Author's response to reviews

Title: The cost determinants of routine infant immunization services: a meta-regression analysis of six country studies

Authors:

Nicolas Menzies (nick.menzies@gmail.com)
Christian Suharlim (csuharlim@mail.harvard.edu)
Fangli Geng (fag686@mail.harvard.edu)
Zachary Ward (zward@hsph.harvard.edu)
Logan Brenzel (logan.brenzel@gatesfoundation.org)
Stephen Resch (sresch@hsph.harvard.edu)

Version: 1 Date: 31 Jul 2017

Author’s response to reviews:

Itemized response to reviewer comments for: “The cost determinants of routine infant immunization services: a meta-regression analysis of six country studies” Manuscript ID BMED-D-17-00486

REVIEWER #1

1. Thanks for the opportunity to review this manuscript. The authors use an econometric approach that I am not familiar with, and so my comments are related to better understanding what the authors did and how to interpret the tables, etc. It would be useful for the authors to provide more guidance in the paper for this, if possible.

Thank you for these comments. We have made edits to the paper to more clearly explain the estimation approach, described below under individual items below.

2. Abstract: Suggest reviewing and revising the conclusion. This paper uses a sophisticated statistical method to estimate cost functions. Yet the conclusion to the abstract isn’t really highlighting the paper's results. Rather the conclusion that cost estimates are higher than historical values, could be deduced from earlier papers from the EPIC study. Furthermore, the analysis of determinants does not explicitly include a price variable on the RHS, or an explicit measure of an expanded vaccine schedule—so the conclusion isn't directly tied to the results (albeit there is number of doses on RHS—so yes, more doses delivered). In addition, language such as ‘average costs were tightly linked to service volume’ is vague. It would be great to focus
on how these methods and findings can advance our understanding of immunization costs and how this information can be used to guide program and policy in the future.

We have revised the abstract conclusion (p3) to place greater emphasis on the regression results. The conclusion now reads:

“Conclusions: We identified multiple features of sites and their operating environment that were associated with differences in average unit costs, with service volume being the most influential. These findings can inform efforts to improve the efficiency of service delivery and better understand resource needs.”

3. Page 5 and Table S3—methods indicates data collected on above site costs, but then Table S3 indicates that these higher level program support costs are omitted from the Total Cost estimate. Authors should explain why these are omitted under section 2.3.

For the average cost estimates shown in Table 2 (p14) we present results both with and without higher-level program support costs. For the regression analyses we excluded higher-level program support costs as these costs are unlikely be explained by the site-level determinants considered as explanatory variables. We have added text in the relevant sections to clarify these points:

Section 2.5 (p7, paragraph 1): “Average cost estimates are reported with and without higher-level program support (‘above site-level costs’) included.”

Section 2.6 (p7, paragraph 2): “We excluded higher-level program support (‘above site-level costs’) from this part of the analysis, as these costs are unlikely be explained by the site-level determinants considered as explanatory variables.”

4. Authors should reference again the other EPIC papers where these data are better explained (references 8-11).

We have referenced these studies again in Section 2.2 (p5, paragraph 3), noting that the results we report will differ slightly from these earlier publications due to data cleaning and standardization undertaken to prepare the data for analysis:

“Data cleaning and standardization resulted in minor differences with earlier studies country-level results [8-11].”

5. Authors should explicitly indicate which vaccines and target populations are included in the total cost estimates in the body of the paper. Is it all the vaccines listed in Table S1 or a subset of the vaccines for children 0-12 months? Is TT for pregnant women included? Do the total costs reflect different vaccine schedules across countries? This information is critical for sections 2.4, 2.5 and 2.6.
The analysis considered all immunization services provided to children 0-12 months of age. This approach was taken to improve the standardization across the different country-level studies. We have edited the beginning of section 2.3 to clarify these points (p6, paragraph 1):

“Total site-level economic costs were estimated from a program perspective using conventional methods, including all activities undertaken to provide routine immunization services to children aged 0-12 months. We focused on the -12 year old age group to allow greater standardization across the data collected through individual country-level studies.”

We provide the full vaccination schedule in Table S1 (appendix, page 2) as we believe this provides the reader greater context. We have added a footnote to the table noting that “The analysis considered the subset of these schedules delivered to children 0-12 months of age.”

6. If different vaccine schedules influence total costs across countries, is it necessary to include an indicator of the vaccine schedule to capture economies of scope resulting from incorporation of vaccines for children older than 0 to 12 months, and vaccines for adults, etc?

We agree with the reviewer that this form of economies of scope would be useful to investigate, but were unable to do so in this sample. As the analysis only included 6 countries, we have limited power to investigate determinants at the country-level rather than site-level. We included one country-level fixed effect in the regression equations (per capita GDP), and this will be acting as a proxy for many factors that vary with income level, including prices and vaccine schedule. We have made this explicit in Section 2.6 (p8, paragraph 2):

“We also included log per capita GDP in 2011(log(GDP)) as a crude index of inter-country price differences and other factors that vary with income level.”

7. In the discussion, the authors note that the expanded vaccine schedules contribute to inter-country differences—not clear to me, why there is no RHS variable to account for this?

We agree with the reviewer that this would be an interesting variable to investigate, however as we only has data for six countries we did not have the statistical power to investigate factors that varied at the country-level. We have edited the discussion to describe this point (p21, paragraph 2):

“Per capita GDP was found to have a positive relationship with costs per dose. This likely reflects differences in price levels between countries, but could also be related to the many other factors that vary with country income level. As we only had 6 countries represented in this sample we were unable to decompose the effect of these country-level factors.”

8. Section 2.6 Regression analysis—overall I think the authors have to do a better job explaining the implications of using a Bayesian approach in this paper, as it affects how the tables are presented and interpreted compared to standard parametric statistics and Bayesian approaches.
We have revised Section 2.6 to provide more guidance on the implications of the estimation approach (p7, paragraph 2):

“We used a Bayesian hierarchical regression model to combine data from different country studies and account for the multi-stage sample design, with country- and province-level random effects. Using a hierarchical regression model allowed for sources of variation at site-, province-, and country-level, and provided a framework for synthesizing data across countries. As we used a Bayesian approach, the uncertainty measures included in the results section (such as intervals provided around point estimates) represent posterior probabilities conditional on priors, likelihood and regression model, unlike traditional confidence intervals.”

9. Suggest that authors clearly define the dependent variable as total cost on page 7, line 28, before describing explanatory variables (this should have been described fully in section 2.3—taking note of point 3 above). Just to make sure that reader is aware reference to cost means 'Total cost' to distinguish from presentation of unit costs.

The dependent variable is introduced at the beginning of this section (2.6). We have edited this text to make this clearer (p7, paragraph 2):

“We explored site-level cost variation by regressing the log of total costs incurred at each site against several explanatory variables.”

10. Did author check correlation coefficients for RHS variables? Seems likely that some are highly correlated with one another. No mention of multicollinearity of RHS variables in discussion.

RHS variables were correlated to a certain extent, but multicollinearity was not a problem. If it was it would have been evident through implausibly large effect sizes and associated standard errors.

11. Par 1, page 10—since I'm not familiar wit the statistical approach of WAIC, it was very difficult for me to evaluate the results.

We have revised section 2.6 to describe the information provided by WAIC (p7, paragraph 1):

“Similar to AIC, WAIC is a statistic that measures the extent to which a model is able to explain variation in the data, while penalizing unnecessary model complexity. A lower WAIC implies a better fitting model.”

12. I would say in this section to first provide the 5 models and provide the reader with a bit more rationale to why these 5 models.

We have moved the description of the regression models from the results to Section 2.6 (p8, paragraph 3):
“We fit a series of progressively more inclusive regression models including these explanatory variables. Model 1: Intercept plus country- and province-level random effects. Model 2: Model 1 plus log(doses) and log(doses) squared. Model 3: Model 2 plus log(GDP) and site characteristics. Model 4: Model 3 plus features of the operating environment. Model 5: Model 4 plus country-level random effects for log(doses).”

13. Then also explain better using first differences using results from 2 of the 5 model specifications.

The first differences were calculated using only one of the regression models (model 5). We have edited Section 3.2 to describe the reasoning for this (p16, paragraph 2):

“…and first differences were calculated to demonstrate the implications of the regression results (Table 4), based on the best fitting regression model by WAIC (model 5).”

14. I'm not familiar with the output elasticity definition. Page 10, line 17 "We calculated output elasticity as the percentage increase in service volume for a 1% increase in costs, "- shouldn't the output elasticity be a 1% change in output results in an x percent increase in total costs? Maybe this has to do with the Bayesian specification.

The definition of output elasticity we use is conventional for the economics literature (for example https://www.jstor.org/stable/25602782) and unrelated to the Bayesian estimation approach. We have edited the relevant text to clarify this definition (p11, paragraph 1):

"We calculated output elasticity as the percentage increase in outputs (service volume) for a 1% increase in inputs (total costs)…"

15. I have not see sensitivity analysis referred to in econometrics—typically this is used in economic evaluation literature, but this doesn't seem to fit here—the authors aren't testing uncertainty in a parameter estimate, rather isn't the sensitivity analysis just alternative regression specifications?

Thank you for this suggestion, we have changed the title of this section to “ALTERNATIVE REGRESSION SPECIFICATIONS” (p19).

16. Page 10, line 46—should that read, "We anticipated that Total costs of service delivery would be increasing at higher coverage levels." That is total costs, not unit costs?

The current text is correct. We have edited this section to clarify why we had hypothesized increasing unit costs (essentially, the very high costs of reaching the last child) (p12, paragraph 1):

“We anticipated that unit costs of service delivery would be increasing at higher coverage levels, due to high marginal costs of reaching the very last members of the target population. However…”
17. Table 3—what are the figures in parenthesis? Are they standard errors?

This is correct. We have added this to the footnote of the table (p15).

18. The foot note to table 3 would be better included in the methods section—need to explain how to evaluate goodness of fit, and why there are no indicators of significance for Bayesian regression estimates.

We have added text to the methods section to better describe the goodness-of-fit statistic (Item 11 above). Measures of significance (probability statements about the distribution of results) are possible with Bayesian methods. However, we have not included ‘p-value’ type measures in the table as we believe these can lead to an inappropriate binary interpretation of results (ie effect is present vs. effect is not present).

19. Page 17—Par 1, lines 12-24—this needs to be explained much more clearly and walk the reader through what the output elasticity means (give example for one country—i.e. a 1% increase in x % results in an ? % in Y…)

We have added text to Section 3.2 to help readers understand the implications of these elasticity results (p18, paragraph 1):

“This implies that, across the entire sample, a site with 10% higher costs would have on average 14% (12, 17) higher service volume, controlling for other effects.”

20. Also, help reader get from output elasticity to percentage increases in unit costs shown on row 18/19. Not clear to me how this was calculated.

These results were obtained from the fitted regression equations, in the same way that the first differences were calculated. We have added text to the end of Section 2.6 in the Methods to describe the approach (p11, paragraph 1):

“Using the fitted regression equations we compared predicted unit costs (total costs divided by total doses) across the range of service delivery volume observed in each country, controlling for other factors.”

21. Section 3.3, clearly indicate that looking at these predictors on total costs, not unit costs.

We have edited Section 3.3 to clarify the dependent variable in these regressions (p18, paragraph 2):

“We estimated regression models for total site-level costs using the subset of sites with data available on these predictors, and in each case identified a statistically discernable positive relationship with costs (Models 6-10, Table S4).”

However, what is not currently clear in the text is that since the regression models controlled for service volume, these coefficient estimates (p16, paragraph 2) and first differences reported in
section 3.2) apply equally to total costs and average costs. Text has been added to section 3.2 clarifying this:

“As the regression model included service volume, the first differences calculated for other predictors apply to both the cost per dose and total site-level costs, for sites of equal service volume.”

22. Can authors use clear language to explain the 'dedication index'. i.e. say an increase in the percent of time staff worked on immunization was associated with ….

We have edited Section 2.6 to provided a clearer description of the dedication index (p8, paragraph 4):

“We also created an index for the average fraction of time staff spent working on immunization (Dedication index, not available for Uganda) to describe the extent to which staff were committed exclusively to immunization activities, as compared to being spread across multiple service areas. This was used as a measure of economies of scope for labor.”

We have also added a parenthetical to Section 3.3 summarizing this information (p19, paragraph 1):

“A 25 percentage point increase in the dedication index (extent to which staff were dedicated exclusively to providing immunization services) was associated with”

23. Section 3.4 page 18, line 58/59—I wonder if including log catchment population in main regression is stronger than including hospital RHS variable—seems more intuitive that this would increase total costs.

Log catchment population and the hospital variable represent two separate constructs – catchment population size and facility size/complexity respectively. As the definition of a hospital can vary across countries we preferred the ‘no. inpatient beds’ variable as a measure of facility size/complexity, but this was not available for all countries and so we only considered this in the ‘additional predictors’ section.

24. Did authors consider using variables such as frequency of stock-outs or facility wastage rates for a key expensive vaccine for a quality measure?

We attempted to investigate these variables but were not able to calculate them for a sufficient number of sites to include them in the analysis.

25. Section 4. Discussion—Appreciate that the authors note the different estimates from previous EPIC country studies (reference 7). Very interesting that for three countries, estimates were very similar, while for Moldova, Uganda and Zambia, they were quite a bit different. Would be really useful to readers to unpack this more other than just attributing to data cleaning and refinement to cost and outcomes data. Was it measurement error? Or methodological differences? Or some combination?
We undertook a careful reexamination of the data with the original study teams to understand the source of these differences. As noted in the text, there were a number of minor differences that came from the data cleaning process. However, larger differences came for the different approaches taken to calculating unit costs – whether these were calculated as mean(TC/TQ), or alternately as mean(TC)/mean(TQ), where TC and TQ are vectors of site-level total costs and total doses respectively. Due to the close relationship between service volume and unit costs, and the wide range in service volume in each country, these different approaches yield very different cost estimates. For this reason we elected to present both approaches in Table 2, and remark on these in the text (Section 3.1, p13, paragraph 1):

“Cost per outcome estimates calculated as a simple average across sites were 15% to 96% higher than values weighted by service volume, implying that sites with higher service volume had lower costs per outcome.”

26. Page 19—Par 3 is the first time authors refer to 'statistically significant' relationships'. This terminology was noticeably missing in results section, and my understanding is that using Bayesian prevents assessing statistical significance. Suggest rewording to be consistent with results section and method.

Bayesian methods do not prevent probability statements being made. However, the phrase ‘statistically significant’ is closely tied to frequentist methods, and to avoid confusion we use the term ‘statistically discernable’. We have edited the text to make the phrasing consistent, and provided a definition in Section 2.6 (p11, paragraph 1):

“We present equal-tailed 95% credible intervals to describe the uncertainty in these results, and use the term ‘statistically discernable’ to describe situations where these intervals exclude no effect.”

27. I would appreciate a discussion on advantages/disadvantages of the choice of functional form/method—since using this Bayesian approach seems new for the health economics literature. If not, then perhaps make that clearer in methods section. Also, would appreciate a section on caveats or limitations that includes discussion of possible multicollinearity of RHS variables and omitted variables—that is things that are truly hard to measure that likely influence total costs and estimates.

We have edited the Methods section to provide greater detail on the Bayesian hierarchical estimation approach (described under item 9). We have also edited the discussion section to elaborate these limitations (p22, paragraph 3):

“Many of the findings are consistent with multiple explanations (including omitted variable bias) … Other limitations include the exclusion of predictors that could not be calculated for a sufficient number of sites to be considered in the analysis (for example, vaccine stock-outs and wastage rates), or that primarily varied at the country-level (for example, input prices and vaccine schedules), and for which we had minimal power to investigate. While we attempted to create explanatory variables with consistent interpretations, it is likely some of these constructs varied between countries, such as the definition of outreach services, or hospital status.”
28. Page 5, line 29/30—suggest referring to ‘above site costs’—this is a recent focus of global health costing, so great that these data include ‘above site costs’.

We have edited the text to make this link explicit (p7, paragraph 1):

“Average cost estimates are reported with and without higher-level program support (‘above site costs’) included.”

29. Table S3—Should the definition refer to children 0-12 months, rather than 0-12 years? A lot of vaccines in that schedule for adults (pregnant women) and older kids.

This is correct, thank you for catching this error (now corrected).

30. Page 7—operating environment is a convenient use of words, but I don’t see any benefit for lumping, geographic setting, ANC coverage and access to health care using one consolidated term. It isn’t intuitive, where the individual variables do make sense when referred to on their own.

In this section, the sentence structure distinguishes service volume, other site characteristics, and features of the operating environment. We believe it is useful to organize the explanatory variables in this way, but are open to a different phrasing. The operating environment does not represent a term in the regression equation (these are described in the parentheticals).

31. Page 11, make sure to refer to USD

We have edited the beginning of Section 3.1 (p12, paragraph 2) to spell out ‘US dollars’ before using the term ‘USD’ subsequently.

32. Page 18, what does ‘top-coded estimate’ of DTP3 coverage mean?

We have edited the text to explain this term where it first appears in Section 2.7 (p12, paragraph 1):

“We omitted DTP3 coverage from the main analysis, but undertook sensitivity analyses with (i) reported DTP3 coverage top-coded at 100% (ie revised to a value of 100% where the original value was >100%)…”

33. Page 18—authors refer to ‘statistically significant positive relationship’—did they want to use consistent terminology in previous section ‘statistically discernable positive relationship’?

Thank you for catching this – we have revised the text to use the term ‘statistically discernable’ throughout.

REVIEWER #2
34. The authors put considerable work into standardizing immunisation cost data from 6 countries in order to estimate cost functions. The presentation of methods and results is well written and clear, and the authors are careful to interpret the results within the limitations of the data. A few comments for the authors to consider.

Thank you for these kind comments.

35. The data are cross-sectional for each country, and the cost functions then likely represent short-run average costs curves (as far as I can reason it out). To some extent this depends on whether we consider each health facility a "plant" (i.e., production units) with its own short run average cost curve, or whether we consider each country a "plant" with their own short-run average cost curve. To the extent that the authors try to measure facility 'coverage' in sensitivity analysis, they are implying that each facility should be a plant. I am not 100% sure, but I would think that this implies that the random-effects should be applied at the facility level (rather than higher levels). Since the authors use higher level random effects, they are implying that countries/provinces are "plants", but the presentation of the results does not couch the interpretation of these models in ways that help me distinguish what may or may not be long vs. short run cost curves, and I am simply confused as to what is being measured by the various coefficients.

As the reviewer correctly notes, the study data are cross-sectional and observational. For this reason our ability to identify causal relationships (including the short- and long-run cost curves faced by individual sites) is weak. We have added text to the limitations section to make this explicit (p22, paragraph 3):

“Many of the findings are consistent with multiple explanations (including omitted variable bias), and should be viewed as hypothesis generating rather than confirming any specific causal relationship. For this reason, the total and average cost curves we estimate for each country (Figure 1) demonstrate how costs vary between sites of different size, but additional (and potentially unwarranted) assumptions are required to interpret these as the cost curves that would be experienced by sites attempting to increase service volume.”

The random effects are used to allow for variation at country- and province-level that is not captured by other predictors. If these random effect were not included the coefficients for the fixed effects could be biased, and standard errors would be artificially low. No random effects are included for the sites as any unexplained variation at the site level is included in the regression residual.

36. This concept of long vs. short run is not mentioned in the manuscript, nor the implications of measuring short-run vs. long-run costs curves. Typically for public health programs, the primary goal is to provide high coverage, whether or not the coverage level is optimally efficient from a unit cost point of view. Rather, the goal often likely is to have a given level of coverage (or target level of coverage) and from there determine what the most efficient method of production. It seems to that the long-run average cost curve would be most useful for these kinds of decisions. Simply recommending larger sites (since the data on coverage at site level is inconclusive) does not strike me as overall terribly useful.
We agree with the reviewer that long-run average cost curves would be useful, yet these results cannot be estimated using our data. Our results are purely descriptive, and we take care not to provide a recommendation of larger sites.

37. In any event, I think the paper would benefit from at least some theoretical discussion at the beginning of what it is exactly they are trying to estimate, and what information/decisions are informed by what they are trying to estimate. This may be largely an abridged rehash of intermediate microeconomics concepts, but I think it would be useful for most readers of a public health journal trying to understand the implications of this research.

We have edited the Background Section to elaborate how the results should be interpreted (p4, paragraph 2):

“We use these data to describe country- and site-level variation in routine immunization costs, and identify systematic costs differences related to site operating characteristics. Given the observational nature of the data, the relationships we estimate are purely descriptive. However, the larger sample size allows us to provide a fine-grained description of how costs vary between similar sites, which in turn can suggest potential approaches for improving the efficiency of service delivery and allow a more precise understanding of resource needs.”

38. The authors should check the language after they have done this. For example, Page 17, lines 7-8 (& page 19 line 41) may need to be "increasing returns to scale" rather than "economies of scale" (the latter being a long term concept). Or maybe not.

Thank you for pointing this out. Based on the discussion on this item and previous ones, it appears the text is not explicit enough about the descriptive nature of the analysis, and that the term “economies of scale’ might lead the reader to incorrectly infer this is a causal relationship. We have edited the text to remove ‘economies of scale’ are replace this with a more explicit description (for example, abstract; “In each country, higher service volume was strongly associated with lower average costs.”)

39. With respect to the conclusion that there are strong economies of scale (or what have you), and strong evidence for this, I am not completely sure that I agree with the authors. The sign of the variable "log(doses) sq" changes between models 4 and 5, which seems to imply switching whether the overall function has a negative vs. positive slope (at least across the output observed)?

The change in sign on the log(doses) sq affects the curvature of the cost curve. However the magnitude of this term (whether positive or negative) is sufficiently small relative to the main effect (log(doses)) that the average cost curves for each country are monotonically declining over the full range of the data (this is most easily observed in the log-scale plots in Figure S2).

40. Missing in discussion: Overall variation in Ghana is quite large for a given number of doses (especially under 10,000 doses) as compared to Benin (visual inspection of Figure 1 is a bit difficult because of the differences in the units on the axis). Further, in all countries except Moldova, sites with <5,000/<10,000 (depending on the country) total doses had unit costs that
were as low as sites with high numbers of total doses, indicating that low unit costs are possible in all but the lowest service volume sites - it would seem like understanding the differences in average costs at low volume sites might be a priority for future research. The authors could expand this a bit. How much of this variation [at low service volume sites] disappears once the other factors controlled for in the regression are accounted for? How much resources could be saved if high cost / low volume sites could find ways to reduce their unit costs similar to other low volume sites? I think this kind of data - while a bit of speculation - would highlight the extent to which this type of research might be useful / valuable.

We have edited the discussion text to discuss these issues. Of note, the higher variability in average costs in small sites does not necessarily imply they should be targeted for intervention. Even though larger sites show lower absolute variation in unit costs, the fact that these are multiplied by much greater patient numbers means that the total cost savings from improved efficiency could be greater in these large sites (p20, paragraph 3):

“For most countries there were many small sites with high average costs, and these small sites exhibited substantial variation in unit costs. The reduction in unit costs associated with higher service volume was only minimal for sites at the upper end of the service volume distribution, and these large sites exhibited only minor variation in unit costs. While this suggests greater reductions in average costs might be achieved through efforts to improve efficiency in small sites, it is not clear that these sites should be prioritized, as large total cost reductions might be possible with only small reductions in unit costs at large sites.”

(p23, paragraph 1):

“Finally, the results for Ghana showed noticeably greater unexplained variation compared to the other countries (Figure 1, Figure S2), yet the reasons for this are unclear.”

41. Table 3: should explicitly state what the numbers in parenthesis represent (presumably the standard error)

Thank you for catching this omission. The appropriate text (“Values in parentheses represent standard errors”) has been added to the footnotes (p15).

42. Total costs model 5: It is not clear to me if the random effect is for the intercept or a country-level random slope for log(doses)?

All models include random intercepts at country- and province-level. In addition, model 5 allows random country-level slopes for log(doses). This is shown in Table 3 (Section ‘Random effects included’), and in text added to Section 2.6.

REVIEWER #3

43. I checked the statistical methods part. The models are sensible and they are described sufficiently, although it would be even better if Stan code would be included.
Thank you for these comments.

44. The inference part is not described sufficiently. Based on the reference I assume Stan version 2.8.0 was used. This could mentioned explicitly. The authors should report the options used to run Stan: the number of chains, the number of warm-up iterations, and the number iterations after warm-up. The authors should report results from the convergence diagnostics. It would be sufficient to report whether 1) all Rhat values were below 1.1, 2) zero divergences, and 3) typical and the smallest effective sample sizes neffs. All these can easily obtained from the Stan output.

We have edited the appendix text to include these details (appendix, p7, paragraph 2):

“Regression models were estimated using an adaptive Hamiltonian Monte Carlo algorithm [7] as implemented by the Stan software package, version 2.16.2 [9, 10]. For each regression model we ran 4 chains for 5000 iterations (2500 warm-up, 2500 sampling), obtaining 10000 posterior draws used for subsequent analyses. For the models shown in the main analysis (Models 1 to 5 in Table 3), Rhat values were all equal to 1.0 (Rhat values above 1.0 indicate poor chain convergence), there were zero divergences (divergences are an indicator of poor sampling), and the minimum effective sample size for model parameters was 1329, and typical values were >2000 (low values would indicate imprecision in parameter estimates due to Monte Carlo error). Processing of data and results were undertaken in R version 3.3.3 [11].”

The Stan references have also been updated (Stan version 2.16.2 was used).

45. WAIC is used sensible for the model comparison (Note that the Stan development is now recommending to use of leave-one-out cross-validation instead of WAIC. See R package loo and Vehtari, Gelman and Gabry (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. In Statistics and Computing, 27(5):1413-1432).

Thank you. We will investigate loo for future applications.

46. Since there are different versions of WAIC, the authors should report which version was used. If some R package (like loo) was used, it would be sufficient to report if the default choice in such package has been used.

We used the operationalization of WAIC provided by Vehtari and Gelman (WAIC and cross-validation in Stan, 2014). We have added this reference to the text (appendix, p7, paragraph 1).

47. The difference in WAIC values for models 3 and 4 is likely to be non significant (the authors write modest difference). Other differences are likely to be significant. In the appendix there are also many models with non-significant differences. Otherwise the results of the statistical analysis are reasonably reported and conclusions are sensible based on the results.

Thank you for these comments. We have replaced the word ‘modest’ with ‘insignificant’ in the main text (p16, paragraph 1).