ORIGINAL RESEARCH

# The Dutch Hospital Standardised Mortality Ratio (HSMR) method and cardiac surgery: benchmarking in a national cohort using hospital administration data versus a clinical database

S Siregar,[1] M E Pouw,[2] K G M Moons,[3] M I M Versteegh,[4] M L Bots,[3] Y van der Graaf,[3] C J Kalkman,[2] L A van Herwerden,[1] R H H Groenwold[3]

## ABSTRACT

**Objective** To compare the accuracy of data from hospital administration databases and a national clinical cardiac surgery database and to compare the performance of the Dutch hospital standardised mortality ratio (HSMR) method and the logistic European System for Cardiac Operative Risk Evaluation, for the purpose of benchmarking of mortality across hospitals.

**Methods** Information on all patients undergoing cardiac surgery between 1 January 2007 and 31 December 2010 in 10 centres was extracted from The Netherlands Association for Cardio-Thoracic Surgery database and the Hospital Discharge Registry. The number of cardiac surgery interventions was compared between both databases. The European System for Cardiac Operative Risk Evaluation and hospital standardised mortality ratio models were updated in the study population and compared using the C-statistic, calibration plots and the Brier-score.

**Results** The number of cardiac surgery interventions performed could not be assessed using the administrative database as the intervention code was incorrect in 1.4–26.3%, depending on the type of intervention. In 7.3% no intervention code was registered. The updated administrative model was inferior to the updated clinical model with respect to discrimination (c-statistic of 0.77 vs 0.85, p<0.001) and calibration (Brier Score of 2.8% vs 2.6%, p<0.001, maximum score 3.0%). Two average performing hospitals according to the clinical model became outliers when benchmarking was performed using the administrative model.

**Conclusions** In cardiac surgery, administrative data are less suitable than clinical data for the purpose of benchmarking. The use of either administrative or clinical risk-adjustment models can affect the outlier status of hospitals. Risk-adjustment models including procedure-specific clinical risk factors are recommended.

## INTRODUCTION

A valid comparison of outcomes between hospitals or healthcare providers (benchmarking) requires adjustment for severity of the health condition of patients and the performed interventions, often referred to as case-mix differences.[1–3] For this purpose prediction models have been developed to estimate risk-adjusted outcomes across hospitals. Most of these models are based on routinely collected administrative hospital data. For example, the hospital standardised mortality ratio (HSMR), first developed by Jarman in 1999 for the UK, is a risk-adjusted mortality rate calculated using prediction models based on administrative data.[4] Because administrative data are collected for other purposes, they are easily available, and thus the use of these data for benchmarking is cheap and requires relatively little extra effort.

However, administrative databases are often criticised for being inaccurate, incomplete and containing limited information.[5–9] As a consequence, comparisons of risk-adjusted outcome rates between healthcare providers that are based on administrative database data might be unreliable, leading to unjustified criticism. For that reason clinical databases with corresponding clinical prediction models have been developed (eg, European System for Cardiac Operative Risk Evaluation (EuroSCORE) and Society of Thoracic Surgeons risk models in cardiac surgery) that include multiple clinical predictors for mortality.[10–12] The EuroSCORE is a prediction model that was specifically designed to predict the risk of operative mortality related to cardiac surgery using 18 demographic and risk factors. The EuroSCORE can thus be used to adjust for differences in case mix in the comparison between healthcare providers. Models based on clinical risk factors are claimed to have a better predictive performance, resulting in improved risk adjustment, and enable valid comparison of outcomes across centres.[5–7 13] The downside is that clinical databases are more expensive; they comprise information that is obtained by active data collection by dedicated individuals and thus require continuous maintenance. Previous studies have not come to a conclusive answer to the question if clinical risk factors are necessary for adequate risk adjustment. Some concluded that administrative data are sufficient to enable benchmarking, whereas others show a clear inferiority and insufficiency when compared with clinical data.[6–8 13–19]

The aim of our study was to analyse whether a risk adjustment model based on administrative data allows for adequate benchmarking in cardiac surgery. Using a nationwide cohort of cardiac surgery patients, we assessed the accuracy of an administrative database and the predictive

▸ Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/heartjnl-2013-304645).

[1]Department of Cardio-Thoracic Surgery, University Medical Centre Utrecht, Utrecht, The Netherlands
[2]Department of Anesthesiology, University Medical Center Utrecht, Utrecht, The Netherlands
[3]Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands
[4]Department of Cardio-Thoracic Surgery, Leiden University Medical Center, Leiden, The Netherlands

**Correspondence to**
Dr Sabrina Siregar, Department of Cardio-Thoracic Surgery, University Medical Centre Utrecht, Utrecht 3508GA, The Netherlands; s.siregar@umcutrecht.nl

SS and MEP contributed equally to this study.

Received 30 July 2013
Revised 26 September 2013
Accepted 16 October 2013

**To cite:** Siregar S, Pouw ME, Moons KGM, *et al*. *Heart* Published Online First: [*please include* Day Month Year] doi:10.1136/heartjnl-2013-304645

performance of administrative models in comparison with a clinical database and the clinical EuroSCORE model.[20]

## METHODS

### Data

EuroSCORE and administrative variables of a national cohort of cardiac surgery patients in The Netherlands have been collected in two separate databases: (1) The adult national cardiac surgery database of the Netherlands Association of Thoracic Surgery (NVT) and (2) The National Hospital Discharge Registry (HDR) of The Netherlands.[20–22]

#### The adult national cardiac surgery database of the Netherlands Association of Thoracic Surgery

This clinical database has a national coverage with participation of all 16 centres performing cardiac surgery in The Netherlands.[20] All patients undergoing cardiac surgery excluding trans catheter aortic valve implantation, circulatory assist devices and pacemakers, are included in the database. Ten out of 16 cardiac centres participated in our study, in which 34 229 consecutive procedures were performed between 1 January 2007 and 31 December 2010. Procedures with incomplete data were excluded (N=218, 0.6%), resulting in 34 011 procedures for further analyses. The dataset consisted of predictors for mortality as listed in table 1, defined according to the EuroSCORE.[10] The EuroSCORE was developed to estimate the operative risk of mortality related to cardiac surgery (within 30 days and/or during the same hospital admission).[11] In this study, the EuroSCORE was used to estimate the risk of inhospital mortality.

#### The Hospital Discharge Registry

The HDR contains administrative data of all 10 hospitals included in this study. The dataset consists of patient characteristics and admission details such as age, comorbidity, sex and urgency of admission. For interventions the International Classifications of Health Interventions coding system is used and for diagnoses the International Classification of Disease-9.[23] The Dutch HSMR method is based on the HDR database and uses 50 risk-adjustment models, each for one specific group of diagnoses. The models estimate the risk of mortality for patients with a diagnosis belonging to the specific diagnose group.[21]

#### Linkage of datasets

In order to compare the HDR and the NVT databases and the models based on them, information on cardiac surgery

**Table 1** Variables recorded in the administrative database (HDR) and the clinical database (NVT)

| Administrative variables | N (%) N=26 178 | OR in updated model | Clinical variables | N (%) N=26 178 | OR in updated model |
|---|---|---|---|---|---|
| Age <25 years (categories of 5 years up to >85 years) | 66.5 (± 10.7) | Reference 0.29–1.01 | Age (continuous) | 66.6 (±10.7) | 1.06*** |
| Female sex | 7714 (29.5) | 1.44*** | Female sex | 7714 (29.5) | 1.33*** |
| Acute myocardial infarction | 1899 (7.3) | 1.25 | Recent myocardial infarction (<90 days) | 3191 (12.2) | 1.57*** |
| Congestive heart failure | 696 (2.7) | 4.11*** | LVEF 30–50%<br>LVEF <30% | 4165 (15.9)<br>1324 (5.1) | 1.69***<br>2.95*** |
| Pulmonary disease | 623 (2.4) | – | Pulmonary disease | 3019 (11.5) | 1.79*** |
| Renal disease | 293 (1.1) | 3.62*** | Serum creatine >200 μmol/L | 464 (1.8) | 2.79*** |
| Urgency | 3292 (12.6) | 2.20*** | Emergency operation | 1317 (5.0) | 2.38*** |
| Peripheral vascular disease | 551 (2.1) | 2.17*** | Extracardiac arteriopathy | 3202 (12.2) | 1.83*** |
| Cerebral vascular accident | 241 (0.9) | 2.75*** | Neurological dysfunction | 780 (3.0) | 1.26 |
| Peptic ulcer | 51 (0.2) | 4.36*** | Previous cardiac surgery | 1709 (6.5) | 2.78*** |
| Social economic status | | | Systolic pulmonary pressure >60 mm Hg | 606 (2.3) | 1.97*** |
| Lowest | 5450 (16.5%) | Reference | | | |
| Below average | 5379 (16.3%) | 0.89 | Active endocarditis | 216 (0.8) | 1.45 |
| Average | 4999 (15.1%) | 0.79* | Unstable angina | 15 776 (6.0) | 1.95*** |
| Above average | 5801 (17.5%) | 0.70** | Critical preoperative state | 983 (3.8) | 2.51*** |
| Highest | 4541 (13.7%) | 0.95 | Ventricular septal rupture | 47 (0.2) | 3.93*** |
| Unknown | 6925 (20.9%) | | | | |
| Year of discharge | | | Other than isolated CABG | 11 809 (45.1) | 3.43*** |
| 2007 | 6829 (20.6%) | reference | Thoracic aortic surgery | 1258 (4.8) | 2.75*** |
| 2008 | 6697 (20.2%) | 1.04 | | | |
| 2009 | 6941 (21.0%) | 0.83 | | | |
| 2010 | 5711 (17.3%) | 0.69*** | | | |
| Unknown | 6917 (20.9%) | | | | |
| Admission from | | | | | |
| Home | 19 907 (60.2%) | reference | | | |
| Nursing home | 145 (0.4%) | 3.41*** | | | |
| General hospital | 4952 (15.0%) | 1.27** | | | |
| Academic centre | 1174 (3.5%) | 1.38* | | | |
| Unknown | 6917 (20.9%) | | | | |

For dichotomous variables the number of patients and percentage of total population is reported; for continuous variables the mean and standard deviation. *p<0.05; **p<0.01; ***p<0.001.
LVEF, left ventricular ejection fraction.

interventions was required from both databases. Therefore, the HDR and NVT databases were linked to identify similar records. The HDR and NVT databases contain anonymised data, meaning no directly identifying information is stored. Records from both databases were linked to the municipal registries based on date of birth, gender and zip code, and were subsequently linked to each other. The linkage was performed by Statistics Netherlands and is described in previous publications.[20 24 25] The linkage of datasets is illustrated in the flow chart shown in figure 1. In total 26 178 (77%) records from the NVT database could be linked to a record in the HDR database and were used for further analyses. The predicted mortality according to the logistic EuroSCORE did not differ between the linked and the non-linked population (median 3.7%). Reasons for failed linkage were: the HDR record could not be linked to the municipal registries or no HDR record existed for the specific intervention (18.7%), the NVT record could not be linked to the municipal registries (2.7%) or no administrative model was available for the record (1.6%). The linkage of the HDR database to the municipal registries caused most linkage failure, as only four digits (out of six) of the zip code were available in the HDR database.

### Comparison of data between the NVT and HDR databases: intervention and inhospital mortality

The type of intervention and the outcome inhospital mortality were compared between the registries. Considering the fact that the NVT and the HDR registries use other risk factors for risk adjustment, these were not compared. The NVT database was used as the reference for the type of intervention, because this information is collected by the surgeons themselves. The HDR database was used as the reference for inhospital mortality, as the date of mortality is extracted directly from the up-to-date municipality registers. The comparison of inhospital mortality between both databases was performed on patient level (as opposed to intervention level), to avoid persons being counted multiple times for mortality.

### Comparison of risk-adjustment models

#### The administrative and clinical model

The Dutch HSMR method (models based on administrative data) and the logistic EuroSCORE (model based on clinical data) were applied in their original form to our study population, to predict the risk of inhospital mortality in our study population.[10 21] These models will subsequently be called Administrative.1 and Clinical.1. Existing risk-adjustment models can be updated to a new study population. Updated models are adjusted to the characteristics of that population and are likely to show improved generalisability.[26] There are several methods to update a risk-adjustment model.[26] As cardiac surgery interventions are incorporated in multiple Dutch HSMR models (ie, several diagnosis groups), one model for cardiac surgery was constructed using stepwise backward selection based on Akaike's Information Criterion.[27] This means that the intercept and the coefficients of all included covariates were estimated again in our study population and only relevant risk factors were included in the updated model. To update the EuroSCORE model, the intercept and the coefficients of all included covariates were also estimated again in our study population. This resulted in the updated models Administrative.2 and Clinical.2. The models can be updated even more thoroughly by inclusion of interaction terms, in order to maximise risk adjustment in our study population. Thus, first-order interaction terms between all covariates were added to the updated models, resulting in the models Administrative.3 and Clinical.3.[27]

#### Comparison of model performance

The predictive performance of a risk-adjustment model is quantified by means of calibration and discrimination. Discrimination refers to the ability of a model to differentiate

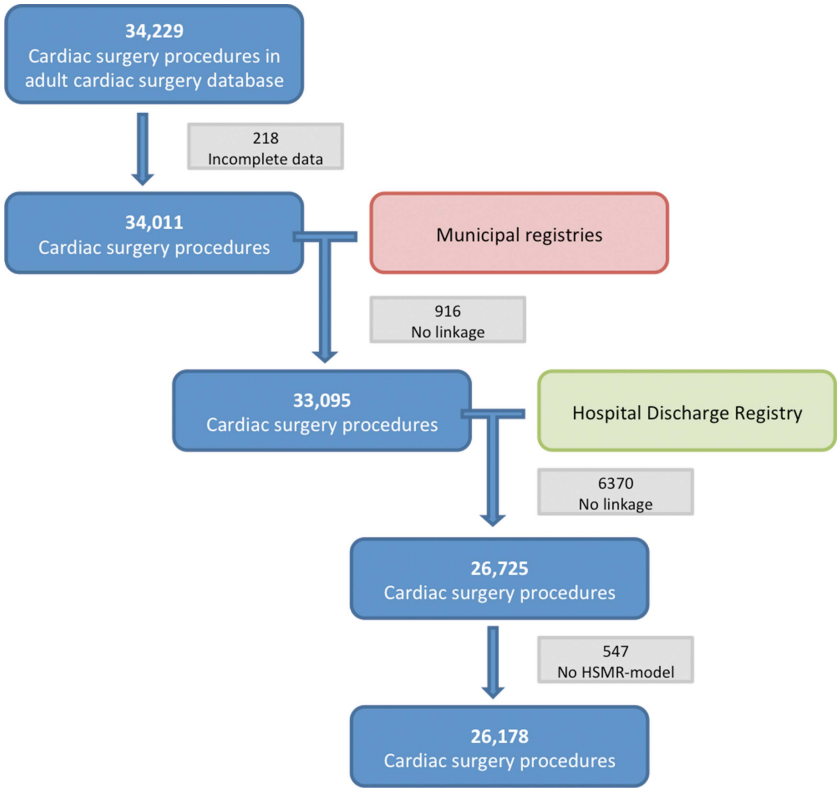


**Figure 1** Flow chart of data flow. Data from the adult cardiac surgery database (Netherlands Association for Cardio-Thoracic Surgery) was linked to municipal registries and the Hospital Discharge Registry. In total 26 178 cardiac surgery procedures were included for further analyses. HSMR, hospital standardised mortality ratio.

between subjects with and without the outcome and depends on the variables included in the model. The discrimination of the models was quantified using the area under the ROC -curve, which is equivalent to the c-statistic. The 95% CI of the c-statistic and the difference between two c-statistics was tested using DeLong's test.[28]

The calibration of a risk model refers to the ability of a model to predict how many patients will have the outcome. The calibration was assessed using calibration plots and the Brier Score. The Brier Score measures model accuracy on patient level by squaring and summing the difference between the predicted and the observed outcome per patient. The method by Redelmeier was used to estimate the 95% CI of the Brier Score and test the difference between two Brier scores.[29]

### Benchmarking

In this study, benchmarking is performed by calculating the standardised mortality ratio (SMR) for all hospitals. The SMR is calculated by dividing the observed mortality with the expected mortality within a hospital. SMRs of the administrative and clinical models were compared. Centres with a SMR for which the 95% CI did not cover the value 1 were considered to be outliers. The 95% CI of the SMRs were estimated using the method described by Breslow and Day.[30]

All analyses were performed in R V.2.15.[31]

## RESULTS

### Risk factor coding

The risk factors in the linked subset from the administrative and clinical databases are presented in table 1. Mean age was 66.6 years (±10.7) and 29.5% of patients were female. A comparison of the prevalence of risk factors could not be made, as the definitions differed between the administrative and the clinical database.

### Number of cardiac interventions performed (by type of intervention)

In total 14 300 (54.6%) isolated CABG procedures were performed according to the NVT database. Other frequently performed interventions were: aortic valve replacement with or without concomitant CABG (12.1% and 8.3%, respectively) and mitral valve repair with or without concomitant CABG (3.1% and 2.7%, respectively). The proportion of isolated CABG, isolated aortic valve replacement, isolated mitral valve repair and isolated mitral valve replacement, which was coded with the correct main intervention code in the HDR ranged from 64.6% to 92.2% (table 2). The intervention code in the HDR was missing in 1923 (7.3%) procedures. As a result, the number of cardiac surgery interventions could not be accurately assessed using HDR data.

### Inhospital mortality

Inhospital mortality in the HDR database is derived from the municipal registries which are highly accurate. In the NVT database 42 of 762 (5.5%) patients who died during hospital stay were not coded as such and the other way around, 36 of 25 005 (0.1%) survivors were incorrectly coded as inhospital mortality during the same hospital admission.

### Calibration of the administrative models and the clinical models

Calibration of the risk models is shown in figure 2. The original models (Administrative.1 and Clinical.1) were poorly calibrated. Administrative.1 underestimated the risk of mortality, whereas Clinical.1 overestimated the risk of mortality. Updating improved calibration of both models, as the difference between observed and predicted mortality became smaller. However, in all model pairs the Brier Score for the administrative models remained significantly higher in comparison with the clinical models, indicating inferior calibration of the administrative model (table 3). The maximum Brier score in this data was 3.0%. Rescaling of the Brier Score on a scale from 0% to 100% would result in a score of 93.8% for Administrative.3 and 87.8% for Clinical.3.

Results were comparable in the subgroup analyses on isolated CABG procedures (figure 2 and table 3), where the maximum Brier score that was possible in these data was 1.3%.

### Discrimination of the administrative models and the clinical models

Discrimination of the models is shown in figure 3. The c-statistics of the administrative models (0.756–0.788) are substantially lower than that of the clinical models (0.838–0.846), indicating inferior discrimination of the administrative models ($p<0.001$ for all three model pairs). Updating of the administrative model did not improve the discrimination (figure 3).

### The effect on benchmarking

The effect of the use of administrative versus clinical models on benchmarking is shown in figure 4. The majority of SMRs calculated using the original administrative model was higher than 1, which indicates that the model underestimated the risk of mortality. For the original clinical model the opposite was found: the model overestimated the risk of mortality.

Updating of models resulted in better predictions on hospital level (SMRs closer to 1). However, a considerable difference was found between the updated administrative versus the updated clinical models, for example in hospital B and hospital C (figure 4).The mean difference in SMR for Admistrative.1 versus Clinical.1 was 1.13 (range 0.23–2.08), 0.12 (range 0.004–0.37) for Administrative.2 versus clinical.2, and 0.11 (range 0.001–0.43) for Administrative.3 versus Clinical.3.

**Table 2** Comparison of intervention type and inhospital mortality

| | | Hospital discharge registry (administrative data) | | |
|---|---|---|---|---|
| NVT database (clinical data) | | Correct main intervention code | Incorrect main intervention code | No code |
| Intervention type | | | | |
| Isolated CABG | 14 300 (100%) | 13 185 (92.2%) | 197 (1.4%) | 918 (6.4%) |
| Isolated AoV replacement | 3157 (100%) | 2461 (78.0%) | 457 (14.5%) | 239 (7.6%) |
| Isolated MV repair | 820 (100%) | 625 (76.2%) | 134 (16.3%) | 61 (7.4%) |
| Isolated MV replacement | 316 (100%) | 204 (64.6%) | 83 (26.3%) | 29 (9.2%) |

AoV, aortic valve; CABG, coronary artery bypass grafting; MV, mitral valve; NVT, Netherlands Association for Cardio-Thoracic Surgery.

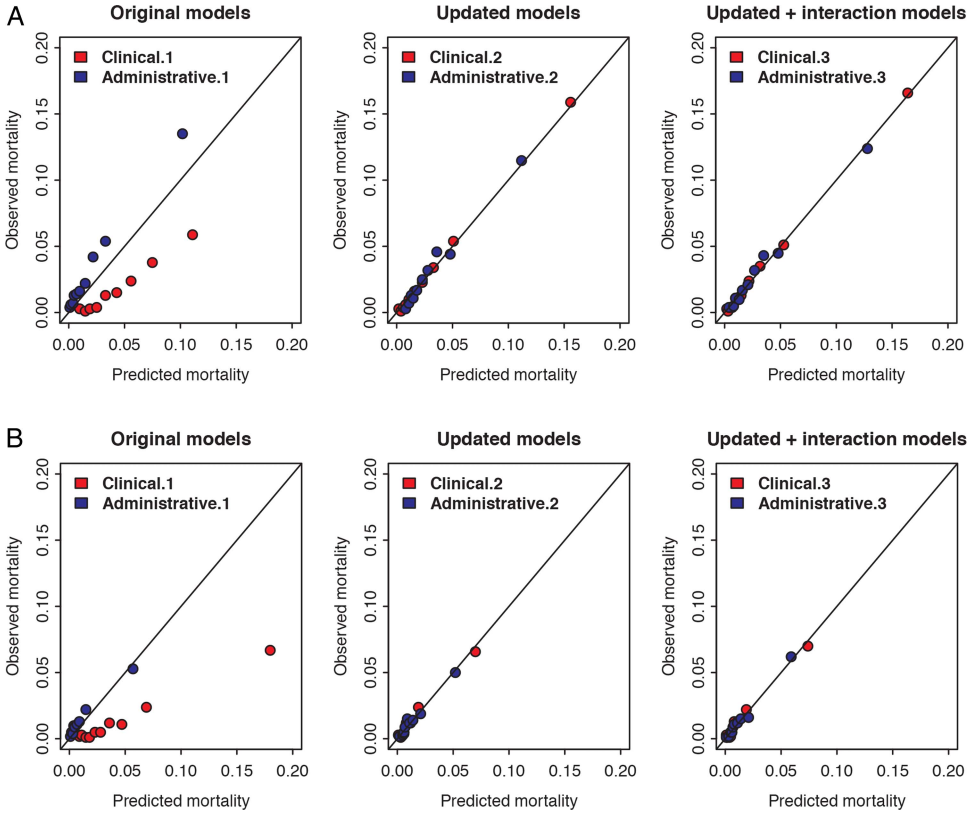


**Figure 2** Calibration plot of the three clinical models and the three administrative models. The calibration plots of the clinical models are depicted in red and the calibration plots of the administrative models in blue. Panel A: models fitted on all cardiac surgery. Panel B: models fitted on isolated coronary artery bypass grafting procedures.

The SMRs calculated using the clinical and administrative models yielded different outliers. Hospital C and hospital J changed outlier status when either the updated model Administrative.3 or Clinical.3 was used. The analyses using only isolated CABG surgery yielded comparable results as those based on all cardiac surgery data (figure 4).

## DISCUSSION

### Principle findings

This study compared (1) data accuracy in the administrative HDR database to that in the clinical cardiac surgery database of the Netherlands Association of Cardio-Thoracic Surgery (NVT) and (2) the predictive performance of administrative models to that of the clinical EuroSCORE model.

The reported intervention code in the administrative database was incorrect in up to 26%, depending on the type of surgery. As a result, the number of cardiac surgery interventions could not be accurately assessed.

After updating of the models to our data, the calibration of the administrative model was inferior to that of the clinical model. The importance of this shortcoming is marked by the identification of other outliers when used for benchmarking of hospitals.

### Why models based on administrative data have inferior calibration and discrimination

When developing a risk prediction model, the first logical step is to consider which variables could be predictors for the outcome. However, administrative models are limited to the routinely collected variables, which might not necessarily be the strongest predictors. In our study, several strong predictors for mortality (shown in table 1) were not available in the administrative database. The other way around, administrative risk

**Table 3** Brier Score of the three clinical models and the three administrative models, for all cardiac surgery and for only isolated coronary artery bypass surgery

| | All cardiac surgery | | | Isolated CABG surgery | | |
|---|---|---|---|---|---|---|
| **Brier scores** | **Administrative** | **Clinical** | **p Value difference** | **Administrative** | **Clinical** | **p Value difference** |
| Original models | 2.9% (2.8–3.0) | 3.0% (2.8–3.2) | 0.093 | 1.3% (1.2–1.5) | 1.4% (1.1–1.7) | 0.030 |
| Updated models | 2.9% (2.7–3.1) | 2.7% (2.5–2.9) | <0.001 | 1.3% (1.1–1.4) | 1.2% (1.0–1.3) | <0.001 |
| Updated+interaction terms | 2.8% (2.6–3.0) | 2.6% (2.5–2.8) | <0.001 | 1.2% (1.1–1.4) | 1.2% (1.0–1.3) | 0.026 |

Brier scores range from 0 to a value depending on the prevalence of the outcome. The maximum Brier score that was possible in this data was 3.0% for all cardiac surgery and 1.3% for isolated CABG. A lower Brier score indicates better calibration. Brackets denote 95% CIs.
CABG: coronary artery bypass grafting.

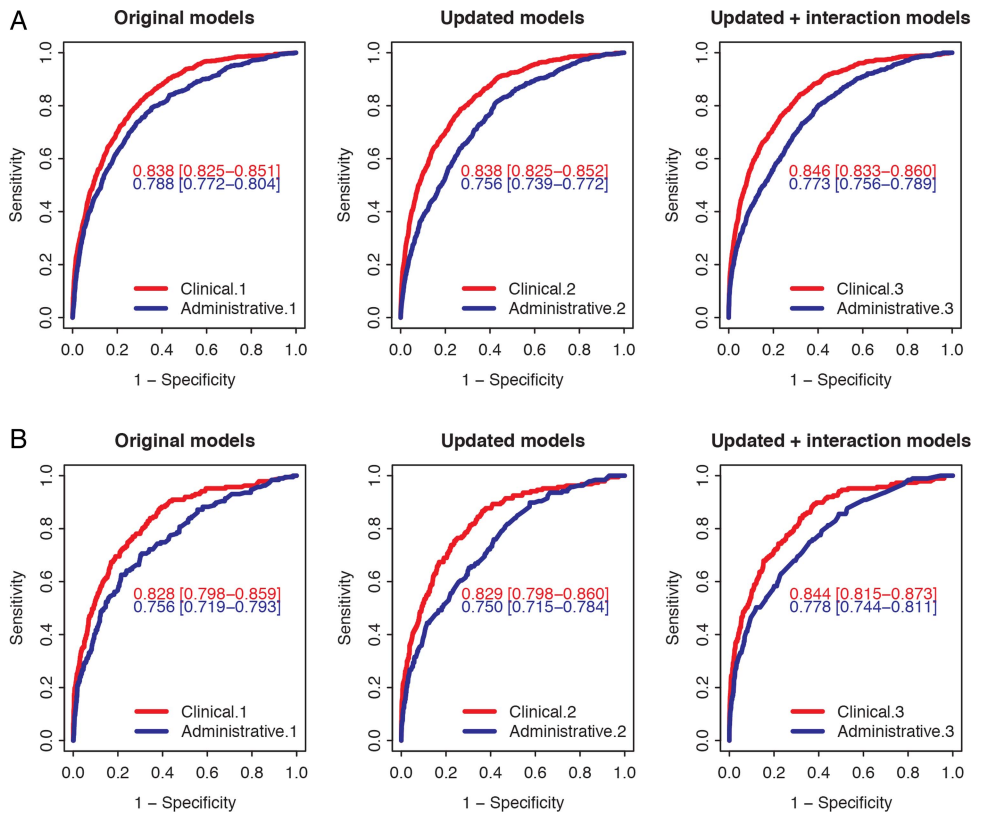


**Figure 3** Area under the ROC-curve of the clinical and the administrative models for the prediction of inhospital mortality. The ROC curves of the clinical models are depicted in red and the ROC curves of the administrative models in blue. Panel A: models fitted on all cardiac surgery. Panel B: models fitted on isolated coronary artery bypass grafting procedures.

factors that were strongly associated with mortality had a low prevalence in our study population. This is likely to have affected the calibration and discrimination of the administrative models. Previous studies reported that much of the predictive performance of risk models is derived from a relatively small number of clinical variables and the predictive performance of administrative models could be improved with the addition of a limited number of clinical variables.[7 13 19 32 33]

### Why administrative data are inferior to clinical data for benchmarking purposes

The requirements of a risk-adjustment model depend on its goal. For benchmarking an adequate calibration is required: the model should adequately predict the expected mortality rate in a hospital. It can be seen as a scale that should weigh correctly. The performance of a scale mainly depends on its ability to weigh a (kilo) gram. If this feature is adequate, but the weighing is off par, the scale can be reset to zero to adjust it to any new situation. Similarly, the performance of a model depends on the strength of the predictors in the model (ie, discrimination), as the model can be recalibrated to update it in time or to make it suitable for a new population. It follows from the aforementioned that the inferior discrimination of administrative models (in comparison with clinical models) will result in inferior calibration. It is shown in this study that this could very well affect the outlier status of a hospital.

### Other issues in the use of administrative data

There are other reasons why the HDR database with routinely collected data turned out to be unsuitable for analyses of outcomes in cardiac surgery. First, for a considerable number of records in our study population the intervention code was incorrect, unspecified (eg, "cardiac surgery") or missing. Consequently, the number of cardiac surgery interventions performed could not be reliably assessed. Previous studies have also reported discrepant counts of operations in administrative data versus clinical data.[8 17 34]

Inaccurate coding could be attributed to the fact that data were collected by persons who were not actively involved in the clinical care and thus were dissociated from clinical information that could be necessary for correct reporting of data.[35] In addition, occasionally not all interventions and diagnoses are recorded. Also, admission and discharge dates are collected, instead of dates of intervention. This has been reported before as an important reason for variance in cardiac surgery volumes between administrative and clinical databases.[17]

Furthermore, the HSMR method uses administrative models for specific diagnose codes. However, in cardiac surgery analyses of outcomes is performed by intervention type, as risk is considered to be mainly related to the performed intervention.

### Implications for practice

The use of administrative data has many advantages over the use of clinical data. The data are routinely collected and stored, making them cheaper and readily available. However, the apparent benefits should be carefully weighed against the limitations and drawbacks of administrative models, when compared with clinical models.

Public benchmarking in general can be dangerous in the sense that the general public cannot be expected to understand the limitations and the prerequisites under which the results should be interpreted. The limitations are more pronounced for

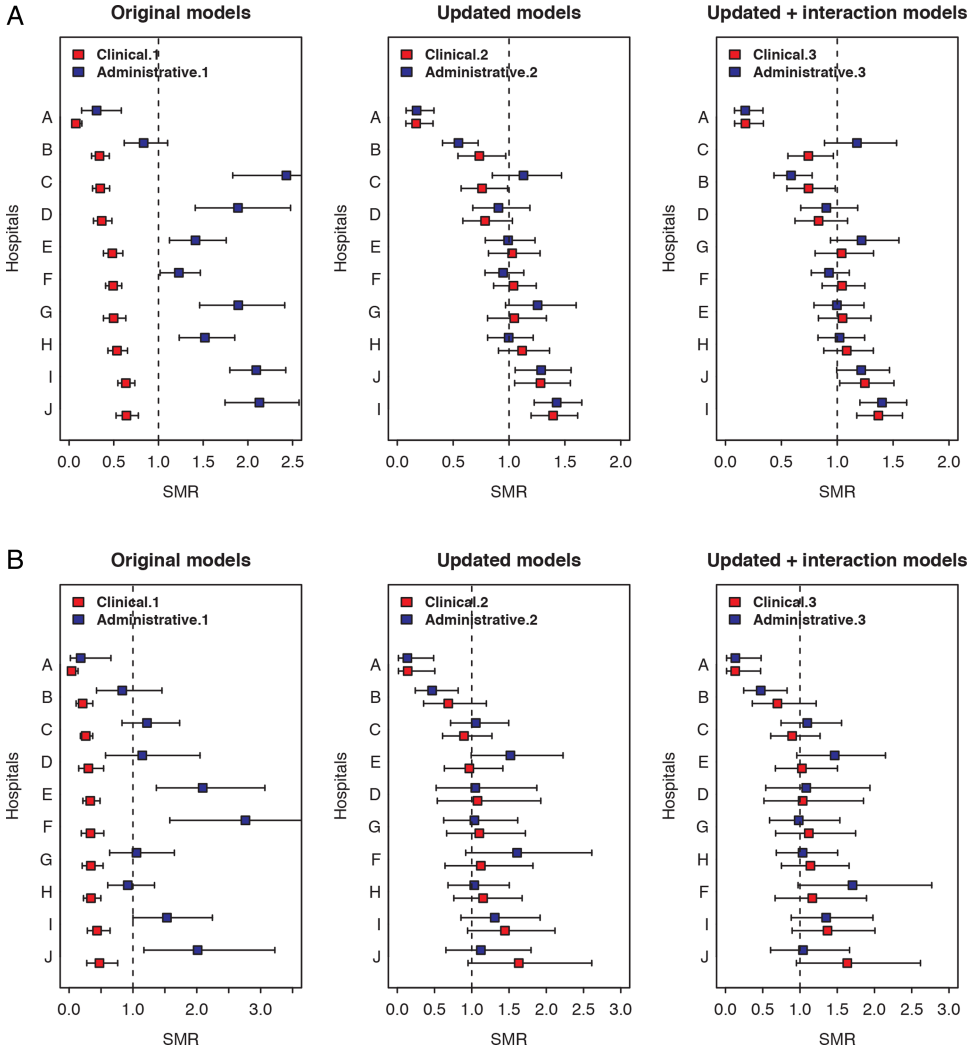


**Figure 4** Benchmarking using standardised mortality ratio (SMR) calculated by the clinical models and the administrative models. The SMRs of the clinical models are depicted in red and the SMRs of the administrative models in blue. Panel A: models fitted on all cardiac surgery. Panel B: models fitted on isolated coronary artery bypass grafting procedures.

administrative data. This is particularly important because benchmarking could have far-reaching consequences when known to healthcare consumers, the media, health insurance companies or governmental bodies. In this context, development of models with a high predictive performance, which might include clinical risk factors, should be strived for at all times.[36] If clinical data are already collected, their availability for benchmarking should be encouraged.

On the other hand, clinical data appeared to have an evident weakness as well. The outcome inhospital mortality was misclassified in nearly 6% of the records in the clinical database used in this study. For outcomes such as vital state and readmissions, administrative databases were highly accurate, as information was derived from municipal registries. Administrative data sources could be used to verify outcomes data, thus complementing clinical databases. In this way, the strengths of both types of data are combined in order to optimise benchmarking in healthcare.[37]

The findings in this study are likely to hold true for populations other than cardiac surgery patients and in other countries in the world. Most probably, other specific surgical interventions, such as for example oesophageal or hepatobiliary surgery, also require adjustment for risk factors not commonly included in administrative databases. Consequently, benchmarking in those populations will result in similar issues as encountered in this study.

## Possible limitations

These analyses were based on data from 10 out of 16 cardiac surgery centres in The Netherlands. In general, the population of the six hospitals not participating in this study did not differ from the study population with regards to age, sex and the median logistic EuroSCORE. However, it is unknown if the results with regards to data accuracy are generalisable to all centres.

Second, the sensitivity of the linkage between the clinical and the administrative database was 77%. Although we did not find a difference in the overall risk profile between the linked and non-linked records, we do acknowledge that a substantial part of the total population was excluded from the analyses. We have no reason to believe that administrative models would perform any differently in the non-linked records or that data accuracy was better in the non-linked records. The conclusions of our study are thus unlikely to be affected by this limitation.

The goal of this study was to assess the accuracy of administrative data and the predictive performance of the accompanying models. As such, it was not our intention to design a new model for risk prediction in cardiac surgery. Thus, we chose to stay in line with the methods used to construct the original models and refrain from further sophisticated methods such as hierarchical modelling and shrinkage of coefficients.

The outcome in this study is inhospital mortality. Several publications have previously shown why mortality at fixed time

intervals is a more appropriate measure in outcomes evaluation. We acknowledge the limitations of this outcome and we are aware that mortality is one of the several indicators that can be used to measure quality, but certainly not the only one. For the purpose of our study, we have no reason to believe this has affected our results, as the clinical and the administrative models were fitted on this outcome.

## CONCLUSION

Although there are advantages to the use of administrative models for benchmarking in cardiac surgery, their calibration and discrimination (and thus performance in benchmarking) is inferior to that of clinical models. The use of either an administrative or a clinical model may affect the outlier status of hospitals. Therefore, in specific populations such as cardiac surgery, the use of prediction models including clinical risk factors is recommended.

### Key messages

**What is already known about this subject**

- Administrative data are inexpensive to collect and easily accessible, but the accuracy of coding remains questionable and it is known to contain limited information on patient condition and severity of disease.
- Clinical data do not have most of these problems and have a good capability of predicting outcomes after cardiac surgery.
- The hospital standardised mortality ratio (HSMR) is an increasingly used method for healthcare benchmarking using administrative data.

**What does this study add**

- The number of cardiac surgery interventions performed could not be accurately assessed in routinely collected administrative data.
- For cardiac surgery, administrative models (developed according to the Dutch HSMR method) have inferior predictive performance when compared with a clinical model (European System for Cardiac Operative Risk Evaluation).
- The use of either administrative or clinical risk-adjustment models can affect the outlier status of hospitals when benchmarking is performed.
- Risk-adjustment models including procedure-specific clinical risk factors are recommended.

**How might this impact on clinical practice**

- The findings in this study might stimulate healthcare providers and policy makers to use clinical data for the purpose of provider profiling.
- Administrative data should be used for outcomes such as mortality and readmissions, in addition to the clinical risk factors.
- The conclusions of this study help to clarify the limitations of the HSMR method in specific patient populations, such as cardiac surgery.

**Contributors** Contributors: SS and MEP wrote the statistical analysis plan, cleaned and analysed the data, and drafted and revised the paper. RHHG, wrote the statistical analysis plan, supervised the statistical analyses and made thorough critical revisions of the paper. MLB, YvdG and CJK supervised the statistical analyses and made thorough critical revisions of the paper. MIMV and LAvH monitored data collection in the Netherlands Association for Cardio-Thoracic Surgery database, provided the data and made thorough critical revisions of the paper.

**Funding** The Department of Cardio-Thoracic Surgery UMC Utrecht has received financial support from the Netherlands Association of Cardio-Thoracic Surgery to cover part of the first author's salary.

**Competing interests** All authors have completed the Unified Competing Interests form at http://www.icmje.org/coi_disclosure.pdf (available on request from the corresponding author) and declare: no support from any organisation for the submitted work; no financial relationships with any organisations that might have an interest in the submitted work in the previous 3 years; no other relationships or activities that could appear to have influenced the submitted work.

**Ethics approval** The Hospital Discharge Registry is national statistical data, made available by National Statistics Netherlands. The data from The Netherlands Association for Cardio-Thoracic Surgery database was collected in the 10 participating centres and sent to National Statistics. National Statistics performed the linkage to the Hospital Discharge Registry as a Trusted Third Party. The data provided for this study was fully anonymised; it was not possible to identify the individuals from the information provided. Considering the aforementioned, approval from the ethics committee was not required and not obtained.

**Provenance and peer review** Not commissioned; externally peer reviewed.



## REFERENCES

1 Iezzoni LI. *Risk adjustment for measuring healthcare outcomes*. Ann Arbor, Mich: Health Administration Press, 1994.
2 Shahian DM, Normand SL, Torchiana DF, *et al*. Cardiac surgery report cards: comprehensive review and statistical critique. *Ann Thorac Surg* 2001;72:2155–68.
3 Heijink R, Koolman X, Pieter D, *et al*. Measuring and explaining mortality in Dutch hospitals; the hospital standardized mortality rate between 2003 and 2005. *BMC Health Serv Res* 2008;8:73.
4 Jarman B, Gault S, Alves B, *et al*. Explaining differences in English hospital death rates using routinely collected data. *BMJ* 1999;318:1515–20.
5 Bohensky MA, Jolley D, Pilcher DV, *et al*. Prognostic models based on administrative data alone inadequately predict the survival outcomes for critically ill patients at 180 days post-hospital discharge. *J Crit Care* 2012;27:422.e11–21.
6 Brinkman S, Abu-Hanna A, van der Veen A, *et al*. A comparison of the performance of a model based on administrative data and a model based on clinical data: effect of severity of illness on standardized mortality ratios of intensive care units. *Crit Care Med* 2012;40:373–8.
7 Hannan EL, Racz MJ, Jollis JG, *et al*. Using Medicare claims data to assess provider quality for CABG surgery: does it work well enough? *Health Serv Res* 1997;31:659–78.
8 Shahian DM, Silverstein T, Lovett AF, *et al*. Comparison of clinical and administrative data sources for hospital coronary artery bypass graft surgery report cards. *Circulation* 2007;115:1518–27.
9 Glance LG, Dick AW, Osler TM, *et al*. Accuracy of hospital report cards based on administrative data. *Health Serv Res* 2006;41:1413–37.
10 Nashef SA, Roques F, Michel P, *et al*. European system for cardiac operative risk evaluation (EuroSCORE). *Eur J Cardiothorac Surg* 1999;16:9–13.
11 Roques F, Michel P, Goldstone AR, *et al*. The logistic EuroSCORE. *Eur Heart J* 2003;24:881–2.
12 Shahian DM, O'Brien SM, Filardo G, *et al*. The Society of Thoracic Surgeons 2008 cardiac surgery risk models: part 1–coronary artery bypass grafting surgery. *Ann Thorac Surg* 2009;88(1 Suppl):S2–22.
13 Geraci JM, Johnson ML, Gordon HS, *et al*. Mortality after cardiac bypass surgery: prediction from administrative versus clinical data. *Med Care* 2005;43:149–58.
14 Bratzler DW, Normand SL, Wang Y, *et al*. An administrative claims model for profiling hospital 30-day mortality rates for pneumonia patients. *PLoS One* 2011;6:e17401.
15 Gordon HS, Johnson ML, Wray NP, *et al*. Mortality after noncardiac surgery: prediction from administrative versus clinical data. *Med Care* 2005;43:159–67.
16 Hall BL, Hirbe M, Waterman B, *et al*. Comparison of mortality risk adjustment using a clinical data algorithm (American College of Surgeons National Surgical Quality Improvement Program) and an administrative data algorithm (Solucient) at the case level within a single institution. *J Am Coll Surg* 2007;205:767–77.
17 Mack MJ, Herbert M, Prince S, *et al*. Does reporting of coronary artery bypass grafting from administrative databases accurately reflect actual clinical outcomes? *J Thorac Cardiovasc Surg* 2005;129:1309–17.

18 Parker JP, Li Z, Damberg CL, *et al*. Administrative versus clinical data for coronary artery bypass graft surgery report cards: the view from California. *Med Care* 2006;44:687–95.
19 Ugolini C, Nobilio L. Risk adjustment for coronary artery bypass graft surgery: an administrative approach versus EuroSCORE. *Int J Qual Healthcare* 2004;16: 157–64.
20 Siregar S, Groenwold RH, Versteegh MI, *et al*. Data Resource Profile: Adult cardiac surgery database of the Netherlands Association for Cardio-Thoracic Surgery. *Int J Epidemiology* 2013;42:142–9.
21 Jarman B, Pieter D, van der Veen AA, *et al*. The hospital standardised mortality ratio: a powerful tool for Dutch hospitals to assess their quality of care? *Qual Saf Healthcare* 2010;19:9–13.
22 Dutch Hospital Data. http://www.dutchhospitaldata.nl. Utrecht, The Netherlands.
23 World Health Organization. http://www.who.int. 2012.
24 Centraal Bureau voor de Statistiek. http://www.cbs.nl. Den Haag, The Netherlands.
25 Vaartjes I, Hoes AW, Reitsma JB, *et al*. Age- and gender-specific risk of death after first hospitalization for heart failure. *BMC Public Health* 2010;10:637.
26 Toll DB, Janssen KJ, Vergouwe Y, *et al*. Validation, updating and impact of clinical prediction rules: a review. *J Clin Epidemiol* 2008;61:1085–94.
27 Steyerberg EW. *Clinical prediction models*. New York: Springer Science+Business Media, 2009.
28 DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988;44:837–45.
29 Redelmeier DA, Bloch DA, Hickam DH. Assessing predictive accuracy: how to compare Brier scores. *J Clin Epidemiol* 1991;44:1141–6.
30 Breslow NE, Day NE. Rates and rate standardization. In: Heseltine E, ed. *Statistical Methods in Cancer Research: Vol. II—The Design and Analysis of Cohort Studies*. Lyon: International Agency for Research on Cancer, 1987:48–80.
31 *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2011.
32 Jones RH, Hannan EL, Hammermeister KE, *et al*. Identification of preoperative variables needed for risk adjustment of short-term mortality after coronary artery bypass graft surgery. The Working Group Panel on the Cooperative CABG Database Project. *J Am Coll Cardiol* 1996;28:1478–87.
33 Simms AD, Reynolds S, Pieper K, *et al*. Evaluation of the NICE mini-GRACE risk scores for acute myocardial infarction using the Myocardial Ischaemia National Audit Project (MINAP) 2003–2009: National Institute for Cardiovascular Outcomes Research (NICOR). *Heart* 2013;99:35–40.
34 Aylin P, Bottle A, Majeed A. Use of administrative data or clinical databases as predictors of risk of death in hospital: comparison of models. *BMJ* 2007;334:1044.
35 Zhan C, Miller MR. Administrative data based patient safety research: a critical review. *Qual Saf Healthcare* 2003;12(Suppl 2):ii58–63.
36 Herrett E, Shah AD, Boggon R, *et al*. Completeness and diagnostic validity of recording acute myocardial infarction events in primary care, hospital care, disease registry, and national mortality records: cohort study. *BMJ* 2013;346:f2350.
37 Manda SO, Gale CP, Hall AS, *et al*. Statistical profiling of hospital performance using acute coronary syndrome mortality. *Cardiovasc J Afr* 2012;23:546–51.