Main Manuscript

# Introduction

Health technology appraisals (HTA), of the type pioneered by NICE in the UK, commonly involve constructing a form of mathematical model, known as decision-analytic models (DAMs), in order to help decide between mutually exclusive options. They do this by simulating the consequences of adopting each of the options with regard to both cost and health outcome, and often over a lifetime horizon of many decades. The outputs produced by such models have helped influence reimbursement decisions which can be worth tens or hundreds of millions of pounds; they can close or open very large healthcare markets to companies, and affect the quality and range of packages of treatment and care available to patients and physicians. Reimbursement decisions based in part on the outputs of DAMs can have complex and important spillover effects, setting precedent for how later forms of health technology are evaluated by a given healthcare provider, seeding or poisoning an international consensus among providers about whether or not a particular treatment should on offer.

The outputs of DAMs can therefore have very important medical, social and financial outcomes. Of course, DAM outputs depend on DAM inputs – i.e. the data used to produce model parameters - and DAM processes - the model logic used to convert the inputs into outputs. A principle at the heart of the production of DAMs in NICE-style HTA is that the data are uncertain, and this uncertainty should itself be modelled by conducting probabilistic sensitivity analysis (PSA). The model inputs cannot usually be represented as single points but as distributions of values drawn from statistical models. Variation in the model inputs, reflecting modeller uncertainty about the true values of the associated parameters, is propagated through the model logic, and each time leads to a slightly different answer. This leads to a distribution of estimates of the possible consequences of each option, forming a bivariate ‘cloud’ of estimated outputs which may, or may not, then affect decision uncertainty.

When producing DAMs there is often an interplay between input and logic: the number of discrete states used to model a disease or treatment path will determine the number of parameters needed by the model, and therefore the sources of data that are sought for such parameterisation. Conversely, the availability of data, and level of aggregation at which data are reported, can shape the logic of the model: for example, a modeller may know from extensive consultation with clinicians that the condition their DAM models is usually categorised as having seven distinct disease severity states; however, the only data the modeller has identified reports the consequences (health and financial) of each state reports it at a different level of aggregation - say four states, or ten states - and so the DAM’s logic becomes reshaped around this schema.

Another related constraint on the model logic is that a range of different parameters – related to health consequence, financial consequence, and transition probabilities – are required for each state, and these parameters are seldom available from the same source of data, or reported at the same level of aggregation. For example, treatment costs may be reported over three mutually exclusive categories but health related quality of life (HRQL) reported over five. In order to make use of both cost and HRQL data some form of mapping between states is needed: in the example described, the either the costs estimates need mapped from three to five states, or the HRQL estimates need to be mapped from five to three states.

The aim of this paper is to describe a process of mapping parameters between states that, in keeping with the principle of PSA, does not misrepresent the additional uncertainty involved in doing so. This process can best be described as *stochastic mapping*, and involves applying some additional PSA-style statistical simulation in order to appropriately characterise the additional uncertainty involved in the mapping process. This stochastic mapping could either be thought of as an additional ­pre-processing stage used to derive inputs for the DAM; or as an expansion of the DAM’s logic forward, in the direction of the inputs, by an additional stage. Both ways of thinking about the process are functionally equivalent, although the former way of conceiving the process makes it clearer that the outputs generated by the pre-processing step can be re-used in other DAMs.

Figure 1 Two equivalent ways of thinking about the process of stochastic mapping described in this paper. The top subfigure conceives of the stochastic mapping as a discrete pre-processing step for certain inputs to the DAM. The bottom subfigure conceives of the mapping as an additional component within the DAM logic.

This paper presents two examples of stochastic mapping: the first in which parameters, available at a seven state level of disaggregation, are mapped onto a three state schema; the second in which the seven states are mapped onto a five state schema. Both examples were developed when creating a DAM for an HTA used to simulated the consequences of adopting a range of different diagnostic strategies for helping to determine whether patients with atrial fibrillation (AF) should be prescribed one of a number of possible oral anticoagulants (OACs), which reduce the risk of AF-related stroke, but can have severe side effects, including a risk of intracranial heamorrhage (ICH) whose consequences can be as severe as stroke.

The aim of this paper is two-fold: firstly...

The approach was developed and applied in the context of a health technology assessment for the National Institute for Health Research (NIHR). The next section will briefly describe the HTA, the economic model developed, the relevant outcome values required by the model and the data available to populate it. The description is intended to provide a motivating example and context to the development of the approach; however, it should be noted that the approach has applications much broader than the model in the HTA.

## Background and motivating example

A health economic model was needed which estimated the long-term outcomes of patients with atrial fibrillation (AF), a progressive condition which elevates stroke risk, under a range of treatment regimens. The specific technology under evaluation was transthoracic echocardiography, an imaging technology which can help identify a subgroup of AF patients with a particularly high risk of stroke, so that they can be prescribed oral anticoagulants (OACs) to reduce stroke risk at an earlier stage. The model comprises a short-term diagnostic stage, in which an initial population of AF patients is divided into four diagnostic subgroups: false positives, false negatives, true positives and true negatives; the first two categories are forms of incorrect diagnoses, and the latter two are forms of correct diagnoses. Each diagnostic subgroup has a different long-term consequence to the healthcare provider (cost) and the patient (health), and so the long-term consequences for each of the four subgroups needed to be estimated in order for the downstream effects of the diagnostic technology to be evaluated.

An important risk that needed to be incorporated in the long-term section of the model was that of major haemorrhages which can occur as a consequence of taking OACs. A proportion of major haemorrhages are intracranial, and a proportion of these intracranial haemorrhages can cause permanent brain damage, which can be as disabling in their effects as the strokes which OACs are taken in order to prevent. Modelling the long-term consequences of taking OACs, and the avoidable risks which result from a false positive diagnosis, therefore needs the consequences of OAC-induced haemorrhage to be modelled appropriately.

## Description of model

The model allowed stroke risk to vary with age and gender, in line with the CHADS2 and CHADS2VASc clinical stroke risk algorithms [REF]. HRQoL was also allowed to vary with age and gender, using estimates from X&Y. [REF] Additionally, the model allowed patients to experience multiple clinical events at different ages, such as a stroke then later an intracranial haemorrhage.. Because of this, HRQoL multipliers rather than HRQoL decrements were applied when an event occurred, as they allow some adjustment of HRQoL event penalties to account for age, gender and clinical background. A multiplier will also never lead a a negative HRQoL estimate, which is likely to be inappropriate for most conditions.

## Estimates needed

Rather than use four structurally different models for each of the four subgroups, a single model structure was used, but with some parameters - relating to stroke risk, risk of major haemorrhage, and ongoing cost of treatment - set differently for each subgroup. The model distinguishes between different levels of stroke severity and major haemorrhage severity, each level having different short-term and long-term health and cost consequences.

Each discrete health state and event needs to have more than one type of output value associated with it. In the model these outcome values included health consequences (the HRQoL multiplier), and event costs, event risk multipliers such as changes in the risk of stroke as a result of using an OAC were incorporated. Both health effects and costs are divided into instantaneous effects, which are applied at the time of the event, and ongoing effects, which continue either until the next event and/or death. Additionally, there needs to be a range of estimates for each output value in order to appropriately represent uncertainty as part of probabilistic sensitivity analysis (PSA).

Estimates of the costs of strokes were identified for two discrete stroke severity levels: independent strokes, in which the patient retains a higher level of functioning, and dependent strokes, which are more severe. For this reason, the model used these two stroke categories. By contrast, two sources of estimates were identified for the HRQoL effects of stroke: estimates from a relatively old study, in which HRQoL was reported separately for dependent and independent level stroke; and estimates from a newer study, in which HRQoL was reported by modified Rankin Scale (mRS), a clinical scale with six discrete levels (mRS levels 0 to 5). By convention, mRS levels 0 to 2 correspond to an independent stroke, and levels 3 to 5 to a dependent stroke.

## Applications of the mapping approach

### First application of mapping approach

The first application of the approach described here was to allow the newer data for HRQoL impacts of stroke to be used within the model, despite it using a larger number of categories. The newer estimates were considered preferable because it was believed that the older estimates would not be representative of current health outcomes following stroke, which are likely to have improved due to improvements in stroke management. [REFs?] However, using the newer data, which was presented at a different level of disaggregation involves mapping from the more to the less disaggregated number of states. Doing this in a way which appropriately accounts for the additional uncertainty introduced from mapping is the primary purpose of the approach presented in this paper.

### Second application of mapping approach

The second application of the approach was to make use of data on the same patient population in order to provide consistent estimates of the HRQoL estimates of ICHs as well as strokes. This involved making an additional set of assumptions about the mapping between categories. In particular how mRS categories for assessing stroke impact link to Glasgow Outcome Scale (GOS) states for categorising traumatic head injury. Further details are provided below.

# Methods

## Overview of approach

The aim of the approach is to map from a larger number to a smaller number of states in a way that adequately represents the additional uncertainty involved in doing so. The larger number of states will be referred to as the unrestricted system, and the smaller number as the restricted system.

There are two main sources of uncertainty which the approach captures. The first source of uncertainty is uncertainty in the population mean associated with each state in the unrestricted system, which results from sampling error. The second source of uncertainty is uncertainty about the true proportion of the overall population who are members of each state within the unrestricted system, which impacts on how the population should be split within the restricted system, and the mean value associated with each state within the restricted system. Both of these sources of uncertainty will be discussed in more detail below.

## Types of information required

Uncertainty about the true population mean in each state within the unrestricted system is what is presented in summary statistics which report sample means and standard errors for each state within the unrestricted system. Because of the central limit theorem, this information can be used to parameterise Normal distributions from which distributions of expected values (EVs) can be drawn. Each of these EV distributions is then sampled repeatedly sampled from later in the process.

Due to the finite samples, there is uncertainty about the true proportion of a population who are within each state in the unrestricted system. To represent this uncertainty within a model, the sample sizes in each state need to be reported. These are then used to parameterise a Dirichlet distribution, which can be repeatedly sampled from in order to produce distributions of estimated proportions within each unrestricted system state that reflects sampling uncertainty.

## Intermediate stages

After the two forms of data described above are identified and used to parameterise appropriate statistical distirbutions, the next stage of the process is the sample and combine these distributions in order to produce simple quasi-individual level simulations, draws from the distributions, which are then grouped according to the mapping scheme linking the unrestricted to the restricted system. The aim of this intermediate stage is produce observations which the researcher can map onto the restricted system as if the individual level data were available. Although closed form solutions are in many cases available, the intuition of this approach is simple and can be used even in cases where closed form solutions are not possible or readily apparent.

The use of statistical simulation and the way draws from distributions are combined with draws from other distributions according to pre-specified rules is similar to the numerical simulation approaches often used in Bayesian models. Gelman & Hill (REF) refers to the use of 'fake data simulation' of the type described here as an 'informal Bayesian' approach. For that reason, estimates of uncertainty and variability within this paper are presented as credibility intervals (CrIs) rather than confidence intervals (CIs).

Within the approach, the outputs of some distributions are combined with the outputs from other distributions to form the inputs to other distributions. It is conceptually helpful to think of these distributions as 'upstream' or 'downstream' relative to other distributions, and easiest to see these dependencies graphically, in figures 1 and 2.

For example - references to nodes

Combining outputs from a number of distributions to form inputs to another distribution, is conceptually similar to sampling repeatedly from posterior distributions in formal Bayesian models. An arbitrarily large number of draws from the downstream distribution are determined in this approach. In this paper 10,000 draws were used; structural sensitivity analysis, presented in appendix XXXX, showed little difference in the credible interval estimates between using 10,000 draws and 100,000 draws, suggesting that 10,000 draws represents a fair compromise between speed and stability of estimates.

## First application of approach: stroke

As stated earlier, the first application of the approach is to map, with uncertainty, from the unrestricted system of mRS, to the restricted system of: dead; dependent stroke; and independent stroke. This section will first introduce the mRS categories, and how they are considered to relate to the dependent and independent health states; i.e. what assumptions are made about how the unrestricted system maps onto the restricted system. It will then look at the data source used to provide information about mean HRQoL in the unrestricted set and sample frequencies reported in each state. It will discuss the clinical trial which the reported summarise using the mRS system. Comparison will then briefly be made with another source of data which reports HRQoL values using the restricted category system used in the model, in order to demonstrate why in this case mapping from the unrestricted system is preferable.

The mRS is the 'unrestricted system' in both applications of the approach. It is a commonly used measure of disability or dependence in daily activities following a stroke, and was introduced in its current form by van Swieten at al in 1988, which adapted the original scale introduced in a paper by J Rankin, published in 1957 [REFS]. The mRS is a seven level discrete ordinal scale, with scores ranging from 0 to 6 inclusive, and more severe disability and dependence indicated by higher scores. The mRS has good inter-rater reliability, which along with its widespread use suggests it is an effective means of capturing variations in long-term health outcomes following a stroke. [REFs]

The unrestricted-to-restricted category mapping assumptions made in the first application are shown in Table 2. It is assumed that an independent stroke outcome corresponds to an mRS state 0, 1 or 2, and that a dependent stroke outcome corresponds to an mRS state 3, 4, and 5. Death is mRS state 6, and was assumed throughout to have a utility value of 0.

[Table 2  about here]

A graphical representation of use of the approach in the first application is shown in Figure 1.

[Figure 1 about here]

It will then look at the data source used to provide information about mean HRQoL and sample frequencies reported in each state. It will discuss the clinical trial which the reported summarise using the mRS system.

The Rivero-Arias paper used data from the Oxford Vascular Study (OXVASC). [REF] OXVASC is a large scale population-based cohort, initiated in 2002, involving almost 100,000 individuals registered in Oxfordshire, England.(8)  The source paper used 1,283 patients from this study, recruited between April 2002 and March 2007, who had suffered either stroke or transient ischemic attack (TIA). These patients were followed-up for up to 24 months. The condition of the patients was assessed using the disease specific measure of the mRS, as well as the EuroQoL 5 Dimension (EQ-5D) tool. Based on this, the EQ-5D utilities associated with each state were estimated. (2)

Rivero-Arias reported that, of the 1,283 patients who had a stroke within the Oxford vascular study (OXVASC) cohort, 24.8% (319 / 1,283) were dead within 24 months. [REF] Of those who survived, mRS scores following the stroke was graded according to the modified Rankin Scale (mRS) 24 months after the event in 425 patients.(2) In order to avoid having to also model a survival function using very limited data, the patients who died of stroke were assumed to have died within the first 24 months, and the 24 month state was assumed to be the patient’s long-term condition, with no additional variation in stroke-related HRQoL effect after this period.

[Table 1 about here]

The ordinary least squares (OLS) based mean estimates for the utility associated with each state, combined with the standard deviations around these mean estimates, were also reported in the source paper.[j12]   The numbers used from the source paper to estimate the distribution of patients in different mRS categories are presented in Table 1.

HRQoL estimates were not collected for all patients in the Rivero-Arias study, and so the HRQoL estimates were based only on those where the data were collected. These estimates were not adjusted when used in this approach, and so the assumption that these estimates are representative of those for whom the data were not collected was implicitly made; this is known as the missing completely at random (MCAR) assumption. (9)

Comparison will then briefly be made with another source of data which reports HRQoL values using the restricted category system used in the model, in order to demonstrate why in this case mapping from the unrestricted system is preferable.

## Second application of approach: ICH

Because of the simulation model and specific clinical problem which motivates the approach described here, it was considered useful to be able to use estimates reported in Rivero-Arias to also produce estimates of the HRQoL consequences of intracranial haemorrhages. This is because both ICHs and strokes involve disruptions in blood flow within the brain, with qualitatively similar potential long-term consequences in terms of disablement, HRQoL impairment, and treatment costs. By making use of the estimates from Rivero-Arias again, estimates for both the HRQoL impairments of strokes and ICHs are based on the same patient population. Treatment with OACs reduces the risk of the former but increases the risk of the latter, and so using data from the same population for estimates of the HRQoL consequences of both event types means differences between estimates of event types will not be an artefact of differences in the patient population. However, such estimates will in part depend on the mapping assumptions made between the mRS and GOS systems. The definition of the GOS states, along with the mRS states they were assumed to correspond to, are shown in table XX.

TABLE X - GOS and mRS states

The mapping assumptions made were that GOS 5 (‘good recovery’) corresponds to mRS states 0 or 1, that GOS 4 (‘moderately disabled’) corresponds to mRS states 2 or 3, and that GOS 3 (‘severely disabled’) corresponds to mRS states 4 or 5. GOS 2 is defined as a 'persistent vegetative state', and was assumed to have the same HRQoL as a GOS 1 ('dead'), which corresponds to mRS 6 ('dead'). From the perspective of the simulation model, GOS 2 is distinct from GOS 1, however, in that different, and higher, instantaneous and ongoing costs apply.

Because the estimates of HRQoL in each GOS state depends on the mapping assumptions between mRS and GOS which have been made, a structural sensitivity analysis was performed in which slightly different mapping assumptions were made. Further details and results of this are presented in appendix XXX, and briefly summarised in the results section.

### Graphical representation of process

A graphical model of the approach as used in the second application is shown in figure 2.

[Figure 2 about here]

Additional discussion with reference to specific nodes, and use of 'downstream' and 'upstream' terminology

# Results

## Results of first application

Table 4 below shows the mean simulated proportions in the dead, independent and dependent stroke state, together with 95% credible intervals (CrIs) as well as mean simulated utility multipliers associated with each of the states, also with 95% CrIs. The simulation suggests that approximately one quarter of patients die as a result of a stroke, around one fifth are left in a dependent state, and the remainder are left in an independent state. Being in a dependent state leads, on average, to slightly more than a halving of the patient’s quality of life, whereas being in an independent state leads to quality of life reducing by around one fifth compared with patients whose strokes had no lasting effect (mRS 0).

[Table 4 about here]

## Results of second application

The utility multipliers associated with different GOS states, based on the simulation approach described above, are shown in Table 5.

[Table 5 about here]

### Sensitivity Analysis - what if different mappings were assumed?

…

# Discussion

## Key Findings

This paper has shown how relatively simple statistical simulation based approaches can be used to map from an unrestricted to a restricted system. The simulation based approach ensures that the additional parameter uncertainty involved in mapping to the restricted system, compared with using data already presented in the restricted system level, is captured as additional variance in the distribution of estimates, and so the level of parameter and potentially decision uncertainty generated by the mapping process is not understated.

The motivating example for using the approach was a health economic model in which cost data were reported aggregated by the restricted state category, and the modeller was required to make a choice when populating these states with associated HRQoL estimates, between estimates already disaggregated at the restricted system level, but which are based on older data, and results based on newer data but which were reported at the unrestricted system level. The approach made it easier to make use of newer clinical data.

The flexibility and generalisability of the approach was demonstrated by showing that it could also be applied to use the same clinical data source to populate two series of health states: those related to different stroke outcomes, and those related to different ICH outcomes. This meant that recent and relevant clinical data from a single population could be used to simulate the effects both of strokes and ICHs.

The flexibility of the approach was further demonstrated through the production of a sensitivity analysis in which a different set of mapping assumptions about the relationship between mRS and GOS categories. This involved only making very minor changes to the code used in the primary analysis, showing how the effects of the mapping assumptions made can be tested relatively simply.

The clinical as opposed to methodological key finding comes from noting the differences between stroke HRQoL estimates for independent and dependent stroke that are based on the XXX, and those based on mapping from YYY. These suggest that the estimates based on the older data may no longer be representative of expected HRQoL outcomes following a dependent stroke, and so should not be used.

## Possible Mechanism for Findings

The approach described here increases the ranges of sources of data which can be used to parameterise attributes associated with discrete model states. The motivating case provided in the first application was to make use of a more recent source of clinical data, where results were reported at a different level of disaggregation than those used in the simulation model. This is just one of a wide range of possible uses for this approach.

The estimates produced by the approach are a reflection of both the structural and distributional assumptions made. These assumptions form the rules which are applied to generate the results, and so are by definition the mechanism for the findings. The need to explicitly state assumptions and use simulation methods to demonstrate their logical implications is a strength of the approach, as the assumptions and results can then be easily checked and scrutinised by interested parties.

Methodologically, the estimates produced result from a combination of the data used and the assumptions made, as with all quantitative models. Although assumptions had to be made regarding, for example, the choice of statistical model and unrestricted-to-restricted system mapping, these assumptions are made explicit in this approach, and the dependence of the results on these assumptions can be assessed through structural sensitivity analyses. This is in contrast to simpler unrestricted-to-restricted mapping processes, such as simply aggregating the frequencies of unrestricted categories which map onto a single restricted category, which do not account for the additional uncertainty involved in performing such a mapping.

## Comparisons with previously published research

The key clinical finding is the discrepancy between the HRQoL estimates in the dependent stroke category using the mapping approach and newer data, compared with those directly reported but using older data. Our estimated multipliers are very similar to those presented in Dornan et al.,(12) for independent strokes but somewhat higher than those reported in that paper for dependent strokes. This is largely due to the distribution of mRS states within the Independent Stroke and Dependent Stroke categories, which for both categories of stroke are weighted towards less severe mRS states. In the case of dependent strokes (mRS 3-5), for example, only around 4% were the worst category (mRS 5), which has an estimated EQ-5D multiplier near zero, while three-quarters were in the least worst category (mRS 3), which has an estimated EQ-5D multiplier over 0.5. Such a finding could be explained by assuming that patient outcomes following stroke have improved due to better short term treatment and intervention. This highlights the importance of making use of more recent data where possible, given that healthcare systems change and improve. To do otherwise may be to misrepresent the costs and clinical consequences of modern treatment regimens for particular conditions. [j40]

## Limitations

The approach described here could be described as 'informal Bayesian', in that it combines multiple statistical distributions and assumptions ('priors') in order to derive simulated draws of an unknown derived and downstream distribution ('posteriors'). The 'posterior' distributions are produced by sampling from the 'prior' distributions a large but pre-determined number of time. However, the approach is not fully Bayesian in that estimates of the parameters are not themselves updated using, for example, a Gibbs sampling process applied repeatedly following structural specification of the model. The comparatively quick-and-dirty approach used here has the advantage of simplicity, but a more formal treatment of the estimation of parameters may be preferable in some contexts.

The approach here includes a structural sensitivity analysis which shows the effect of making a different assumption about how mRS categories map to GOS categories. However a formal comparison and synthesis of results produced by different mapping assumptions has not been undertaken. Estimates resulting from different mapping assumptions could be combined using a meta-model framework, in which results are weighted by, for example, results from expert elicitation

No way of quantitatively assessing veracity of results produced as no IPD available. Without assessment using individual patient data, the approach cannot be externally validated in terms of capacity of the approach to recover the true attribute estimates associated with different health states. Closed form solutions exist and may be preferable...

## Recommendations for further research

* Formal testing with real or simulated IPD
* Formalisation of approach using, for example, hierarchical models using WINBUGS

## Conclusion

The approach described here is relatively easy to implement, intuitive in interpretation, and allows a decision analytic model to make use of a broader range of data to populate state attributes.