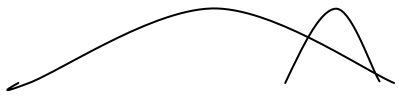
Comments from authors

Comments from Sophie

#	Page	Comment	Response
1	1	p. 1. Intro is very clearly written but need to expand a lot on the motivation. We aim to compare to see which is the best. The frequent lack of PD data is a limitation. IF PL data is available – this is recommended. If not PL available this often done – poss could be improved on.	I'd like to defer this until the rest of the paper has been finalised. **Action**: ***Revisit when the other sections are finalised.***
2	1	Any other examples where we might want monotonicity?	**As with comment # 1**
3	3	Often would be the case that estimates of U1 and U2 come from different patients. What would you recommend if this was the case & PL data was available?	I think there should be something on this in the discussion. The situation for which method 10 would be ideally suited would be one where there are, say, 1000 patients, but most of them (say 900, or even 950) are in the better state (U1). This means the SEs for U2 are wide compared with the SEs for U1. The SEs for U2 could be so wide that they overlap the U1 SEs, and the upper estimates of U2 are higher than the upper estimates of U1, as in the illustration below.  It would clearly be erroneous to 'reward' the worse health state with some estimates of the mean that are higher than the better health state, and so in this case method 10 could be used even where there are patient level data available. **Action**: ***Authors to discuss whether they agree with this argument, and if so produce a paragraph for the discussion section describing it.***
4	4	I'm not convinced of this [using normals]	Now addressed as Betas used instead. **Action**: ***No further action required***

5	4	Unclear to me. Do you mean U1 =U2 if samples not monotonic. And Not obvious to me why we need to consider both of these?	Yes. This is what I mean. However there are two ways to do this: one method replaces U1 with U2, the other replaces U2 with U1. I'm trying to describe this algorithmically. **Action**: ***is any action required?!***
6	5	Refs? [to "Methods three, four, five and six have also been observed in economic models, as they are relatively easy to implement"]	Matt's initial view was that providing refs might count as 'naming and shaming' and so should be avoided. However I'd be OK to include one or two refs if other authors want it enough **Action**: ***Matt to reconsider whether to provide refs to this point; other authors to say whether they consider having references in support of this to be vital, and if so to suggest some refs to me.***
7	7	I'd be tempted to include column in table 2 with this info. [about whether a method *cannot* violate monotonicity]	I've added a † symbol to indicate this in table 2. **Action*****: no further action needed.***
8	8	Suggests that a diff hypothetical sample may be of interest	Agreed. The data have been changed by Kate. The estimates produced by different methods are now more different to each other **Action*****: no further action needed.***
9	10	"Need fig 2 on here for comparison" (re fig 4, the scatterplots.	It's now included. **Action*****: no further action needed.***
10	12	I would present 1 hypothetical dataset as results. But I would rerun with some other datasets as this could strengthen conclusion.	I'm ambivalent about this. I did something along these lines for method 2 but it didn't change the bottom line, and was a fair amount of effort. I'd rather have this flagged as an avenue for further research/limitation. **Action*****: Authors to decide if further analyses are needed for this paper or whether this should be something mentioned in limitations/further research.***
11	12	Only looks wrong because hypothetical data isn't perfectly correlated, but it could be.	I think hypothetical data were outcomes are perfectly correlated wouldn't be a useful hypothetical data. **Action:** ***no further action needed?***
13	13	I would go with another hypothetical sample.	*See comments to 10.*
14	13	I think OK to just use two [U1 and U2, rather than 3: U0, U1, U2]	I think this is a valid limitation as the original poster used three states. Also I'm not sure whether the analytic solution is fully appropriate for 3+ states. (Because of covariances between U0 and U2.) **Action:** ***Authors to discuss and agree on whether this is something we should mention as a limitation. Kate to clarify***

			whether the analytic solution for the difference method would be appropriate for 3+ states as well as 2 states.
15	13	But need to say whether methods are suitable to > 2 [states, such as U0, U1, U2]	I think this relates to comment 14. **Action:** ***Authors to discuss this point alongside comment 14.***
16	13	Suppose -0.6 to 1, actually [regarding plausible range of HRQoL values]	I've added another paragraph in the limitations The final limitation is relatively simple to address. Because worse-than-death health states exist and are evaluated in some economic evaluations, it may be inappropriate to use estimates directly from a Beta distribution which is bounded within the range 0 to 1. This problem could be easily addressed by rescaling the output from the Beta distributions from the range 0 to 1 to the range -0.594 to 1, in the case of EQ-5D. [References needed.] **Action:** ***Authors to comment on this paragraph, and suggest edits of it and references.***
17	14	i.e. clinical opinion important [in response to paragraph: In cases where the means for U1 and U2 are very close together and the standard errors are large, it is important to ask how confident we are about the validity of the monotonicity assumption, and how implausible it would be that the mean HRQoL in U1 is actually lower than the mean HRQoL in U2. If we are confident that the monotonicity assumption is correct, then method ten should be able to ensure that this relationship is represented in the PSA.	I think this is an important point. The main argument I think the paper should make is that we have developed a method which ensures that monotonicity cannot be violated in PSA runs, but using this method thoughtlessly means making a strong assumption whether the modeller realises or not. This method should only be applied where there is a very strong clinical belief that the two variables really are monotonically related. In cases where clinicians aren't convinced that HRQoL in state A is really worse on average than in state B, then perhaps it shouldn't be used, and something like independent sampling may be more appropriate. **Action:** ***Authors to discuss whether they agree with this argument and then agree on how this should be worded in the paper.***
18	14	Method for implementation of method 10 in excel may increase usage & citations	Agree fully. **Action:** ***Sophie to produce easy-to-use and simple Excel workbook for using this***

			method.

Comments from Nick

#	Page	Comment	Response
1	1	Which journal are you going for? Should all the below go in the 'background' section?	MDM. Could we defer decisions about intro until after other sections have been finalised? **Action**: ***Revisit when the other sections are finalised.***
2	2	I think we need some references in here somewhere. Throughout the paper we say we've observed these methods but we've never said where.	**See Sophie comment # 6**
3	4	Is there any particular reason why you did this? Standard texts would probably say to use a beta, or a transformation with a lognormal or gamma. This seems fairly important to me, since we are saying how it should be done, and really it's unlikely you'd ever recommend to use a normal dist for utility data. I presume this wouldn't make much difference to the results, but we should say something about it I think.	**See Sophie comment # 4**
4	6	Bit condescending!	**Action**: ***NL to suggest alternative phrasing.***
5	6	Ok, but should mention that if utilities are less than 0 this is inappropriate as the upper value will be limited at less than 1.	This refers to text that is no longer in the manuscript but the point is still valid. The response to Sophie comment #14 discusses this issue briefly by adding another paragraph to the limitations section. **Action**: ***NL to review the paragraph produced in response to Sophie comment #14 and decide: 1) whether this, with some edits, addresses this point too; 2) whether something should be mentioned in the methods section too.***
6	7	Various comments	**No longer relevant as about approach for method 10 which is no longer used.** **Action**: ***None.***
7	13	Would it [three states] make 10 more problematic too?	**See Sophie comment #14.**
8	14	What about where utility decrement is used with a lognormal or gamma distribution (allowing utilities to range from – infinity to 1)? This and the beta are what are talked about in standard texts for utilities.	I think this is largely addressed by the changes, though something in implications for research/limitations about the lognormal and gamma distributions should be mentioned too. I've added another sentence for the

		Would it be possible to use all the methods with these different distributions?	implications for research section, but this section needs developing again. Action: ***NL/other authors to suggest new text addressing this point in the implications for practice and/or limitations section.***
9	14	Not totally sure how "new" this is. Think I used something pretty similar in a model about 7 years ago. The new bit might be the way of finding the beta parameter values, but the idea of using a "difference modelling" approach definitely isn't new.	I'd like to keep the 'new' term unless someone can point to an existing reference showing where it's been used before. At the very least doing this should inspire a peer reviewer to say "No it's not new" and suggest a reference to make the point! If the method is already in existence, why hasn't it been used before? Action: ***NL to provide reference(s) to where this/similar approach has been used previously.***
10	15	A general issue would be that in the absence of IPD, it may often be hard to say for certain that monotonicity should exist if CIs overlap. For example, a more progressed health state might be one in which patients no longer take toxic drugs, so utility may actually not be any lower. I know we allude to this a bit in the "implications for practice", but really there might be cause to not use methods that ensure monotonicity if the CIs overlap at all – often that might be more reasonable than seeing CIs that overlap and saying "no, we think there is monotonicity so we're going to use this approach". I think it might be best to recommend that people report whether there was monotonicity in their data, and maybe present a variance-covariance matrix.	I think this links in with some other points. Perhaps the response to Sophie's comment #3 relates to this. I'd like to have a final round commenting on the most recent changes to the methods/results, then address this afterwards collectively. Action: ***Defer until after methods/results section changes have been commented on by all authors.***
11	15	I think we might dismiss method 1 a bit too readily. The fact is, it will give us an unbiased estimate of the mean, which makes it better than methods 3, 4, 5 and 6 (which we should very strongly discourage – I think this should be a main message of the paper), and it clearly deals with uncertainty better than 2 and in this eg, 7 and 8. Which leaves only 1, 9 and 10 as options. As an idea, if we ran method 1 more times (say 10000 times) would it give more similar CIs to the IPD, with an	Action: ***Revisit this after comments back on changes to methods/results section.***

		unbiased mean? Could monotonicity just be dealt with with method 1 and more PSA simulations?	
13	15	Method 10 looks good, but I'd be a little concerned that it actually seems to slightly underestimate uncertainty compared to the truth. Is there any reason why this might be? Would a lower N stop this? I thought figures 6 and 7 were interesting – is there any explanation for these? See my comments on the numerical optimisation algorithm – I'm not sure if we can definitively say this is what should be done as there must be lots of different ways to apply this 'difference' approach. A different method for estimating N might give results that look better?	**Action**: ***Revisit this after comments back on changes to methods/results section.***