# Structure

## Introduction

### Para 1: Why we started (Question answered)

In constructing the economic models used to estimate the cost-effectiveness of a new healthcare technology within the NICE Single Technology Assessment process, it is common for modellers to need to make use of summary data to parameterise their model. In some cases these summary data relate to variables known to be associated with each other in specific ways. Of particular interest are examples where one has estimates of the utility associated with degrees of disease severity. In such cases, it appears erroneous to assume that the mean patient utility associated with a more severe disease state should be higher than that of a less severe disease state. Instead, it should be assumed that a *monotonic relationship* exists between these variables. Ideally, all inputs used within economic models should be represented by statistical distributions, which accurately reflect the modellers’ uncertainty about the true value of input parameters. If modellers use summary statistics to populate these inputs naively, however, by assuming independence between variables known to be linked in this way, then in some instances probabilistic sensitivity analyses will produce severity-associated utility estimates that cross over, suggesting that worse disease states have higher utilities or vice versa. Such estimates thus lack face validity, and should be considered erroneous due to their failure to incorporate knowledge about the monotonic association between the variables. We have observed a number of models submitted to peer review or used in part of the NICE Single Technology Appraisal process where, in the context of summary data, the methodologies used for handling of monotonicity has been erroneous. [References] The aim of this paper is to consider how monotonicity can be better handled in these circumstances, to offer practical guidance and indicate useful areas for further research.

### Para 2: What we did (“In this study we...”)

Our research and conclusions resulted from an iterative three stage process. In the first stage, a ‘long list’ of ten possible methodologies for handling monotonicity was produced. A fictitious dataset was developed, involving thirty patients reporting utility values associated with the disease at both a moderate and a severe state. The performances of these ten methods were assessed using this dataset, and from this a ‘short list’ of methods with greatest plausibility was developed, whose statistical properties were then considered further within stage two. A third stage of the process involved considering how dependent our conclusions are on the choice of fictitious data generated, and exploring in more detail the implications of relaxing one of the assumptions made in generating these data.

## Method: What did we do?

### Para 1

In stage one, a simulated data set was developed. A hypothetical disease was investigated where the patients could have either moderate or severe disease. A dummy data set was constructed simulating the responses from 30 patients who provided utility values for both states. Patients were assumed always to rate the utility in the moderate health state (U1) higher than that in the severe state (U2). The assumed data are shown in Figure 1, and it assumed these data are reported as summary statistics as shown in Table 1.



Figure A plot of the simulated data set

|  |  |  |
| --- | --- | --- |
|  | **U1** | **U2** |
| Mean | 0.60 | 0.55 |
| 95% confidence interval of mean | 0.555 to 0.644 | 0.506 to 0.594 |

Table The assumed available summary data

### Para 2

When modellers are generating multiple estimates for using within probabilistic sensitivity analyses (PSA) using these summary data, the key monotonicity condition that must hold is that an estimate of U2 should always be equal to or lower than a corresponding estimate for U1. More formally, if there are M runs within the PSA, and the subscript i defines predicted values from the ith run, then U1i ≥ U2i for all i ; where M is the total number of PSA samples. A violation of monotonicity would result in the distribution of predicted values U1-U2 produced from the PSA including some negative values.

### Para 3

The long list of ten approaches considered for handling the U2 < U1 monotonicity relationship are described in Table 2 below. The mean, standard deviation, maximum and minimum values of U1 – U2 were recorded for each, having performed 1000 Monte Carlo simulations. As this was a simulation study these results were compared against 1000 sets of 30 bootstrap samples (with replacement) from the paired data, which had a mean difference of 0.05, a maximum difference of 0.073 and a minimum difference of 0.036. It is acknowledged that these ‘true’ values would not be known where only summary data were provided.

|  |  |  |
| --- | --- | --- |
| **Method Number** | **Name** | **Method Description** |
| 1 | Independent Sampling | For each of the M PSA runs, take one draw from U1 and one draw from U2 independently (i.e. assume no covariance between U1 and U2) |
| 2 | Quantile Matching/  Number Seed Recycling | For each of the M PSA runs, use the same random number seed when drawing a sample from U2 and U1. (This is equivalent to selecting the same quantile from both distributions.) |
| 3 | Upward Replacement | For each of the M PSA runs:  Stage 1: draw a sample from U2  Stage 2: draw a sample from U1  Stage 3: Check if the value of U1 drawn is less than the corresponding value of U2 drawn. If it is, then replace the value of U1 with the U2 value. |
| 4 | Downward Replacement | For each of the M PSA runs:  Stage 1: draw a sample from U1  Stage 2: draw a sample from U2.  Stage 3: Check if the value of U2 drawn is greater than the corresponding value of U1 drawn. If it is, then replace the value of U2 with the U1 value. |
| 5 | Upward Resampling | For each of the M PSA runs:  Stage 1: draw first from U1.  Stage 2: draw from U2.  Stage 3: Check if the value of U1 is less than U2. If it is, then go back to Stage 2 (i.e. resample). If not, then stop. |
| 6 | Downward Resampling | For each of the M PSA runs:  Stage 1: draw first from U2.  Stage 2: draw from U1.  Stage 3: Check if the value of U2 is greater than U1. If it is, then go back to Stage 2 (i.e. resample). If not, then stop. |
| 7 | AIVM Covariance | Assume that the covariance between U1 and U2 is equal to the average of the individual variances of the means (AIVM) of U1 and U2.  If assuming this covariance implies that the correlation between U1 and U2 is greater than 1, then instead select the covariance between U1 and U2 which implies a correlation of 1 between U1 and U2. |
| 8 | Lower bounded Covariance Retrofitting | Select the minimum value of a covariance between U1 and U2 such that the two following conditions are met:  Condition 1: U1 – U2 > 0 for all PSA runs.  Condition 2: The covariance between U1 and U2 is greater than AIVM.  If this implies that the correlation between U1 and U2 is greater than 1, then instead use the covariance value associated with a correlation of 1. |
| 9 | Upper Bounded Retrofitting | Methodology 8 but where the second condition is that the covariance between U1 and U2 is less than AIVM. |
| 10 | Beta Distribution Difference Modelling | Use a derived distribution of U1, called U1(\*), rather than U1 itself.  U1(\*) is defined as equal to U2 + Δ, where Δ is drawn from a Beta distribution. The parameters of the Beta distribution are selected following discussion with a clinician, and such that the difference between U1(\*) and U1 are small. |

Table Summary of the long list of ten methodologies considered

### Para 4

For methodologies 8 and 9 it was assumed that a plausible confidence interval could be defined by setting a covariance that led to the minimum difference being marginally greater than 0. These were 0.0004 and 0.00084 respectively. For methodology 10, it was assumed that the maximum difference would be set following a discussion with a clinician. For illustrative purposes a Beta (1,19) was used which was assumed to provide a reasonable estimation of the difference. Methodology 2 is broadly equivalent to pairing the quantiles from the estimated distributions of U1 and U2 within PSA runs, matching the lowest estimate of U1 with the lowest estimate of U2, the second lowest estimate of U1 with the second lowest estimate of U2, and so on. For this reason, quantile-pairing was not considered as a separate strategy at this stage.

|  |  |  |
| --- | --- | --- |
| Criterion Number | Criterion Name | Criterion Description |
| 1 | Monotonicity | The condition that U2i < U1i for all i.  This criterion can be subdivided into  1a) strict monotonicity: it is analytically impossible for monotonicity to be violated with this method  1b) relaxed monotonicity: monotonicity may occasionally be violated, but the proportion of occasions where this occurs is judged to be small enough that this is not of concern. |
| 2 | Bias in means minimised | The expectation of the sampled values of U1 (or U2) used in the PSA should, on the average, be equal to the expectation of U1 (or U2) that was reported. |
| 3 | Bias in variances minimised | The variance of the sampled values of U1 (or U2) used in the PSA should, on the average, be similar to the variance of U1 (or U2) that was reported (or that were derived from what was reported where confidence intervals were reported). |
| 4 | Uses available information | Where both U1 and U2 each have means and variances (or confidence intervals) reported, then all of this information should be incorporated when generating PSA. |
| 5 | Additional information, including clinical experience, not necessary | The method should be able to do more with the existing summary data metrics presented, rather than requiring additional information. |

|  |  |  |
| --- | --- | --- |
|  | U1 | U2 |
| U1 | 0.01535 | 0.96716 |
| U2 | 0.01438 | 0.01440 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | U1 | U2 | U3 |
| U1 | 0.01718904 | 0.9778667 | 0.9567961 |
| U2 | 0.01588622 | 0.01535431 | 0.9671625 |
| U3 | 0.01505169 | 0.01437985 | 0.0143972 |

### Para 5

Within stage two, the long list of ten candidate methodologies was reduced to a short list of two approaches which, given the hypothetical dataset, satisfied the twin aims of ensuring monotonicity was not violated, but also that the modellers uncertainty about the true difference between the two quantities U2 and U1 is not mis-represented. The performance and statistical properties of the short list of distributions was assessed by extending the simulated dataset to include a third health state, mild disease, whose associated utility U0 can be assumed to relate monotonically U1 and U2 such that U2 < U1< U0.

### Para 6

Within stage three, the properties of the hypothetical datasets were reevaluated to identify whether the decision to reject one of the simpler approaches in the long-list resulted from an artefact of the hypothetical data generating process, and so whether in cases where the real data are unlike the hypothetical data, the simpler approach could be considered.

### Para 7

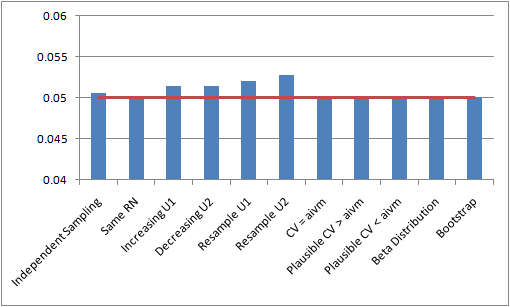
Based on this iterated process of simulation and evaluation, a number of contingent conclusions, practical recommendations and suggested avenues for further research were developed. These are presented in the conclusion and discussion section at the end of the article.

## Results: What did we find?

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 1 | 2 | 3 | 4 | 5 |
| 1 | No. | Yes | Yes | Yes | Yes |
| 2 | Usually. Depends on U1 and U2 | Yes | No/sometimes. Depends on U1 and U2. | Yes. | Yes |
| 3 | Always | No. Produces biased estimates. | ??? | Yes. | Yes |
| 4 | Always | No. Produces biased estimates. | ??? | Yes. | Yes |
| 5 | Always | No. Produces biased estimates. | ??? | Yes. | Yes |
| 6 | Always | No. Produces biased estimates. | ??? | Yes. | Yes |
| 7 | Sometimes. Depends on U1 and U2 | Yes. | ??? | Yes. | Yes |
| 8 | Yes, until covariance cannot be increased further due to implying correlations > 1 | Yes. | ???? | Yes. | Yes |
| 9 | As with 8 | Yes. | ???? | Yes. | Yes |
| 10 | Always. | Yes. | ???? | No/Maybe. | No. Maybe/Yes if approach developed. (See discussion) |
|  | | | | | |

### Para 1

Given the summary data available, it is known that, as the mean value of U1 is 0.60, and the mean value of U2 is 0.55, the mean difference between the two values should be 0.05. If a methodology produces a set of predicted values for this difference between U2 and U1 which has a mean value significantly different to 0.05, then this indicates the methodology produces biased estimates of the distributions, and so should be rejected as an appropriate method for this reason. The estimated differences produced by 1000 bootstraps of the ten methods is shown in Figure 2 below. It is seen that the methodologies which artificially adjust either U1 or U2 (methodologies 3 to 6) introduce bias by inflating the expected differences.

Figure The estimated mean differences between U1 and U2 according to the ten methodologies

### Para 2

A second condition that the methodologies should satisfy is that they should not misrepresent uncertainty about the variation of the differences between U1 and U2. This is a distinct issue to under-reporting the variance associated with either U1 or U2 independently as, in principle, the distributions U1 and U2 could each have equal variance. One could then simply define U2 as U1 plus the mean value of the difference between U2 and U2 (in this case 0.05). Although from this it might appear that variability has not been underreported, the strong assumption involved in defining U2 in this way, such that all values within U2 move in lock step with corresponding values of U1, means that the representation of uncertainty used within the PSA is unrealistic, reducing the validity of the estimates produced by the economic model.

### Para 3

Figure 3 below reports the mean estimated differences between U1 and U2 produced by each of the ten candidate methodologies. Where the minimum of these differences is below zero, it is demonstrated that monotonicity has been violated on some occasions as some estimates for U2 are greater than corresponding estimates for U1. Where the differences between the mean, minimum and maximum estimated differences are minimal or non-existent, then the appropriateness of the methodology should be doubted due to the lock-step issue mentioned above. It is seen that using the same random number or setting a CV value equal to AIVM does not allow the uncertainty to be captured in the Monte Carlo sampling. Methods 3-6, which were shown to produce a bias mean estimate of the difference, also provided a greatly inflated estimate of the uncertainty.

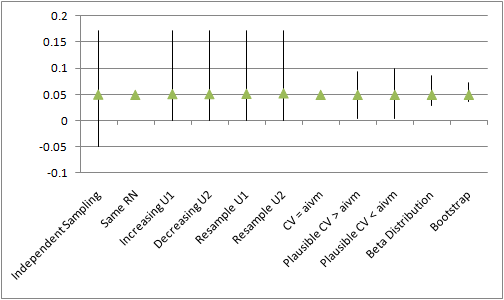


Figure The mean, maximum and minimum differences between U1 and U2

### Para 4

Based on the above analyses, all but three approaches were contingently rejected due to either violating monotonicity (method 1), producing biased estimates of mean differences (methods 3,4, 5, and 6), or misrepresenting the difference between U1 and U2 as zero (methods 2 and 7). The three approaches judged assessing further within stage two were thus methods 8, 9 and 10.

### Para 5

The differences in the non-rejected methods were further explored with the addition of a third health state (mild disease – U0) to the dummy data set. The value for U0 was greater than for U1. The summary data for U0 was a mean of 0.65 and a 95% confidence interval of 0.603 – 0.697. The three retained methods were adapted to the three-variable condition as follows:

* 8) Using independent Beta distributions to describe the difference between U0 and U1 and U1 and U2. Both Beta distributions were assumed to be (1,19).
* 9) Using a Beta distribution to describe the difference between U0 and U2, (assumed to be 16,144) and a second Beta distribution (assumed to be 70,70) to indicate where the value of U1 was estimated to lie within this range.
* 10) Selecting co-variance parameters (U0-U1; U0-U2; and U0-U3). Three different sets of parameters were chosen to provide a positive correlation between U0-U1 and U1 to U2, a negative correlation and a small correlation.

The values chosen to populate these models are illustrative only and assumed to be reasonable representations of the differences between U0, U1 and U2.

Particular attention was paid to the potential for correlations between the values in each methodology. The suitability of each method for incorporating clinical belief, such as that if the difference between U0 and U1 were larger than average, the difference between U1 and U2 was also likely to be larger (or would be smaller) than average, was assessed.

### Para 6

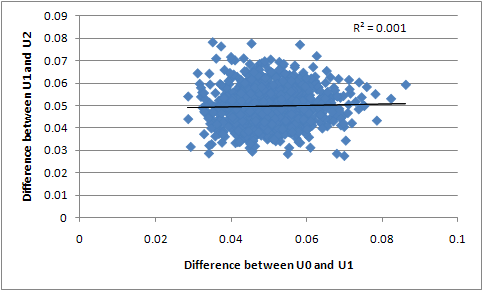


Figure Methodology 8

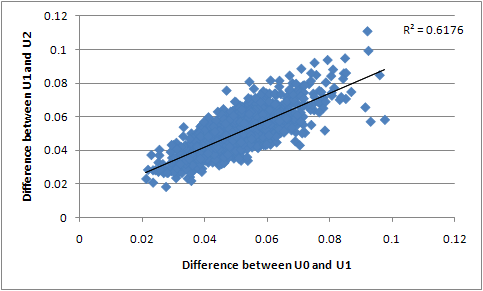


Figure Methodology 9

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |

Figure Methodology 10

### Para 7

After reviewing the results presented above, it was noted by one author that the decision to reject approach 3, use of the same number seed, may be contingent on the type of simulated data used, which may not be generalisable to other forms of summary data. This was because the simulated dataset made the assumption that the confidence intervals of the standard errors of the two distributions were approximately equal. In cases where the standard errors were not equal, it was considered that this finding may not hold. In order to test this concept more comprehensively, a follow-up simulation was created, whereby two normal distributions were generated a fixed difference (one unit) apart, and the ratio of the variances of the two distributions was varied systematically. For each of these variance ratios, the effect of quantile matching (effectively the same as using the same random number stream for both distributions) on two metrics was estimated, by comparing quantile-matched with independently sampling from the two distributions. These two metrics are: the proportion of draws where monotonicity was violated, shown in Figure 7, and the estimated variance of the difference between the two component distributions, shown in Figure 8.

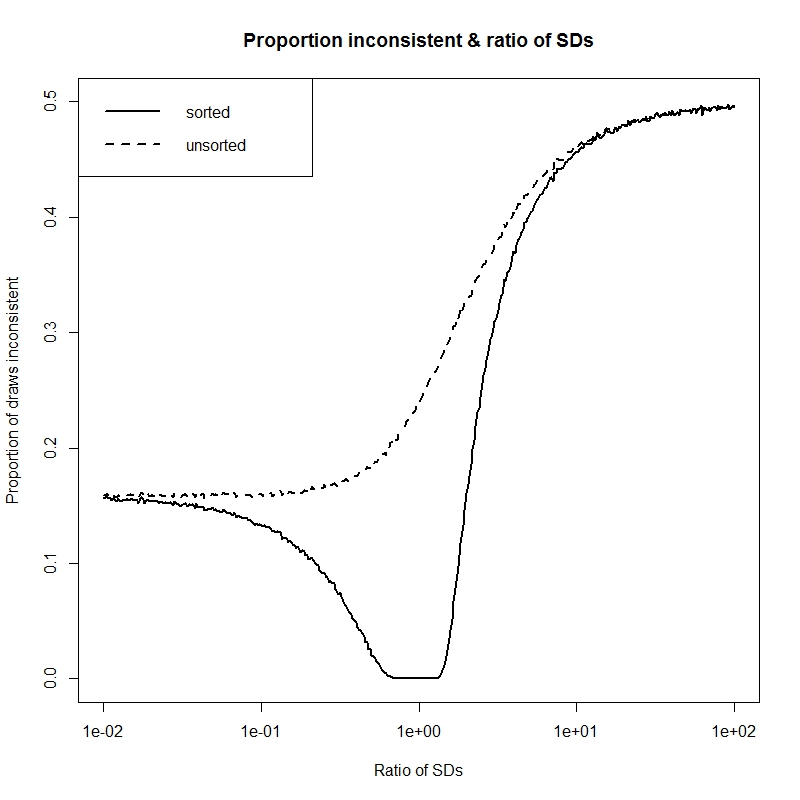


Figure Proportion of inconsistent draws (violating monotonicity) as a function of ratio of SDs of distributions

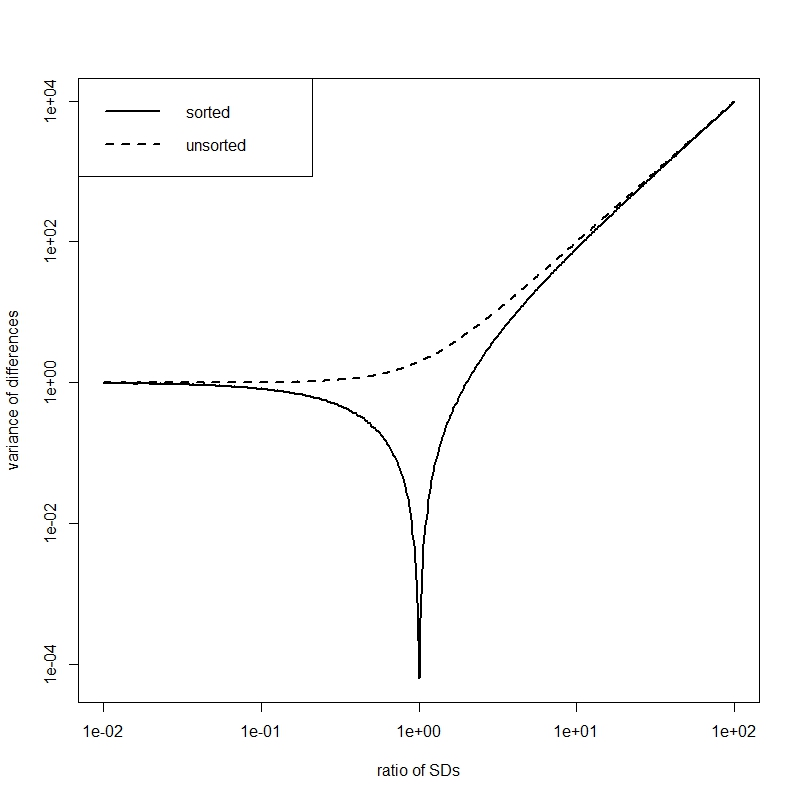


Figure Variance of differences between distributions as a function of ratio of SDs of distributions

It is seen that there is only zero predicted variance of the difference between U1 and U2 in cases where the variance of both U1 and U2 are assumed to be equal, as was the case for the hypothetical distribution used to assess the ten methodologies described above. In cases where the variances of the two component distributions are unequal, as summarised by the ratio of the variances, the ratio of the predicted differences between the two distributions increases rapidly. However, for all variance ratios, the range of predicted differences between the distributions is significantly less than for the base case where the two distributions are drawn independently and not paired with each other. A trade-off in conditions appears to exist, such that larger predicted variances of differences exist as the ratio of the variances moves away from unity, but as this happens so the proportion of paired draws that violate the monotonicity assumption increases as well. There appears to be a small but potentially usable ‘sweet spot’, a series of ranges of ratios of variance, where both non-zero predicted variance of difference exists, and where there is also no violation of monotonicity. Where summary statistics for two monotonically paired distributions suggest that the ratio of variances is likely to be within this acceptable range, then this simpler approach could be considered.

## Discussion:

### Para 1: Summary of what found

The analyses presented here confirm the rationale for performing the analyses: simply drawing estimates for parameters that are known to be monotonically paired from independent distributions is liable to produce paired estimates for these parameters that violate the monotonicity assumption, and so producing a series of scenarios within probabilistic sensitivity analyses which lack clinical face validity. Trying to correct these monotonicity violations by simply resampling from the distributions, or forcing a predicted value which violates the monotonicity assumption, can however produced biased estimates of the mean value of the parameters, and so should also not be done. Approaches involving fixing paired estimates, within a given PSA draw, of one parameter so as to be equal to a paired value of another draw plus or minus an increment analytically determined to be positive, appears to be the type of solution required. However, the correct way to do this remains to be determined. A very naïve approach would be simple add on the mean difference of the two distributions onto one distribution. However, doing this would lead to some of the information available to the modeller being disregarded, and as the point of acknowledging monotonicity is to incorporate one additional piece of information in our models this approach thus appears counterproductive, as well as misleading in terms of how it represents the relationship between parameters and our degree of knowledge or ignorance of this. Drawing parameters for both distributions jointly in a way that assumes nonzero covariance between variables necessarily leads to a reduction in the degree of monotonicity reported. However doing this requires imposing another assumption about the relationship between variables for which clear clinical or epidemiological evidence may be lacking. At the limit, this approach converges upon the lock-step approach described and dismissed previously.

[Found that most approaches considered had disadvantages.

Assuming independence produces violations of monotoncitiy

Selective resampling produces bias

Same random number seed produces underestimate of differences (but contingent results)

Appropriate approaches involve setting correlation either explicitly var-covar matrix) or implicitly (beta distributions). However these require greater clinical input and/or more technical expertise, and involve adding new assumptions which are difficult to verify.]

Those methods that were not rejected [at stage 1] involved judiciously selecting a co-variance parameter or expressing a belief in the difference in the utilities via a statistical distribution (in this case a Beta distribution). Both of these approaches would need the input of a clinical expert in order to produce a confidence interval around the difference that would be seen as plausible.

It is seen that a number of methods that have been used in economic evaluations are not appropriate when monotonicity must be upheld.

Those methodologies that provide a bias expectation in the difference will also provide a biased estimation of the incremental cost-effectiveness ratio and should thus be avoided.

Those methods which are known to under (or over) estimate the uncertainty will produce answers from value of information analyses that are incorrect, in addition to providing the decision maker with inappropriate certainty in the adoption decision.

Two broad methodologies were not rejected. The parameters for the co-variances and the parameters for the Beta distribution were selected through a trial and error process (prior to the results of the bootstrapped analysis being known). Ideally a Bayesian approach should be undertaken where a clinical expert would provide certain characteristics that a distribution would demonstrate, with the analyst attempting to replicate this.

Where only two variables were considered there was little difference between the recommended methods. However, where there were three variables and a prior belief that there would be correlation between these the co-variance method was observed to have an advantage in replicating clinical reality.

### Para 2: Shortcomings

[Implications of research dependent on assumed validity of simulation approach. Have already addressed this shortcoming through stage three.]

[Within research particular approach to parameterising distribution was informal, and so difficult to justify externally]

### Para 3: How relates to other findings

[Not sure… ask Matt/Nick…]

### Para 4: Implications for Research

It is critical, when the amount of information available to modellers about input distributions is already limited, not to discard useful and usable information where possible. In effect, when the modeller knows that two variables are monotonically linked in a particular way, but nonetheless chooses to draw the two distributions independently within the PSA runs, the modeller is doing just this, because they are not incorporating this information into the model. However, as the results above have shown attempting to incorporate this knowledge in an inappropriate way can itself be problematic. For example, simply discarding draws which violate the monotonicity assumption can produce biases in the estimated results, disregarding and contradicting the modellers’ knowledge about the true mean sample values of a parameter. Likewise, other approaches can indirectly disregard other knowledge about the range of estimated differences between two or more distributions, by forcing predicted values to move in lock-step. In effect this can result in predictions disregarding one form of information – that of the variance of each separate distribution – in order to incorporate another source of information (that the variables are monotonically associated).

[Good method should avoid throwing away relevant information about the relationship parameters. This includes both formal information such as reported SE and mean values, and also implicit information such as knowledge that values are correlated.]

[Example of this: throwing away SE estimates when estimating 2nd value]

The follow-up simulation indicated that a fairly simple approach, using the same random number seed; or equivalently taking equal numbers of independent samples from each distribution, then matching them by quantile, can produce estimates with satisfactory properties when particular conditions hold. However, in the equal or near equal variance situation they do not. It remains a matter of modeller judgement whether the drop in …

[Standardised approach to parameterising Beta distributions.

[Classicist maximum likelihood estimation]

Simulate two (e.g.) lognormal or normal distributions independently using summary data.

Use lower distribution directly

Use upper distribution using additive fashion.

Estimate a DELTA distribution in which an additional (e.g.) Beta distribution is parameterised according to a rule involving model fit to simulated data.

PSA run estimates for upper distribution is equal to corresponding PSA run estimates for lower distribution plus draw from DELTA distribution.

Issue is how to parameterise DELTA distribution intelligently.

Classical statistical method

Maximum likelihood estimation using e.g. Newton-Ralphson approach

Or more computationally intensive approach.

Or minimising maximum absolute error; squared error etc

Bayesian approaches

Non-normal distribution of errors on parameters

MCMC, Metropolis Hastings Algorithms etc

Other smaller potential issue is to ensure mechanism for simulating values from distribution does not bias results.

### Para 5: Implications for practice

[Depends how much of an issue one considers violation of monotonicity assumption to be.

Less of an issue when not looking at expected values.]

[Make sure to test whether monotonicity is a problem in practice if not in theory.

[Where SEs are small relative to differences between means of distributions, even a naïve approach assuming distributions are independent could produce no violation of monotonicity on average within 1000 PSA runs]

[Need to be sure that the monotonicity assumption is valid in the particular context considered.

For example, although perhaps one can assume increasing disease severity should definitely be associated with decreased patient utility, it might not make as much sense to assume costs are also monotonic. It may be that the most severe stages of a disease are cheaper to treat than less severe stages, as fewer effective treatment options exist.

[It could also be that many clinicians assume two formally defined disease states are actually very similar, and so it would not be contrary to their expectations if the utility value of a ‘more severe’ state were actually about the same or higher on average than that of the ‘less severe’ state.

[Also far less likely to always be true when considering predicted rather than expected values, as it’s very probable that on some occasions at least some patients with a more severe disease will have higher utility than at least some patients with a less severe condition.]

When presented with summary data and with a belief that monotonicity must apply, a judicious selection of the covariance parameters or of the distributions for the differences appears appropriate. The former strategy is likely to be preferential if there are more than two parameters and there is some belief of correlation between the variables.

### Para 6

# Notes

* Each paragraph should start with a clear message (a ‘topic sentence’)
* Try to do each sentence in one go (for consistency)

# Editing

## Micro-editing

### Is the information correct?

### Are requirements stated in ‘Instructions to authors’ met?

### Is the English clear and simple?

### Is the grammar and spelling correct?

## Macro-editing

### Is there a clear message?

#### Is there a clear message?

#### Is the message worth giving?

#### Is the message proven?

#### Where does the message appear

### Is the market appropriate?

### Is the structure appropriate?

#### Does it follow IMARD structure?

#### Are paragraphs clearly written?

### Is the tone appropriate?

## Yellow marker test

### Highlight most important sentences

#### Are the first sentences of paragraphs highlighted? (Otherwise meaning may be buried)