PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Belimumab: a technological advance for SLE patients? Report of a systematic review and meta-analysis
AUTHORS	Kandala, Ngianga-Bakwin; Connock, Martin; Grove, Amy; Sutcliffe, Paul; Mohiuddin, Syed; Hartley, Louise; Court, Rachel; Cummis, Ewen; Gordon, Caroline; Clarke, Aileen

VERSION 1 - REVIEW

REVIEWER	Peter Watson Statistician MRC Cognition and Brain Sciences Unit 15 Chaucer Road Cambridge UK CB2 7EF I have no conflicting interests with the research presented in this study.
REVIEW RETURNED	11-Apr-2013

THE STUDY	There appear to be many analyses and response variables without any particular one being of primary interest. I have a concern given the heterogeneity of the ethnicity (page 10 second paragraph) and the small implied number of studies (page 3, results, first sentence) of generalisability of the results and representativeness to other populations. The degree of between study heterogeneity could be stated using I^2 and, if not already, accounted for in deriving pooled estimates. Other aspects of the results and figures could be described in greater depth (see comments below) including labelling and captioning of all the figures and more clearly linking the results in the text to those in the figures and stating which analyses are used to produce the results plotted in the figures. A couple of references on meta-analysis that may be of use I put in the comments below which may be worth adding to the bibliography.
RESULTS & CONCLUSIONS	It is not clear how the unnumbered and uncaptioned figures (pages 27-29) relate to the results (pages 8-9) and if, and how adjusted, the pooled odds ratios quoted on page 8 (first paragraph lines 8-9) relate to the binary responses in Table 3 (page 32). I think the lack of statistical significance of both inter-study heterogeneity of effect sizes (page 10 first paragraph) and the confidence intervals of the figures mostly containing values ('1' for odds ratios and '0' for mean differences) suggesting no group differences could be down to limited power and possibly low sample sizes. This weakness may be mitigated by the number of point estimates suggesting a (hopefully clinically meaningful) benefit of the belimumab treatment but this needs to be motivated in the text.

	There are limitations (conclusions section on page 3) concerning 'hidden confounders' and interpretability of pooled estimates. It is not clear to me (see later comments) if this is 'merely' downplaying a pooled estimate and usefulness of a meta-analysis as the (limited number, three, of) populations being pooled are so different from each other or, more seriously, if there could be possible uncontrolled differences in clinically meaningful characteristics (confounders) between the placebo and treatment groups in one or more studies which would render any differences between the groups problematic to interpret as they could be simply due to factors other than the belimumab treatment. There is some mention of stratified randomisation on page 7 (start of second paragraph) but no details of what factors were used as stratifers.
GENERAL COMMENTS	Belimumab: a technological advance for SLE patients? Report of a systematic review and meta-analysis. bmjopen-2013-002852. This study compares a group using a new treatment for multi-organ auto-immune disease, Belimumab, with a placebo group by, on pages 8 and 9, obtaining confidence intervals for odds ratios (for a series of binary responses) plotted in the figures on page 27 and mean differences (for continuous ones) plotted in the figure on page 28 and reports a meta-analysis on page 9 for each of five (?) outcomes which look at the group effect which I suspect may be plotted in the figures on page 29. I, unfortunately, found the description (on pages 8 and 9) and presentation of the results (in the figures on pages 27-29) confusing and imprecise making it difficult to marry together the description of the results in the text and the confidence intervals plotted in the figures. The structure of the data being analysed needs to be fleshed out in the body of the text to help understanding of the results e.g. I am not sure if the meta-analyses are pooling across different studies or different subgroups within a single study or precisely what the SLE in the title of this paper stands for (it presumably is an abbreviation?) In particular the figures on pages 27-29 were not numbered or captioned which made it more difficult to know which analyses and effect sizes (odds ratio or 'mean difference') they were referring to and, in particular, which is the Figure 6 listed as corresponding to the meta-analyses reported briefly in the second paragraph on page 9. There is also an effect size called the 'hazard ratio' in a figure on page 27 which does not seem to be defined in the text. There are a lot of responses (listed both within the figures and represented by these different figures on pages 27-29 and also mentioned as a basis for various meta-analyses in the first sentence of the second paragraph on page 9). It is not clear to me if the results of analyses of these separate responses are being presented or discussed separately or together. Page 6. I think it makes more sense grammatically to say at the end of the first sentence of the 'Statistical analysis' paragraph on page 6 that odds ratios and mean differences 'were calculated for binary and continuous outcomes respectively'. Two reviewers are mentioned on page 6 under 'inclusion criteria' as assessing inclusion of studies. Was this assessment done independently by the two raters and, if so, could a kappa statistic, or alternative, be quoted to show inter-rater agreement?

	Pages 6 and 8. The statistical analysis on page 6 mentions 'unadjusted odds ratios'. Adjusted odds ratios are then presented (fifth line from bottom of first paragraph on page 8) but it doesn't mention in either sentence what these odds ratios are adjusted for or how or why both unadjusted and adjusted odds ratios are used. If its ok to use unadjusted odds ratios why adjust them? Pages 6 and 8. Is the 'mean difference' reported in the 'Statistical analysis' paragraph on page 6 and in the second paragraph on page 8 a standardised group one such as Cohen's d if you are wishing to compare results for different responses which may have different scales? Pages 6 and 9. I would like to see in the meta-analysis (second paragraph on page 9) the value of I^2 and any associated p-value, which was used (as stated on page 6 in the second last paragraph labelled 'statistical analysis') to test for the heterogeneity of effect size as this is an important test given that the degree of study heterogeneity is referred to throughout this paper. There are rules of thumb for small, medium and large values of I^2 that could be used. A value of 0% indicates no observed heterogeneity, 25%-49% is low heterogeneity, 50%-74% is moderate and 75% and above is large (Higgins et al, 2003). You could also mention in the statistical analysis paragraph on page 6 if you used a Der Simonian pooled estimate for the effect sizes or a fixed effect one such as the Mantel-Haenszel estimate for odds ratios in the meta-analysis as you found (page 9) little or no between study variation. Page 7. Second paragraph, line 1 mentions 'stratified randomisation' was used. What factors were stratified for and for what factors were the arms 'well balanced'? Page 8. I am not clear from the results on pages 8 and 9 how we should go about interpreting the confidence intervals in the figures on pages 27 to 29. Confidence intervals for odds ratios 'pooled across trials' are presented in the first paragraph (line 6) on page 8 but these are not graphed in the figures on pages 27 and 28 and I am not sure how these tie in with the confidence intervals in the figures. Are the results on lines 8-9 of the first paragraph on page 8 pooling odds ratios across all the binary variables mentioned in Table 3 (page 32) in BLISS-52 and BLISS-76 and is a pooled odds ratio interpretable when pooling over apparently different tests? The 'pooled across trials' implies some meta-analysis may have been performed to yield these pooled odds ratios. Page 9. The first line of the second paragraph on page 9 implies that a meta-analysis is performed on each of at least five different responses (as meta-analyses usually pool over trials measuring effect sizes using the same response and groups) and there is a mention of figure 6 which is the last figure in the paper presumably the one on page 29 yet I can't see six separate plots here. I would also have expected to see a confidence interval for a pooled effect size at the base of each of the forest plots corresponding to the meta-analysis of each response. Page 9. The last sentence of the first paragraph on page 9 mentions there were various causes of death but does not mention what these were which I would have thought would be of interest in giving a background to the data. I am not sure if the 'study level' referred to in the first sentence of the second paragraph on page 9 refers to

	separate subgroups within studies or, the usual pooling unit of pooling in meta-analyses, separate studies. Pages 9 and 29. I don't see any mention of a funnel plot to test and adjust for any possible publication bias. This analysis, at least, is usually performed and plotted routinely in meta-analyses including those submitted to this journal. Other tests can also be used – see, for example, Peters et al. (2010). Pages 9, 27-29. Page 9 implies a meta-analysis has been performed and, in light of this, I was surprised to see the size of the point estimates in the middle of all the confidence intervals plotted in the figures on pages 27-29 looking the same size as these usually differ in size as they are proportional to the weighting given to the studies in the meta-analysis to construct a pooled estimate. I also think, therefore, for the forest plot(s) you could add in a column by the plot showing the value of the weights used to confirm the studies had a similar weighting used in constructing the pooled estimate. Page 10. The first paragraph mentions that there was no heterogeneity found (across the studies or subgroups?) in the BLISS-52 trial but, counterintuitively, the racial background and ethnicity of participants 'varied considerably' and concludes there should be heterogeneity which confuses the conclusion and makes one start to doubt the tests of heterogeneity that have been used in this analysis as basis for obtaining pooled estimates. I am not sure if the conclusion (page 10 first line of first paragraph) that the benefits of belimumab are 'greater across the board' is warranted looking at the confidence interval plots on pages 27-29 since most of these intervals contain either an odds ratio of one or a zero difference which both correspond to no difference. One might possibly argue that, ignoring variances, the bulk of the point estimates, comprising odds ratios and mean group differences, are benefitting the use of the treatment, belimumab, but this needs to be carefully argued in the light that few of them are statistically significant and given the acknowledged heterogeneity (on page 10) which the authors may wish to account for if they have not done so already in obtaining pooled effect sizes despite the 'usual' tests of these not flagging this which may be due to lack of power from heterogeneity across only three studies being tested. Pages 27, 28 and 29. The figure(s) containing the forest plots need to be numbered and captioned. Is it necessary to both plot and quote the confidence intervals for group differences in these figures. Would simply plotting these confidence intervals be enough? The plot on page 28 plots hazard ratios (as opposed to rates?) in the 'time to event' figure which are, generally, not the same as odds ratios. The hazard ratios should be defined in the text but I can't see any mention of hazard ratios anywhere else in the paper (e.g. in the statistical analysis paragraph on page 6 or in the results sections on pages 8 and 9). The study does not explicitly state on page 9 in the meta-analysis results section how many trials are being pooled to obtain pooled effect sizes in the meta-analyses although elsewhere (for example on page 3, first line in first paragraph) three trials are mentioned and two 'relevant trials' (page 2 second bullet point under 'strengths and limitations'). Usually one has sufficient numbers of studies being pooled to make any results generalizable across different types of

	study to different populations. I mention this, as three trials, if this is the number used, does not seem very many for a meta-analysis particularly one where there is considerable between study heterogeneity at least in ethnicity (as already noted in the first paragraph on page 10), and as some of the plots in the figures on pages 27-29 only contain four rows (and then assuming one would be pooling BLISS-52 and BLISS-76 whose pooling might be questionable given separate confidence intervals are presented for these in the fifth last row from the end of the first paragraph on page 8). On page 3 (in the conclusions paragraph) the fourth line states generalizability of 'pooled results should be viewed with caution' and lines 5 and 6 mention possible 'hidden confounders'. Is this saying that the pooled studies may have differed from one another in many respects (confounders) and/or is it saying there are so many possibly uncontrolled confounders of clinical relevance in these group comparisons that we are looking at group differences (the belimumab treatment group vs the placebo group) that could be due to other clinically meaningful confounding factors which differ between the treatment and placebo groups? The latter could be a serious drawback to interpretability of any results whereas the former would, at least, preclude an interpretable pooled estimate since we would be averaging over such disparate (and few) populations which rather undermines the usefulness of a meta-analysis. References Higgins, JP, Thompson SG, Deeks JJ and Altman DG (2003). Measuring inconsistency in meta-analyses. BMJ, 327, 557-560. Peters, J.L., Sutton, A.J., Jones, D.R. and Abrams K.R. (2010). Assessing publication bias in meta-analyses in the presence of between-study heterogeneity. Journal of the Royal Statistical Society A, 173(3), 575-591 There is an on-line copy of this paper at http://onlinelibrary.wiley.com/doi/10.1111/j.1467-985X.2009.00629.x/full.

REVIEWER	Ricard Cervera, MD, PhD, FRCP Head, Department of Autoimmune Diseases Hospital Clínic Barcelona, Catalonia, Spain Statement: I have no competing interests with the authors of this manuscript.
REVIEW RETURNED	25-Apr-2013

THE STUDY	The supplemental documents do not contain information that should be better reported in the manuscript.
GENERAL COMMENTS	This is an interesting systematic review and meta-analysis of the randomized controlled trials of belimumab in patients with systemic lupus erythematosus. The study was well designed, the results are interesting and the manuscript is well written with well balanced discussion.

VERSION 1 – AUTHOR RESPONSE

Reviewer 1:
Reviewer: Peter Watson
Statistician

MRC Cognition and Brain Sciences Unit
15 Chaucer Road
Cambridge
UK
CB2 7EF

I have no conflicting interests with the research presented in this study.

1. There appear to be many analyses and response variables without any particular one being of primary interest. I have a concern given the heterogeneity of the ethnicity (page 10 second paragraph) and the small implied number of studies (page 3, results, first sentence) of generalisability of the results and representativeness to other populations. The degree of between study heterogeneity could be stated using I^2 and, if not already, accounted for in deriving pooled estimates. Other aspects of the results and figures could be described in greater depth (see comments below) including labeling and captioning of all the figures and more clearly linking the results in the text to those in the figures and stating which analyses are used to produce the results plotted in the figures.

Reply: We would like to thank the reviewer for his comments on our manuscript. We have now carefully checked the manuscript to account for these comments and suggestions.

The reviewer is right, although the main primary outcome to determine the effectiveness of belimumab was the Responder Index (SRI) at week 52, we also examined other outcomemeasures for the three RCTs that evaluated belimumab effectiveness e.g. examining the SLE Responder Index (SRI) at week 76. We also included those outcomes identified by the belimumab investigators in their protocol as “major secondery and other outcomes”. We have now clearly identified the primary outcome designated in the RCTs (namely SRI at 52 weeks), we have stated that this is also our primary outcome, and have included text to explain the origin of this novel outcome measure as developed between the FDA and the belimumab trialists.

We have attempted to highlight more explicitly that our manuscript concerns the generalizability of pooled results and that these should be viewed with caution. We noted that population heterogeneity; geography and / or variation in trial conduct may be influence results; we have removed reference to “hidden confounders”. Although formal tests for statistical heterogeneity were negative, BLISS-52 results were systematically more favourable for all measured outcomes.
These elaborations on the interpretation of our results are found mainly in lines:
89-95; 197-201; 261-271.

2. A couple of references on meta-analysis that may be of use I put in the comments below which may be worth adding to the bibliography.
Reply: We have added the Higgins reference as suggested; the reference for publication bias has not been added because it was not possible to ascertain if there was publication bias with only two RCTs; we have added text to this effect (line 154) the reference to the Cochrane Handbook (number 21) was therefore considered sufficient. Ref 21 (page 317) recommends at least 10 studies would be required for analysis of small study bias and we have been guided by this.

3. It is not clear how the unnumbered and uncaptioned figures (pages 27-29) relate to the results (pages 8-9) and if, and how adjusted, the pooled odds ratios quoted on page 8 (first paragraph lines 8-9) relate to the binary responses in Table 3 (page 32). I think the lack of statistical significance of both inter-study heterogeneity of effect sizes (page 10 first paragraph) and the confidence intervals of the figures mostly containing values ('1' for odds ratios and '0' for mean differences) suggesting no group differences could be down to limited power and possibly low sample sizes. This weakness may be mitigated by the number of point estimates suggesting a (hopefully clinically meaningful) benefit of the belimumab treatment but this need to be motivated in the text.

Reply: We have revised and numbered captions of figures and relate them clearly to the results section as suggested. Lines 214-218 explain how the results depicted in figure were derived; lines 227-230 explain how the adjusted odds ratios were derived / reported. As for the weakness of the study as mentioned by the reviewer, the reviewer makes an important point. The reviewer indicates that confidence intervals “suggesting no group differences” might be attributable to lack of power is of course probable, however the modest effect size (small benefit of belimumab) is a major contributory factor. Due to the scarcity of RCTs evaluating the effectiveness of belimumab, we restricted our study to the available evidence. The three RCTs combined investigated 2133 SLE patients, which may be a good sample size for this type of rare condition (e.g. the SLE Rituximab trial, the only other major recent trial for SLE, recruited 184 patients into two arms).

4. There are limitations (conclusions section on page 3) concerning 'hidden confounders' and interpretability of pooled estimates. It is not clear to me (see later comments) if this is 'merely' downplaying a pooled estimate and usefulness of a meta-analysis as the (limited number, three, of) populations being pooled are so different from each other or, more seriously, if there could be possible uncontrolled differences in clinically meaningful characteristics (confounders) between the placebo and treatment groups in one or more studies which would render any differences between the groups problematic to interpret as they could be simply due to factors other than the belimumab treatment. There is some mention of stratified randomisation on page 7 (start of second paragraph) but no details of what factors were used as stratifers.
Reply: We have attempted to clarify these issues. We have removed the phrase “hidden confounders” and have explicitly considered the influence of geographical / ethnic / trial conduct differences between the BLISS trials by first pointing to the systematic difference in results between B52 and B76 (lines: 236-242; 262-272) and by alluding to the ethnic / geographical data presented in Table2 and Figure3; lines 262-272) . We now provide explicit information about the stratification undertaken in the BLISS trials and the use of strata in adjusting results reported in the published accounts (lines 182-184; 227-229). The limitation we mentioned is not only limited to the general applicability of the nature of meta-analysis but also to real limitations due to confounders such as the geographic location and the ethnicity where the studies were conducted.

5.This study compares a group using a new treatment for multi-organ auto-immune disease, Belimumab, with a placebo group by, on pages 8 and 9, obtaining confidence intervals for odds ratios (for a series of binary responses) plotted in the figures on page 27 and mean differences (for continuous ones) plotted in the figure on page 28 and reports a meta-analysis on page 9 for each of five (?) outcomes which look at the group effect which I suspect may be plotted in the figures on page 29.

Reply: Please see the method section of the paper. We performed a meta-analysis of two randomized controlled trials (RCTs) of belimumab against placebo or best supportive care. To improve clarity we have edited the figure captions and Method sections. The Meta-analysis figure (Figure 6) shows the results of random effects meta- analysis of the two BLISS trials for each of 14 outcomes designated by belimumab trialists as primary or major secondary or “other major” outcomes. For convenience of viewing we combined the results for different types of outcome into a single figure (binary, time to

event and continuous) using Excel.

6. I, unfortunately, found the description (on pages 8 and 9) and presentation of the results (in the figures on pages 27-29) confusing and imprecise making it difficult to marry together the description of the results in the text and the confidence intervals plotted in the figures. The structure of the data being analysed needs to be fleshed out in the body of the text to help understanding of the results e.g. I am not sure if the meta-analyses are pooling across different studies or different subgroups within a single study or precisely what the SLE in the title of this paper stands for (it presumably is an abbreviation?)

Reply: We have defined SLE in the title and text. We have clarified the results and figures presented to explain that the pooling was across different studies (the two RCTs). We also present within-study results for the primary outcome according to different geographical subgroups (lines 262-272). With many outcomes and sub-groups analysis it became difficult for the reader we consider that we have improved the paper in this regard.

7. In particular the figures on pages 27-29 were not numbered or captioned which made it more difficult to know which analyses and effect sizes (odds ratio or 'mean difference') they were referring to and, in particular, which is the Figure 6 listed as corresponding to the meta-analyses reported briefly in the second paragraph on page 9. There is also an effect size called the 'hazard ratio' in a figure on page 27 which does not seem to be defined in the text. There are a lot of responses (listed both within the figures and represented by these different figures on pages 27-29 and also mentioned as a basis for various meta-analyses in the first sentence of the second paragraph on page 9). It is not clear to me if the results of analyses of these separate responses are being presented or discussed separately or together.
Reply: Thank you for these comments. We have now explained the derivation of the hazard ratio results (lines 236-242). We have attempted to explain why so many outcome measures exist for SLE (lines 88 to 95) and how this led to the development of the SRI measure. These results are mainly, but not exclusively, discussed together since the most noticeable feature common to all is the better performance of belimumab in B52 relative to B76 (lines 262 265; 307-313).

8. Page 6. I think it makes more sense grammatically to say at the end of the first sentence of the 'Statistical analysis' paragraph on page 6 that odds ratios and mean differences 'were calculated for binary and continuous outcomes respectively'. Two reviewers are mentioned on page 6 under 'inclusion criteria' as assessing inclusion of studies. Was this assessment done independently by the two raters and, if so, could a kappa statistic, or alternative, be quoted to show inter-rater agreement?

Reply: We have added modified the sentence as suggested and clarified the independence and tasks of the two reviewers (lines 138-145).

9. Pages 6 and 8. The statistical analysis on page 6 mentions 'unadjusted odds ratios'. Adjusted odds ratios are then presented (fifth line from bottom of first paragraph on page 8) but it doesn't mention in either sentence what these odds ratios are adjusted for or how or why both unadjusted and adjusted odds ratios are used. If its ok to use unadjusted odds ratios why adjust them?

Reply: We have now clarified the use of adjusted and unadjusted odds ratio to make it clear to the reader why both were presented (lines 214-218; 227-229)

10. Pages 6 and 8. Is the 'mean difference' reported in the 'Statistical analysis' paragraph on page 6 and in the second paragraph on page 8 a standardised group one such as Cohen's d if you are wishing to compare results for different responses which may have different scales?

Reply: The mean difference' reported in the 'Statistical analysis' in paragraph 6 and 8 is mean difference' reported in the BLISS RCTs. Each outcome used the same assessment tool in both trials and "standardized mean difference" such as Cohen's d was not appropriate.

11. Pages 6 and 9. I would like to see in the meta-analysis (second paragraph on page 9) the value of I^2 and any associated p-value, which was used (as stated on page 6 in the second last paragraph labelled 'statistical analysis') to test for the heterogeneity of effect size as this is an important test given that the degree of study heterogeneity is referred to throughout this paper. There are rules of thumb for small, medium and large values of I^2 that could be used. A value of 0% indicates no observed heterogeneity, 25%-49% is low heterogeneity, 50%-74% is moderate and 75% and above is large (Higgins et al, 2003). You could also mention in the statistical analysis paragraph on page 6 if you used a Der Simonian pooled estimate for the effect sizes or a fixed effect one such as the Mantel-Haenszel estimate for odds ratios in the meta-analysis as you found (page 9) little or no between study variation.

Reply: We now explain that we used the random effects method of DerSimonian Laird (line 156) to pool effect sizes. We anticipated heterogeneity so a random effects model was more appropriate in this case than the fixed effects model. We have now displayed (in Figure 6) the value of I^2 and the associated p-value as suggested, and we have tightened the text so the lack of statistical heterogeneity refers specifically to binary and time to event outcomes.

12. Page 7. Second paragraph, line 1 mentions 'stratified randomisation' was used. What factors were stratified for and for what factors were the arms 'well balanced'?
Reply: We have now clarified this (lines 227-229). Baseline balance included values for: proteinuria, disease duration, gender, race, IgG, autoantibody, and complement levels , baseline SLEDAI and PGA scores, BILAG organ domain involvement and SLICC Damage Index score; we have now included this information in the caption to figure 5.

13. Page 8. I am not clear from the results on pages 8 and 9 how we should go about interpreting the confidence intervals in the figures on pages 27 to 29. Confidence intervals for odds ratios 'pooled across trials' are presented in the first paragraph (line 6) on page 8 but these are not graphed in the figures on pages 27 and 28 and I am not sure how these tie in with the confidence intervals in the figures. Are the results on lines 8-9 of the first paragraph on page 8 pooling odds ratios across all the binary variables mentioned in Table 3 (page 32) in BLISS-52 and BLISS-76 and is a pooled odds ratio interpretable when pooling over apparently different tests? The 'pooled across trials' implies some meta-analysis may have been performed to yield these pooled odds ratios.

Reply: In the figures referred to (on pages 27 to 29) ORs are unadjusted (now explained more clearly lines 214-218). Additionally we have explained that the BLISS trial journal articles and the manufacturer's submissions to the FDA and to NICE used a logistic regression model (individually for each trial in the journal articles, and after pooling populations in the case of the submissions to the approval authorities; lines -229; 227 and 276-279).

14. Page 9. The first line of the second paragraph on page 9 implies that a meta-analysis is performed on each of at least five different responses (as meta-analyses usually pool over trials measuring effect sizes using the same response and groups) and there is a mention of figure 6 which is the last figure in the paper presumably the one on page 29 yet I can't see six separate plots here. I would also have expected to see a confidence interval for a pooled effect size at the base of each of the forest plots corresponding to the meta-analysis of each response.

Reply: The text referring to Figure 6 has been clarified (lines 282-285). The figure has been redrawn

and figure caption improved to correct for errors and improve clarity for the reader.

15. Page 9. The last sentence of the first paragraph on page 9 mentions there were various causes of death but does not mention what these were which I would have thought would be of interest in giving a background to the data. I am not sure if the 'study level' referred to in the first sentence of the second paragraph on page 9 refers to separate subgroups within studies or, the usual pooling unit of pooling in meta-analyses, separate studies.
Reply: We have now included the causes of death (lines 250-251). The term "study level" was used to distinguish the results presented from those in the manufacturer's submission to the FDA in which IPD from the two BLISS trials was pooled prior to logistic regression analysis; hopefully this is now clear from the text (lines 281-287)

16. Pages 9 and 29. I don't see any mention of a funnel plot to test and adjust for any possible publication bias. This analysis, at least, is usually performed and plotted routinely in meta-analyses including those submitted to this journal. Other tests can also be used – see, for example, Peters et al. (2010).
Reply: The reviewer makes a potentially important point here. We did not include a formal test of small study bias because there are too few trials evaluating the effectiveness of Belimumab (see line 154 with accompanying reference 21) . We believe, a test of publication bias in this context may not be useful.

17. Pages 9, 27-29. Page 9 implies a meta-analysis has been performed and, in light of this, I was surprised to see the size of the point estimates in the middle of all the confidence intervals plotted in the figures on pages 27-29 looking the same size as these usually differ in size as they are proportional to the weighting given to the studies in the meta-analysis to construct a pooled estimate. I also think, therefore, for the forest plot(s) you could add in a column by the plot showing the value of the weights used to confirm the studies had a similar weighting used in constructing the pooled estimate.

Reply: All points are the same size because each refers to the pooled estimate for a particular outcome, not to a single study estimate given a specific weight in the analysis. We hope the text and figure caption and redrawn figure now make this clearer.

18. Page 10. The first paragraph mentions that there was no heterogeneity found (across the studies or subgroups?) in the BLISS-52 trial but, counterintuitively, the racial background and ethnicity of participants 'varied considerably' and concludes there should be heterogeneity which confuses the conclusion and makes one start to doubt the tests of heterogeneity that have been used in this analysis as basis for obtaining pooled estimates. I am not sure if the conclusion (page 10 first line of first paragraph) that the benefits of belimumab are 'greater across the board' is warranted looking at the confidence interval plots on pages 27-29 since most of these intervals contain either an odds ratio of one or a zero difference which both correspond to no difference. One might possibly argue that, ignoring variances, the bulk of the point estimates, comprising odds ratios and mean group differences, are benefitting the use of the treatment, belimumab, but this needs to be carefully argued in the light that few of them are statistically significant and given the acknowledged heterogeneity (on page 10) which the authors may wish to account for if they have not done so already in obtaining pooled effect sizes despite the 'usual' tests of these not flagging this which may be due to lack of power from heterogeneity across only three studies being tested.

Reply: We performed the I squared test for statistical heterogeneity between the two BLISS trials used in the meta analyses and found low values for all outcomes. But we believe that there are other sources of heterogeneity (geographical, trial conduct etc.) which have exerted a systematic influence on the outcomes, the major indicator of this influence being the consistently superior performance of

one trial compared to the other across multiple outcomes. Hopefully the new text (e.g. lines 262 -272) explains this more clearly. The fact that BLISS 76 outcomes almost always fail to reach statistical significance is now brought out more clearly (e.g. lines 253-260); even though the fact that the primary outcome in BLISS 76 was satisfied on extending observation to 76 weeks eliminates the statistical significance of the SRI. While the lack of statistical significance may be attributable to some extent to lack of power it is also clear that effect sizes in BLISS 76 are modest.

19. Pages 27, 28 and 29. The figure(s) containing the forest plots need to be numbered and captioned. Is it necessary to both plot and quote the confidence intervals for group differences in these figures? Would simply plotting these confidence intervals be enough?

Reply: We have now numbered and captioned the plots.
Figures 3 and 4 present the comparison (intervention versus control) separately for the two BLISS trials because this highlights the fact that BLISS 52 always gives a better result for belimumab than does BLISS 76. We think the CIs are necessary because again they highlight the difference between the trials.

20. The plot on page 28 plots hazard ratios (as opposed to rates?) in the ‘time to event’ figure which are, generally, not the same as odds ratios. The hazard ratios should be defined in the text but I can’t see any mention of hazard ratios anywhere else in the paper (e.g. in the statistical analysis paragraph on page 6 or in the results sections on pages 8 and 9).
Reply: Thank you for the comments. We have now explained the hazard ratios in lines 236-242.

21. The study does not explicitly state on page 9 in the meta-analysis results section how many trials are being pooled to obtain pooled effect sizes in the meta-analyses although elsewhere (for example on page 3, first line in first paragraph) three trials are mentioned and two 'relevant trials' (page 2 second bullet point under 'strengths and limitations'). Usually one has sufficient numbers of studies being pooled to make any results generalizable across different types of study to different populations. I mention this, as three trials, if this is the number used, does not seem very many for a meta-analysis particularly one where there is considerable between study heterogeneity at least in ethnicity (as already noted in the first paragraph on page 10), and as some of the plots in the figures on pages 27-29 only contain four rows (and then assuming one would be pooling BLISS-52 and BLISS-76 whose pooling might be questionable given separate confidence intervals are presented for these in the fifth last row from the end of the first paragraph on page 8).
Reply: We performed the meta-analyses using outcomes from two randomized controlled trials (the two BLISS trials). (This is now more explicit in lines 281-282, and in the caption to the figure 6). We explain why the L02 trial was only used in assessing safety outcomes on lines 194-198. Problems in interpreting what is represented in the figures have been addressed in figure captions and with more explicit description in the body of the text.(lines 212-218 and 237-243).

22. On page 3 (in the conclusions paragraph) the fourth line states generalizability of ‘pooled results should be viewed with caution’ and lines 5 and 6 mention possible 'hidden confounders'. Is this saying that the pooled studies may have differed from one another in many respects (confounders) and/or is it saying there are so many possibly uncontrolled confounders of clinical relevance in these group comparisons that we are looking at group differences (the belimumab treatment group vs the placebo group) that could be due to other clinically meaningful confounding factors which differ between the treatment and placebo groups? The latter could be a serious drawback to interpretability of any results whereas the former would, at least, preclude an interpretable pooled estimate since we would be averaging over such disparate (and few) populations which rather undermines the usefulness of a meta-analysis.

Reply: Perhaps, it was not clear in the previous version of the paper. We have removed reference to “hidden confounders” and have clarified the conclusion section (especially lines 361-365). Please also see our reply in point 4 above.

References
Higgins, JP, Thompson SG, Deeks JJ and Altman DG (2003). Measuring inconsistency in meta-analyses. BMJ, 327, 557-560.

Peters, J.L., Sutton, A.J., Jones, D.R. and Abrams K.R. (2010). Assessing publication bias in meta-analyses in the presence of between-study heterogeneity. Journal of the Royal Statistical Society A, 173(3), 575-591 There is an on-line copy of this paper at http://onlinelibrary.wiley.com/doi/10.1111/j.1467-985X.2009.00629.x/full.

Reviewer 2
Ricard Cervera, MD, PhD, FRCP
Head, Department of Autoimmune Diseases
Hospital Clínic
Barcelona, Catalonia, Spain

Statement: I have no competing interests with the authors of this manuscript
Reviewer: This is an interesting systematic review and meta-analysis of the randomized controlled trials of belimumab in patients with systemic lupus erythematosus. The study was well designed, the results are interesting and the manuscript is well written with well balanced discussion.

Reply: We would like to thank the reviewer for his kind comments on our manuscript.

VERSION 2 – REVIEW

REVIEWER	Peter Watson Statistician MRC Cognition and Brain Sciences Unit 15 Chaucer Road Cambridge CB2 7EF I have no competing interests with this research.
REVIEW RETURNED	06-Jun-2013

THE STUDY	I'd like to see a bit more in the text about motivating the meta-analyses given that only two studies are combined. Perhaps motivate by saying that despite having only two studies they are relatively large ones. I also think the implications of not adjusting for publication bias could be discussed in the text perhaps in a sentence or two just to reassure the readers that we are not overestimating the effect sizes in this paper.
RESULTS & CONCLUSIONS	I think the results could be more logically described in the text by emphasising the consistency of the effect sizes across different responses and also motivating the results in Figure 6 (page 47) which combine the two BLISS trials which appear to have very similar effect sizes. I wasn't sure if the effect sizes from the two BLISS trials were meant to be regarded as behaving differently so as to be looked at separately or if they were sufficiently similar that they could be summarised using a combined BLISS effect size.

GENERAL COMMENTS	Belimumab: a technological advance for SLE patients? Report of a systematic review and meta-analysis. Bmjopen-2013-002852.R1. The authors have taken on board a lot of the comments in my earlier review including defining a primary outcome Page 6 (lines 207-208 on page 6) and in using random effects methods (end of first full paragraph on page 9) to compute effect sizes and their confidence intervals. I think whereas there have been some improvements there are still some queries (hopefully minor ones!) that could be addressed in the text to make the meta-analyses in this paper read more compellingly. I do hope the authors find these comments useful. Page 5. The authors provide a reference suggesting there are three studies considered for inclusion in the meta-analyses (from line 166 on page 5 and Table 1 on page 48) but only two of these, the BLISS-52 and BLISS-76, appear to be actually used in the meta-analyses (lines 3-5 of second full paragraph on page 6). Two does strike me as a rather limited number of studies for combining meta-analysis. I wondered if this had implications for the magnitude of sampling error regarding the accuracy of the combined effect sizes (comparing Belimumab with Placebo as given in Figure 6 on page 47) as these would be based upon only two samples. Hunter and Schmidt (2004) warn that with a small total sample size which typically results with small numbers of studies there can be high sampling error which limits the accuracy of the effect sizes? It is also intuitive that if the number of studies is too small, the resulting effect sizes can be unstable, and vary depending on which studies are included. It might be useful for the authors, therefore, to motivate the results from these meta-analyses to reassure on the accuracy of the obtained effect sizes especially as, the authors mention in their reply to my point 18 (page 39) the effect sizes which mostly favour Belimumab plotted in Figures 4 and 5 are also modest. You could probably get around this by firstly acknowledging the low number of studies but then plugging the fact in the text that although you only have a few studies they are relatively large (page 25 line 1 of paragraph 3) totalling over 2000 people although actually counting up the frequencies in Table 2 (page 49) for the BLISS-52 and BLISS-76 which appear to be used in all of the meta-analyses I get N=1697 used in the meta-analyses but this is still seems to me to be a large total sample. Page 8. On lines five and six (=lines 261-262) in the first full paragraph it is mentioned correctly (judging from the confidence intervals in Figures 4 and 5 (pages 45 and 46)) that BLISS-52 is superior to BLISS-76? (I'd personally say more specifically in the text on page 8 that BLISS-52 is 'consistently better over a range of different test responses' to show the difference between Belimumab and Placebo is in the same direction for all tests) but the confidence intervals for the two trials do overlap quite substantially for nearly all of the the test responses so are these differences between BLISS-52 and BLISS-76 of a magnitude to be clinically relevant given the differences in effect sizes do not appear to be statistically significantly different? Page 21. The authors report (third line of paragraph 3=line 157) there were too few studies to effectively test for publication bias. Although I can appreciate this might preclude such a test surely the lack of a test and possible publication bias adjustment suggests that

	the effect sizes reported in this paper are still liable to reviewer bias and error which must at least lead to a caution against interpretation of the effect sizes. Page 24. The fourth last line of page 24 mentions that the confidence intervals in Figure 6 (page 47) represent combined trials for the two BLISS trials which suggests Figure 6 is combining results in Figures 4 and 5 which compute separate confidence intervals for the two BLISS trials (BLISS-52 and BLISS-76). This appears at variance with lines 261-262 on page 8 (lines 4-5) in the first full paragraph which mention that BLISS-52 is superior to BLISS-76 in having a higher effect size. This latter point would suggest one is combining two effect sizes from trials which have different effect sizes to obtain a mean which does not seem intuitive. I wonder if you could motivate the combined BLISS confidence intervals plotted in Figure 6 more especially as the results in Figure 6 look more compelling than the individual BLISS trial plots in Figures 4 and 5. I think the narrative of the paper could be: we have three studies and compare two of these (BLISS-52 and BLISS-76) in Figures 4 and 5 which appear to have similar effect sizes (especially given there is no between trial heterogeneity found – first line of page 9). We then combine these two studies to obtain a Figure 6 which shows effect sizes which exhibit clear differences in favour of Belimumab both in statistical and clinical terms. Is this correct? I was not sure if you want to discuss these studies individually (as in Figures 4 and 5 and page 8 lines 4-5 of first full paragraph) or pooled as in Figure 6 (last four lines of page 8 to end of first four lines of page 9). I think also the second paragraph in the discussion is misleading in saying one study shows statistical significance and one doesn't for the primary response (SRI) in that the plotted confidence intervals in Figure 4 (first two plotted page 45) appear to show the differences in OR effect sizes is small between the two BLISS trials. I also think the BLISS-76 trial has a 95% confidence interval (second one plotted in Figure 4 on page 45) which appears not to include one but contrastingly is quoted on the second last line of the second paragraph of the discussion on page 9 as containing one (0.92-1.87) and being non-significant. Pages 25, 45-46. Figures 4 and 5. These figures do not appear to contain overall effect sizes ie pooled over the studies included in the meta-analysis since from page 25 (first line of third paragraph) the B52 and B76 represent two of the three RCTs involved in the meta-analysis. I wondered, as this is a meta-analysis, if you also thought of adding in the pooled confidence interval (pooling B52 and B76) for the effect sizes in Figures 4 and 5 below the trials being pooled to gve these pooled estimates as appears more logical and instructive and is customary in meta-analyses (these pooled confidence intervals seem to be given in a separate Figure 6 on page 47 instead)? I assume they do since each row appears to correspond to a different test but this isn't explicitly stated in the figures. Page 5. Could you elaborate in what way the outcome measures are 'adjusted' as described on line 158 in the statistical analysis section? Is this to do with including stratification factors as mentioned on lines 233-234 in the second paragraph on page 7? Pages 5 (line 155) & Page 46. Figure 5. I assume MD in column 3 represents standardised means differences ie SMD. Standardisation

	could also be added to line 155 in the first line of the statistical analysis section on page 5 which mentions mean differences. Reference Hunter JE and Schmidt FL (2004) Methods of meta-analyses. Correcting error and bias in research findings. Sage:Thousand Oaks.

VERSION 2 – AUTHOR RESPONSE

Reviewer 1:
Reviewer: Peter Watson
Statistician

MRC Cognition and Brain Sciences Unit
15 Chaucer Road
Cambridge
UK
CB2 7EF

I have no conflicting interests with the research presented in this study.

I'd like to see a bit more in the text about motivating the meta-analyses given that only two studies are combined. Perhaps motivate by saying that despite having only two studies they are relatively large ones. I also think the implications of not adjusting for publication bias could be discussed in the text perhaps in a sentence or two just to reassure the readers that we are not overestimating the effect sizes in this paper.

Reply: We have added the suggested wording in the statistical analysis section as shown below (see p5):
”There were too few studies for an analysis of publication bias.21 Although our thorough search found no further studies, we cannot completely rule out that any method for combining the two trials may result in an over-estimate or under-estimate of effect sizes due to publication bias.”

I think the results could be more logically described in the text by emphasising the consistency of the effect sizes across different responses and also motivating the results in Figure 6 (page 47) which combine the two BLISS trials which appear to have very similar effect sizes. I wasn't sure if the effect sizes from the two BLISS trials were meant to be regarded as behaving differently so as to be looked at separately or if they were sufficiently similar that they could be summarised using a combined BLISS effect size.

Reply: Thank you for the suggestion. We combined the two BLISS outcomes to assess the impact of each trial (effect sizes for individual trials versus a pooled estimate using a different method to that employed by the manufacturer in their submission to the FDA and to NICE) and to demonstrate the consistency of the results. The text has now been made clearer as suggested.

The authors have taken on board a lot of the comments in my earlier review including defining a primary outcome Page 6 (lines 207-208 on page 6) and in using random effects methods (end of first full paragraph on page 9) to compute effect sizes and their confidence intervals.
I think whereas there have been some improvements there are still some queries (hopefully minor ones!) that could be addressed in the text to make the meta-analyses in this paper read more

compellingly. I do hope the authors find these comments useful.

Reply: We thank the reviewer for the thoughtful comments and suggestions, which have been inserted.

Page 5. The authors provide a reference suggesting there are three studies considered for inclusion in the meta-analyses (from line 166 on page 5 and Table 1 on page 48) but only two of these, the BLISS-52 and BLISS-76, appear to be actually used in the meta-analyses (lines 3-5 of second full paragraph on page 6). Two does strike me as a rather limited number of studies for combining meta-analysis. I wondered if this had implications for the magnitude of sampling error regarding the accuracy of the combined effect sizes (comparing Belimumab with Placebo as given in Figure 6 on page 47) as these would be based upon only two samples. Hunter and Schmidt (2004) warn that with a small total sample size which typically results with small numbers of studies there can be high sampling error which limits the accuracy of the effect sizes? It is also intuitive that if the number of studies is too small, the resulting effect sizes can be unstable, and vary depending on which studies are included. It might be useful for the authors, therefore, to motivate the results from these meta-analyses to reassure on the accuracy of the obtained effect sizes especially as, the authors mention in their reply to my point 18 (page 39) the effect sizes which mostly favour Belimumab plotted in Figures 4 and 5 are also modest. You could probably get around this by firstly acknowledging the low number of studies but then plugging the fact in the text that although you only have a few studies they are relatively large (page 25 line 1 of paragraph 3) totalling over 2000 people although actually counting up the frequencies in Table 2 (page 49) for the BLISS-52 and BLISS-76 which appear to be used in all of the meta-analyses I get N=1697 used in the meta-analyses but this is still seems to me to be a large total sample.

Reply: We agree with the referee that there are limits to the accuracy of the effects size. However, the meta-analytic “psychometric method” of Hunter and Schmidt allows for adjustment / correction of methodological flaws in constituent studies but in practical terms focuses on study size limitations (as indicated by the reviewer). Here we attribute methodological flaws to geographical differences / study conduct (rather than study size), therefore we have retained our original meta-analytic procedure, which is far more commonly employed for pharmacological interventions for medical conditions than are Hunter-Schmidt methods.

As suggested we have inserted on the manuscript in the sample sizes used in the meta-analysis and it should be noted that we have only meta-analysed results for the license indications (10mg/kg monthly) (see p9 1st paragraph).
Please note that in the manuscript (Line 310), we state that the 2000 patients referred to by reviewers applies to all the three belimumab trials and for all doses.

Page 8. On lines five and six (=lines 261-262) in the first full paragraph it is mentioned correctly (judging from the confidence intervals in Figures 4 and 5 (pages 45 and 46)) that BLISS-52 is superior to BLISS-76? (I’d personally say more specifically in the text on page 8 that BLISS-52 is ‘consistently better over a range of different test responses’ to show the difference between Belimumab and Placebo is in the same direction for all tests) but the confidence intervals for the two trials do overlap quite substantially for nearly all of the the test responses so are these differences between BLISS-52 and BLISS-76 of a magnitude to be clinically relevant given the differences in effect sizes do not appear to be statistically significantly different?

Reply: We have inserted in the manuscript the reviewer’s suggested wording (“over a range of tests responses) see page 8. In the discussion, we have already implied an overlap in the confidence interval for the outcomes from the two BLISS trials.

Page 21. The authors report (third line of paragraph 3=line 157) there were too few studies to effectively test for publication bias. Although I can appreciate this might preclude such a test surely the lack of a test and possible publication bias adjustment suggests that the effect sizes reported in this paper are still liable to reviewer bias and error which must at least lead to a caution against interpretation of the effect sizes.

Reply: We have inserted the appropriate text as suggested in the statistical methods section (see p5)

Page 24. The fourth last line of page 24 mentions that the confidence intervals in Figure 6 (page 47) represent combined trials for the two BLISS trials which suggests Figure 6 is combining results in Figures 4 and 5 which compute separate confidence intervals for the two BLISS trials (BLISS-52 and BLISS-76). This appears at variance with lines 261-262 on page 8 (lines 4-5) in the first full paragraph which mention that BLISS-52 is superior to BLISS-76 in having a higher effect size. This latter point would suggest one is combining two effect sizes from trials which have different effect sizes to obtain a mean which does not seem intuitive. I wonder if you could motivate the combined BLISS confidence intervals plotted in Figure 6 more especially as the results in Figure 6 look more compelling than the individual BLISS trial plots in Figures 4 and 5.

Reply: The reviewer may be right. However, we combined the two BLISS outcomes to assess the impact of each trial (effect sizes for individual trials versus a pooled estimate using a different method to that employed by the manufacturer in their submission to the FDA and to NICE) and to demonstrate the consistency of the results. The text has now been made clearer as suggested.

I think the narrative of the paper could be: we have three studies and compare two of these (BLISS-52 and BLISS-76) in Figures 4 and 5 which appear to have similar effect sizes (especially given there is no between trial heterogeneity found – first line of page 9). We then combine these two studies to obtain a Figure 6 which shows effect sizes which exhibit clear differences in favour of Belimumab both in statistical and clinical terms. Is this correct? I was not sure if you want to discuss these studies individually (as in Figures 4 and 5 and page 8 lines 4-5 of first full paragraph) or pooled as in Figure 6 (last four lines of page 8 to end of first four lines of page 9).

Reply: We wished to discuss the trials individually and after pooling by the two methods (meta-analysis and the manufacturer logistic regression) in order to highlight for the readers the impact of geographical differences between the two trials, which potentially have clinical implications for different populations receiving belimumab.

I think also the second paragraph in the discussion is misleading in saying one study shows statistical significance and one doesn't for the primary response (SRI) in that the plotted confidence intervals in Figure 4 (first two plotted page 45) appear to show the differences in OR effect sizes is small between the two BLISS trials. I also think the BLISS-76 trial has a 95% confidence interval (second one plotted in Figure 4 on page 45) which appears not to include one but contrastingly is quoted on the second last line of the second paragraph of the discussion on page 9 as containing one (0.92-1.87) and being non-significant.

Reply: Please see page 9. We state that the primary outcome was reached in both trials. The second last line of the second paragraph of the discussion on page 9 "as containing one (0.92-1.87) and being non-significant" is not contradictory since this refers to the SRI result at week 76 and this is not the primary outcome (the text states that the primary outcome is SRI at week 52).

Pages 25, 45-46. Figures 4 and 5. These figures do not appear to contain overall effect sizes ie

pooled over the studies included in the meta-analysis since from page 25 (first line of third paragraph) the B52 and B76 represent two of the three RCTs involved in the meta-analysis. I wondered, as this is a meta-analysis, if you also thought of adding in the pooled confidence interval (pooling B52 and B76) for the effect sizes in Figures 4 and 5 below the trials being pooled to gve these pooled estimates as appears more logical and instructive and is customary in meta-analyses (these pooled confidence intervals seem to be given in a separate Figure 6 on page 47 instead)? I assume they do since each row appears to correspond to a different test but this isn't explicitly stated in the figures.

Reply: Figures 4 and 5 were not obtained as results of the meta-analysis. They represent the individual results for the two trials. The results were presented in this way to highlight the consistently better point estimate for all test responses in BLISS 52. We consider this would be masked if a MA pooled estimate is also included in the figures.

Page 5. Could you elaborate in what way the outcome measures are 'adjusted' as described on line 158 in the statistical analysis section? Is this to do with including stratification factors as mentioned on lines 233-234 in the second paragraph on page 7?

Reply: Lines 187-189 states the stratification factors used at randomisation (for BLISS trials) and Line 231-233 indicates that these factors were used to adjust the outcomes by logistic regression.

Pages 5 (line 155) & Page 46. Figure 5. I assume MD in column 3 represents standardised means differences ie SMD. Standardisation could also be added to line 155 in the first line of the statistical analysis section on page 5 which mentions mean differences.

Reply: Please note that MD refers means differences not the SMD (standardised means differences). Please note that Figure 5 x-axis is labelled MD not SMD.

Reference

Hunter JE and Schmidt FL (2004) Methods of meta-analyses. Correcting error and bias in research findings. Sage:Thousand Oaks.