PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Do patients and health care providers have discordant preferences about which aspects of treatments matter most? Evidence from a systematic review of Discrete Choice Experiments
AUTHORS	Harrison, Mark; Milbers, Katherine; Hudson, Marie; Bansback, Nick

VERSION 1 - REVIEW

REVIEWER	Matthew Quaife London School of Hygiene and Tropical Medicine, UK
REVIEW RETURNED	03-Nov-2016

GENERAL COMMENTS	This manuscript presents a systematic review of studies comparing the preferences of health care providers and patients for treatments. The paper aims to describe the characteristics of studies assessing the concordance of stated preferences as captured through discrete choice experiments (DCEs). This is an interesting review which uses a systematic review methodology to addresses a critical issue in patient centred care. I particularly appreciated the careful discussion of variation in measurement around preference concordance, and the results are generally presented clearly and discussed well. I think it is an important study for publication. I have a few substantive comments. I would have appreciated more discussion of the “so what” and further discussion of the implications of preference divergence would be useful. If a physician is reading, what should they take away from this study? For example, the description of implications of not describing heterogeneity in preferences is strong (p17), however a similar analysis of the broader implications of inconsistent preferences would be useful. Feasibly, the patient focus on process would invoke greater consideration of side-effect profiles, for example, whilst a physician outcome focus could affect how information is conveyed by practitioners. What are the potential implications of discordance as it stands, and how could preferences be made more concordant? For the mainly non-DCE audience of this journal, I think that this would be useful. Secondly, the risk of bias from included studies is not explicitly considered. Although reported in the PRISMA statement, the criteria used for assessing potential bias is somewhat unclear. Clearly some thought has gone into the quality of included studies, for example piloting among relevant groups and the extent to which survey design differed between patients and physicians. However, it would be useful if this could be formally presented – perhaps in a bias assessment grid. Additionally, the quantitative synthesis could be

	broken down into subgroups of “higher” and “lower” quality (for example) to explore if concordance was stronger in better DCE designs. One would expect for concordance to be higher if choice tasks were the same, experiments piloted in both populations, etc. When discussing the different versions of the study were used “in patients and health care provider groups” p.16, more detail on how versions were different would be useful. Was it simply a different DCE design in different groups? If so, how do papers justify this? If there are any systematic differences between physician and patient DCE tasks across the literature it would be useful to detail here. What were the framing differences and in what direction could they have biased results? The findings of the concordance section of results is quite brief, and could be expanded. In addition, although included tables do a good job at describing the characteristics of included studies, it would be useful to see a single synthesising table of included studies – perhaps in supplementary material if no space in the main text. Where discussed in the text, studies are referenced comprehensively and well but it is a little confusing to try and identify studies and their characteristics through citations. The approach taken to synthesis concordance findings (figure 2) is a nice application of one previously used in the literature (and I appreciate the discussion of limitations), but I would have expected a more thorough interpretation of the results – what do the absolute/relative differences in figure 2 exactly mean? For the weighted average lines, the methods state “we simply took the weighted average of this score by attribute classification” – is this weighted across the whole sample or separately for each classification? Perhaps representing this information in equation form in the methods section may make it more intuitive. The issue of binary concordance (or not) is brought up in the abstract and the final paragraph of discussion, but is not discussed much in between. It is a useful point, but I wonder if the authors can suggest a way of operationalizing this approach? Or perhaps discuss what it really means for some (perhaps low-value) attributes to be discordant but one (highly valued) attribute to be concordant? Are there any suggestions from these data that indicate how we can make preferences more concordant? Or be aware of where discordancy might be particularly important? Finally, some page numbers in the PRISMA report do not correspond to where items appear in the text. For example, “Provide a general interpretation of the results in the context of other evidence, and implications for future research.” Points to pages 14-5, yet these pages are simply the results and do not discuss other evidence and implications (as are included in the discussion). Minor comments P5. L15 no reference given for preference diagnosis being thought to be as important as medical diagnosis, which I am not sure is a strong claim P6 l9-10 concordance between actual choices and hypothetical choices is still under investigation, and no reference is provided for

	this claim here. P9 table 1 “>10 attributes” and “not reported” not required P10 line 13/4 list how many of neither used literature review v regulatory documents. Did all “neither” studies do this? Figure 2: the text below the figure is not clear – the order of attributes is not consistent with reference numbers or the order in which they appear in the table so are not easy to follow. I wonder if citations might be better by the figure attributes themselves, though acknowledge this may be a referencing software limitation P 16 line 17 “concordance seems to differ according to the individuals involved in making the choice” – would remove the word “seems”. Referring to an example from the results may make this clearer The first paragraph of the discussion speaks in generalities and could be strengthened by integration of specific results, particularly the quantitative synthesis. Referencing of van Helvoort-Postulart 2009 does not justify the description of DCEs as the gold standard for understanding preferences (p6 l1)

REVIEWER	Axel Mühlbacher Hochschule Neubrandenburg, Germany
REVIEW RETURNED	09-Nov-2016

GENERAL COMMENTS	The paper focuses on an interesting topic and raises the interesting question of discordant preferences of patients and care providers. However, the present study shows some deficiencies and needs revisions prior to publication. General: Several grammar, spelling and spacing issues. The manuscript needs additional proof-reading. Abstract: “Concordance and discordance varied within studies according to the type of attribute being considered.” This sentence is ambiguous. Please clarify when there is concordance and disconcordance. Strengths and limitations of this study: The first bullet point is not a strength of the study but rather a strength of DCEs in general. I would recommend deleting this point. Background: On p. 5, l.51-55 the authors critizise that recent reviews and DCEs only focus on differences between patient and health care provider preferences at the aggregate level, without taking “the true problem of heterogeneous preferences” as well as improvement of “the decision quality at the individual level”(p.6,l.16) into account. However, the research question or the study objective of the current review is never clearly defined. What is the aim of the study? Are you trying to analyze differences on the individual level? Please clarify. Methods / Systematic search: p. 6, l.38 “The search was validated by checking that all references from two previous systematic reviews

	involving discrete choice experiments". In general, this might be a reasonable means of validity check. Nevertheless, reference 20 is not a review that focuses on differences between patients and other stakeholders. Harrison et al. reviewed DCEs that included a risk attribute. Hence, the assumption that the search can be validated by checking that their references are included must be questioned (or explained in more detail). I would recommend to validate the search by using reviews that either had the same research question or that included DCEs in health care in general, e.g., Clark et al. 2014, de Bekker-Grob et al. 2012 or others. Methods / Systematic search: p. 6, l.51 "when relevant DCEs have been published in health20,21" Can you please clarify why you use references when describing your inclusion criteria? Methods / Systematic search: p. 7, l.23 It seems as if references 13 and 25 cite the same paper. Methods / Data synthesis: The authors state that comparing DCEs is limited by differences in scales. This is correct. Accordingly, authors apply an approach to synthesize data and compare coefficients. However, the results of this approach are not shown in the manuscript. Please add as this is fundamental for understanding the results section. Results / Systematic review summary: This is a good and detailed description of the selected studies. In order to better understand the results and the limitations of the studies it would be very helpful to state when the selected studies were published. The inclusion criteria include a time span from 1995 – 2015. Results / Table 1: "Number of attributes" states 43 DCEs in 38 studies. This is not described in the main text. Can you please explain this difference? Results / Study sample and framing: p. 12, l.38-43 "The dominant framing of the question where the instruction was the same was to pick between the option, with no specific framing of who they were making the decision for reported." This sentence is rather difficult to follow. Please restructure and specify which implications this might have in terms of your research question. Results / Findings of concordance: In the "Data synthesis" (p. 8) section the authors describe their approach to synthesise data in order to compare different DCEs. They actually never present these results. However, the results section only includes qualitative statements such as "The greatest discordance between patients and health care providers were for mortality". Since individual studies are referenced it seems as if these statements based on the analysis of individual DCEs and not on the described synthesis and comparison of coefficients. The manuscript would clearly benefit from a table showing the results of the synthesis in order to support the statements in this section. Discussion: p. 16, 1st paragraph: This section is rather general and does not really discuss the presented results. Do you have any explanation why concordance or discordance of patient and health care professional preferences varies across attributes used or according to the individuals involved?

	Discussion: p. 16, 3rd paragraph: “There is a need to try to identify groups of patients and health care providers with similar preferences using latent class methods.” There clearly is a demand for latent class methods. However, one has to keep in mind that latent class methods in this research area are rather new. The review includes studies beginning in 1995. This is a potential limitation in this case and should be discussed. Discussion: p. 17, 2nd paragraph: Secondly, the way we synthesized coefficients between and across studies required assumptions that are known to be problematic.” You actually never present any coefficients, but rather rely on qualitative statements.

REVIEWER	David Hailey University of Wollongong Australia
REVIEW RETURNED	02-Jan-2017

GENERAL COMMENTS	This systematic review of studies that used discrete-choice experiments (DCEs) to compare patient and health care provider preferences for health care interventions provides a useful overview of the area. The literature search and data extraction are adequately described. The data synthesis appears to be a reasonable, pragmatic approach that is appropriate given the context of the review. Comments for consideration: 1. PAGE 2. Suggest splitting the sentence in Objectives. 2. The Results and Conclusions in the abstract cover concordance issues but there is very little on the methodology of the DCEs that were reviewed. Inclusion of some of the points made in the Discussion would strengthen the presentation. 3. P 3. The first point does not relate to a strength of the study, and limitations are not mentioned. 4. P 8, 23-35. Some of the numbers given in the text do not match those in Table 1. 5. Table 1 is detailed and lengthy, and some of the material is also presented in the text. It might be shortened, for example by removing the section on country, which could go in an appendix if needed. 6. P10, 48. Suggest include reference to Table 1 at the end of the sentence 7. P 11, 3-5 I wonder whether this sentence is necessary 8. ,17 Table 1 appears to indicate 12 of 24 studies. 9. It would be worth commenting on the high proportion of studies where survey development was not described 10. , 25-27. With this sentence in the text there seems no point in having the same information in Table 1. (p10, 3-9) 11. , 44-54. Suggest split this long sentence 12. P 12 ‘some studies’ referred to; was that a single study (ref 50)? 13. P 14, 37. What is meant here by ‘fairly consistent’? 14. P 14-15 Rather brief details are presented of findings on concordance; it would be of interest to have further information on this area of the review. 15. P 15, 21-53. The findings presented here seem to be incomplete. Eight studies are referred to in the text but only two of these are included in the nine studies cited in Figure 2. The findings

	are of interest but a full presentation is needed. 16. P 16 The Discussion includes points on some limitations of the reviewed studies and makes useful recommendations for future DCEs. As suggested above, reference to these might be made in the abstract.

VERSION 1 – AUTHOR RESPONSE

We thank the reviewers for their excellent suggestions regarding our manuscript. We have addressed each of the points raised and have provided a discussion of the many changes we made to our manuscript as a result. We are grateful for these suggestions and believe the changes have improved the manuscript.

Reviewers' Comments to Author:

Reviewer: 1
Reviewer Name: Matthew Quaife
Institution and Country: London School of Hygiene and Tropical Medicine, UK Please state any competing interests or state 'None declared': None declared

Please leave your comments for the authors below This manuscript presents a systematic review of studies comparing the preferences of health care providers and patients for treatments. The paper aims to describe the characteristics of studies assessing the concordance of stated preferences as captured through discrete choice experiments (DCEs). This is an interesting review which uses a systematic review methodology to addresses a critical issue in patient centred care. I particularly appreciated the careful discussion of variation in measurement around preference concordance, and the results are generally presented clearly and discussed well. I think it is an important study for publication.

I have a few substantive comments. I would have appreciated more discussion of the "so what" and further discussion of the implications of preference divergence would be useful. If a physician is reading, what should they take away from this study?

1.1. For example, the description of implications of not describing heterogeneity in preferences is strong (p17), however a similar analysis of the broader implications of inconsistent preferences would be useful. Feasibly, the patient focus on process would invoke greater consideration of side-effect profiles, for example, whilst a physician outcome focus could affect how information is conveyed by practitioners. What are the potential implications of discordance as it stands, and how could preferences be made more concordant? For the mainly non-DCE audience of this journal, I think that this would be useful.

RESPONSE: We appreciate the reviewers request for additional discussion of implications and agree that this will add extra interest for the non-DCE audience. We have added the following to the discussion:
"The implications of our findings are that the health care that people want often is not the same as the health care that health care providers think that people want. This lack of concordance suggests that decisions which involve significant trade-offs (preference-sensitive care), there is a role for eliciting people's preferences and values about health care through tools like decision aids to enable health care providers to offer the most appropriate options69 or to match health care providers and patients with similar preferences and values."

1.2. Secondly, the risk of bias from included studies is not explicitly considered. Although reported in

the PRISMA statement, the criteria used for assessing potential bias is somewhat unclear. Clearly some thought has gone into the quality of included studies, for example piloting among relevant groups and the extent to which survey design differed between patients and physicians. However, it would be useful if this could be formally presented – perhaps in a bias assessment grid. Additionally, the quantitative synthesis could be broken down into subgroups of "higher" and "lower" quality (for example) to explore if concordance was stronger in better DCE designs. One would expect for concordance to be higher if choice tasks were the same, experiments piloted in both populations, etc.

RESPONSE: We generated a crude quality score by assigning binary indicators of whether a study had reported generating attributes for the DCE in all groups (1= yes, 0=no), piloted the DCE (1= yes, 0=no), piloted the DCE in all groups (1= yes, 0=no), and gave the respondents the same choice (1= yes, 0=no). This gave us a crude, unweighted, indicator of quality on a scale from 1 to 5. Using these categories we were able to classify 6 (16%) studies as low quality, 13 (34%) as medium quality and 19 (50%) as high quality.

Comparing the concordance of patient and health care provider preferences for different types of attributes showed that differences between concordance of preferences outcome measures seemed to be affected by the 'quality' of studies, but did not show any conclusive evidence for structure or process attributes. Given the low number of 'low quality' studies and the arbitrary nature of the quality scoring we are reluctant to attempt to draw any firm conclusions about the relationship between quality of studies and their finding of concordance of patient and health care provider preferences.

[Figure uploaded for review purposes]

1.3. When discussing the different versions of the study were used "in patients and health care provider groups" p.16, more detail on how versions were different would be useful. Was it simply a different DCE design in different groups? If so, how do papers justify this? If there are any systematic differences between physician and patient DCE tasks across the literature it would be useful to detail here. What were the framing differences and in what direction could they have biased results?

RESPONSE: We have attempted to provide more detail in section 3.4 in response to this comment and the point raised by reviewer 2. We have also added some additional comments in the discussion to reflect the point you are making:

"Where versions differed, this was primarily in the perspective respondents were asked to take when indicating their preferences: some were asked to choose from their own perspective, while in others the perspective of patients and perspective of health care providers was different within the same study. For example, patients might be asked to consider their own preferences, while health care providers were asked to try to predict the preferences of their patient. Even in studies that provided the same instructions to both groups, often it was unclear whether the health care provider should be considering their own preferences, the preferences of a patient, or some other preference. Consequently it is unclear whether the results should be expected to be concordant or discordant, and whether the implications of discordant preferences are important. Only a small number of studies actually provided DCEs with different attributes to different groups of respondents."

1.4. The findings of the concordance section of results is quite brief, and could be expanded. In addition, although included tables do a good job at describing the characteristics of included studies, it would be useful to see a single synthesizing table of included studies – perhaps in supplementary material if no space in the main text. Where discussed in the text, studies are referenced comprehensively and well but it is a little confusing to try and identify studies and their characteristics through citations.

RESPONSE: We have added a single synthesizing table of the key characteristics of the studies we have included in the review in the supplementary material.

1.5. The approach taken to synthesis concordance findings (figure 2) is a nice application of one previously used in the literature (and I appreciate the discussion of limitations), but I would have expected a more thorough interpretation of the results – what do the absolute/relative differences in figure 2 exactly mean? For the weighted average lines, the methods state "we simply took the weighted average of this score by attribute classification" – is this weighted across the whole sample or separately for each classification? Perhaps representing this information in equation form in the methods section may make it more intuitive.

RESPONSE: We have added additional information to describe how the 'concordance score' we calculate is derived and how it is interpreted. In doing this we realized we had made a small error in the calculation of the concordance scores (we were previously dividing by the total number of attributes) and we have updated figure 1 accordingly). We thank the reviewer for this comment which we believe has led to greater clarity in the method and enabled us to discover this minor error. Section now reads:

We attempted to synthesize coefficients derived from each study to observe patterns in attribute types where there was more or less concordance between patients and health care providers by developing a concordance score. Comparing coefficients from DCEs is challenging and limited by differences in the variance scale where separate DCEs are used in patients and health care providers within each study, and different DCEs between studies.28 We follow an approach previously used29,30 where we crudely estimate the relative importance of each attribute (based on the classification described above) by dividing the range of coefficients for each attribute by the sum of all coefficient ranges within a DCE, to provide the rank of importance of the attribute within that study. We then compared the difference in the rank of importance for an attribute between patient and health care providers. Since different studies have different numbers of attributes, we then divided the differences in the rank of importance of an attribute by the number of other attributes within the DCE to provide a concordance score on a common scale (where 0 = perfect concordance of rank importance, -1 indicates that the patients rank the attribute that the The implications of our findings are that the health care that people want is often not the same as what health care professionals think people want. This lack of concordance suggests that for decisions which involve significant trade-offs (preference-sensitive care), there is a role for eliciting people's preferences and values about their health care options, potentially through tools like decision aids, so health care professionals can offer the most appropriate options69 or to match health care providers and patients with similar preferences and values. believes is most important, as the least important, +1 indicates that a health care provider ranks the attribute that the patient believes is most important, as the least important). Finally, we simply took the weighted average of this score across all studies by attribute classification and present these in a figure.

1.6. The issue of binary concordance (or not) is brought up in the abstract and the final paragraph of discussion, but is not discussed much in between. It is a useful point, but I wonder if the authors can suggest a way of operationalizing this approach? Or perhaps discuss what it really means for some (perhaps low-value) attributes to be discordant but one (highly valued) attribute to be concordant? Are there any suggestions from these data that indicate how we can make preferences more concordant? Or be aware of where discordancy might be particularly important?

RESPONSE: The general point from this review is that without information on heterogeneity, it is difficult to make any specific suggestions as all the data is in aggregate form. The aggregate data can hide important heterogeneity, but does give some overall description of the problem. One approach to operationalizing heterogeneity is through the use of latent class models of combined patient and

health professional DCE data, and finding the proportion of patients and health professionals that fit in each latent class.

1.7. Finally, some page numbers in the PRISMA report do not correspond to where items appear in the text. For example, “Provide a general interpretation of the results in the context of other evidence, and implications for future research.” Points to pages 14-5, yet these pages are simply the results and do not discuss other evidence and implications (as are included in the discussion).

RESPONSE: Thank you for highlighting, we have updated the PRISMA statement to reflect this comment and the other suggestions of the reviewers that have added to our manuscript.

Minor comments

P5. L15 no reference given for preference diagnosis being thought to be as important as medical diagnosis, which I am not sure is a strong claim

RESPONSE: We have now cited the Mulley BMJ paper which describes this concept and makes the claim. In the text we do not make any statement about the strength of this claim, however it seems plausible given evidence that we discuss in the second paragraph.

P6 l9-10 concordance between actual choices and hypothetical choices is still under investigation, and no reference is provided for this claim here.

RESPONSE: We have added citations and acknowledged the ongoing exploration of this area.

P9 table 1 “>10 attributes” and “not reported” not required

RESPONSE: Deleted as suggested

P10 line 13/4 list how many of neither used literature review v regulatory documents. Did all “neither” studies do this?

RESPONSE: Added extra detail:
Those that reported generating attributes using neither respondent groups most often used literature reviews alone47,49, or literature reviews in conjunction with expert opinion38,43,66, information from regulatory requirements9, or product labelling45 to inform attributes.

Figure 2: the text below the figure is not clear – the order of attributes is not consistent with reference numbers or the order in which they appear in the table so are not easy to follow. I wonder if citations might be better by the figure attributes themselves, though acknowledge this may be a referencing software limitation

RESPONSE: Thank you for highlighting the lack of consistency in these references, we have now corrected these reference numbers, and the examples we describe in text, to ensure that they are consistent with the numbers cited in the main section of the manuscript.

P 16 line 17 “concordance seems to differ according to the individuals involved in making the choice” – would remove the word “seems”. Referring to an example from the results may make this clearer

RESPONSE: We actually think this sentence is obsolete given the previous and subsequent sentences so have removed the sentence to aid clarity.

The first paragraph of the discussion speaks in generalities and could be strengthened by integration of specific results, particularly the quantitative synthesis.

Referencing of van Helvoort-Postulart 2009 does not justify the description of DCEs as the gold standard for understanding preferences (p6 l1)

RESPONSE: We have changed the sentence to reflect the current state of thinking around DCEs, which also relates to the previous point around the validity of DCEs. Now reads:

"Discrete choice experiments (DCEs) have become an established tool in economic evaluation and decision making14 and for understanding preferences and predicting choices14,15 due to their ability to break down and value different components of treatments and services (whether these are processes, structures or outcomes16–18) as well as identify the trade-offs people make between these different components19,20"

Reviewer: 2
Reviewer Name: Axel Mühlbacher
Institution and Country: Hochschule Neubrandenburg, Germany Please state any competing interests or state ‘None declared’: None declared

Please leave your comments for the authors below The paper focuses on an interesting topic and raises the interesting question of discordant preferences of patients and care providers. However, the present study shows some deficiencies and needs revisions prior to publication.

2.1. General: Several grammar, spelling and spacing issues. The manuscript needs additional proof-reading.

RESPONSE: We have proof-read the manuscript as suggested and have fixed the problems we have identified.

2.2. Abstract: “Concordance and discordance varied within studies according to the type of attribute being considered.” This sentence is ambiguous. Please clarify when there is concordance and disconcordance.

RESPONSE: We agree this is important information to include in the abstract. We have rewritten the abstract to include as much detail as possible, whilst remaining within the 300 word limit, in the results and conclusions to describe the patterns of concordance we saw by attribute classification.

2.3. Strengths and limitations of this study: The first bullet point is not a strength of the study but rather a strength of DCEs in general. I would recommend deleting this point.

RESPONSE: Deleted as suggested.

2.4. Background: On p. 5, l.51-55 the authors critizise that recent reviews and DCEs only focus on differences between patient and health care provider preferences at the aggregate level, without taking “the true problem of heterogeneous preferences” as well as improvement of “the decision quality at the individual level”(p.6,l.16) into account. However, the research question or the study objective of the current review is never clearly defined. What is the aim of the study? Are you trying to analyze differences on the individual level? Please clarify.

RESPONSE: We have updated the aim of the paper as follows:
"The aim of this paper is to review studies which elicit both patient and health care provider preferences for health care interventions using DCEs, specifically to (1) review the methodology of DCEs to evaluate similarities, differences and rigour of their designs, and specifically whether comparisons are made at the aggregate level or account for individual heterogeneity and (2) quantify the extent to which they demonstrate concordance of patient and health care provider preferences."

2.5. Methods / Systematic search: p. 6, l.38 “The search was validated by checking that all references from two previous systematic reviews involving discrete choice experiments”. In general, this might be a reasonable means of validity check. Nevertheless, reference 20 is not a review that focuses on differences between patients and other stakeholders. Harrison et al. reviewed DCEs that included a risk attribute. Hence, the assumption that the search can be validated by checking that their references are included must be questioned (or explained in more detail). I would recommend to validate the search by using reviews that either had the same research question or that included DCEs in health care in general, e.g., Clark et al. 2014, de Bekker-Grob et al. 2012 or others.

RESPONSE: On the reviewers suggestion we have validated our search strategy based on the review by de Bekker-Grob 2012 and found that we included all studies identified by this review and the search strategy employed in this review. We have updated the references to reflect this.

2.6. Methods / Systematic search: p. 6, l.51 “when relevant DCEs have been published in health20,21” Can you please clarify why you use references when describing your inclusion criteria?

RESPONSE: These references are from studies that have shown that few DCEs, if any exist before this date, or have used similar inclusion criteria. These references support the decision not to search before 1995.

2.7. Methods / Systematic search: p. 7, l.23 It seems as if references 13 and 25 cite the same paper.

RESPONSE: These are different papers but present broadly similar information so we agree that one is sufficient to support our sentence. We cite the most recent of the two references now.

2.8. Methods / Data synthesis: The authors state that comparing DCEs is limited by differences in scales. This is correct. Accordingly, authors apply an approach to synthesize data and compare coefficients. However, the results of this approach are not shown in the manuscript. Please add as this is fundamental for understanding the results section.

RESPONSE: The results of this approach are shown in figure 2 and discussed in text in the manuscript in section ‘3.7 Findings of concordance.’ This approach is based on comparison of coefficients, using methods outlined in section ‘2.3 Data Synthesis’, but we actually present a score of concordance. We think the use of the words coefficients may have contributed to some misunderstanding of what we were presenting, we have added additional detail to the methods and changed the use of the word ‘coefficient’ to ‘score of concordance’ when discussing these results.

2.9. Results / Systematic review summary: This is a good and detailed description of the selected studies. In order to better understand the results and the limitations of the studies it would be very helpful to state when the selected studies were published. The inclusion criteria include a time span from 1995 – 2015.

RESPONSE: We have added this detail to paragraph 1 of the results section under Systematic Review Summary:
“The 38 papers we included were published between 2004 and 2015, and the majority (71%) were

published in the period between January 2010 and July 2015."

2.10. Results / Table 1: "Number of attributes" states 43 DCEs in 38 studies. This is not described in the main text. Can you please explain this difference?

RESPONSE: In our review we found five studies which used separate DCEs with different attributes for patients and health care providers. These were counted separately in the table and described in a footnote: Five studies9,35,39,54,55 included separate DCEs for the HCP and non-HCP populations; the numbers of attributes for each DCE were entered independently.

We have now added text to the main body of the paper as well to ensure that this does not cause confusion:
"Five studies included two different DCEs and attributes are included from both versions9,35,39,54,55."

2.11. Results / Study sample and framing: p. 12, l.38-43 "The dominant framing of the question where the instruction was the same was to pick between the option, with no specific framing of who they were making the decision for reported." This sentence is rather difficult to follow. Please restructure and specify which implications this might have in terms of your research question.

RESPONSE: We have split this sentence up and reworded now reads:
In studies giving the same instructions to both groups, the question asked respondent to pick between the alternative options provided, but did not provide any specific framing about of who the respondent should assume they were making the decision for35–38,41–43,45,49,50,53,54,56,57,64,66. One study did, however indicate that the health care providers were asked to choose the option with the biggest global benefit, for themselves52.

2.12. Results / Findings of concordance: In the "Data synthesis" (p. 8) section the authors describe their approach to synthesise data in order to compare different DCEs. They actually never present these results. However, the results section only includes qualitative statements such as "The greatest discordance between patients and health care providers were for mortality". Since individual studies are referenced it seems as if these statements based on the analysis of individual DCEs and not on the described synthesis and comparison of coefficients. The manuscript would clearly benefit from a table showing the results of the synthesis in order to support the statements in this section.

RESPONSE: The results and comparisons are shown in the figure and the text refers to the strength of concordance displayed in the figure. Our data synthesis shows patterns of concordance, and while we have now better described the calculation of the coefficient, do not want to focus on the number itself but the pattern that emerges. We have rewritten the description of these results in text to provide extra detail, as well as adding additional detail about the calculation, interpretation and presentation of results throughout the manuscript. We have also summarized the findings in the abstract, which relates to one of the comments below.

2.13. Discussion: p. 16, 1st paragraph: This section is rather general and does not really discuss the presented results. Do you have any explanation why concordance or discordance of patient and health care professional preferences varies across attributes used or according to the individuals involved?

RESPONSE: We have signposted which parts of the review our summary of results relates to and have added extra detail as suggested. We have also provided additional detail in response to this comment, and the comment of reviewer 1 in the following paragraph to add additional discussion of this point. One key limitation that we have provided in interpreting these differences is that results are

only provided at the aggregate level, so it is difficult to understand more about the reason for discordance and concordance. We have some ideas that might explain the discordance, but felt we wanted to leave this paper as descriptive.

2.14. Discussion: p. 16, 3rd paragraph: “There is a need to try to identify groups of patients and health care providers with similar preferences using latent class methods.” There clearly is a demand for latent class methods. However, one has to keep in mind that latent class methods in this research area are rather new. The review includes studies beginning in 1995. This is a potential limitation in this case and should be discussed.

RESPONSE: We have added to the limitations section to reflect this comment:
“As latent class analysis is a relatively new method in the analysis of DCEs, the period covered by our review may predate any increase in published studies applying these methods to understand heterogeneity of preferences within respondent groups. However, there is a need to try to identify groups of patients and health care providers with similar preferences in future DCEs, and opportunities to reanalyze data collected in previously published DCEs to understand preference heterogeneity using these methods.”

2.15. Discussion: p. 17, 2nd paragraph: Secondly, the way we synthesized coefficients between and across studies required assumptions that are known to be problematic.” You actually never present any coefficients, but rather rely on qualitative statements.

RESPONSE: These results are presented in Figure 2, and we have now added in the text in response to the reviewer’s previous comment. The statements in quotation marks are examples of the types of attributes classified under the Donabedian framework. If this has given a misleading impression that these are qualitative statements, we apologize.

Reviewer: 3
Reviewer Name: David Hailey
Institution and Country: University of Wollongong, Australia Please state any competing interests or state ‘None declared’: None declared

Please leave your comments for the authors below This systematic review of studies that used discrete-choice experiments (DCEs) to compare patient and health care provider preferences for health care interventions provides a useful overview of the area. The literature search and data extraction are adequately described. The data synthesis appears to be a reasonable, pragmatic approach that is appropriate given the context of the review.

Comments for consideration:

3.1. PAGE 2. Suggest splitting the sentence in Objectives.

RESPONSE: Split into two sentences as suggested.

3. 2. The Results and Conclusions in the abstract cover concordance issues but there is very little on the methodology of the DCEs that were reviewed. Inclusion of some of the points made in the Discussion would strengthen the presentation.

3. 3. P 3. The first point does not relate to a strength of the study, and limitations are not mentioned.

RESPONSE: Have deleted point 1 as suggested also by reviewer 2, but incorporated some into point

2 and emphasized that this is both a strength and limitation now. We have also added an acknowledgement of the assumptions we have made in synthesizing results as a limitation.

3.4. P 8, 23-35. Some of the numbers given in the text do not match those in Table 1.

RESPONSE: Thanks for highlighting, have corrected these.

3.5. Table 1 is detailed and lengthy, and some of the material is also presented in the text. It might be shortened, for example by removing the section on country, which could go in an appendix if needed.

RESPONSE: Removed as suggested, as well as some other details in this table mentioned by other reviewers

3.6. P10, 48. Suggest include reference to Table 1 at the end of the sentence

RESPONSE: Added as suggested

3.7. P 11, 3-5 I wonder whether this sentence is necessary

RESPONSE: Deleted as suggested

3.8. ,17 Table 1 appears to indicate 12 of 24 studies.

RESPONSE: Table 1 is correct, text corrected. Thanks for highlighting.

3.9. It would be worth commenting on the high proportion of studies where survey development was not described

RESPONSE: This is not a high number in terms of attribute generation, although we agree that this was not presented clearly, but it was for piloting, which is more clear. We have sought to improve clarity of this section:

"The groups that were used to generate attributes and pilot surveys varied. 13 (34%) of the studies that reported their attribute generation sought input from people representative of all groups who would be asked to complete the DCE31–35,53,54,56–61, and 13 (54%) of the studies that reported the piloting in their study piloted the survey in all respondent groups31,40,42,45,47,50–52,57,62,60,61,63. There were only five studies that reported having generated their attributes and piloted their survey in all groups of respondents31,57,62,60,61.

In the 25 studies that did not report having generated attributes using input from all respondent groups, there was an equal split between those that generated attributes using only health care providers (n=7)39,40,42,50–52,64, non-health care providers (n=9)36,37,44,46,48,62,63,65,66, or neither (n=7)9,38,43,45,47,49,67. Those that reported generating attributes using neither respondent groups most often used literature reviews alone47,49, or literature reviews in conjunction with expert opinion38,43,67, information from regulatory requirements9, or product labelling45 to inform attributes. Two studies did not report that attributes had been developed in groups representative of the intended respondents; one study reported that attributes and levels were chosen by the authors55 (Lee) and the other did not provide any detail41."

3.10. 25-27. With this sentence in the text there seems no point in having the same information in Table 1. (p10, 3-9)

RESPONSE: Agree, this has been removed from the table.

3.11. , 44-54. Suggest split this long sentence

RESPONSE: Agree, this sentence has been revised

3.12. P 12 ‘some studies’ referred to; was that a single study (ref 50)?

RESPONSE: Yes, we have now changed the wording changed to reflect this.

3.13. P 14, 37. What is meant here by ‘fairly consistent’?

RESPONSE: Agree, was ambiguous. Have changed to “The predominance of mixed concordance and discordance conclusions appear to be consistent irrespective of the methods used to test for concordance.”

14. P 14-15 Rather brief details are presented of findings on concordance; it would be of interest to have further information on this area of the review.

RESPONSE: We have provided additional detail on the findings of concordance in line with this comment and the comments of the previous two reviewers. The manuscript now includes additional detail in the method, results and discussion section, as well as updating the figure. We believe that we have provided findings to the extent that they are supported by the data we had.

15. P 15, 21-53. The findings presented here seem to be incomplete. Eight studies are referred to in the text but only two of these are included in the nine studies cited in Figure 2. The findings are of interest but a full presentation is needed.

RESPONSE: Thank you for highlighting this issue. In this section we cite examples of the type of attribute we are describing in figure 2, with the examples intended to add some context to the classification of attributes. However we realize that the examples we were using in text and in the figure were drawn from different sources which is confusing. In addition to this issue, reviewer 1 noted that our citation numbers were different in the text and the figure. We have revised these statements in text and the figure and believe that this will avoid this source of confusion.

16. P 16 The discussion includes points on some limitations of the reviewed studies and makes useful recommendations for future DCEs. As suggested above, reference to these might be made in the abstract.

RESPONSE: We have rewritten the abstract in response to this comment and those of reviewer 2. We have added in the key recommendations to the extent that the word limit of the abstract permits.

VERSION 2 – REVIEW

REVIEWER	Matthew Quaife London School of Hygiene and Tropical Medicine, UK
REVIEW RETURNED	22-Feb-2017

GENERAL COMMENTS	Appreciate the authors' considered responses - no further comments.

REVIEWER	David Hailey

	University of Wollongong Australia
REVIEW RETURNED	20-Feb-2017

GENERAL COMMENTS	Previous comments on the manuscript have been suitably addressed and appropriate changes made. No further comments.