Author's response to reviews

Title: National average household concentrations of PM2.5 from solid cookfuel use: Results from measurements and modeling in India for the Global Burden of Disease -2010 assessment

Authors:

Kalpana Balakrishnan (kalpanasrmc@ehe.org.in)
Santu Ghosh (santu@ehe.org.in)
Bhaswati Ganguli (bgstat@gmail.com)
Sankar Sambandam (sankars@ehe.org.in)
Nigel Bruce (sankars@ehe.org.in)
Douglas F Barnes (barnesdf@gmail.com)
Kirk R Smith (krksmith@berkeley.edu)

Version: 2 Date: 30 June 2013

Author's response to reviews: see over
Reviewer 1

Review

1. Is the question posed by the authors new and well defined? Yes

Major Compulsory revisions

2. Are the methods appropriate & well described, and are sufficient details provided to replicate the work?

Reviewer Query 2a:
Authors state that “additional details of methods is being submitted as separate publication”; the details are required for methodology used in this study especially dates and years of measuring PM2.5, and conduct of questionnaire; along with further details of validation methods for the model. The cooking method as well as type of fuel; place of cooking might change by months/season; hence necessary to include the times of data collection for each method (e.g. whether all states were sampled at the same time? PM2.5 measurements/validation were collocated and questionnaire was conducted at the same time? It is stated that the sampling strategy involved stratification based on type of kitchen and fuel; hence all such data were collected prior to stratification? The methods need to be specified. The number of households sampled differs from abstract and methods 600 or 617; needs to be consistent.

Response 2a: Section 1a in "Methods" describing sampling strategy, household selection and recruitment has been amended to include additional details requested by the reviewer. The revised text is provided below.

The sentence “additional details of methods” refers to “additional details of measurement results” that were collected during the same field exercise but are the subject of another analysis being submitted as a separate publication.

The inconsistency in reported sample size between the abstract and the methods section has been corrected to reflect the correct sample size of 617. The detail pertaining to model validation is provided in Response 2b.

Revised text (modified text is highlighted within the original paragraphs) Pages 7,8

Between November 2004 and January 2005, 617 households in four geographically and culturally distinct states (Central-Madhya Pradesh (MP), South-Tamil Nadu (TN), North-Uttaranchal (UA) and East-West Bengal (WB)) of India, were recruited for household measurements. The choice of states was made primarily to provide a representative basis for the model. Selection of households across the country to generate a representative, measured national level estimate was not feasible on account of financial and logistic constraints.

Multi-stage sampling was used to randomly select two districts from each state and three villages from each district. Approximately 25 households were selected by stratified random sampling based on fuel and kitchen type, in each village resulting in around 150-155 households from each
Each village encompassed as many as several hundred households. To select the study households, the field team first conducted a rapid assessment of all households in the village. The team members went to each household and asked several short questions, including ones about primary fuel type and kitchen type. After the completion of the rapid assessment, a stratified random sample – based on fuel and kitchen type – of twenty five households was drawn. The following day, these households were invited to participate in the study. Urban households could not be included (we elaborate on this, in the discussion section).

Informed consent was obtained from all study households prior to any assessments. The protocols for measurements were approved by the human subjects committees of Sri Ramachandra University and The University of California, Berkeley. All household level assessments including questionnaire administration and air pollution measurements were performed shortly after recruitment and simultaneously in the four states using four field teams. Field teams were trained jointly by the core investigators prior to deputing the teams for field work. A manual containing standard operating procedures was provided to all field team members for respective data collection tasks. Field data collection was completed between December 2004 and March 2005.

**Reviewer Query 2b:** Provide additional detail on model validation

**Response 2b:**

*We had earlier reported standard error estimates from bootstrapping in support of our model accuracy. Although we reported correlation between predicted and measured estimates of kitchen concentration, this was not done using a 'k-fold' cross validation approach. At the suggestion of both reviewers, we performed a k-fold cross validation, excluding households from each of the 24 villages (~25 households) sequentially. The 24-fold cross-validation (using the log transformed 24 hr kitchen concentration dataset) provided an overall correlation coefficient of r=0.56 between*
modeled and measured values (Figure 3 Top panel has been revised to show the results from the 24 fold cross validation).

Revised text is provided below

Revised text (modified text is highlighted within the original paragraphs) Page 15

Stage 4. Assessing model accuracy through k-fold cross validation and bootstrapping

We applied cross validation and bootstrapping methods to estimate the accuracy of models developed in earlier stages. We first performed a k-fold cross validation for the household model (described in Stage 2) by excluding households from each of the 24 villages (~25 households) sequentially. The 24-fold cross-validation (using the log transformed 24 hr kitchen concentration dataset) provided an overall correlation coefficient between modeled and measured values.

Bootstrapping was then used to estimate the standard error of prediction for the national level model (described in Stage 3). To compute the bootstrapping standard error of the kitchen area PM$_{2.5}$ estimates, we first generated 200 constructed datasets (replicates) of PM$_{2.5}$ as

$$\log(P\hat{M}_{2.5}) \sim \text{Normal}(\mu = X\hat{\beta}, \sigma^2 = \hat{\sigma}^2);$$

where $X$ refers to the vector of all the predictors in a household. Each constructed dataset was required to be of the same size as the original data based on estimated parameters and empirical predictors. The model was applied on each of the 200 constructed datasets (the estimates started to converge after application on 100 replicates and was doubled to allow an additional margin for stability) to obtain the empirical standard deviations of each parameter along with error variance. We used the empirical standard deviation of error variance, considered to be the standard error to obtain the bootstrapping standard error of predicted PM$_{2.5}$ concentrations.

Reviewer Query 2c: Earlier in methods (Stage 2) it is mentioned “...we first developed........in relation to these variables...” which variables? (variables are specified a paragraph later?); deem clarification. Further clarification is needed whether any living room estimates were done or not? It is presumed that only kitchen area concentrations were measured and no where specifically stated. It is presumed that paired kitchen - living area
measurements were developed through modeling; along the same numbers as kitchen area measurements? About 600?

Response 2c: Listing of variables has been provided within the same para. Also, details on the numbers for paired kitchen living measurements have been added. Revised text provided below

Revised text (modified text is highlighted within the original paragraphs) Pages 10,11

Questionnaires were administered in all study households to collect information on a range of household level variables. This primarily included physical variables likely to directly influence household concentrations such as fuel type, kitchen location, stove type, ventilation, fuel quantity and cooking duration. Information on indicators of other sources of indoor emission of particulate matter were also captured by recording use of solid fuels for heating, indoor smoking, number of hours without electricity (indicative of use of kerosene based lamps for lighting) and use of incense or mosquito coils. We focused on direct physical variables found in national surveys around the world, including India and had no access to indirect indicators such as ethnicity and behavioral variables that would be likely to affect exposure, such as meal type.

We first developed models to estimate kitchen area concentrations (from measurements conducted in 617 households) in relation to these variables. Most household variables related to cookfuel use are likely to directly influence kitchen area concentrations, with living area concentrations in turn, being influenced by respective kitchen area concentrations. We therefore developed regressions equations for the relationship between kitchen and living area concentrations (from valid paired measurements available in 451 households) in order to be able to derive the living area from kitchen concentrations. We describe the procedures for modeling the kitchen and living area concentrations separately in greater detail below.

Reviewer Query 2d: Gravimetric method use is referred to a 2004 publication by the same author; needs clarification whether it was the same data as this study or not. In addition (Methods Stage 1; b, para 4) ... 20% of what samples were paired with blanks;
needs further clarification.

**Response 2d:** The 2004 publication concerns another study and has been referred only for details of methods. The gravimetric dataset used in this study was collected separately. Gravimetric measurements were collected from 96 households (~15% of the total sample of 617). 20% of this sample (n=18) were paired with blanks. The revised sentences in the section are provided below

Revised text (highlighted in yellow within original paragraphs)

Revised text (Page 9)

Particle coefficients were derived for each instrument in the field through co-location of UCB-PATs monitors and gravimetric samplers in around 15% of households (n=96).

Revised text (Page 10)

Gravimetric PM$_{2.5}$ samples were collected using methods published previously [8]. Briefly, samples were collected using a BGI triplex cyclone (scc1.062, Waltham, MA) in portable constant-flow SKC pumps (Model 224-PCAR8, SKC, Eighty-Four, PA, USA) equipped with a 37-mm diameter Teflon filter (pore size 0.45 µm also supplied by SKC) at a flow rate of 1.5 l/min. Filters were weighed using a Thermo Cahn C – 34 Microbalance (Thermo Scientific, Waltham, MA, USA) at Sri Ramachandra University and a Mettler Toledo-MT5 balance (Mettler, Greisensee, Switzerland) at The Energy Research Institute in New Delhi. Both balances operated at a resolution of 0.1 µg and were used according to the same standard operating procedure. All filters were conditioned in a temperature and relative humidity controlled room before weighing. Approximately, twenty percent of the gravimetric samples (collected from 96 households) were paired with field blanks (n=18); none of the pre- and post- field blank weights differed by greater than 0.003 mg.

**Reviewer Query 3.** Are the data sound and well controlled?

Data presented in sufficient details for mean concentrations of PM2.5 in tables by states and other variables & authors have stated the limitations. The number of samples in model need to be specified in tables. The numbers if table 1 are in 400s; and in table 3 in 600s. The ns
need to be specified clearly. The PM2.5 concentrations as seen in tables imply a not normal
distribution; whether log or any other transformation was done or not? If done whether the
distribution changed to normal? Perhaps not; that’s why data points between 5th and 95th
percentile were used; which dataset is shown in tables ? For bootstrapping method reason
for replicating (only) 200 datasets; is to be justified.

Response 3: Table 1 includes data from households with kitchen and living concentrations
lying between the 5th and 95th percentile. The N’s are more clearly specified as suggested
to avoid ambiguity. The legend for Table 1 has been modified to reflect the dataset used for
the reported results.

A Box-Cox procedure was used to decide on the optimal transformation of the dependent
variable for use in the regression equation after exclusion of data between 5th to
95th percentile as has been mentioned in the text (page 17).

Table 3 provides results from the model as applied to rural and urban solid fuel using
households in the NFHS dataset to provide a state level estimate. The N from the
measurement dataset is not shown. The state level population is provided for allowing an
assessment of relative sampling weights in the national level estimate.

The bootstrapping estimate of standard error started to converge after using 100 replicates
and we doubled this number to allow an extra margin for stability. This has been indicated
in the section, as suggested (Page 14).

Reviewer Query 4. Does the manuscript adhere to the relevant standards for reporting and
data deposition?

Background does not provide sufficient information about reported PM2.5 concentrations
from other settings especially in reference to GBD 2010. Else tables, figures and maps are
fine; sample numbers and dates are required in methods as well as in tables. Further
comparison is not provided for PM2.5 estimates from other studies in the similar setting/s
and other settings.

Response 4: Although GBD 2010 did not use quantitative measurement data on household
air pollution reported from other studies (it relied instead on results from the India modeling
exercise presented through this paper), a brief description has been added to the
background section and a table (Table 4) providing a more detailed comparison of estimates
from the present study to other studies in India/other regions has been added.

Reviewer Query 5a. Are the discussion and conclusions well balanced and adequately
supported by the data?
Other studies used as references; are not referred to adequately; whether done in India or
in other settings. How can author justify that their estimates are valid? Given that results
of gravimetric method are not shown. Authors state limitations for cross sectional study but
do not provide dates/seasons/year for the measurements.

Response 5a: As mentioned in response 4 we have added a table for comparison to other
studies. Date /season and year have been provided. The gravimetric data set was used
primarily to validate the UCB measurements.
Reviewer Query 5b. Only methodological aspects are discussed; meaning the authors are not comfortable with the national level data based on their methods? It is to be specified whether the study was done for theoretical purpose or for public health practice perspective?

Response 5b: The methodological aspects are discussed to both provide strength to the approaches used and provide a robust justification for the analytical process that underlies the GBD results. The study was done both to provide a methodological approach for burden of disease assessments and for application of results in public health practice at the country level, especially in the planning of intervention efforts. As more measurement datasets become available from other countries, approaches developed in the study are likely to find wider applications at the country/regional level. Based on the suggestion from the reviewer, we have explicitly added sentences to reflect the public health practice focus. Revised text provided below

Revised Text (Highlighted in yellow within original paragraphs) Page 25

Although in need of further refinements, the model shows substantial promise to be able to generate household concentration estimates due to cooking fuel in rural households that may be aggregated to estimate population exposures at the state or national level in India. The predictive power for estimating concentrations in an individual household is modest, but at the state and national level in India, it provides substantial improvement over simple binary metrics such as solid versus non-solid cookfuel use, commonly used as exposure indicators, in HAP studies. Such a population level estimate was essential to allow a linkage to the IERs in conducting the more sophisticated CRA analyses for the GBD-2010[4]. The model estimates also add considerable strength of evidence for the need to scope and implement effective public health intervention efforts at the state and national levels. With the average concentrations experienced in households being significantly higher than health based air quality guideline values, the results from the study indicate the need for achieving substantive exposure reductions at the population level.

In the 30 years since the first set of solid cookfuel related exposure studies in rural households of developing countries were reported [33], progress on developing good models that are
sophisticated enough to capture the heterogeneity while relying on easy to collect indicators has been slow, with only a few recent studies making significant contributions[11,12]. We hope the results presented in the study spur additional efforts to validate as well as develop newer models to address the complexities of exposure reconstruction for household air pollution at individual, local, national and global scales. Routine integration of measurement efforts with national surveys such as NFHS, LSMS and DFHS would not only allow additional refinements in the model for national/state level estimates in the future, but also allow the use of such models in monitoring and evaluation of public health efforts directed towards intervention for HAP.

Reviewer Query 5c: As stated in discussed; the inability to perform personal level measurements; how can thisbe a limitation? I could not find this objective anywhere in manuscript or was it the intentionto imply personal level exposure?

Response 5c: We did not intend to present this as a limitation resulting from our inability to perform measurements as part of the study. But, since personal measurements were not part of the study objectives, the outputs do not provide an opportunity to describe how the measured/estimated household concentrations may be related to personal exposures for sub-groups of population such as women, men or children. We also have added text to reflect how this gap was closed in the GBD exercise using results from this study, that provided only household concentrations (Page 21 and 22).

Reviewer Query 6. Do the title and abstract accurately convey what has been found? Title appears to be fine (pl see below)

Abstract: Given the title; (GBD 2010 assessment) background needs to specify whether the exercise was performed for theoretical purpose (modeling data) or for public health practice considerations.

Has been added
Number of states in India need to be mentioned too.

Has been added
In Methods;year and dates of data collection in 4 states is to be stated specifically.

Has been added
Results ; predictorsof PM2.5 concentrations require specific mention of; (which) fuel type , kitchen type (asdefined?), method of ventilation, geographical location of..., or whether increasing or decreasing cooking duration. PM2.5 estimates correlated between which respectivemeasures?

Was difficult to include in the abstract while meeting word count restrictions
Along with national estimates perhaps state level estimates need to bementioned too.

Was difficult to include estimates from 29 states while meeting word count restrictions
Conclusion; please clarify the specific future refinements that will allow application to this? Or other populations.

Has been added

Revised Abstract

Modeling national average household concentrations of PM$_{2.5}$ from solid cookfuel use for the Global Burden of Disease -2010 assessment: Results from cross-sectional assessments in India

Background

Recent progress in global burden of disease (GBD) methodologies that use integrated–exposure–response curves for combustion particles required the development of models to quantitatively estimate average household (HAP) levels experienced by large populations. Such models are also needed for framing country–level public health intervention efforts.

Objectives

We aimed to develop a model to provide national and state level (for 29 states) estimates for average household concentrations of PM$_{2.5}$, from solid cookfuel use in India.

Methods

We monitored 24-hr household concentrations of PM$_{2.5}$, in 617 rural households from 4 states in India on a cross-sectional basis between December 2004 and March 2005. A log-linear regression model that estimates household concentrations as a function of multiple, independent household variables available in national household surveys was developed, which was then used to generate national / state-level estimates using The Indian National Family and Health Survey (NFHS-3, 2005).

Results

The measured mean 24-hr concentration of PM$_{2.5}$ in solid cookfuel using households ranged from 163 $\mu$g/m$^3$ (95% CI: 143,183; Median 106; IQR:191) in the living area to 609 $\mu$g/m$^3$ (95% CI:
547,671; Median: 472; IQR: 734) in the kitchen area. Fuel type, kitchen type, ventilation, geographical location and cooking duration were found to be significant predictors of PM$_{2.5}$ concentrations in the household model. k-fold cross validation showed a fair degree of correlation ($r=0.56$) between modeled and measured values. Extrapolation of the household model to all solid cookfuel using households in India, covered by NFHS-3, resulted in a modeled estimate of 450µg/m$^3$ (95% CI: 318,640) and 113µg/m$^3$ (95% CI: 102,127) for national average 24-hr PM$_{2.5}$ concentrations in the kitchen and living areas respectively.

**Conclusions**

The model affords substantial improvement over commonly used exposure indicators such as “percent solid cookfuel use” in HAP disease burden assessments, by providing the first estimates of national and state HAP levels experienced in India while informing exposure estimates used in the GBD-2010 exercise. The estimates indicate the need for achieving substantive exposure reductions at the population level. Routine integration of measurement efforts with national surveys would allow use of such model estimates in framing, monitoring and evaluation of intervention efforts.
**Reviewer’s report II**

**Title:** National average household concentrations of PM2.5 from solid cookfuel use: Results from measurements and modeling in India for the Global Burden of Disease -2010 assessment  
**Version:** 1  
**Date:** 26 April 2013  
**Reviewer:** Darby Jack

**Reviewer’s report:**

This is an important paper and should be published, both because it makes significant methodological contributions, and because it provides a valuable window into the analytical process that underpins the GBD/CRA results. The question is well posed, the data is well matched to the question, and the methods are sound.

Given that a central contribution of this paper is to document the analysis that led to the exposure estimates that were actually used in the CRA, my comments mostly focus on the presentation, validation and discussion of the modeling approach. All of these fit in the category of discretionary revisions, though I do feel that #1 is quite important.

1. I would like to see a cross validation analysis of the core regression model. A k-fold approach seems appropriate in this context (with groups defined by villages). Without validation results, the reader cannot evaluate the predictive power of the model. This is my primary concern with the paper as it now stands.

*We had earlier reported standard error estimates from bootstrapping in support of our model accuracy. Although we reported correlation between predicted and measured estimates of kitchen concentration, this was not done using a ‘k-fold’ cross validation approach. As suggested, we performed a k-fold cross validation, excluding households from each of the 24 villages (~25 households) sequentially. The 24-fold cross-validation (using the log transformed 24 hr kitchen concentration dataset) provided an overall correlation coefficient of r=0.56 between modeled and measured values (Figure 3 Top panel has been revised to show the results from the 24-fold cross validation).*

**Stage 4. Assessing model accuracy through k-fold cross validation and bootstrapping**

We applied cross validation and bootstrapping methods to estimate the accuracy of models developed in earlier stages. We first performed a k-fold cross validation for the household model (described in Stage 2) by excluding households from each of the 24 villages (~25 households).
sequentially. The 24-fold cross-validation (using the log transformed 24 hr kitchen concentration
dataset) provided an overall correlation coefficient between modeled and measured values.

Bootstrapping was then used to estimate the standard error of prediction for the national level
model (described in Stage 3). To compute the bootstrapping standard error of the kitchen area
PM$_{2.5}$ estimates, we first generated 200 constructed datasets (replicates) of PM$_{2.5}$ as

$$\log(PM_{2.5}) \sim \text{Normal}(\mu = X\hat{\beta}, \sigma^2 = \hat{\sigma}^2);$$

where $X$ refers to the vector of all the predictors in a

household. Each constructed dataset was required to be of the same size as the original data
based on estimated parameters and empirical predictors. The model was applied on each of the

200 constructed datasets (the estimates started to converge after application on 100 replicates and
was doubled to allow an additional margin for stability) to obtain the empirical standard
deviations of each parameter along with error variance. We used the empirical standard
deviation of error variance, considered to be the standard error to obtain the bootstrapping
standard error of predicted PM$_{2.5}$ concentrations.

2. The purpose and implications of the bootstrapping procedure described in the
last paragraph on page 11 is unclear. The paragraph refers to imputation, but it’s
unclear what variables are imputed at this stage of the analysis. If the imputation
mentioned on page 11 is for cooking hours and ventilation (p 13) then it makes
sense to move the paragraph to that section. Either way, this paragraph needs to
be clarified.

_The purpose of the bootstrapping has been clarified as suggested by moving the
text(relevant to bootstrapping)to a separate section describing model validation._

_This modification has been incorporated in page 14 of the revised manuscript_

3. I was struck by the lack of behavioral and socioeconomic variables (as
opposed to physical) variables in the predictive model. For example, variables
like ethnicity and wealth likely predict which foods are cooked, and thus
concentrations. The authors note that variables that did not predict pollution were
dropped from the model – perhaps and socioeconomic variables dropped out?
This should be addressed in the text.

*We addressed the reviewer suggestion by including the following text*

*Revised text (modified text is highlighted within the original paragraphs) Page 10*

Questionnaires were administered in all study households to collect information on a range of household level variables. *This primarily included physical variables likely to directly influence household concentrations such as fuel type, kitchen location, stove type, ventilation, fuel quantity and cooking duration. Information on indicators of other sources of indoor emission of particulate matter were also captured by recording use of solid fuels for heating, indoor smoking, number of hours without electricity (indicative of use of kerosene based lamps for lighting) and use of incense or mosquito coils. We focused on direct physical variables found in national surveys around the world, including India and had no access to indirect indicators such as ethnicity and behavioral variables that would be likely to affect exposure, such as meal type.*

4. The authors need say more about why they chose to model living area concentrations as function of kitchen concentrations (eqn 2), instead of predicting them from household characteristics (using a model analogous to eqn 1).

*We added the following text to justify the need to model living area concentrations as a function of kitchen concentrations.*

*Revised text (modified text is highlighted within the original paragraphs) Page 11*

We first developed models to estimate kitchen area concentrations (from measurements conducted in 617 households) in relation to these variables. *Most household variables related to cookfuel use are likely to directly influence kitchen area concentrations, with living area*
concentrations in turn, being influenced by respective kitchen area concentrations. We therefore developed regressions equations for the relationship between kitchen and living area concentrations (from paired measurements available in 451 households) in order to be able to derive the living area from kitchen concentrations. We describe the procedures for modeling the kitchen and living area concentrations separately in greater detail below.

5. Additional discussion of the role of the authors’ analysis in the CRA estimates would strengthen the paper. Were the exposure values reported on page 16 the actual values used in the CRA analysis, or were they modified before use? In particular, were any adjustments made to account for the (potentially large) differences between kitchen concentrations and 24-hour personal exposure? How was the analysis carried out in countries that lack the household concentration database used here?

*We have added additional text to clarify the role of results obtained from this study in GBD 2010. The modified text from the relevant section is provided below*

*Revised text (modified text is highlighted within the original paragraphs) Page 22*

Generation of a population level exposure estimate for use in Integrated Exposure Response (IER) curves in GBD-2010 assessment: As mentioned in the introduction, recent refinements in burden of disease assessment methodologies for GBD-2010 require a quantitative estimate of population exposure to be able to use IERs for relative risk estimation of various disease endpoints associated with ambient and household air pollution. The generation of a national level estimate for India fulfilled this important requirement, while providing an approach for application in other countries. *India had some of the largest measurement datasets available together with national level survey information. GBD 2010 therefore used the household concentration estimates reported in this paper together with estimated ratios between daily*
average personal exposures and kitchen concentration from available published studies to arrive at personal exposure estimates for population subgroups including women, men and young children. Exposure estimates obtained thus, were used in IERs developed for estimation of relative risks for acute lower respiratory infections in children, ischemic heart disease (IHD), stroke, and lung cancer in GBD 2010[4].

6. The paper should include some discussion of how season affects household concentrations. In some parts of India, there are pronounced seasonal patterns in the background levels of particulate pollution. At minimum, the paper should describe when (which months) the exposure measurements were made, and, to the extent possible, should discuss how the timing of exposure monitoring might have affected the results.

Has been added in page 22

7. The conclusion should include a discussion of how household survey questionnaires (e.g., DHS; LSMS; NFHS) can best capture data that would be useful for the kinds of predictive modeling of household air pollution exposures described here.

Has been added in page 25

8. Table 2 would be easier to read if it listed variable names rather than subscripted betas.

Has been modified as suggested

Level of interest: An article of importance in its field
Quality of written English: Needs some language corrections before being published
Statistical review: No, the manuscript does not need to be seen by a statistician.
Declaration of competing interests: None