Project Name: Use of Bayesian Techniques in Randomized Clinical Trials: A CMS Case Study
Project ID: STAB0508

Disposition of Comments

Table 1: Invited Peer Reviewer Comments

Reviewer[1]	Section[2]	Reviewer Comments	Author Response[3]
Peer1	General	This is an excellent and thoughtful assessment of the value of Bayesian methods to CMS policymaking. The report is generally readable and often eloquent. The issues raised are important ones; particularly as both public and private-sector policymakers frequently face critical tasks of synthesizing disparate sources of information into coherent policy choices. The example of the implantable cardioverter-defibrillator is entirely apt and is likely to be repeated. This report provides both an excellent historical summary of the ICD coverage "story," as well as a highly illustrative example of how data from the numerous ICD clinical trials could have been used in a Bayesian approach to better inform the CMS coverage decisions of 2003 and 2005.	We thank the reviewer for their comment.

Peer1	General	In terms of potential general improvements to the report, I first of all wonder if the authors would consider reducing the number and complexity of their tables and figures. While they have done well to strive for completeness, there is a tremendous volume of information in the tables/figures to digest. Perhaps some of the more detailed tables (e.g. Table 7) could be reserved for the Appendix, and some of this tabular data could be presented graphically? Perhaps some of the graphs (e.g., Fig. 11) could be simplified to line graphs? I am also left wondering about the nuts-and-bolts aspect of how CMS, FDA, and other agencies would actually use Bayesian methods to achieve the suggested aims proposed by the authors. As mentioned in the text, some of the barriers to implementation of Bayesian analyses involve access to all relevant sources of data, expertise in the relevant statistical methods and software, consensus regarding prior distributions, and consensus regarding interpretation of posteriors. Assuming CMS agreed to implement Bayesian analyses for future "high stakes" coverage decisions, how (politics aside) might such a process work?	We agree with the reviewer that the information in Chapter 5 is more complex than the rest of the report. We felt that this level of complexity was needed to accurately portray the use of Bayesian techniques in the CMS context. Based on previous feedback we have moved much of the details from the case study to the Appendix and a statistical manuscript. In the current Chapter 5 we attempt to ease the burden on the reader by including "key points" and clinical questions/answers. Following reviewers' suggestions however, we have now further revised/simplified our figures (for example, the Kaplan-Meier curves are now lines, the orientation of the figures with estimates and CIs of hazard ratios are now all vertical). We also agree with the reviewer's concerns about the nuts-and-bolts aspect of how CMS will use Bayesian methods to achieve the suggested aims. The purpose of the report was to provide an overview of the Bayesian approach and its application to the CMS policymaking context. We will share the reviewer's comments with CMS and welcome further discussions with CMS and stakeholders about next steps in possible implementation of Bayesian approaches in the coverage process.
Peer1	General	The authors are to be congratulated for a superb report.	We thank the reviewer for their comment.

Peer1	Executive Summary	The E.S. Results (2 pages) have too much detail compared to the E.S. Methods (5 sentences), thus the Results appear somewhat out of context. In particular, the detailed ICD simulation results (p.2-3) need to either be accompanied with more details in the E.S. Methods, or they should be greatly abbreviated here. Some of the Executive Summary's conclusions appear a bit too strong. For example, the authors seem to be saying that subgroup analyses are always compromised by small sample sizes and the tendency toward excessive post-hoc subgroup testing (p.4). This is not universally true, although these are frequent problems with subgroup analyses. Some of the authors' recommendations are likewise a bit vague. For example, how can an investigator tell if a subgroup effect is "likely to be strong" (p.4)? How can a policymaker tell when trial-based data are "sufficient" (p.5)? There is also Bayesian jargon used (e.g. "the assumed priors," p.5) that isn't defined for the reader until p.6.	Being sensitive to our policymaking audience, we wanted the executive summary not to burden the reader with the technical methods described in Chapter 5 and the Appendix concerning the case study and simulations. We therefore purposely provide a brief description of the methods focusing instead on the findings. As noted above, our goal in the executive summary was to transmit general principles, rather than discuss exceptions and details of implementation. We acknowledge that there is additional future research needed by CMS and others in terms of implementing Bayesian approaches into the CMS policymaking process.
Peer1	Chap 1	This is an extremely well-written, accessible introduction to Bayesian analysis, particularly for a clinical audience. I will consider using this chapter (with permission) in teaching these methods to our Masters students.	We thank the reviewer for their comment. Use of the chapter for teaching is encouraged once a final version of the report has been published.
Peer1	Chap 2	This section begins well but seems to meander a bit at the end. The last paragraph seems to state a number of obvious and general truths that aren't particularly relevant to why use of Bayesian analysis would be good for CMS.	We felt that this last paragraph discusses important topics of stakeholder engagement and transparency and have therefore left it unchanged.
Peer1	Chap 3	The clarity of this chapter could be improved by standardizing the organization and style of the 4 sub-chapters on literature themes. The first, "Advantages and Disadvantages …," reads smoothly as a narrative literature review. The second, "Use of Bayesian Techniques …," seems to digress on page 41 into an overly detailed description of the Bayesian vs. Frequentist debate. The style also departs from the objectivity of the first sub-chapter to a more editorializing perspective (e.g., the statements in favor of using skeptical priors on pp. 47-48). The third sub-chapter barely references the literature at all (2 footnotes), and reads more like a tutorial, although it is well written. The final sub-chapter returns to the style of an objective narrative literature review. I'd favor the use of a single style throughout, preferably the style used in sub-chapters 1 and 4. Numbered subheadings might further help the reader navigate.	We recognize that the organization and style of the 4 subsections in Chapter 3 are not uniform, but we judged that the differences were appropriate to the material being reviewed in each subsection and the intended audience.

Peer1	Chap 4	This is a clear and concise description of the ICD coverage "story." It may be worthwhile adding on page 68 that Medicare's coverage of Cardiac Resynchronization Therapy defibrillators (CRT-D) in 2005 reduced the number of NYHA Class IV CHF patients who would not be covered for defibrillator implantation by CMS.	We now include this additional detail in the ICD "story" on page 39.
Peer1	Chap 5 - Appendix	Chapter 5/Appendix. While this is a great example and it is nicely presented in both Chapter 5 and the Appendix, I'm a little concerned with the overall strategy of combining primary prevention (e.g. MADIT-II, SCD-HeFT) and secondary prevention (e.g. CASH, AVID) ICD trials. On clinical grounds, one could argue that these are entirely distinct patient populations and thus the findings from one group of trials would not be expected to inform the findings in the other group. From a policy perspective, CMS was not necessarily wrong to focus their attention on the primary prevention trial data only (were they?) when deliberating on the coverage expansions of 2003 and 2005. I understand that Bayesian methods relax the assumptions of homogeneous effects across trials, but if these are 2 fundamentally different patient populations, it's not clear how even Bayesian methods would permit analytic aggregation (i.e., wouldn't all the priors derived from the secondary prevention trials be non-informative for primary prevention trials)? I certainly could be missing a key point (maybe Bayesian analysis really enables the grouping of proverbial apples and oranges), but if so I would like to see a clearer description of the authors' non-intuitive decision to aggregate these trial data in this manner. Would statistical tests for homogeneous effects—thee bread-and-butter of classical meta-analysis—be inappropriate in a Bayesian context? If so, why?	When planning our simulations and case study we discussed within our team and with our technical expert panel the advantages and disadvantages of combining the secondary and primary prevention data. It was felt that the analysis of the complete data set with the four prognostic characteristics would provide substantial differentiation of the patient population and exploration of the similarities and differences among the trials. We agree, however, that exploring the patient-level data taking into account the information about the prevention type is interesting. We therefore now include discussion of these findings (supporting our approach of combining the data from all trials) on page 48 (with reference to a more extended treatment in the Appendix).

Peer1	Chap 6	Similar to my comments on the Executive Summary, I think some of these findings need to be tempered a bit. While I agree with the general principles stated here, occasionally a subgroup effect should be acted upon by policymakers if the welfare issues to the affected sub-population are profound. For example, if a prescription drug was shown in a randomized clinical trial to have 50 times the expected mortality rate in a particular subgroup with an interaction effect p-value < 0.001, it may be highly appropriate for the manufacturer and FDA to restrict the use of the drug in that subgroup, rather than deem the evidence "exploratory." I have the same comments here as above in the Executive Summary regarding the use of the words "strong" and "sufficient" in findings 3 and 4. These are highly subjective assessments, and I think it would be more helpful for policymakers if the authors quantified what these terms mean operationally. For example, should CMS never accept a subgroup analysis that has a non-informative prior (i.e., not "strong")? Does "sufficient" mean that the null hypothesis is excluded with some probability (i.e., outside the posterior 95% credible interval)?	As noted above, our goal in the executive summary and Chapter 6 was to transmit general principles, rather than discuss exceptions and details of implementation. We acknowledge that there is additional future research needed by CMS and others in terms of implementing Bayesian approaches into the CMS policymaking process.
Peer1	Pg 59	"A costly intervention to the Medicare community" – do you mean costly to payers?	We have modified the bullet to clarify that the ICD is potentially a costly intervention to the Medicare program (page 34).
Peer1	Pg 108	"Ventricular tachycardi" is missing an "a"	The correction has been made (page 64).
Peer1	Fig 1 & 3-10	Figures 1 and 3-10 appear to have been done in low-resolution graphics. It would be better for presentation purposes if these were upgraded to the quality of Figures 2 and 11-14.	We have edited Figures 1-10 to improve resolution/clarity.
Peer1	Tables 4-7	Tables 4-7 and Appendix Tables A1-A4 appear to be identical, as are Appendix Tables A23, A25, A26 and A27 duplicates of Tables 8-11. Some of the Figures are duplicated, too. Is this duplication necessary (is the Appendix designed to stand alone)? Also, Tables 4 and A1 both would be more legible with fewer significant digits.	The Appendix is designed to stand alone and so we have left the duplicative figures/tables.
Peer1	Minor Comment	The 2nd and 3rd sentences of the Key Point on page A-166 seems to be overly simplistic and outside the scope of this paper. I'm sure the authors would agree that there are lots of complicated reasons, beyond differences in patient prognosis, why randomized trials often have dissimilar outcomes to observational registries. Probably better to omit these sentences.	We have omitted these sentences from the final report (pages 46 and A-116).

Peer1	NOTE	I'm a little surprised to see analyses of ICD/NCDR data included in this report (p. 81, p. A-165, Table 10, Appendix Table A26). I'm a member of the Research and Publications Subcommittee for the American College of Cardiology/National Cardiovascular Data Registry (ACC/NCDR), which oversees use and publication of analyses of ICD/NCDR data. I may be misinformed, but I do not recall any of the authors on this report being principal investigator of an approved project to using ACC/NCDR data. I do recall that Dr. Sana Al-Khatib of DCRI is an approved PI, but she is not listed as a co-author here, and I don't think that her NCDR-approved project was for the purpose of this technology assessment. In any event, publication of analyses using ICD/NCDR data must be approved by the ACC/NCDR prior to submission for publication. Have the appropriate data use and publication permissions been obtained?	We are sorry for the confusion. The data cited in the report was obtained directly from CMS and reflects the Medicare patients within the ICD ACC/NCDR registry rather than the larger ACC/NCDR ICD registry. We now clarify this in the text on page 46 and throughout the document when we reference the registry participants. Note that because the ACC/NCDR registry does not currently contain long-term outcomes, we did not use the registry data for our case study but instead used the MUSTT registry participants.
Peer2	General	Overall I found this to be an extremely well-written report presented in a very scholarly and well-balanced fashion. The authors have skillfully avoided entering into the realm of polemics and have rather elegantly demonstrated both the benefits and disadvantages of a Bayesian analysis compared to the more frequently employed Frequentist paradigm.	We thank the reviewer for their comment.
Peer2	General	I have no major concerns about this report although I did find the example in Chapter 5 somewhat difficult to follow, related primarily to formatting issues. It is [difficult to] smoothly and seamlessly read this chapter when there are constant references to figures, tables or appendix tables without any accompanying page numbers. This makes for very disjointed reading and impinged on the clarity of the example, especially compared to the earlier chapters. Also the apparent desire to keep the example as simple as possible by moving much of the methods to the appendix had the paradoxical effect of rendering the example more difficult to follow. I would personally favor integrating the current appendix directly into Chapter 5.	We thank the reviewer for their comment. In previous drafts of the report, we included the Appendix material in the main text. However, readers of these previous versions found that the additional detail was distracting and so we've moved it to the Appendix.
Peer2	Pg 17, ln 2	The word "**better**" might be more appropriately replaced with the word "more".	The suggested change has been made (page 10).
Peer2	Pg 18, ln 12	I would add "…shrunk toward **the mean of the posterior distribution**, **in this case** toward the null value of 0.	The suggested insertion has been made (page 10).
Peer2	Pg 19, ln 6	Sequential meta-analysis is also routinely referred to as a cumulative meta-analysis and this could be mentioned.	The suggested change has been made (page 11).

Peer2	Pg 19, ln 9	Strong beliefs – this may be a limitation to Bayesian analysis but of course this depends on the origins of these strong beliefs. If they come from well-randomized clinical trials then this would not be considered a worse case scenario.	We now clarify on page 11 that we assume that strong beliefs are primarily based on intuition, rather than objective information such as a previous meta-analysis, and that data-based priors are superior to opinion-based priors.
Peer2	Pg 20, ln 22	…and similar examples of statistical esoterica – this has a rather pejorative connation and a better word may be desirable.	We have removed this phrase as suggested (page 12).
Peer2	Pg 21, ln 9	… given the observed data **and the prior information.**	The suggested change has been made (page 13).
Peer2	Pg 82, ln 1-5	Not sure of the origin of these points.	On pages 46 and A-116, we now differentiate between what we observed and what we concluded based on our analyses.
Peer2	Pg 84, ln 1-5	It is unclear what the two prior beliefs represent.	The two priors represent beliefs of no treatment effect. Both priors are centered around no treatment effect. We describe prior 2 as being more informative in the sense that it places heavier mass around no treatment effect. We now clarify this in the text (page 47).
Peer2	Pg 85, ln 14	The authors have taken a hazard ratio of 0.8 to indicate clinical significance. The justification for choosing one arbitrary dichotomous cut-point for clinical significance merits perhaps more reflection and discussion. Indeed, one of the disadvantages of the frequentist paradigm is the dichotomous $p < 0.05$, and one of the advantages of a Bayesian approach is the possibility, without a type one error penalty, to look at several different cut-points. Perhaps at least for the overall results the results with various cut-points of HR 0.9, 0.8 and 0.7 might be illuminating.	We based the hazard ratio threshold for clinical significance of 0.8 on feedback from our technical expert panel. We now clarify this and provide for the reader results using other cut-points as well, namely, 0.7, 0.8 and 0.9 (pages 45 and 48).
Peer2	Table 6	A table of p values is bound to ignite the ire of some (not only Bayesians) and is of limited decision making utility. Perhaps at least for a small subset the effect size & 95 CI should be reported.	We now include for each cell in Table 6 the hazard ratio and 95% confidence interval (unadjusted for multiple testing) comparing survival by treatment in the subgroups of interest. Missing entries indicate unavailable data for the particular subgroup. Entries highlighted in red indicate significant results at the unadjusted significance level of 5%.
Peer2	Table 7	I would like to see the cumulated number of patients and events for each of the 48 subgroups. This could also be presented in Table 9.	We agree that this information is useful to the reader and have modified Table 7 to include the overall number of patients and events.

Peer3		I applaud the authors and CMS for addressing an interesting and timely topic with an in-depth report that attempts to tailor itself to a non-technical audience. The report has several overall strengths including the tutorial on Bayesian methods, literature review, application to an area of interest to CMS, and provision of a detailed statistical appendix. As one might expect, the strengths are also potential weaknesses given the length of the report and the tendency to gloss over what might be important details. Nevertheless, the overall result is quite interesting and relevant. I expect that it will be a useful document for CMS and ultimately for the clinical trials and policy communities in larger context.	We thank the reviewer for their comment.
Peer3	Page 2	Simulation studies are discussed in part by saying "the often have low power to detect differential treatment effects". There is a general lack of rigor in the document when discussing this concept. The essential point is that studies designed to detect main effects (as almost all are) will have little power to detect treatment-covariate interaction. Sometimes the report refers to this concept explicitly and in other places it is termed "differential treatment effect". I think it would be helpful to refer it always as treatment-covariate interaction. Furthermore, it should be generally acknowledged and articulated that main effect designs will necessarily leave minimal power for detection of these (or other) interactions. As a very general rule, it takes roughly four times the sample size to detect interactions compared to main effects, assuming that we are discussing effects of approximately the same magnitude. Obviously, large magnitude effects can be detected more easily.	As suggested, we have globally replaced references to "differential treatment effects" with "treatment-covariate interaction."

Peer3	Page 3 and 4	In addition, it might be helpful distinguish between qualitative and quantitative treatment covariate interactions. As a general rule, we would be interested in interactions that are large in magnitude whether they are qualitative or quantitative. However, small quantitative interactions (those interactions that have the same direction but quantitative differences in magnitude) are generally of little or no interest. This is because there is no therapeutic implication. If treatment X is better in both subsets, it is the recommended treatment even if treatment X is slightly better for one subset than it is for the other. In contrast, qualitative interactions (interactions that show treatment helping one subset but harming another subset) always carry therapeutic importance. Treatment X is appropriate for one subset but treatment Y is appropriate for the other. These qualitative interactions may require more power to detect because the statistical test can less efficient. I think the report and the methodology in general might need to be more respectful of these concepts because of the therapeutic relevance and the fact that we may not need to fuss very much over many quantitative interactions.	We have now added a discussion of heterogeneity and concepts of clinical and statistical significance to the Tutorial (page12).
Peer3	Page 3	Another concept introduced on Page 3 that is of potential concern is the use of patient-level versus aggregate data. It is stated that “the analysis of aggregate data may be more sensitive to priors”. I can image that this may be the case, however there may be additional issues with the analysis of aggregate data. In particular, the analysis of aggregate data can represent a type of “ecological fallacy” if the underlying means are subject to confounding. The analysis whether frequentist or Bayesian would then represent a kind of incomplete analysis of covariance, and can be biased or even yield the wrong algebraic sign. In any case, there may be serious pitfalls with aggregate data.	We include a discussion of confounding and its potential impact to findings in Chapter 6, item 7 (page 51).

Peer3		The tutorial on Bayesian methodology will likely be appreciated by many consumers of this report. I don't think it is necessary to draw differences between frequentist and Bayesian methods that emphasize potential friction or controversies. Nevertheless, I think it would be helpful for non-statistical audiences to appreciate some of the fundamental inferential differences that may be consequential for the way that they think about interpretation of data and policy decisions. A key in my opinion is the difference between the frequentist and Bayesian perspective on the parameters of the underlying data model. To the frequentist, parameters are fixed constants of nature. In this case, sample variation must be expressed in the familiar but somewhat awkward framework of hypothetical repetitions of an identical experiment. This is the awkwardness that causes us to misinterpret confidence intervals and p-values. To the Bayesian, model parameters are random variables, and therefore sampling variation is manifest as probability distributions for those parameters. This view also requires the Bayesian to specify probability distributions for parameters. Hence, the notion of "credible intervals". Although these are superficially analogous to confidence intervals and p-values, the difference is quite fundamental and important to understand. It is useful not to place value judgments on these differences but helpful to articulate clearly the different inferential frameworks that are required for Bayesian versus frequentist analysis, and in particular to explain the different perspectives on model parameters.	We thank the reviewer for their comment and agree with their view that the Tutorial was vague about some of the more technical points. As the reviewer mentions, the Tutorial's goal was to be appreciated by the consumers of this report (policymakers and regulators) and therefore the vagueness of the Tutorial regarding certain technical points was intentional.

Peer3		Another fundamental issue that I believed the report in general and the tutorial specifically does not deal with effectively is the prior distribution. The report does a fine job of illustrating generic types of prior distributions including uninformative, skeptical, and optimistic. No definition is offered for the later concept of "genuine priors", but this may not be essential. My concern deals with the construction of prior distributions in the absence of firm statistical evidence. (I don't think there would be wide-spread disagreement among statistical experts if prior distributions are constructed from actual data). An essential question is whether or not subjective belief in the absence of actual data is appropriately represented by a probability distribution. This issue is close to the heart of the general criticisms about Bayesian methodology. I don't know that the report deals with it directly enough. Absent data, is a probability distribution the appropriate way to represent our ignorance regarding a model parameter?	As indicated we do describe "genuine" priors on page 33 of the literature review. In the Tutorial we attempted to keep our discussion of priors and Bayesian approaches as simple in possible to allow the Tutorial to be understood by the target audience. We do emphasize now the role of objective and subjective priors on page 8 of the Tutorial.
Peer3		I also found myself wondering a little bit about the precision with which non-informative priors are described. The confidence interval approach to the discussion of prior distributions is a useful one. However, I need some additional convincing that a wide confidence interval could be taken literally as a "non-informative prior". I imagine the confidence interval as being approximately a normal distribution whereas a non-informative prior should be an improper uniform distribution. I don't know if this point is essential but I would want to be sure that our non-statistical colleagues are not misled.	We have removed the use of the confidence interval in the discussion of non-informative priors and now include a brief note about such priors having an interval over the entire real line (page 10).
Peer3		It may also be helpful to point out that apart from the specific differences that I have mentioned above; every frequentist procedure has an essentially equivalent Bayesian one. To me, this means that a frequentist based conclusion can be shown to be essentially the same as a Bayesian conclusion where the Bayesian statistician uses a prior distribution that the frequentist was seemingly unaware of. In a crude sense, the frequentist might be seen as using a prior distribution without recognizing it.	We thank the reviewer for their comment.

Peer3		The literature review seems quite useful, although I personally was not too impressed with the formal aspects of the Medline search. The reference list is most helpful. The specific problems of subgroup analysis seem quite appropriate illustrating the differences between Bayesian and frequentist approaches. The general strengths of Bayesian methodology in this setting make it a useful example. The discussion regarding fixed effect models, random effects models, and random effects models with information from outside the study is a helpful framework. I was not convinced by the discussion regarding biological heterogeneity. The presence or absence of such heterogeneity is an unknown characteristic of nature. In a very real sense, the approach to heterogeneity is based on a pure assumption and not surprisingly it appears that our ability to deal with the consequences of that assumption is in part affected by the methodology. The reason this issue is so problematic is because the assumption of heterogeneity is derived only partly from what we think we know about biology and derived much too influentially by fashions of the day, including politics. For example, we might be looking for heterogeneity on the basis of non-biological constructs masquerading as biological ones. Race for example is at best a surrogate and may not be a biological construct whatsoever. In many cases, sex is also not a relevant biological factor. These issues aside, I would want to be sure that the methodology chosen does not yield a forgone conclusion derived from pure assumption.	Although we acknowledge the reviewer's comments, we note that we are not arguing that biological heterogeneity is always present, but rather that it is often part of the philosophical rationale behind the Bayesian approach. We have also expanded our discussion of heterogeneity in the Tutorial (page12).
Peer3		The discussion on the resistance to the use of Bayesian methods was most helpful. I was unfamiliar with many of the concepts there. I would agree wholeheartedly with the list of items mentioned on Page 56.	We thank the reviewer for their comment.

Peer3		The extensive discussion of the implantable cardioverter defibrillator is clearly an appropriate illustration of Bayesian methodology. However, I found myself struggling to extract the most relevant lesions from the example. Although this section of the report reflects a scholarly approach that might be useful for peer-reviewed journal, I wonder if for this audience, the report should be restructure in a way to present the implications and lessons is a more digestible format. I would like to see the data displays be more friendly and informative. Most people will not try to digest the tables, especially the ones (inappropriately) full of p-values. Some of the figures have odd anomalies including being essentially obscured by censoring indicators in the case of survival curves, and the curious switching between horizontal and vertical formats for confidence interval plots. I am fearful that much more time will be required to make the technical details of this report less voluminous and more accessible to the intended audience.	We acknowledge that the case study is at a different level of detail than the other chapters, but thought that this was needed in order to convey accurately the use of Bayesian approaches. We tried, however, to simplify the chapter for readers through the inclusion of "key points" and specific clinical questions and answers. In addition, following reviewers' suggestions, we have revised several tables and figures. Table 6 now has point and interval estimates for the hazard ratios for the subgroups of interest. Kaplan-Meier figures now omit the censoring indicators. Moreover, figures with estimates of hazard ratios in the combined analysis now all use the same orientation.
Peer3	Pages 3 and 4	Another general issue with the report is illustrated by some of the bullets and conclusions here. There is the concept that observed interactions may not be the same across all trials. This implies that we may need to account for random effects in interaction estimates or three-way interactions (the third factor being an effect of the trial). It would seldom be worthwhile to design for this effect. I wonder if the general concept of designing studies to detect the interactions of interest as opposed to simply designing analysis should be discussed at a deeper level. The design question touches not only on our ability to detect interactions but the potential bias is in registry data that are discussed briefly on Page 3. It is well known that patients in registries may carry different prognosis from those studied in clinical trials. What may under-appreciated is that treatment effect estimates that derive from registries are typically biased (perhaps confounded by indication), whereas relative treatment effect estimates from randomized trials are probably more likely to generalize across patient subsets.	We acknowledge the reviewer's comment concerning the use of Bayesian approaches in the design of clinical trials. We now clarify (on page 1) that although we address (and acknowledge the importance of) the use of Bayesian approaches both for clinical trial design and analysis, we focus our report on their use for clinical trial analysis, as this was most applicable to the CMS policymaking context.

[1] Peer reviewers are not listed in alphabetical order.

[2] Page and line numbers refer to the draft report.

[3] Page and line numbers refer to the final report.