Prognosis and prognostic research: Developing a prognostic model

Patrick Royston,[1] Karel G M Moons,[2] Douglas G Altman,[3] Yvonne Vergouwe[2]

In the second article in their series, **Patrick Royston and colleagues** describe different approaches to building clinical prognostic models

[1]MRC Clinical Trials Unit, London NW1 2DA

[2]Julius Centre for Health Sciences and Primary Care, University Medical Centre Utrecht, Utrecht, Netherlands

[3]Centre for Statistics in Medicine, University of Oxford, Oxford OX2 6UD

Correspondence to: P Royston pr@ctu.mrc.ac.uk

Accepted: 6 October 2008

Cite this as: *BMJ* 2009;338:b604
doi: 10.1136/bmj.b604

The first article in this series reviewed why prognosis is important and how it is practised in different medical settings.[1] We also highlighted the difference between multivariable models used in aetiological research and those used in prognostic research and outlined the design characteristics for studies developing a prognostic model. In this article we focus on developing a multivariable prognostic model. We illustrate the statistical issues using a logistic regression model to predict the risk of a specific event. The principles largely apply to all multivariable regression methods, including models for continuous outcomes and for time to event outcomes.

The goal is to construct an accurate and discriminating prediction model from multiple variables. Models may be a complicated function of the predictors, as in weather forecasting, but in clinical applications considerations of practicality and face validity usually suggest a simple, interpretable model (as in box 1).

Surprisingly, there is no widely agreed approach to building a multivariable prognostic model from a set of candidate predictors. Katz gave a readable introduction to multivariable models,[3] and technical details are also widely available.[4-6] We concentrate here on a few fairly standard modelling approaches and also consider how to handle continuous predictors, such as age.

Preliminaries

We assume here that the available data are sufficiently accurate for prognosis and adequately represent the population of interest. Before starting to develop a multivariable prediction model, numerous decisions must be made that affect the model and therefore the conclusions of the research. These include:

- Selecting clinically relevant candidate predictors for possible inclusion in the model
- Evaluating the quality of the data and judging what to do about missing values
- Data handling decisions
- Choosing a strategy for selecting the important variables in the final model
- Deciding how to model continuous variables
- Selecting measure(s) of model performance[5] or predictive accuracy.

Other considerations include assessing the robustness of the model to influential observations and outliers, studying possible interaction between predictors, deciding whether and how to adjust the final model for overfitting (so called shrinkage),[5] and exploring the stability (reproducibility) of a model.[7]

Box 1 | Example of a prognostic model

Risk score from a logistic regression model to predict the risk of postoperative nausea or vomiting (PONV) within the first 24 hours after surgery[2]:

Risk score= −2.28+(1.27×female sex)+(0.65×history of PONV or motion sickness)+(0.72×non-smoking)+(0.78×postoperative opioid use)

where all variables are coded 0 for no or 1 for yes.

The value −2.28 is called the intercept and the other numbers are the estimated regression coefficients for the predictors, which indicate their mutually adjusted relative contribution to the outcome risk. The regression coefficients are log(odds ratios) for a change of 1 unit in the corresponding predictor.

The predicted risk (or probability) of PONV=1/(1+$e^{-\text{risk score}}$).

Box 2 | Modelling continuous predictors

Simple predictor transformations intended to detect and model non-linearity can be systematically identified using, for example, fractional polynomials, a generalisation of conventional polynomials (linear, quadratic, etc).[6 27] Power transformations of a predictor beyond squares and cubes, including reciprocals, logarithms, and square roots are allowed. These transformations contain a single term, but to enhance flexibility can be extended to two term models (eg, terms in log *x* and x^2). Fractional polynomial functions can successfully model non-linear relationships found in prognostic studies. The multivariable fractional polynomial procedure is an extension to multivariable models including at least one continuous predictor,[4 27] and combines backward elimination of weaker predictors with transformation of continuous predictors.

Restricted cubic splines are an alternative approach to modelling continuous predictors.[5] Their main advantage is their flexibility for representing a wide range of perhaps complex curve shapes. Drawbacks are the frequent occurrence of wiggles in fitted curves that may be unreal and open to misinterpretation[28 29] and the absence of a simple description of the fitted curve.

This article is the second in a series of four aiming to provide an accessible overview of the principles and methods of prognostic research

Selecting candidate predictors

Studies often measure more predictors than can sensibly be used in a model, and pruning is required. Predictors already reported as prognostic would normally be candidates. Predictors that are highly correlated with others contribute little independent information and may be excluded beforehand.[5] However, predictors that are not significant in univariable analysis should not be excluded as candidates.[8-10]

Evaluating data quality

There are no secure rules for evaluating the quality of data. Judgment is required. In principle, data used for developing a prognostic model should be fit for purpose. Measurements of candidate predictors and outcomes should be comparable across clinicians or study centres. Predictors known to have considerable measurement error may be unsuitable because this dilutes their prognostic information.

Modern statistical techniques (such as multiple imputation) can handle data sets with missing values.[11 12] However, all approaches make critical but essentially untestable assumptions about how the data went missing. The likely influence on the results increases with the amount of data that are missing. Missing data are seldom completely random. They are usually related, directly or indirectly, to other subject or disease characteristics, including the outcome under study. Thus exclusion of all individuals with a missing value leads not only to loss of statistical power but often to incorrect estimates of the predictive power of the model and specific predictors.[11] A complete case analysis may be sensible when few observations (say <5%) are missing.[5] If a candidate predictor has a lot of missing data it may be excluded because the problem is likely to recur.

Data handling decisions

Often, new variables need to be created (for example, diastolic and systolic blood pressure may be combined to give mean arterial pressure). For ordered categorical variables, such as stage of disease, collapsing of categories or a judicious choice of coding may be required. We advise against turning continuous predictors into dichotomies.[13] Keeping variables continuous is preferable since much more predictive information is retained.[14 15]

Selecting variables

No consensus exists on the best method for selecting variables. There are two main strategies, each with variants.

In the full model approach all the candidate variables are included in the model. This model is claimed to avoid overfitting and selection bias and provide correct standard errors and P values.[5] However, as many important preliminary choices must be made and it is often impractical to include all candidates, the full model is not always easy to define.

The backward elimination approach starts with all the candidate variables. A nominal significance level, often 5%, is chosen in advance. A sequence of hypothesis tests is applied to determine whether a given variable should be removed from the model. Backward elimination is preferable to forward selection (whereby the model is built up from the best candidate predictor).[16] The choice of significance level has a major effect on the number of variables selected. A 1% level almost always results in a model with fewer variables than a 5% level. Significance levels of 10% or 15% can result in inclusion of some unimportant variables, as can the full model approach. (A variant is the Akaike information criterion,[17] a measure of model fit that includes a penalty against large models and hence attempts to reduce overfitting. For a single predictor, the criterion equates to selection at 15.7% significance.[17])

Selection of predictors by significance testing, particularly at conventional significance levels, is known to produce selection bias and optimism as a result of overfitting, meaning that the model is (too) closely adapted to the data.[5 9 17] Selection bias means that a regression coefficient is overestimated, because the corresponding predictor is more likely to be significant if its estimated effect is larger (perhaps by chance) rather than smaller. Overfitting leads to worse prediction in independent data; it is more likely to occur in small data sets or with weakly predictive variables. Note, however, that selected predictor variables with very small P values (say, <0.001) are much less prone to selection bias and overfitting than weak predictors with P values near the nominal significance level. Commonly, prognostic data sets include a few strong predictors and several weaker ones.

Modelling continuous predictors

Handling continuous predictors in multivariable models is important. It is unwise to assume linearity as it can lead to misinterpretation of the influence of a predictor and to inaccurate predictions in new patients.[14] See box 2 for further comments on how to handle continuous predictors in prognostic modelling.

Assessing performance

The performance of a logistic regression model may be assessed in terms of calibration and discrimination. Calibration can be investigated by plotting the observed proportions of events against the predicted risks for groups defined by ranges of individual predicted risks; a common approach is to use 10 risk groups of equal size. Ideally, if the observed proportions of events and predicted probabilities agree over the whole range of probabilities, the plot shows a 45° line (that is, the slope is 1). This plot can be accompanied by the Hosmer-Lemeshow test,[19] although the test has limited power to assess poor calibration. The overall observed and predicted event probabilities are by definition equal for the sample used to develop the model. This is not guaranteed when the model's performance is evaluated on a different sample in a validation study. As we will discuss in the next article,[18] it is more difficult to get a model to perform well in an independent sample than in the development sample.

Various statistics can summarise discrimination between individuals with and without the outcome event. The area under the receiver operating curve,[10 20] or the equivalent *c* (concordance) index, is the chance that given two patients, one who will develop an event and the other who will not, the model will assign a higher probability of an event to the former. The *c* index for a prognostic model is typically between about 0.6 and 0.85 (higher values are seen in diagnostic settings[21]). Another measure is R^2, which for logistic regression assesses the explained variation in risk and is the square of the correlation between the observed outcome (0 or 1) and the predicted risk.[22]

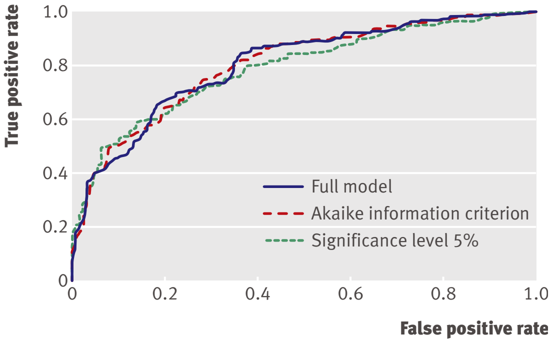

Fig 1 | Receiver operating characteristic (ROC) curves for three multivariable models of survival with kidney cancer

Example of prognostic model for survival with kidney cancer

Between 1992 and 1997, 350 patients with metastatic renal carcinoma entered a randomised trial comparing interferon alfa with medroxyprogesterone acetate at 31 centres in the UK.[23] Here we develop a prognostic model for the (binary) outcome of death in the first year versus survived 12 months or more. Of 347 patients with follow-up information, 218 (63%) died in the first year.

We took the following preliminary decisions before building the model:

- We chose 14 candidate predictors, including treatment, that had been reported to be prognostic
- Four predictors with more than 10% missing data were eliminated. Table 1 shows the 10 remaining candidate predictors (11 variables).
- WHO performance status (0, 1, 2) was modelled as a single entity with two dummy variables
- For illustration, we selected the significant predictors in the model using backward elimination with the Akaike information criterion[17] and using 0.05 as the significance level. We compared the results with the full model (all 10 predictors)
- Because of its skew distribution, time to metastasis was transformed to approximate normality by adding 1 day and taking logarithms. All other continuous predictors were initially modelled as linear
- For each model, we calculated the *c* index and receiver operating curves and assessed calibration using the Hosmer-Lemeshow test.

Table 1 shows the full model and the two reduced models selected by backward elimination using the Akaike information criterion and 5% significance. Positive regression coefficients indicate an increased risk of death over 12 months. None of the three models failed the Hosmer-Lemeshow goodness of fit test (all $P>0.4$).

Two important points emerge. Firstly, larger significance levels gave models with more predictors. Secondly, reducing the size of the model by reducing the significance level hardly affected the *c* index. Figure 1 shows the similar receiver operating curves for the three models. We note, however, that the *c* index has been criticised for its inability to detect meaningful differences.[24] As often happens, a few predictors were strongly influential and the remainder were relatively weak. Removing the weaker predictors had little effect on the *c* index.

An important goal of a prognostic model is to classify patients into risk groups. As an example, we can use as a cut-off value a risk score of 1.4 with the full model (vertical line, fig 2) which corresponds to a

Table 1 | Selected predictors of 12 month survival for patients with kidney cancer. Estimated mutually adjusted regression coefficient (standard error) for three multivariable models obtained using different strategies to select variables (see text)

Predictor*	Full model	Akaike information criterion	5% significance level
WHO performance status 1 (versus 0)	0.62 (0.30)	0.55 (0.29)	0.50 (0.2)
WHO performance status 2 (versus 0)	1.69 (0.42)	1.62 (0.41)	1.55 (0.40)
Haemoglobin (g/l)	−0.45 (0.08)	−0.44 (0.08)	−0.39 (0.08)
White cell count (×10^9/l)	0.12 (0.05)	0.13 (0.05)	0.13 (0.05)
Transformed time from diagnosis of metastatic disease to randomisation†	−0.29 (0.10)	−0.30 (0.10)	−0.27 (0.09)
Interferon treatment	−0.61 (0.27)	−0.61 (0.26)	−0.58 (0.26)
Nephrectomy	0.39 (0.29)	0.44 (0.28)	—
Female sex	−0.57 (0.29)	−0.56 (0.28)	—
Lung metastasis	−0.36 (0.28)	—	—
Age (per 10 year)	−0.07 (0.13)	—	—
Multiple sites of metastasis	−0.09 (0.36)	—	—
Intercept	6.54 (1.63)	5.70 (1.29)	4.99 (1.22)
C index	0.80 (0.02)	0.80 (0.02)	0.79 (0.02)

We assumed linear effects of continuous predictors. Details of the distribution of each candidate predictor have been omitted to save space.
*Binary variables are coded 0 for no, 1 for yes.
†log(days from metastasis to randomisation + 1).

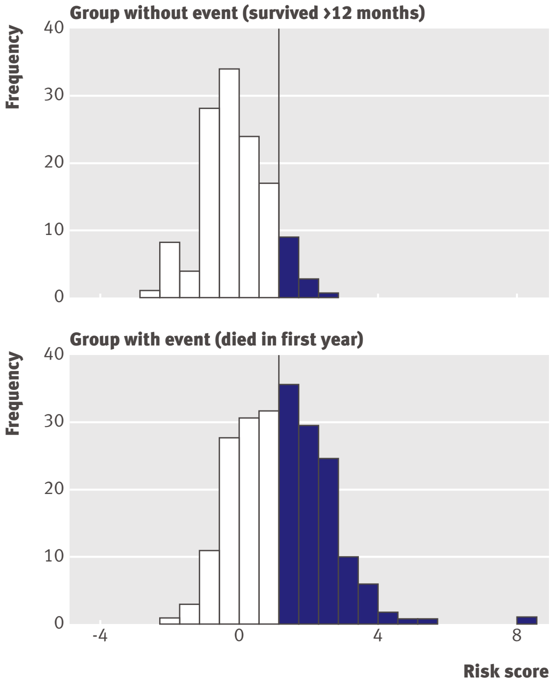

Fig 2 | Distribution of risk scores from full model for patients who survived at least 12 months or died within 12 months. The vertical line represents a risk score of 1.4, corresponding to an estimated death risk of 80%. The specificity (90%) is the proportion of patients in the upper panel whose risk is below 1.4. The sensitivity (46%) is the proportion of patients in the lower panel whose risk score is above 1.4

Table 2 | Multivariable model for 12 month survival in the kidney cancer data, based on multivariable fractional polynomials for continuous predictors and selection of variables at the 5% significance level

Predictor*	Coefficient (SE)
WHO performance status 1 (versus 0)	0.55 (0.29)
WHO performance status 2 (versus 0)	1.67 (0.41)
1/(haemoglobin/10)^2	3.86 (0.73)
Time from diagnosis of metastatic disease to randomisation (years)	−0.55 (0.23)
Interferon treatment	−0.64 (0.26)
Intercept	−2.21 (0.52)
C index	0.78 (0.02)

Predicted risks can be calculated from the following standard formula (as in box 1): risk score =−2.21+0.55 (if WHO performance status =1)+1.67 (if WHO performance status =2)+3.86/(haemoglobin/10)^2 −0.55×(time from diagnosis of metastatic disease to randomisation)−0.64 (if on interferon treatment). Predicted risk =1/(1 + $e^{-\text{risk score}}$).
*Binary variables are coded 0 for no or 1 for yes.

predicted risk of 80%. Patients with estimated risks above the cut-off value are predicted to die within 12 months, and those with risks below the cut-off to survive 12 months. The resulting false positive rate is 10% (specificity 90%) and the true positive rate (sensitivity) is 46%. The combination of false and true positive rates is shown in the receiver operating curve (fig 1) and more indirectly in the distribution of risk scores in fig 2. The overlap in risk scores between those who died or survived 12 months is considerable, showing that the interpretation of the *c* index of 0.8 (table 1) is not straightforward.[24]

Continuous predictors were next handled with the multivariable fractional polynomial procedure (see box 2) using backward elimination at the 5% significance level. Only one continuous predictor (haemoglobin) showed significant non-linearity, and the transformation 1/haemoglobin^2 was indicated. That variable was selected in the final model and white cell count was eliminated (table 2).

Figure 3 shows the association between haemoglobin concentration and 12 month mortality, when haemoglobin is included in the model in different ways. The model with haemoglobin as a 10 group categorical variable, although noisy, agreed much better with the model including the fractional polynomial form of haemoglobin than the other models. Low haemoglobin concentration seems to be more hazardous than the linear function suggested.

In this example, details of the model varied according to modelling choices but performance was quite similar across the different modelling methods. Although the fractional polynomial model described the association between haemoglobin and 12 month mortality better than the linear function, the gain in discrimination was limited. This may be explained by the small number of patients with low haemoglobin concentrations.

Discussion

We have illustrated several important aspects of developing a multivariable prognostic model with empirical data. Although there is no clear consensus on the best method of model building, the importance of having an adequate sample size and high quality data is widely agreed. Model building from small data sets requires particular care.[5 9 10] A model's performance is likely to be overestimated when it is developed and assessed on the same dataset. The problem is greatest with small sample sizes, many candidate predictors, and weakly influential predictors.[5 9 10] The amount of optimism in the model can be assessed and corrected by internal validation techniques.[5]

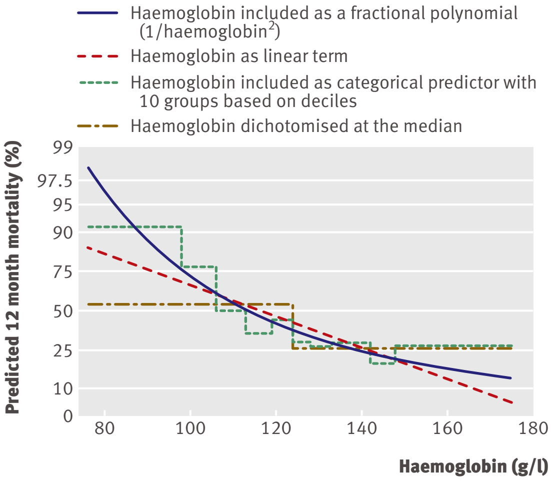

Fig 3 | Estimates of the association between haemoglobin and 12 month mortality in the kidney cancer data, adjusted for the variables in the Akaike information criterion derived model (table 1). The vertical scale is linear in the log odds of mortality and is therefore non-linear in relation to mortality

SUMMARY POINTS

Models with multiple variables can be developed to give accurate and discriminating predictions

In clinical practice simpler models are more practicable

There is no consensus on the ideal method for developing a model

Methods to develop simple, interpretable models are described and compared

Developing a model is a complex process, so readers of a report of a new prognostic model need to know sufficient details of the data handling and modelling methods.[25] All candidate predictors and those included in the final model and their explicit coding should be carefully reported. All regression coefficients should be reported (including the intercept) to allow readers to calculate risk predictions for their own patients.

The predictive performance or accuracy of a model may be adversely affected by poor methodological choices or weaknesses in the data. But even with a high quality model there may simply be too much unexplained variation to generate accurate predictions. A critical requirement of a multivariable model is thus transportability, or external validity—that is, confirmation that the model performs as expected in new but similar patients.[26] We consider these issues in the next two articles of this series.[18 21]

Funding: PR is supported by the UK Medical Research Council (U1228.06.001.00002.01). KGMM and YV are supported by the Netherlands Organization for Scientific Research (ZON-MW 917.46.360). DGA is supported by Cancer Research UK.

Contributors: The four articles in the series were conceived and planned by DGA, KGMM, PR, and YV. PR wrote the first draft of this article. All the authors contributed to subsequent revisions. PR is the guarantor.

Competing interests: None declared.

1 Moons KGM, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: What, why and how? *BMJ* 2009;339:b375.

2 Van den Bosch JE, Kalkman CJ, Vergouwe Y, Van Klei WA, Bonsel GJ, Grobbee DE, et al. Assessing the applicability of scoring systems for predicting postoperative nausea and vomiting. *Anaesthesia* 2005;60:323-31.

3 Katz MH. Multivariable analysis: a primer for readers of medical research. *Ann Intern Med* 2003;138:644-50.

4 Royston P, Sauerbrei W. Building multivariable regression models with continuous covariates in clinical epidemiology—with an emphasis on fractional polynomials. *Methods Inf Med* 2005;44:561-71.

5 Harrell FE Jr. *Regression modeling strategies with applications to linear models, logistic regression, and survival analysis*. New York: Springer, 2001.

6 Royston P, Sauerbrei W. *Multivariable model-building: a pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables*. Chichester: John Wiley, 2008.

7 Austin PC, Tu JV. Bootstrap methods for developing predictive models. *Am Stat* 2004;58:131-7.

8 Sun GW, Shook TL, Kay GL. Inappropriate use of bivariable analysis to screen risk factors for use in multivariable analysis. *J Clin Epidemiol* 1996;49:907-16.

9 Steyerberg EW, Eijkemans MJ, Harrell FE Jr, Habbema JD. Prognostic modeling with logistic regression analysis: in search of a sensible strategy in small data sets. *Med Decis Making* 2001;21:45-56.

10 Harrell FE Jr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* 1996;15:361-87.

11 Donders AR, van der Heijden GJ, Stijnen T, Moons KG. Review: a gentle introduction to imputation of missing values. *J Clin Epidemiol* 2006;59:1087-91.

12 Little RJA, Rubin DB. *Statistical analysis with missing data*. 2nd ed. New York: John Wiley, 2002.

13 Altman DG, Royston P. The cost of dichotomising continuous variables. *BMJ* 2006;332:1080.

14 Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. *Stat Med* 2006;25:127-41.

15 Blettner M, Sauerbrei W. Influence of model-building strategies on the results of a case-control study. *Stat Med* 1993;12:1325-38.

16 Mantel N. Why stepdown procedures in variable selection? *Technometrics* 1970;12:621-5.

17 Sauerbrei W. The use of resampling methods to simplify regression models in medical statistics. *Appl Stat* 1999;48:313-29.

18 Altman DG, Vergouwe Y, Royston P, Moons KGM. Validating a prognostic model. *BMJ* 2009;338:b605.

19 Hosmer DW, Lemeshow S. *Applied logistic regression*. 2nd ed. New York: Wiley, 2000.

20 Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982;143:29-36.

21 Moons KGM, Altman DG, Vergouwe Y, Royston P. Prognosis and prognostic research: application and impact of prediction models in clinical practice. *BMJ* 2009;338:b606.

22 DeMaris A. Explained variance in logistic regression. A Monte Carlo study of proposed measures. *Sociol Methods Res* 2002;31:27-74.

23 Ritchie A, Griffiths G, Parmar M, Fossa SD, Selby PJ, Cornbleet MA, et al. Interferon-alpha and survival in metastatic renal carcinoma: early results of a randomised controlled trial. *Lancet* 1999;353:14-7.

24 Pencina MJ, D'Agostino RB Sr, D'Agostino RB Jr, Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat Med* 2008;27:157-72.

25 Hernandez AV, Vergouwe Y, Steyerberg EW. Reporting of predictive logistic models should be based on evidence-based guidelines. *Chest* 2003;124:2034-5.

26 Altman DG, Royston P. What do we mean by validating a prognostic model? *Stat Med* 2000;19:453-73.

27 Sauerbrei W, Royston P. Building multivariable prognostic and diagnostic models: transformation of the predictors by using fractional polynomials. *J R Stat Soc Series A* 1999;162:71-94.

28 Rosenberg PS, Katki H, Swanson CA, Brown LM, Wacholder S, Hoover RN. Quantifying epidemiologic risk factors using non-parametric regression: model selection remains the greatest challenge. *Stat Med* 2003;22:3369-81.

29 Boucher KM, Slattery ML, Berry TD, Quesenberry C, Anderson K. Statistical methods in epidemiology: a comparison of statistical methods to analyze dose-response and trend analysis in epidemiologic studies. *J Clin Epidemiol* 1998;51:1223-33.

"The Q word"

Hell hath no fury like a nurse having heard "My, isn't it quiet today," usually by a doctor. According to healthcare folklore, its incantation will provoke the "shift from hell." Being a man of science, not superstition, I can see no reason why the phrase should have a malign influence over the workload of staff. I have had the audacity (read misfortune) to use the word on a variety of occasions, and it has never caused the sky to fall in, the earth to open up, or any reversal of fortune.

However, this Freudian slip will instantly single you out as an amateur, turn all your professional relationships sour, and lead to a volley of verbal reprimands from all within earshot. May I suggest fans of "the Q word" should instead use "somnambulistic"—it means the same thing. You will undoubtedly be congratulated on your verbosity, and, because no one will understand what it means, you can say it with much aplomb and without fear of retribution.

David Warriner core medical trainee year 1 in diabetes, Northern General Hospital, Sheffield orange_cyclist@hotmail.com

Cite this as: *BMJ* 2009;338:b1286