NIHR Health Services Research programme

NHS National Institute for Health Research

HSR Protocol - project ref: 09/2000/65
Version: 1
Date: 1 July 2011

Ensuring comparisons of health care providers are fair: risk modelling for quality improvement in the critically ill

Chief investigator	Dr. David Harrison
Sponsor	Intensive Care National Audit & Research Centre (ICNARC)
Funder	HSR programme
NIHR Portfolio number	
ISRCTN registration (if applicable)	N/A

Ensuring comparisons of health care providers are fair: risk modelling for quality improvement in the critically ill

Aims and objectives

The aim of the proposed project is to improve risk prediction models to underpin quality improvement programmes for the critically ill (patients receiving general or specialist adult critical care or experiencing in-hospital cardiac arrest).

This aim will be addressed through the following objectives:

1. To improve current risk prediction models for critically ill patients, to include:

    a) A external validation of current models in critical care units in Scotland;
    b) introduction of new important variables;
    c) improved modelling of interactions between physiological parameters;
    d) improved handling of missing data;
    e) improved modelling of reasons for admission to/diagnosis on admission to critical care; and
    f) modelling of mortality at fixed time-points or of time-to-event outcomes (in place of hospital mortality currently used) through linkage to national data.

2. To develop and validate new risk prediction models for critically ill patients, to include:

    a) models for cardiothoracic critical care;
    b) models for hospital and critical care admissions following in-hospital cardiac arrest; and
    c) models for critical care units admitting low risk patients.

3. Immediate translation of improved risk models into practice, through:

    a) adoption into routine comparative outcome reporting for national clinical audits; and
    b) communication of research output to providers, managers, commissioners, policy makers and academics in critical care.

Background

High quality care is at the centre of the NHS.1 National clinical audit has a key role to play in ensuring quality, particularly in areas where patient choice cannot play a significant part. Recently, the importance of national clinical audit has been reinforced by the establishment of the National Clinical Audit and Patients' Outcomes Programme (NCAPOP) and the Health Quality Improvement Partnership (HQIP) under the guidance of the National Clinical Audit Advisory Group (NCAAG).

The Intensive Care National Audit & Research Centre (ICNARC) is an independent charity that runs national clinical audit programmes to monitor and improve care for the critically ill. Currently, ICNARC co-ordinates two national clinical audits: the Case Mix Programme (CMP) – a national clinical audit for adult critical care; and the National Cardiac Arrest Audit (NCAA) – a national clinical audit for in-hospital cardiac arrest (co-ordinated jointly with the Resuscitation Council (UK)). The CMP has been established for over 15 years and is seen as a flagship national clinical audit. The resulting high-quality clinical database (of over 1 million critical care admissions) has underpinned evaluations of policy and practice in critical care.2 The NCAA has recently been established. Both national clinical audits are underpinned by the need and ability to report accurate risk-adjusted results.

Risk prediction models for adult, general critical care are well established, but on-going improvement work is essential to further improve accuracy.3 In 2006, ICNARC published an MRC-funded validation (Ref: G9813469) of existing models (APACHE II, APACHE III, SAPS II and MPM II) and concluded that there was little difference in performance among the models, but there was scope for further improvement.4 While retaining APACHE II for the purpose of international comparisons, ICNARC developed and validated the ICNARC model,5 which currently underpins the risk-adjusted outcomes reported for the CMP. There are, however, a number of areas where we have identified the potential to improve our modelling.

There is currently no UK risk prediction model for in-hospital cardiac arrest. Initial comparative reporting for the NCAA is based on stratifying patients according to single risk factors. As part of this project, we will develop a risk prediction model for in-hospital cardiac arrests based on the data from the NCAA database.

Methods

3.1 Design and theoretical/conceptual framework

Risk modelling study on existing data

3.2 Sampling

The selection of sites is based on those participating in the Scottish Intensive Care Society Audit Group (SICSAG), the CMP and the NCAA.

The SICSAG is the national clinical audit for adult critical care in Scotland; the SICSAG database has 100% coverage of 24 adult, general critical care units in Scotland. SICSAG data from approximately 30,000 admissions from 2007 to 2009 will be provided by NHS National Services Scotland for external validation of the existing ICNARC risk prediction model (objective 1a).

The coverage of the CMP is extremely high, with 226 critical care units participating including over 90% of NHS adult, general critical care units in

England, Wales and Northern Ireland. Our previous research has established that the fit of risk models deteriorates over time,6 and so we will use the most recent three years' data (from 2008 to 2010) for model development, giving a sample size of over 280,000 admissions (72,000 events) for objective 1b-f, which will be sufficient to consider complex model structures including interaction terms. As the CMP is an ongoing programme, additional data will accrue while the development work is ongoing. At one year into the project, an additional 100,000 admissions will be available for final validation of models in independent data.

Of twenty-seven specialist cardiothoracic critical care units providing Level 3 (intensive) care, five currently participate in the CMP. From 2008 to 2010 we will have a sample size of approximately 6,000 admissions to these units (550 events) for objective 2a. Based on a commonly applied rule of thumb requiring a minimum of ten events per variable,7 this sample size would enable us to consider models with approximately 55 variables. Due to the comparatively small sample size and the low coverage of specialist units, models developed for cardiothoracic critical care are likely to be experimental in nature and to form the basis for a larger, more representative evaluation in the future.

Data collection for the NCAA is currently ongoing in 98 hospitals and recruitment of new hospitals continues. We would therefore anticipate, by the end of 2011, to have data available for approximately 16,000 in-hospital cardiac arrests (2,000 events) for objective 2b.

3.3 Setting/context

Admissions to adult, general and adult cardiothoracic critical care units in NHS acute hospitals (CMP) and individuals experiencing cardiac arrests within NHS hospitals and attended by the in-hospital resuscitation team or equivalent (NCAA).

3.4 Data Collection

The project will utilise data collected for the CMP and the NCAA. These data are collected to precise rules and definitions by trained data collectors and undergo extensive validation, both locally and centrally, for completeness, logicality and consistency. The CMP database has been independently assessed and scored highly by the Directory of Clinical Databases (DoCDat; http://www.icapp.nhs.uk/docdat/) against their ten domains (describing elements of coverage and accuracy). Similar processes for the NCAA were developed building on the knowledge and experience of the successful systems in place for the CMP.

3.5 Data analysis

3.5.1 Measures of model performance

Throughout the project, risk prediction models will be validated for their

discrimination, calibration and overall fit. The following panel of measures will be used to give an overall assessment of model performance. We have previously applied these methods to validate existing risk prediction models,8 and to assess the performance of the ICNARC model.9

Discrimination will be assessed with the c index. The c index is the probability of concordance between outcomes and predictions.10 For binary outcomes (e.g. death), this is the probability that a randomly chosen individual that experienced the event (a non-survivor) will have a higher predicted risk than a randomly chosen individual that did not experience the event (a survivor). This has been shown to be identical to the area under the receiver operating characteristic (ROC) curve.11 Perfect discrimination (i.e. all patients that experience the event have higher predicted risk than all patients that do not experience the event) corresponds to c = 1; discrimination that is no better than chance corresponds to c = 0.5.

Calibration will primarily be assessed graphically by plotting the observed risk against the predicted risk in equal sized groups divided at quantiles of predicted risk. The associated Hosmer-Lemeshow goodness-of-fit statistic, C*, will be calculated, representing a chi-squared test statistic for perfect calibration.12 However, it is noted that this is a test statistic and not a meaningful measure of the degree of miscalibration. The test is highly sensitive to sample size.13 Consequently, statistically significant values of the Hosmer-Lemeshow statistic do not necessarily correspond to important miscalibration.

Cox's calibration regression provides a simple method to quantify the degree of miscalibration of a model.14 Cox suggested fitting the model

$$\text{true log odds} = \alpha + \beta \times \text{predicted log odds}$$

using logistic regression. The value of α represents the calibration at a prediction of 0.5 when $\beta \neq 1$, or calibration more generally when $\beta = 1$. The value of β represents the degree of variability in the predicted probabilities. If $\beta > 1$, the "probabilities show the right general pattern of variation but do not vary enough." If $0 < \beta < 1$, the probabilities vary too much. Perfect predictions correspond to $\alpha = 0$ and $\beta = 1$ (i.e. true log odds = predicted log odds); perfect calibration corresponds to $\alpha = 0$ conditional on $\beta = 1$ ($\alpha = 0|\beta = 1$); and the correct degree of variation corresponds to $\beta = 1$ conditional on the observed value of α ($\beta = 1|\ \alpha$). All of these aspects of the model fit can be tested with hypothesis tests.

Shapiro's R, based on Shapiro's Q, is an overall measure of the accuracy of the model, reflecting both calibration and discrimination.15 R is the geometric mean of the probability assigned to the event that occurred. Perfect predictions correspond to R = 1; poor predictions (when a constant of 0.5 is assigned to every individual) correspond to R = 0.5.

Brier's score, B, was developed in relation to meteorological forecasts; it is an overall measure of the accuracy of predictions.16 B is the mean square error between outcomes and predictions. Perfect predictions correspond to B = 0; poor predictions (when a constant of 0.5 is assigned to every individual)

correspond to B = 0.25. Brier's score can also be decomposed into components corresponding to the accuracy of the average prediction, the excess variance of predictions and the covariance of outcomes and predictions.

For both Shapiro's R and Brier's score, there is a corresponding approximate R2 measure of 'explained variation' that rescales the respective measure to range between 0 (corresponding to the value when a constant prediction equal to the overall risk in the population is assigned to every individual) and 1 (corresponding to perfect predictions). These R2 measures are known as the entropy-based R2 and the sum-of-squares R2 and have been shown to have good statistical properties from among the many possible R2 measures proposed for logistic regression.17 We will therefore present each R2 measure alongside the respective, untransformed measure.

3.5.2 *Methods to assess model performance*

For each of the proposed changes to existing models and development of de novo models, the revised models will be validated using the above measures both within the development sample and in independent validation data.

Within the development data, model validation will be performed using bootstrap methods. When a model is developed and validated on the same dataset, the apparent performance of the model will be better than its true performance (termed "optimism"). A frequently employed solution to this problem is to split the available data at random into separate development and validation samples. This approach, however, is wasteful of data as the final model will not have been fitted using all available data. Efron proposed an alternative approach using bootstrapping techniques. In this method, repeated bootstrap samples (random samples with replacement from the original data) are taken, the model is re-estimated in each bootstrap sample and the measures of model performance are calculated in both the bootstrap sample and the original data. The average difference between the value of each measure in the bootstrap sample and in the original data provides an estimate of the optimism, and this can be subtracted from the observed measure for the model when fitted in the full dataset.18 However, Efron further demonstrated that the estimate of the optimism obtained by this method is itself optimistic, and so he proposed a refinement to this method (the .632 bootstrap) to adjust for this.18 We will use the .632 bootstrap method for internal validation in this project.

As this project is nested within on-going national audits, data will continue to accrue following the point at which the development dataset is locked for analysis. At the end of each phase of the project, a considerable additional dataset will have accrued, which we will use for independent validation of the models. As new sites also join the audits over time, we will, where possible, also evaluate the models only in data arising from new sites that were not included in the model development to provide an even more independent validation.

3.5.3 *Objective 1: To improve current risk prediction models for critically ill patients.*

Although the ICNARC model has been demonstrated to have better performance among patients admitted to UK critical care units than other risk prediction models, we have identified, through discussions with end users in critical care and experts in health services research and risk modelling, a number of areas in which performance of the model could potentially be enhanced. The areas addressed below represent our current thinking and will be prioritised and possibly added to following discussion at the first meeting of the proposed Expert Group convened to oversee this work.

External validation of the current model in critical care units in Scotland

The 2009 recalibration of the ICNARC model will be externally validated using SICSAG data from 2007 to 2009. Data and dataset definitions in the SICSAG dataset will be closely examined to identify the closest possible map to the data required for the ICNARC model, with any imperfections in the map documented and, where possible, the sensitivity of the model to these assumptions assessed in the CMP dataset.

Introduction of new important variables

Since the development of the ICNARC model and following an extensive literature review, the CMP dataset has undergone two revisions that have introduced additional physiological parameters that have some evidence to indicate that they may have a strong association with outcome, over and above the physiological parameters already included in the ICNARC model: blood lactate;19 and pupil reactivity.20 We will investigate the effect of incorporating these parameters into the ICNARC Physiology Score (the physiological component of the ICNARC model) either in addition to or in place of existing parameters.

Improved modelling of interactions between physiological parameters

All physiological parameters included in the ICNARC Physiology Score are currently assumed to act independently in an additive fashion. This same assumption has been used in most previous risk models. The one exception to this is the APACHE III Acute Physiology Score (which was carried forward unchanged into APACHE IV), which takes account of interactions between PaCO2 and pH and between the different components of the Glasgow Coma Score. These specific interactions were considered when developing the ICNARC model, but were not found to improve the model performance compared with simpler model formulations.

The assumption that all physiological parameters act independently on outcome, while appealing in terms of simplifying the model construction, is physiologically untrue. Consultation with experts in the body's physiological response to critical illness has identified areas in which interactions exist. We will draw up, a priori, a list of potential physiological parameters between

which an interaction would be expected. We will then investigate the effect on model performance of incorporating these interactions within the ICNARC Physiology Score. We will also investigate the potential for interactions between age and physiology, both in terms of individual parameters (e.g. blood pressure) and overall score.

Improved handling of missing data

Physiology data may be missing either because tests are not performed or, more rarely, because the results of tests are not available (e.g. due to missing patient notes). Tests may not be performed because the patient does not need the tests or because the patient deteriorates so rapidly that there is no time to perform the tests prior to death. Following the approach used for previous models, missing values for physiological measurements in the ICNARC model are assumed to be normal (i.e. scoring zero points on the ICNARC Physiology Score) on the assumption that physiological parameters are most likely to be missing because the test was not done as the treating clinician expected that the result would be normal. Advanced methodologies exist for handling missing data (e.g. multiple imputation). However, these methods rely on the assumption that data are “missing at random” that is, conditional on the observed data, the probability that an observation is missing does not depend on the value of the missing observation. This assumption may be considered to be at odds with the expectation that physiological parameters are more likely to be missing when they are normal. However, this may not be the case if the other data that were available to the clinician in deciding that a certain test was not required are also included in the dataset. We will explore the impact of alternative assumptions and modelling strategies for missing data on the performance of model.

Improved modelling of reasons for admission to/diagnosis in critical care

Reasons for admission to critical care are recorded in the CMP Database using the ICNARC Coding Method, a four-tiered (body system (e.g. respiratory)/anatomical site (e.g. lungs)/ physiological or pathological process (e.g. infection)/condition (e.g. bacterial pneumonia)), hierarchical coding method specifically developed for this purpose.21 Currently, coefficients for the ICNARC model are applied at only two tiers of this code – the individual condition (tier 4) or, if insufficient cases available to accurately estimate a coefficient at the condition level, the body system (tier 1). There is scope to improve the overall performance of the ICNARC model by considering the inclusion of coefficients allocated at intermediate levels of detail. The difficulty in this lies in combining together different conditions in a meaningful way. In some instances, incorporating the site (tier 2) may provide additional, important prognostic detail whereas, in others, it may be more important to incorporate the physiological or pathological process (tier 3) across multiple anatomical sites.

Modelling of mortality at fixed time-points or of time-to-event outcomes

Existing risk prediction models for critical care, including the ICNARC model, have been based on an outcome of mortality at hospital discharge. For the ICNARC model, and for our recalibrations of existing models, we improved on previous models by following-up patients to final discharge from acute hospital (rather than discharge from the hospital housing the critical care unit). However, this outcome may still be subject to bias as hospital discharge may not occur at the same time-point in a patient's recovery, in all sites, at all times, due to organisational factors involved in the discharge process such as the availability of rehabilitation or community services. In primary research, the most widely accepted outcomes are those measured at a fixed point in time, such as 30 or 90 days. For routine audit, such outcomes have usually been avoided due to the considerable workload (and therefore cost) of following up the outcomes of patients that have left hospital. However, the increasing ability to link routinely collected datasets may enable this barrier to be overcome. We will use the Medical Research Information Service (MRIS) at the NHS Information Centre to link data from the CMP with death registrations using available patient identifiers (NHS number, date of birth, sex, postcode) to obtain the date of death for patients dying after discharge from acute hospital. Obtaining the exact date of death will enable us to consider models evaluating mortality at a fixed time-point (e.g. 30 days, 90 days, 1 year) or using Cox regression models in place of logistic regression models to predict an outcome of time to death.

As these models will use different outcomes to that currently used for the ICNARC model, it is not appropriate to directly compare the performance of the different models using the performance measures identified above. Rather, the result of this element of the project is the potential to produce a suite of models predicting different outcomes, for which different models may be more appropriate in different settings. One aspect to consider in developing these models will be the improvements in performance from modelling physiology differently in predicting different outcomes versus the improvements in standardisation of using a single fixed physiology score across all models. In selecting a model (or models) to underpin the CMP going forward, issues to consider include: the appropriate time-point at which to assess outcome in order to balance rapid reporting (favouring a shorter timeframe) against finality of outcome (at 30 days 3% of all admissions will still be receiving critical care and a further 17% will not have been discharged from acute hospital); the completeness of data linkage (and consequently of outcome data); and the frequency and speed of data linkage for use in routine reporting.

3.5.4 *Objective 2a: Development and validation of novel models for cardiothoracic critical care*

In the UK, the vast majority of patients undergoing cardiothoracic surgery will subsequently be admitted to a dedicated specialist cardiothoracic critical care unit. Mortality for these patients is very low by comparison with other critically ill patient groups. Previous generation risk prediction models for critical care (e.g. APACHE II, SAPS II, MPM II) have excluded patients undergoing

cardiothoracic surgery from their development. The ICNARC model relaxed this assumption, making use of the introduction of an interaction term between acute physiological derangement and reason for admission to better model the outcome for these patients. However, this approach was still based only on a minority of patients admitted to a general intensive care unit following their cardiothoracic surgery and these patients may not be representative of the population of patients undergoing cardiothoracic surgery in the UK.

Data from admissions to specialist cardiothoracic critical care units currently participating in the CMP will be assessed to identify risk factors for acute hospital mortality. This will form the basis for preliminary investigations as to the improvement in model performance associated with developing a separate, novel model for these patients, as compared with direct application of the ICNARC model.

Later generations of the APACHE models (APACHE III and APACHE IV) have included separate models for patients undergoing coronary artery bypass grafting (CABG) that incorporate variables specific to these patients that are not included in the main models (and not collected in the CMP). Risk prediction models also exist for cardiothoracic surgery (rather than specifically cardiothoracic critical care), and these also incorporate other variables. For the purpose of this project, we will be limited to the variables available within the CMP. However, if the development of specific models for cardiothoracic critical care shows potential, then this work may form the basis for future, prospective research in a larger, more representative sample of cardiothoracic critical care units considering a wider scope of risk factors and for the establishment of a specialist national audit.

3.5.5 *Objective 2b: Development and validation of novel models for in-hospital cardiac arrest*

There is no established model for outcome prediction following in-hospital cardiac arrest. The NCAA includes data on established risk factors and both short-term (e.g. return of spontaneous circulation greater than 20 minutes) and medium-term (e.g. Cerebral Performance Category at hospital discharge) outcomes.

Risk models will be developed using the following process: Candidate variables will be investigated for their completeness and consistency in data collection across sites. The strategy for handling missing data will be informed from the outcome of objective 1d. Categorical variables will be examined to identify rarely used categories and, where relevant and following clinical advice, categories containing fewer than 50 events or fewer than 50 non-events will be collapsed into larger categories prior to analysis. Where more than one variable measures the same underlying concept, the co-linearity between the fields will be explored. Where possible, the best approach to modelling each concept (whether a single variable or a combination of variables) will be selected based on clinical considerations and data quality. Where this is not possible, multiple models based on the different alternative approaches will be compared and the best model selected on statistical

grounds. A full model will then be constructed consisting of all remaining candidate variables. The effect on model performance of combining together similar categories within variables will be explored. A stepwise process will then be employed to establish whether the model can be simplified without detrimental effect on model performance. The least significant variable will be removed from the model, and this will be repeated until no variables remain. At each step, the model will be fitted in development datasets (two thirds of subjects, split at the site level) and validated in validation datasets (the remaining one third of subjects). The split into development and validation datasets will be performed 100 times, and the measures of model performance averaged across these datasets. The performance of the model will be evaluated at each step to establish the best balance between model performance and simplicity. The final selected model will be refitted to the full dataset, and then validated as described above.

3.5.6 *Objective 2c: Development and validation of novel models for critical care units admitting low risk patients*

Traditionally, the care of critically ill patients was divided between 'high dependency units' and 'intensive care units', corresponding primarily to the level of nursing support required by a patient. More recently, the delivery of critical care has attempted to move away from these definitions based on location of care towards more objective criteria based on the support required by an individual patient regardless of their location. This resulted in the definition, by the Intensive Care Society in conjunction with the Department of Health, of Levels of critical care.22 Briefly, Levels 2 and 3 (corresponding approximately with high dependency and intensive care) can be defined as:

Level 2

- Patients receiving pre-operative optimisation
- Patients receiving extended postoperative care
- Patients stepping down from Level 3 care
- Patients receiving support for a single organ system (excluding advanced respiratory support, which is considered Level 3)
- Patients receiving basic respiratory support and basic cardiovascular support (with no other organ support)

Level 3

- Patients receiving advanced respiratory support alone
- Patients receiving support for two or more organ systems (excluding basic respiratory support and basic cardiovascular support, which is considered Level 2)

Models of service delivery and organisation of critical care vary across trusts and sites. While the general trend over recent years has been away from separate high dependency units and intensive care units toward more flexibly configured critical care units, a significant number of sites still deliver care in

separate units. The underlying assumption of current risk prediction models that outcome can be determined from age, acute physiological derangement, prior location, urgency of surgery and reason for admission to critical care may not hold equally across all units due to the potential selection bias introduced by these different models of delivery. To address this, we will apply the same modelling techniques to the same basic underlying variables as in the development of the ICNARC model but applied specifically to critical care units admitting a high proportion of low risk patients, including standalone Level 2 (high dependency) units and those that admit a high proportion of admissions following elective surgery.

3.5.7 Objective 3: Immediate translation of research results into practice

One major advantage of the nesting of this project within two, ongoing, national audits is the opportunity this provides for immediate and widespread adoption of the risk models, developed in this project, into routine use. As soon as models are considered to be final and fully validated, they will be integrated into the regular routine reports received by sites participating in the CMP and the NCAA.

4. Research governance

4.1 Approvals

Both the CMP and the NCAA have support under Section 251 of the NHS Act 2006 for the collection and use of limited patient identifiable data without consent (approval numbers PIAG 2-10(f)/2005 and ECC 2-06(n)/2009). Approval has been obtained from the Privacy Advisory Committee, NHS National Services Scotland, for the use of SICSAG data (reference 53/10).

5. Project management

Day-to-day management of the project will be undertaken by a Study Management Group consisting of the Statistical Research Fellow and the co-investigators. The Study Management Group will meet regularly to monitor the progress of the project against its timelines and milestones.

The project will be overseen by an Expert Group, providing expertise in statistical modelling, national clinical audit, NHS management, clinical input, and service user representation. The Expert Group will be made up of:

- Doug Altman, Professor of Statistics in Medicine, University of Oxford
- Nick Black, Professor of Health Services Research, London School of Hygiene and Tropical Medicine and chair of the National Clinical Audit Advisory Group
- James Carpenter, Reader in Medical and Social Statistics, London School of Hygiene and Tropical Medicine
- Gary Collins, Senior Medical Statistician, University of Oxford
- Maureen Dalziel, ICNARC trustee and service user representative
- Mike Grocott, Consultant in Intensive Care Medicine, Southampton

University Hospitals NHS Trust and Director, The Royal College of Anaesthetists Health Services Research Centre
- Steve Harris, Clinical Research Fellow, London School of Hygiene and Tropical Medicine
- Jon Nicholl, Professor of Health Services Research, University of Sheffield
- Andrew Padkin, Consultant in Intensive Care Medicine, Royal United Hospital Bath NHS Trust
- Graham Ramsay, Chief Executive, Mid Essex Hospital Services NHS Trust

6. Service user/public involvement

ICNARC involves service users at all levels of the organisation, including two trustees. In addition, ICNARC informs its work from the output of two modules funded by ICNARC for the Healthtalkonline website (http://www.healthtalkonline.org), with a further module on organ donation forthcoming.

For this project, we have selected one of our trustees who has a good understanding of the role of risk prediction modelling in performance assessment within the NHS. Maureen Dalziel will take a full and active place on the Expert Group, promoting the patient's perspective. Maureen has personal experience of critical care, having previously been admitted to a critical care unit with severe sepsis.

All involvement of service users in this study will follow the guidelines and recommendations for good practice from INVOLVE (http://www.invo.org.uk).

7. References

1. Department of Health. High Quality Care For All--NHS Next Stage Review Final Report. 2008. London, The Stationary Office.
2. Hutchings A, Durand MA, Grieve R et al. Evaluation of modernisation of adult critical care services in England: time series and cost effectiveness analysis. BMJ 2009; 339:b4353.
3. Moreno RP. Outcome prediction in intensive care: why we need to reinvent the wheel. Curr Opin Crit Care 2008; 14(5):483-4.
4. Harrison DA, Brady AR, Parry GJ et al. Recalibration of risk prediction models in a large multicenter cohort of admissions to adult, general critical care units in the United Kingdom. Crit Care Med 2006; 34(5):1378-88.
5. Harrison DA, Parry GJ, Carpenter JR et al. A new risk prediction model for critical care: the Intensive Care National Audit & Research Centre (ICNARC) model. Crit Care Med 2007; 35(4):1091-8.
6. Harrison DA, Brady AR, Parry GJ et al. Recalibration of risk prediction models in a large multicenter cohort of admissions to adult, general critical care units in the United Kingdom. Crit Care Med 2006; 34(5):1378-88.
7. Peduzzi P, Concato J, Kemper E et al. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol 1996; 49(12):1373-9.
8. Harrison DA, Brady AR, Parry GJ et al. Recalibration of risk prediction models in a large multicenter cohort of admissions to adult, general critical care units in the United Kingdom. Crit Care Med 2006; 34(5):1378-88.
9. Harrison DA, Parry GJ, Carpenter JR et al. A new risk prediction model for critical care: the Intensive Care National Audit & Research Centre (ICNARC) model. Crit Care Med 2007; 35(4):1091-8.
10. Harrell FE, Jr., Califf RM, Pryor DB et al. Evaluating the yield of medical tests. JAMA 1982; 247(18):2543-6.
11. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982; 143(1):29-36.
12. Hosmer Jr DW, Lemeshow S. Goodness-of-fit tests for the multiple logistic regression model. Commun Stat 1980; A9:1043-69.
13. Kramer AA, Zimmerman JE. Assessing the calibration of mortality benchmarks in critical care: The Hosmer-Lemeshow test revisited. Crit Care Med 2007; 35(9):2052-6.
14. Cox DR. Two further applications of a model for binary regression. Biometrika 1958; 45:562-5.
15. Shapiro AR. The evaluation of clinical predictions. A method and initial application. N Engl J Med 1977; 296(26):1509-14.
16. Brier GW. Verification of forecasts expressed in terms of probability. Monthly Weather Review 1950; 75:1-3.
17. Mittlbock M, Schemper M. Explained variation for logistic regression. Stat Med 1996; 15(19):1987-97.

18. Efron B. Estimating the error rate of a prediction rule: improvement on cross-validation. J Am Stat Assoc 1983; 78:316-31.
19. Smith I, Kumar P, Molloy S et al. Base excess and lactate as prognostic indicators for patients admitted to intensive care. Intensive Care Med 2001; 27(1):74-83.
20. Marmarou A, Lu J, Butcher I et al. Prognostic value of the Glasgow Coma Scale and pupil reactivity in traumatic brain injury assessed pre-hospital and on enrollment: an IMPACT analysis. J Neurotrauma 2007; 24(2):270-80.
21. Young JD, Goldfrad C, Rowan K. Development and testing of a hierarchical method to code the reason for admission to intensive care units: the ICNARC Coding Method. Intensive Care National Audit & Research Centre. Br J Anaesth 2001; 87(4):543-8.
22. Intensive Care Society. Levels of Critical Care for Adult Patients. 2009. London, Intensive Care Society.

This protocol refers to independent research commissioned by the National Institute for Health Research (NIHR). Any views and opinions expressed therein are those of the authors and do not necessarily reflect those of the NHS, the NIHR, the HSR programme or the Department of Health.