

Page 9, Par. 5: We agree that we did not adequately summarize the existing literature in this area. We have appreciated your suggestions and enriched our bibliography with some others relevant
papers on the matter in both the Introduction and the Discussion section. Moreover, we have tried to better discuss and compare our results with those from other researchers and the cited paragraph now explains as follows: “In the single-level model, built on the Italian CABG study data, both demographic variables and comorbidities, recognized as important risk factors, were used to adjust outcome estimates. Therefore, the assumption was that any residual differences in outcome between centres should only reflect differences in quality of care. The same algorithm was used to build a hierarchical logistic model. Effect of patient characteristics on outcome (coefficients of factors) are comparable using both the single and the multilevel model, but multilevel SEs result greater than the others.

As other authors have underlined, hierarchical regression models could result in different RAMR from that obtained using conventional logistic regression, even though only patient-level characteristics are used in both models. [32] Actually, in this work some negligible differences between RAMRs obtained by the single and the multilevel approaches were found, but the overall findings of the study remain comparable.

The multilevel analysis showed that 10.1% (ICC) of the differences in the adjusted mortality rates were attributable to differences between centres. This amount of variability explained by the group-level variable is higher than that reported in other studies. Hannan et al, by applying a multilevel model on New York State CABG Registry data, found a percentage of variability not higher than 3.6%, attributable to the hierarchy and indicating only slight intraclass correlation[18]. The 10.1% of variability identified in this work seems to be more similar to the 12.6% identified by Austin et al. in their work on Myocardial Infarction (American Heart Journal, 2003) [32].”

**Page 10, Par. 5:** Concerning your observation on the comparison between our and Hannan’s results, we need to clarify that we used the ICC in order to evaluate the proportion of unexplained variability in mortality accounted for by differences among hospitals. The formula we adopted is that proposed by Snijders and Bosker and is the same used by Hannan in his work on the New York data. In this work, Hannan aimed to assess the predictive performance of models by comparing their results with hospital performance over time and used the ICC statistic only to describe the proportion of variability due to the nested structure of data. He commented the findings as follows: “Table 2 also presents the approximate ICC for the hierarchical models. These range from a low of 0.0156 to a high of 0.0365, indicating only slight intraclass clustering.” However, since the ICC formula (reported in the text, PAGE 7) does not make any reference to time measurements, we think that comparing ICC from different studies is allowed and does not have any negative implications.