A REVIEW OF THE USE OF STATISTICAL REGRESSION MODELS TO INFORM COST EFFECTIVENESS ANALYSES WITHIN THE NICE TECHNOLOGY APPRAISALS PROGRAMME

REPORT BY THE DECISION SUPPORT UNIT

August 2012

Authors:
Ben Kearns, Roberta Ara, Allan Wailoo
School of Health and Related Research, University of Sheffield

Expert Working Group:
Prof Keith Abrams, Professor of Medical Statistics, University of Leicester
Prof Mike Campbell, Professor of Medical Statistics, University of Sheffield
Dr Monica Hernández, Research Fellow in Econometrics, University of Sheffield
Dr Andrea Manca, Senior Research Fellow, University of York
Dr Paul Tappenden, Senior Research Fellow, University of Sheffield

Decision Support Unit,

ScHARR, University of Sheffield, Regent Court, 30 Regent Street, Sheffield, S1 4DA

Tel (+44) (0)114 222 0734

E-mail dsuadmin@sheffield.ac.uk

ABOUT THE DECISION SUPPORT UNIT

The Decision Support Unit (DSU) is a collaboration between the Universities of Sheffield, York and Leicester. We also have members at the University of Bristol, London School of Hygiene and Tropical Medicine and Brunel University.

The DSU is commissioned by the National Institute for Health and Clinical Excellence (NICE) to provide a research and training resource to support the Institute's Technology Appraisal Programme. Please see our website for further information www.nicedsu.org.uk

The production of this document was funded by the National Institute for Health and Clinical Excellence (NICE) through its Decision Support Unit. The views, and any errors or omissions, expressed in this document are of the authors only. NICE may take account of part or all of this document if it considers it appropriate, but it is not bound to do so.

EXECUTIVE SUMMARY

Introduction: Decision analytic models (DAM) used to evaluate the cost-effectiveness of interventions are pivotal sources of evidence used in the NICE Technology Appraisal (TA) process. It is becoming increasingly common for parameter estimates used in the DAMs to be informed by some kind of regression analysis on individual patient level data but there is currently little guidance relating to reporting standards for such inputs.

Objectives: i) To identify the frequency of use of regression models in NICE TA submissions; the parameters they inform, and the amount of information reported to describe and support the analyses. ii) To produce suggestions for guidance on good practice in this area.

Method: A random sample of 79 Appraisal submissions was selected from all appraisals (n=111) issued since the publication of the updated NICE Methods Guide in June 2008. An extensive data extraction form was developed and used to extract information on model formulation, diagnostics, performance, and how the results (and their variability) are fed-into and propagated through the DAM. The focus was on the reporting and transparency of the analyses; we did not seek to make judgements about the appropriateness or otherwise of the analyses.

On completion of the review, our expert working group convened to discuss the results in detail. Recommendations for good practice were drafted together with a checklist for critiquing reporting standards in this area. Consensus and final versions were achieved iteratively through email correspondence.

Results: Of the 79 technology appraisals examined, 47 included at least one regression analysis and a total of 91 separate regression analyses were reported. 56 were de novo analyses provided by the manufacturer/sponsor of the technology (34 from Single Technology Appraisals and 22 from Multiple Technology Appraisals), while the remaining 35 were sourced from existing published literature. Over 50% involved health state utility values with the balance involving health care costs (11%) or probabilities of clinical events (35%).

For the de novo analyses, reporting was poorest around the sample size used, the justification of the type of model estimated, the selection of covariates used, the strategy for identifying the preferred final model and any validation used. Across all the analyses, there was potential for improvement in

the reporting of: the description of the dataset, the model type, the rationale for inclusion of model covariates, the validity of the final model and the uncertainty in the model.

Conclusion: Statistical regression models are in widespread use in NICE TAs yet reporting standards relating to basic information are poor. Whilst some of this may be due to the word limit imposed on TAs, there is still scope for improvement. This is important as increasing levels of reporting transparency enable policy decision makers to have increasing levels of confidence in the resulting estimates of cost-effectiveness. We suggest a series of recommendations that could be used for the minimum reporting requirements for any statistical regression analyses used in a DAM.

CONTENTS

TABLES

1. INTRODUCTION

1.1. BACKGROUND

NICE Technology Appraisals (TAs) are underpinned by the assessment of cost- effectiveness: the estimation of the differences in costs and health benefits between different health technologies. Typically, data from a clinical trial or trials are supplemented with additional data from a variety of sources within a decision analytic model (DAM) that is produced by either the manufacturer of the technology subject to appraisal, an independent assessment centre, or both.

Within these decision models it is becoming increasingly common to find that certain parameters are derived from some kind of regression analysis on individual patient level data. These types of regressions aim to estimate the value of some dependent variable conditional on the values of a set of explanatory variables. Typically, though not exclusively, these are multivariate in nature and may be used to estimate health state utility values, costs or clinical outcomes including risks and time to event.

Little guidance is offered to analysts in the current NICE Guide to the Methods of Technology Appraisal[1] as to how such regression analyses should be reported or how they should be incorporated into decision models. The purpose of this report is to review a selection of previous submissions to NICE appraisals. The review will seek to identify the frequency with which such regression models are used as inputs for cost effectiveness analysis, for which types of parameters and the types of information that are typically reported about them and how they were used. This provided the basis for expert group discussions which produced good practice guidelines in this area. It should be noted that the branch of regression modelling known as survival analysis is excluded from this review. This is because this specific topic is covered in detail in a NICE Technical Support Document[2] though it is recognised that there will inevitably be some overlap. Similarly, we have restricted ourselves to the consideration of the results of analyses of primary data rather than analyses that are based on synthesis of evidence from different sources.

2. METHODS

2.1. SAMPLE

A random sample of 79 Appraisals was selected from all appraisals (n=111) issued since the publication of the updated NICE Methods Guide in June 2008. A total of 56 appraisals undertaken

under the Single Technology Appraisal (STA) process were included and 23 using the Multiple Technology Appraisal (MTA) process.

2.2. REVIEW METHODS

Each TA selected was manually searched by one reviewer to identify all cases where a regression model was used as an input to the cost effectiveness analysis. Where the TA reported using existing regression results from published literature the original articles were retrieved and reviewed.

An extensive data extraction form was developed by the project team. The form aimed to extract information on the approaches taken to model formulation (including the type of model estimated and the choice of covariates), diagnostics (how the model performed particularly with respect to any assumptions made), and performance (including comparisons with other plausible models). We were also interested in how the results of a statistical regression analysis (and variability) are fed-into and propagated through the DAM, this was also included in the data extraction form. An iterative process was followed whereby additional fields were added to the extraction form as the review took place. Relevant data was extracted by two reviewers. While the majority of data was of the form a 'yes/no' response to questions relating to the reporting of information, free text was extracted where it was felt that additional detail would be useful.

The focus of the data extraction was on the reporting and transparency of the analyses undertaken and as presented to the appraisal committee. We did not seek to make judgements about the appropriateness or otherwise of those analyses. Furthermore, we reviewed only the documents provided by those undertaking relevant analyses. We did not review, for example, the electronic versions of the decision models that would allow us to check how the regression analyses were incorporated and whether that was consistent with the written report.

2.3. EXPERT WORKING GROUP

Once the data were collated and summarised, we conducted a workshop attended by all members of our expert working group. The results of the review and suggestions of good practice were discussed in detail. A consensus was formed and used to draft a list of suggested recommendations and a checklist for critiquing the reporting standards used to describe statistical regression models used in TAs. Consensus on the final versions was obtained iteratively using email correspondence.

3. RESULTS

Of the 79 technology appraisals examined, 47 included at least one regression analysis and a total of 91 regression analyses were reported. Of the 91 regressions, 56 were de novo analyses generated to inform the specific DAM (34 from STAs and 22 from MTAs), while the remaining were based on 35 articles from the literature (32 journal articles, 3 Health Technology Assessments).

3.1. SUBJECT OF REGRESSION

Health related quality of life measures were the dependent variable in the majority of the existing (23/35) analyses and about half of the de novo analyses (26/56) (Table 1). When health care costs were the dependent variable (10/91), the statistical models were more likely to be obtained from existing analyses (8/91) reported in the literature. Conversely, where the probability of an event occurring was the dependent variable (32/91), these were more likely to involve de novo analyses (28/91).

Table 1: Subject of regression analysis

	Number (percentage)	
Dependent variable	TA	Literature
	De novo analyses (n=56)	Existing analyses (n=35)
Utilities	26 (46)	23 (66)
Costs	2 (3)	8 (23)
Probability of an event	28 (50)	4 (11)

3.2. DESCRIPTION OF DATASET

Less than half of the de novo analyses described the total dataset either numerically (n=20/56) or graphically (n=3/56) (Table 2). Whenever a de novo analysis included a graphical summary it also included a numerical summary (n=3/56). Conversely, a large proportion of the data used in the existing analyses reported in the literature were described numerically (27/35).

There can be a substantial difference between the number of observations in a dataset and the number of observations used for any specific statistical model due to missing data relating to individual covariates. Both should be described. Similarly, while the actual sample size available for the statistical analysis was reported for all the studies in the existing literature (35/35), just over half (31/56) of the de novo analyses provided the total sample size used. For creating the final

regression model, the sample size was poorly reported in the de novo analyses (17/56) when compared to the existing analyses reported in the literature (24/35).

Table 2: Descriptions of datasets used for estimation

	Number (percentage)	
	TA	Literature
	De novo analyses (n=56)	Existing analyses (n=35)
Data summarised numerically	20 (36)	27 (77)
Data displayed graphically	3 (5)	8 (23)
Total size of dataset available reported	31 (55)	35 (100)
Total size of sample used for final model reported	17 (30)	24 (69)

3.3. SELECTION OF MODEL AND VARIABLES

A rationale was provided for the selection of potential explanatory variables or other aspects of the modelling approach for about a third of the de novo (16/56), and two thirds of the existing (23/35) statistical models drawn from the literature (Table 3). The most common situation in which a rationale was given in the de novo analyses was for probabilities (10/16); either justifying the need for a regression analysis, the methods employed or the source used. Interactions between the explanatory variables were rarely reported (7/91) as having been examined, particularly for the de novo analyses (2/56).

Expert opinion was sometimes used to inform the selection of explanatory variables used in the de novo (6/56) and existing (10/35) statistical models as well as their expected direction of influence. About half of the de novo analyses (30/56) gave no justification for the explanatory variables used in the final model compared to less than a quarter of those reported in the existing literature (8/35). Less than a fifth of the TAs (10/56) reported that a sub-set of possible explanatory variables were used in the final statistical model, based on some form of stepwise selection.

The type of model estimated was explicitly stated in almost all existing analyses (34/35) and over three quarters of de novo analyses (45/56), and the vast majority were linear models estimated using ordinary least squares. The statistical software used was under-reported (22/91) in general and rarely reported for the de novo analyses (9/56).

Table 3: Selection of preferred model and covariates

	Number (percentage)	
	TA	Literature
	De novo analyses (n=56)	Existing analyses (n=35)
Rationale presented	16 (29)	23 (66)
Interactions explored	2 (4)	5 (14)
Selection of covariates in final model		
Rationale given	6 (11)	10 (29)
Stepwise	10 (18)	11 (31)
Explicitly stated no model reduction explored	4 (7)	6 (17)
Single explanatory variable	10 (18)	4 (11)
Not reported?	30 (54)	8 (23)
Model type		
Type of model clearly stated	45 (80)	34 (97)
More than one type of model considered	1 (2)	5 (14)

3.4. MODEL DIAGNOSTICS AND VALIDATION

The data in Table 4 show that in most studies, the coefficient estimates were reported for the final selected model but there were some cases, particularly in the de novo analyses where this did not occur (3/35 of the published analyses, 8/56 of the de novo analyses). Where coefficients were not reported in the published analyses, selected results were stated in the text, and these were used in the submission. Standard errors, confidence intervals or p values were poorly reported in general and were not presented for nearly half of all the de novo analyses (29/56) making it difficult to assess if there were statistically significant relationships between the dependent and explanatory variables.

Summary measures of overall model fit such as R^2 were not provided for two thirds (62/91) of all the statistical models and the reporting rate was even lower for the de novo analyses (11/56). The two de novo analyses that reported a measure based on the mean absolute error (MAE) also reported R^2. Reporting of MAE type statistics was higher in the existing literature (9/35).

Few of the analyses directly compared the observed and predicted values, either graphically (15/91) or numerically (7/91). Only the observed data were presented in 4 of the analyses and only the

predicted data were presented in 7 of the analyses. In 5 other studies, plots or data were presented but it was unclear whether they related to the fitted or observed values.

Few analyses reported carrying out a residual analysis (11/91); this was especially low for de novo analyses (1/56). This de novo analysis was also the only one to consider any other diagnostics for model validation (in this case a test for autocorrelation); methods of model validation were not reported in any of the other de novo analyses (55/56). Reporting rates of residual analyses were slightly higher for the existing literature (13/35).

Table 4: Model fit and diagnostics

	Number (percentage)	
	TA	Literature
	De novo analyses (n=56)	Existing analyses (n=35)
Model reporting:		
Beta coefficients presented	48 (86)	32 (91)
Standard errors (SE) reported	15 (27)	13 (37)
Confidence intervals (CI) reported	13 (23)	7 (20)
P-values	20 (36)	20 (57)
At least 1 of (SE, CI, P-value)	29 (52)	27 (77)
Summary goodness of fit		
R^2	11 (20)	18 (51)
MAE/RMSE/ or similar	2 (4)	9 (26)
Other	0 (0)	7 (20)
None reported	45 (80)	14 (40)
Observed vs. fitted values		
Compared graphically	8 (14)	7 (20)
Compared numerically	3 (5)	4 (11)
Only observed values provided	1 (2)	3 (9)
Only predicted values provided	5 (9)	2 (6)
Unclear what is provided	4 (7)	1 (3)
No information	31 (55)	13 (37)
Other methods of model validation		
Consideration of residuals	1 (2)	10 (29)
Multicollinearity	0 (0)	3 (9)
Other diagnostics	1 (2)	5 (14)
No other methods considered	55 (98)	22 (63)

MAE – Mean Absolute Error, RMSE – Root Mean Squared Error

3.5. MODEL PLAUSIBILITY AND ROBUSTNESS

The plausibility of the estimated coefficients for the models generated were compared with the literature for approximately one quarter (13/56) of the de novo analyses and two thirds (20/35) of the existing analyses (Table 5). While the magnitude and direction of the estimated coefficients were discussed for the majority of the existing analyses (28/35), less than a half of the de novo analyses (25/56) discussed the validity of these.

A check of model robustness by fitting multiple model types to the same data was reported in fewer de novo analyses (6/56) than literature analyses (14/35). Informal comparisons with external data were not reported often (8/91), but the reporting of formal comparisons was more common (16/91).

Table 5: Validity of regression models

	Number (percentage)	
	TA	Literature
	De novo analyses (n=56)	Existing analyses (n=35)
Model plausibility		
Estimates compared with the literature	13 (23)	20 (57)
Estimates checked for face validity	25 (45)	28 (80)
Model robustness		
Comparison of different model types in same dataset (internal)	6 (11)	14 (40)
Informal comparison in different dataset (external)	4 (7)	4 (11)
Formal comparison in different dataset (external)	5 (9)	11 (31)

3.6. USE OF UNCERTAINTY IN THE DAM

When de novo analyses are conducted, the analysts have access to the raw data hence it is possible to capture the full range of uncertainty in the data. However, the majority (31/56) of the DAMs utilising the results of de novo analyses did not report any incorporation of uncertainty in the estimates at all (Table 6). We did not assess the actual implementation of the regression models in the executable models and it may be that this is simply under reported. Probabilistic sensitivity analysis (PSA) was used to incorporate uncertainty in the DAM in less than half (24/56) of the cases. Of these, just 14 reflected the joint uncertainty between the coefficient estimates, including their correlations, and 14 examined the effects on the economic results using univariate sensitivity analyses.

None of the literature analyses reported the variance-covariance matrix, so it was not possible to incorporate any joint uncertainty into a PSA. Where an analysis from the literature was used in a DAM, uncertainty was explored in just over half of cases (19/35). This was usually in the form of a scenario analysis (18/35), of which the majority were using alternative values from the literature

(15/18). In addition, in seven occasions it was reported that the results from the literature were used in a PSA, taking into account the reported uncertainty.

The reporting of uncertainty in the existing literature, as reflected in the DAM, is naturally limited by the reporting of uncertainty in the literature itself.

Table 6: Description of how uncertainty in the regression model is reflected in the DAM

	Number (percentage)	
	TA	Literature
	De novo analyses (n=56)	Existing analyses (n=35)
Uncertainty in data not explored	31 (55)	16 (46)
Scenario Analysis		
Arbitrary alternatives	7 (13)	2 (6)
Best-case / Worst Case	2 (4)	1 (3)
Alternatives from literature	5 (9)	15 (43)
Probabilistic sensitivity analysis		
Arbitrary distribution	4 (7)	0 (0)
Distribution informed by data, but ignoring correlations	6 (11)	9 (26)*
Distribution informed by data, including correlations (if any).	14 (25)	

* Data combined due to insufficient detail published in the literature.

4. DISCUSSION

The main objective of this report is to determine if submissions to the NICE TA programme provide sufficient information to allow recipients of that information, particularly committee members, but also those that critique both industry and assessment group submissions, to reach a judgement as to the suitability of the analysis provided. It is important to note that often analyses are undertaken using datasets that are not in the public domain and rarely are such data provided for those that wish to review or replicate statistical analyses. It should also be noted that word limitations apply to submissions to the NICE TA programme, restricting the amount of information that can be reported. Separate aspects of a regression analyses process were identified, and levels of reporting were presented for these. Comparisons were made between reporting levels in those analyses undertaken

and reported within the TA submission and those which were drawn from previously published literature.

The results shown here suggest that, in general, reporting standards are poor for regression analyses from all sources. In the analyses identified within the review, most aspects of the regression analysis were reported in less than half of the TA submissions. With the exception of confidence intervals for parameter estimates, it was found that reporting levels were always lower in the de novo analyses compared to those appearing in previously published literature. As a minimum it would seem reasonable that the analyses submitted in order to inform NHS decision making should meet the standards required of a peer reviewed publication.

Compared to the reporting levels seen in journal articles, the de novo TA analyses were poor at reporting the following aspects:

- Stating the sample size used for each regression model
- Justifying either the type of model estimated or the selection of covariates by reference to existing literature
- Stating the strategy employed to select the preferred final model
- Any type of formal check for model validation

For both de novo and published analyses, the following areas were identified as having the potential for improved reporting:

- Summarising the dataset and providing a rationale for the type of model employed
- The selection of covariates in the final model. If other models were considered, on what basis was the preferred model selected
- The validity of the model, particularly regarding examination of the distribution of the residuals compared to their assumed distribution(s)
- Variance of parameter estimates, how their uncertainty and associated correlations have been accounted for within the economic model (when possible)

The increased standard of reporting observed in journal articles is to be expected. Before publication journal articles undergo peer review, which may highlight cases of inadequate reporting. Many journals now have either an explicit statistical peer-review or they publish guidelines for the presentation of statistical information; for example the BMJ present two such articles,[3,4] whilst the Uniform Requirements for Manuscripts Submitted to Biomedical Journals states that articles should:

> "*Describe statistical methods with enough detail to enable a knowledgeable reader with access to the original data to verify the reported results.*" (available at http://www.icmje.org/manuscript_1prepare.html)

In contrast, the most detailed guidance (prior to these guidelines) currently available for de novo analyses; the NICE Methods Guide,[1] states that:

> *"As much detail as possible on the data used in the analysis should be provided."(page 44)*

There are some limitations to this review. First, we had difficulties in accessing some of the required manufacturers' submission reports for the MTAs, and did not have access to the full information in some of the MTA and STA reports due to commercial (or academic) in confidence data. However, while the sample size (n=79) used in the study was relatively small, this does represent a substantial proportion of all NICE appraisals and over half (47/79) of the TAs selected included at least one statistical regression model (n=91 in total), illustrating how prevalent they are. We concentrated on evaluating the reporting standards for the statistical regression analyses presented in the TAs and did not seek to make judgements as to the appropriateness of any of the analyses identified. It is hoped that by increasing reporting standards and transparency, the quality of the corresponding analyses will also increase as it will be easier to identify areas for improvement.

5. CONCLUSION

Any evidence used in a reimbursement submission should be clearly and transparently described to ensure that policy decision makers, such as the NICE appraisal committee, and end-users, such as commissioners and clinicians, are confident they have the full facts needed to reach an informed decision. NICE TA guidance is informed by cost-effectiveness analysis using cost per QALY thresholds, inter alia, and as results generated from DAMs can be sensitive to changes in parameter values, it is important that any uncertainty is accurately characterised. Our review clearly shows that there is widespread use of statistical regression models as inputs to DAMs but there is scope for improvement in the reporting basic information such as the sample size used in the final regression. Any lack of transparency in the data, methodologies employed, validity of, and uncertainty in the results limits the confidence that policy decision makers can place in the evidence base.

We suggest a list of recommendations that could be used as the minimum reporting requirements for any statistical regression analyses used in a DAM. The recommendations relate to both the use of de novo analyses and statistical regression models sourced from the published literature.

6. RECOMMENDATIONS

The following section presents recommendations on what should be reported in any regression analysis, along with justifications. The recommendations are broken down into five stages. Attention

is drawn to any unresolved methodological issues. The recommendations are then summarised in a checklist (provided in the Appendix) which could potentially be used by analysts performing the regressions, authors writing a description of the methodologies and results, and anyone wishing to critically appraise the regression models presented. The checklist is an indicator and all items may not be appropriate for all regression analyses.

Pre-modelling considerations

The objectives of the analysis should be explicitly stated, as these will affect the subsequent methods employed.[5] For example, if the objective is to 'map' health utility values calculated from one instrument onto another then this should be stated, along with the variables used and how the analysis will proceed.

In addition to stating the objectives, the use of regression methods to satisfy these objectives should be justified. It is insufficient to state that the model used was reported in a previous HTA submission. Whilst regression methods offer a flexible and powerful framework for analysing data, they rely on certain assumptions and the availability of data of sufficient quality and quantity. Hence in some situations alternative methods of analysis may be more appropriate;[6] for example using contingency tables or simple averages.

The source from which the data were obtained should be stated and the data itself should be described and summarised. The sample size available should always be stated and potential explanatory variables in the dataset should be described (for example, the classification scheme used or unit of measurement) and summarised in sufficient detail. Both numerical and graphical methods should be considered; the former for displaying quantitative features and the latter for qualitative features. Both scatter plots and box-plots are particularly useful ways of displaying the relationship between the dependent and explanatory variables. If interest centres on a treatment effect then histograms with kernel density estimates by treatment group are also useful. Further details on the best-practice for displaying and summarising data may be found in Freeman[7], Freeman, Walters and Campbell,[7] Few[8] and Altman.[9]

It is important that enough details are provided to judge the quality of the data used in the regression analysis, as it will affect the usefulness of any results and hence the confidence that can be placed in them.[10] Possible and actual ranges for the variables used in the regression should be stated, so it is

clear if the results of the regression analysis are to be used to make predictions outside the observed values.

Together the objectives of the analysis and the quality of the data will indicate the regression model(s) that should be considered. These should be explicitly stated, along with any key assumptions that they require. For example, many types of model make the assumption that the outcome (or a transformation of it) has a linear association with any continuous predictor variables.

Arriving at the final model

Given a body of data, there are a wide variety of different ways that a regression model can be built. There are also a wide variety of different recommendations on what authors believe to be the best practice when performing regressions. For example, Harrell et al[11] argue against any form of model reduction, whilst Royston and Sauerbrei[12] prefer it. Nelder[13] recommends parsimonious models, whilst Breiman[14] recommends complex models. One point on which all authors agree is the importance of using common sense and subject-matter knowledge to inform and guide any regression analysis.[15]

Whatever strategy is employed to analyse the data, this should be described with sufficient detail. For example, if a subset of the potential explanatory variables is used, then the criteria for entry and/or removal from the final model should be stated. Any deviations from the norm (such as interactions or polynomials) should also be reported if considered, even if they do not remain in the final model.

It should be noted that there is not always consensus on what should be reported. For example Campbell[16] states that for logistic regression the method used to derive the p-value (e.g. likelihood ratio, Wald or score) should be reported, but Royston and Sauerbrei[12] feel this is largely unimportant. Given the need for balance between reporting sufficient detail and 'swamping' the reader with output, not everything can be reported. As a minimum the use of any non-standard methods, such as robust estimation or bootstrapped confidence intervals, should be reported. In some instances it may be useful to provide additional information, although when this is necessary will be highly context-specific.

When data on multiple covariates are available it is likely that they will have varying degrees of missing data which can lead to models being derived from very different numbers of observations to the original sample size. Hence the actual sample size used for each statistical model should be

reported together with details on how missing data (if any) were handled. For example, if there are 20 potential explanatory variables, each with 10% of their values missing (on average), then using the full model without imputation will retain about 13% of the original observations.[17]

Presentation of the final model

The estimated coefficients for all variables in the final model should be displayed, along with indications of both their uncertainty and the strength of association. The uncertainty can be conveyed using confidence intervals, standard errors or p-values. There are relationships between these measures (i.e. for a known sample size and distribution, any can be derived from any of the others), and at least the confidence intervals or standard errors should be reported. If there are word constraints, the asterisk system can be used to denote significance. The variance-covariance matrix should also be reported, although this could be in an appendix.

Validating the final model

There are a wide range of plots and statistics available for model validation (model criticism). Similarly, there are many different methods by which models may be criticised or validated; different objectives or model types can and do require different methods.

Modelling assumptions can be checked be examining the residuals.[18] If the model fits the data well then its residuals should not have any systematic patterns and they should have a mean value equal to zero. A wide variety of different residual plots can be used to check if the results of the regression analysis are adequate, for more details see Machin, Campbell and Walters.[19] In many cases examination of the fit across subsets of the data is also useful. Plots may be complemented by numerical values of the residuals such as the root-mean-squared-error. It is important that evidence of a residual analysis is provided and it is not sufficient to simply state that an analysis was conducted.

Residual analysis may reveal outlying observations or groups of observations. Numerical diagnostics may be used to quantify the impact of these on the model. These may indicate that certain sub-groups should be modelled separately. It should be stressed that observations should not be removed from an analysis unless it is known that they are genuine anomalous observations. Any deletion should be detailed and justified in the narrative.

In addition to the analysis of residuals, summary measures of goodness of fit may be used, such as information criteria or R^2-type measures. Calculation of these summary measures usually depends on the sample size (or range of data used) and model structure used (such as linear or logistic regression), and so comparisons should only be made between models of the same structure built using the same data. It should be noted that for non-linear models there are multiple different ways to calculate R^2-type measures, with no consensus as to which should be used. In addition Hosmer and Lemeshow[20] state that R^2–type measures do not measure goodness of fit for logistic regression, and should only be used for comparing models. There are a variety of different information criteria (IC) such as Akaike's IC and deviance IC that may be used. For non-linear models goodness of fit measures based on the residuals (such as the deviance) or information criteria are preferred. There are sometimes also measures specific to the model type, such as the Hosmer and Lemeshow test for logistic regression.[20]

Ideally the regression model would also be validated by applying it to external data. This is often not possible, so methods for creating quasi-external data are available. However, it should be noted that there is no consensus on the method for using external (or quasi external) data for validation. For more details see Chatfield[6] and Good and Hardin.[15]

The importance of incorporating common sense and existing knowledge into any analysis has been previously stressed. This should continue after the model has been created, to check if it has face validity, and if its results (predicted values and/or interpretations of the parameters) agree with previously published results. These checks are also assessments of model performance; it may not always be possible to quantify their results in the same way as goodness of fit, but they are just as important (if not more).

Acknowledging and propagating uncertainty in the analysis

An important aspect of a regression analysis is its ability to capture and quantify uncertainty and this information should be propagated through into the DAM. With regards to estimates of parameter coefficients this information is conveyed by their standard errors and their covariances when using classical methods. The variance-covariance matrix should be reported and made available in some form so that it may be used to propagate the uncertainty of a regression analysis through a DAM using PSA.[21,22]

If structural uncertainty is explored, for example using scenario analyses or model averaging,[23] this should also be reported.

Finally, any limitations of the regression model and its range of applications should be noted along with any potential sources of bias. These may have arisen during any of the previous modelling stages. Where feasible, the potential impact of these limitations on the analysis and its use within the DAM should be explored.

7. REFERENCES

1. National Institute for health and Clinical Excellence. Guide to the methods of technology appraisal (updated June 2008). 2008.

2. Latimer, N. NICE DSU Technical Support Document 14: Survival analysis for economics evaluations alongside clinical trials - extrapolation with patient-level data. 2011; available from http://www.nicedsu.org.uk

3. Altman, D.G., Gore, S.M., Gardner, M.J., Pocock, S.J. Statistical guidelines for contributors to medical journals. *British Medical Journal (Clinical Research Ed)* 1983; 286(6376):1489-1493.

4. Lynch, A., Palmer, C. Statistical advice for contributors. 2012; available from http://group.bmj.com/products/journals/instructions-for-authors/statadvice.pdf (accessed April 2012).

5. Cox, D.R., Snell, E.J. Applied statistics: principles and examples. Chapman & Hall, London; 1981.

6. Chatfield, C. The initial examination of data. *Journal of the Royal Statistical Society Series A (Statistics in Society)* 1985; 148(3):214-253.

7. Freeman, J.V., Walters, S.J., Campbell, M.J. How to display data. BMJ Books, Oxford; 2008.

8. Few, S. Show me the Numbers. 2004. California, Analytics Press.

9. Altman, D.G. Statistics and ethics in medical research. VI--Presentation of results. *British Medical Journal* 1980; 281(6254):1542-1544.

10. Hand, D.J. Statistics and data mining: intersecting disciplines. *ACM SIGKDD Explorations Newsletter* 1999; 1(1):16-19.

11. Harrell, F.E., Lee, K.L., Mark, D.B. Tutorial in biostatistics multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine* 1996; 15:361-387.

12. Royston, P., Suaerbrei, W. Multivariable model-building: A pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables. John Wiley & Sons, Chichester; 2008.

13. Nelder, J.A. Statistics, science and technology. *Journal of the Royal Statistical Society Series A (Statistics in Society)* 1986; 149(2):109-121.

14. Breiman, L. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science* 2001; 16(3):199-231.

15. Good, P., Hardin, J. Common mistakes in statistics (and how to avoid them). John Wiley & Sons, New Jersey; 2003.

16. Campbell, M.J. Statistics at square two: understanding modern statistical applications in medicine. BMJ Books, Oxford; 2006.

17. Little, R.J.A., Rubin, D.B. Statistical analysis with missing data. Second ed. Wiley, New York; 2002.

18. Tufte, E. Improving data analysis in political science. *World Politics* 1969; 21(4):641-54.

19. Machin, D., Campbell, M.J., Walters, S.J. Medical statistics. Fourth ed. John Wiley & Sons, Chichester; 2007.

20. Hosmer, D.W., Lemeshow, S. Applied logistic regression. Second ed. John Wiley & Sons, New York; 2000.

21. Briggs, A., Sculpher, M., Claxton, K. Decision modelling for health economic evaluation. Oxford University Press, Oxford; 2006.

22. Stevenson, M., Tappenden, P., Squires, H. Methods for handling uncertainty within pharmaceutical funding decisions. *International Journal of Systems Science (in Press)* 2012.

23. Strong, M., Oakley, J.E., Chilcott, J. Managing structural uncertainty in health economic decision models: a discrepancy approach. *Journal of the Royal Statistical Society Series C (Applied Statistics)* 2012; 61(1):25-45.

8. APPENDIX (Checklist)

PROPOSED CHECKLIST FOR STATISTICAL REGRESSION ANALYSES

Pre-modelling considerations

1. Have the objectives of the analysis been stated?
2. Has the need for a de novo regression analysis been justified?
3. Has the source of the data used been stated? This would include synopses of key study features such as socio-demographic/clinical characteristics and the data collection method.
4. Has the total sample size available been reported?
5. Are sufficient explanations of all variables used provided?
6. Are sufficient numerical and/or graphical summaries provided?
7. Has the quality of data (missing values, outliers, possible bias, etc) been described?
8. Has the type/method of regression model(s) considered been stated/justified?
9. Have any modelling assumptions been stated?
10. Is a convincing rationale given for the inclusion of explanatory variables?

Arriving at the final model

11. Are sufficient details about the computational methods used provided?
12. If more than one model was considered, has justification been given for why the preferred model has been selected?
13. Has the choice of covariates been justified?
14. Is the sample size reported for every model presented?
15. Has the handling of missing values (if any) been described?

Presentation of the final model

16. Are the coefficient estimates provided?
17. Are appropriate measures of uncertainty and significance provided?

Validating the final model

18. Are summary measures of goodness of fit presented?
19. Are details of the results of a residual analysis provided?
20. Has the model is validated on external (or quasi-external) data?
21. Is the plausibility of the modelled predictions and/or coefficients discussed?

22. Are the results compared to the literature and/or other data?

Acknowledging and propagating uncertainty in the analysis

23. Has the method for handling parameter uncertainty been reported?
24. Is sufficient detail given for how parameter uncertainty was handled (e.g. if a variance-covariance matrix is used, is this available in some form?)
25. Is parameter uncertainty appropriately reflected in the DAM?
26. Has any structural (model) uncertainty been explored (in the DAM)?
27. Have the model's limitations been discussed (and explored if possible)?