Original Research Article

ENVIRONMENTAL EPIDEMIOLOGY

OPEN

Estimating long-term average household air pollution concentrations from repeated short-term measurements in the presence of seasonal trends and crossover

Joshua P. Keller[a*], Maggie L. Clark[b]

Abstract. Estimating long-term exposure to household air pollution is essential for quantifying health effects of chronic exposure and the benefits of intervention strategies. However, typically only a small number of short-term measurements are made. We compare different statistical models for combining these short-term measurements into predictions of a long-term average, with emphasis on the impact of temporal trends in concentrations and crossover in study design. We demonstrate that a linear mixed model that includes time adjustment provides the best predictions of long-term average, which have lower error than using household averages or mixed models without time, for a variety of different study designs and underlying temporal trends. In a case study of a cookstove intervention study in Honduras, we further demonstrate how, in the presence of strong seasonal variation, long-term average predictions from the mixed model approach based on only two or three measurements can have less error than predictions based on an average of up to six measurements. These results have important implications for the efficiency of designs and analyses in studies assessing the chronic health impacts of long-term exposure to household air pollution.

Introduction

Exposure to household air pollution constitutes a major morbidity and mortality burden around the world.[1] Stoves for cooking and heating in homes are a major source of household air pollution. Over the last two decades, many studies have been conducted to characterize and mitigate adverse health effects of household air pollution exposure, including improvements to biomass-burning stoves,[2–4] the replacement of biomass-burning stoves with cleaner fuel alternatives,[5–8] and economic programs to facilitate the adoption of cleaner fuel stoves.[9,10] In almost all studies, only a limited number of short-term exposure measurements are made owing to financial cost and participant burden.[11] Typical study measurements are for 24–72-hour duration and the frequency of measures range from one per household[12,13] to a series of three or four measurements aligned across a gestational period[14,15] to measures across a multiyear period.[4,16] Pollutant measurements can be made for room concentrations (typically kitchen) using a stationary monitor or for an individual's personal exposure using a mobile monitor in the breathing zone.

When estimating the health effects of household air pollution exposure, short-term exposure measurements are often used directly with health outcome measurements ascertained at approximately the same time.[5,17,18] Although this is appropriate for investigating acute health effects of exposure, it is suboptimal for investigating chronic exposure effects. Although each individual observation can be unbiased for exposure at that time, variability in single measurements introduces classical measurement error[19] that can bias health analyses of long-term exposure.

Instead of using individual measurements in analyses of chronic exposure, combining the data to estimate the long-term average can reduce the effects of measurement error—in addition to better matching the epidemiologic target of chronic exposure. Depending on the context, the long-term average might be a multiyear, multimonth, or multiweek average. Different averages might further be defined for different conditions, such as changes in stove type.

What this study adds

This article addresses two important topics for the growing number of household air pollution studies that rely on multiple short-term exposure measurements: (1) what statistical approach is best for estimating chronic exposure from short-term household air pollution measurements? and (2) how many measurements are needed? These questions are answered using simulations representing multiple different study designs and in a case study. The results have important implications for the design of future studies and the development of analysis plans for existing studies of chronic exposure.

[a]Department of Statistics, Colorado State University, Fort Collins, Colorado; and [b]Department of Environmental and Radiological Health Sciences, Colorado State University, Fort Collins, Colorado

The research in Honduras was funded by the National Institute of Environmental Health Sciences of the National Institutes of Health under award number ES022269.

The code for conducting the simulation is available in a publicly-available GitHub repository at https://github.com/jpkeller/LTA_Simulation. Readers interested in obtaining the Honduras data should contact the authors about access and confidentiality requirements.

SDC *Supplemental digital content is available through direct URL citations in the HTML and PDF versions of this article (www.environepidem.com).*

Corresponding Author. Address: Department of Statistics, Colorado State University, 1877 Campus Delivery, Fort Collins, CO 80523. E-mail: joshua.keller@colostate.edu (J.P. Keller).

Environmental Epidemiology (2021) 6:e188

Received: 15 September 2021; Accepted 26 November 2021

Published online 20 December 2021

DOI: 10.1097/EE9.0000000000000188

A common approach to estimating long-term averages is calculating household averages,[20,21] which is straightforward computationally and conceptually. A better approach (in terms of lower error) is to predict the long-term average from a linear mixed model that includes a random intercept for each household. Empirically, McCracken et al[22] demonstrated the benefit of using a mixed model compared to using household averages. Predictions of long-term average concentrations from mixed and Bayesian models with random effects have also been used by Grajeda et al[23] and Keller et al,[24] respectively. Most other studies that use linear mixed models use them to study factors related to pollutant concentrations, not to make predictions of long-term average exposures.[25,26]

Two aspects of McCracken et al[22] to consider here are that they compared households with constant stove type and without accounting for the role of a continuous and smooth temporal trend. Seasonal changes, such as the Harmattan[14] in West Africa and rainy and dry seasons in Honduras[27] and Nepal,[28] can affect concentrations, and estimating a long-term average requires accounting for these smoothly varying temporal differences—even in parallel randomized trials. Crossover in stove type, most often seen in stepped-wedge designs,[2,4] allows all households to be measured with multiple stove types.[3] But unlike seasonal trends, which we usually want to average over, it is often the goal to compute different values for changes in stove groups, which requires defining a separate long-term average for each stove type. This was done in Keller et al[24]; here, we provide justification for this approach.

In this article, we address how to best estimate long-term average concentrations in the presence of these temporally varying factors, focusing on the impact of study design (number and timing of measurements) and on the choice of a statistical model. The sections that follow provide (1) a description of the statistical context and assumed data-generating model, (2) a review of different approaches for estimating long-term average concentrations in mixed models with between-subject and time-varying, within-subject variables, (3) a simulation study that demonstrates the effects of different study designs and modeling choices on the error in the predictions, and (4) a case study demonstrating these differences using data from a 3-year stepped-wedge intervention study in Honduras.

Statistical framework

We are interested in estimating the long-term average pollutant concentration for a specific unit. We will refer to this unit as a *household* and describe the concentration as representing household air pollution, but the methodology could apply to measurements of other types of indoor concentrations or an individual's personal exposure. The difference between a household concentration and individual personal exposure is important, but beyond the current scope.

Let X_{ij} denote the measured concentration for household i ($i = 1,\ldots,n$) at visit j ($j = 1,\ldots,J$). We assume that the (possibly log-transformed) value of X_{ij} is a combination of a study-wide average, denoted by the parameter β_0; a time-constant household effect due to unmeasured factors, modeled as $a_i \sim N(0,\sigma_A^2)$; a combination of p measured variables, such as access to electricity, that impacts long-term average concentrations and vary between but not within households, denoted by z_{ir} ($r = 1,\ldots,p$); factors that vary within a household and define different conditions for a long-term average, such as stove type in a crossover trial, denoted by w_{ijk} ($k = 1,\ldots,K$); smooth temporal variation due to seasonality and similar factors, denoted as $f_i(t_j)$; and transient effects, such as day-of-week, denoted by v_{ijs} ($s = 1,\ldots,q$). The assumed data-generating model is as follows:

$$X_{ij} = \beta_0 + a_i + \sum_{r=1}^{p} z_{ir}\beta_r + \sum_{k=1}^{K} w_{ijk}\theta_k + f_i(t_j) + \sum_{s=1}^{q} v_{ijs}\gamma_s + \varepsilon_{ij}, \quad (1)$$

where the final term $\varepsilon_{ij} \sim N(0,\sigma_E^2)$ is the observation error. Using measurements X_{ij} directly in statistical models for a corresponding health outcome will lead to classical measurement error[19] from the variability (ε_{ij} term) in the short-term measurements.

Based on equation 1, the long-term average we wish to estimate for each household is as follows:

$$\mu_i(\boldsymbol{w},\mathcal{T}) = \beta_0 + a_i + \sum_{r=1}^{p} z_{ir}\beta_r + \frac{1}{|\mathcal{T}|}\sum_{t\in\mathcal{T}}\left(f_i(t_j) + \sum_{k=1}^{K} w_{tk}\theta_k\right). \quad (2)$$

The long-term average is defined for a specific time range, denoted by $\mathcal{T}$. This time range might be the entire study duration, a gestational period, or other health-relevant exposure period, but it must be within the temporal range of the data (although perhaps outside the temporal range of observations for a specific unit). The final term in equation 2 averages the temporal trends over this period by summing the values at all time points and then dividing by the number of time points in the period (denoted by $|\mathcal{T}|$). In many settings, it may be desirable to choose $\mathcal{T}$ to symmetrically average over seasonal trends (e.g., including one rainy and one dry season). We write the long-term average $\mu_i(\boldsymbol{w},\mathcal{T})$ as a function of the vector of conditions $\boldsymbol{w}$ (that may vary over time) to indicate a different long-term average for each combination of conditions; if there is only one condition, we can write the long-term average as $\mu_i(\mathcal{T})$.

In the sections that follow, we compare different methods for estimating $\mu_i(\boldsymbol{w},\mathcal{T})$ using the data X_{ij}. The comparison metric is root mean squared error (RMSE), which is defined as $\mathrm{RMSE}(\hat{\mu}_i(\boldsymbol{w},\mathcal{T}),\mu_i(\boldsymbol{w},\mathcal{T})) = \sqrt{E[(\hat{\mu}_i(\boldsymbol{w},\mathcal{T}) - \mu_i(\boldsymbol{w},\mathcal{T}))^2]}$ for the prediction $\hat{\mu}_i(\boldsymbol{w},\mathcal{T})$. The RMSE combines information on bias and variance. In addition to absolute differences in RMSE, we will compare approaches through their relative RMSE, which is the ratio $\mathrm{RMSE}(\hat{\mu}_i(\boldsymbol{w},\mathcal{T}),\mu_i(\boldsymbol{w},\mathcal{T})) / \mathrm{RMSE}(\hat{\mu}_i'(\boldsymbol{w},\mathcal{T}),\mu_i(\boldsymbol{w},\mathcal{T}))$ for two different predicted long-term averages, $\hat{\mu}_i(\boldsymbol{w},\mathcal{T})$ and $\hat{\mu}_i'(\boldsymbol{w},\mathcal{T})$.

Methods for estimating long-term averages

Basic case: no predictors and no time-varying effects

The simplest setting is when there are no predictor variables or time-varying factors. Under these assumptions, the long-term average is $\mu_i = \beta_0 + a_i$ and we have the data-generating model as follows:

$$X_{ij} = \beta_0 + a_i + \varepsilon_{ij}. \quad (3)$$

The household average $\bar{X}_{i\cdot} = \frac{1}{J}\sum_{j=1}^{J} X_{ij}$ is an unbiased estimator of μ_i and has RMSE equal to $\sqrt{\sigma_E^2/J}$. However, $\bar{X}_{i\cdot}$ is inefficient because it does not share any information between households. Alternatively, we can fit a mixed model to equation 3 with a random intercept for each household to account for the a_i term. The predicted average from the mixed model is as follows:

$$\hat{\mu}_i = \frac{\widehat{\sigma_A^2}}{\widehat{\sigma_A^2} + \widehat{\sigma_E^2}/J}\bar{X}_{i\cdot} + \frac{\widehat{\sigma_E^2}/J}{\widehat{\sigma_A^2} + \widehat{\sigma_E^2}/J}\hat{\beta}_0. \quad (4)$$

This prediction, commonly called the empirical best linear unbiased predictor, has optimal[29] (i.e., smallest) RMSE for predicting μ_i. The prediction $\hat{\mu}_i$ in equation 4 is a combination of the household average, $\bar{X}_{i\cdot}$, and the study-wide average, $\hat{\beta}_0$.

The trade-off between these two values is controlled by the shrinkage factor $\frac{\widehat{\sigma_A^2}}{\widehat{\sigma_A^2} + \widehat{\sigma_E^2}/J}$. In the extreme case where there is no between-household difference ($\widehat{\sigma_A^2} = 0$), then $\hat{\mu}_i = \hat{\beta}_0$. In the other extreme in which measurements are made without error ($\widehat{\sigma_E^2} = 0$), then $\hat{\mu}_i = \bar{X}_{i\cdot}$. The mixed model shrinkage approach provides the benefits of both the household-level and shared information.

Impact of between-household predictor variables

Although model 3 provides a convenient illustration, in practice, there are typically measured factors that affect concentrations and vary between households or individuals. Examples include stove type, access to electricity for heating or cooking, physical characteristics of the home, and material assets, which can be associated with economic resources. Including these between-household factors leads to the mixed model:

$$X_{ij} = \beta_0 + a_i + \sum_{r=1}^{p} z_{ir}\beta_r + \epsilon_{ij}. \quad (5)$$

For the model in equation 5, the predicted long-term average is as follows:

$$\hat{\mu}_i = \hat{\beta}_0 + \hat{a}_i + \sum_{r=1}^{p} z_{ir}\hat{\beta}_r, \quad (6)$$

where we suppress the dependence on time interval $\mathcal{T}$ because there are no time-varying factors. The prediction in equation 6 remains optimal compared to the household average approach ($\bar{X}_{i\cdot}$).

If the between-household factors z_{ir} are known, it is always best to include them in the model. However, it can be instructive to consider what happens if z_{ir} is left out of the model (whether because it is unmeasured or because of analyst choice). In this case, σ_A^2 is inflated by a factor of β_r Var(z_{ir}), which would lead to a higher intraclass-correlation, ICC = $\frac{\widehat{\sigma_A^2}}{\widehat{\sigma_A^2} + \widehat{\sigma_E^2}}$. Thus, as more (fixed effect) predictors are added to the model, the ICC will go down and a larger proportion of variance is explained by within-person variation. In general, this means that as the number of fixed effects increases, the benefit of each additional repeated measure is reduced—although not completely; it remains beneficial (i.e., leads to lower RMSE) to make multiple measurements on each household. This also demonstrates how comparisons of ICC across models with different predictors can be misleading.

Impact of seasonal and other time-varying effects

We now consider the impact of smoothly varying seasonal trends on estimating long-term average concentrations. These effects are represented by the $f_i(t_j)$ term in the data-generating model 1. The mixed model for this case is as follows:

$$X_{ij} = \beta_0 + a_i + \sum_{r=1}^{p} z_{ir}\beta_r + \sum_{r=1}^{R} h_r(t_{ij})\eta_r + \varepsilon_{ij}, \quad (7)$$

where $h_r(t_{ij})$ are temporal basis functions, such as splines for calendar time. In some settings, effects of climate may be binned into categories (e.g., "cold/dry," "warm/dry," and "warm/wet" seasons[22]), which amounts to using a piece-wise constant temporal trend. This can simplify interpretation but may not well represent concentrations during transitional periods between seasons. Because of temporal variation in the exposure concentrations, the long-term average now depends upon a chosen time period given as follows:

$$\hat{\mu}_i(\mathcal{T}) = \hat{\beta}_0 + \hat{a}_i + \sum_{r=1}^{p} z_{ir}\hat{\beta}_r + \frac{1}{|\mathcal{T}|}\sum_{t \in \mathcal{T}} h_r(t)\hat{\eta}_r. \quad (8)$$

In practice, averaging over temporal effects is accomplished by making predictions at each time point t in $\mathcal{T}$ and averaging these predictions together.

Other time-varying factors can have short-term, nonsmooth variation, such as weekday differences in cooking patterns or differences in time of day of pollutant measurements. These terms, represented by $\sum_{s=1}^{q} v_{ijs}\gamma_s$, do not affect the predicted long-term average when the values are centered ($\sum_i \sum_j v_{ijs} = 0$), thus the predictor is still equation 8.

A major disadvantage of the household average approach is that it cannot adjust for either type of within-household time-varying factors, leading to considerable bias.

Multiple conditions

Sometimes there are multiple conditions under which a long-term average is desired. This is most likely to occur in crossover or stepped-wedge designs where each household has differing stove assignments at different times. In this case, a mixed model corresponding to (1) is fit. The predicted long-term average is as follows:

$$\hat{\mu}_i(\boldsymbol{w}, \mathcal{T}) = \hat{\beta}_0 + \hat{a}_i + \sum_{r=1}^{p} z_{ir}\hat{\beta}_r + \frac{1}{|\mathcal{T}|}\sum_{t \in \mathcal{T}}\left(h_r(t)\hat{\eta}_r + \sum_{k=1}^{K} w_{itk}\hat{\theta}_k\right). \quad (9)$$

The time period of interest $\mathcal{T}$ may differ for each condition or it may include multiple conditions to be averaged over. For example, in a study investigating exposure–response relationships with a crossover of stove type, it may be of interest to estimate the long-term exposure for the first year under one stove type and the second year under the other stove type. On the contrary, for an accountability study, it may be of interest to estimate the counterfactual of long-term exposure for the full 2-year period under each stove type.

The household average approach could also be used in this case, with separate averages calculated for each observed condition. However, this requires many more observations to achieve the same magnitude of uncertainty, because each condition has fewer measurements. For example, in a study with two measurements per household under each of two conditions, there are four total measurements per household but the RMSE of each condition-specific sample mean is $\sqrt{\sigma_E^2/2}$ and not $\sqrt{\sigma_E^2/4}$. Furthermore, the household average would still be less efficient than using a mixed model fit to the same data.

If there are large differences in concentrations between conditions, such as when comparing households with biomass-burning stoves to households with electric appliances, then it may be necessary to allow for the variance estimates to vary by condition. However, this would primarily impact inference on model parameters and not the point estimates used to predict long-term averages. Additional explorations of the role of condition-specific variance terms are beyond the scope of this article.

Comparison via simulation

To demonstrate the reductions in error offered by different modeling choices in different contexts, we conducted a series of simulations. Each simulation considered a two-group comparison (representing two stove types), but with different sampling strategies (Table 1) and different assumed temporal trends (Figure 1, eTable 1; http://links.lww.com/EE/A169). Designs 1 through 5

Table 1.

Summary of sampling strategies (i.e., designs) in the simulation.

Design	Study type	Study duration	*J*	Description
Design 1	Parallel	4–16 months	2–6	Measurements spaced 3 months apart (i.e., all participants with a visit in each of months 1, 4, 7, etc.). The time for each visit is randomly selected within a one-month window
Design 2	Parallel	16–28 months	2–6	Measurements are spaced 3 months apart, but the initial visit is randomly chosen within the first year (therefore, all months during the year contain visits). The time for each visit is randomly selected within a 1-month window.
Design 3	Parallel	12 months	4	Measurements were made in months 1, 3, 6, and 12. The time for each visit is randomly selected within a five-week window.
Design 4	Parallel	30 months	4	Measurements made in months 1, 3, 6, and 12, except initial visit for each household is randomly chosen within the first 18 months. The time for each visit is randomly selected within a five-week window
Design 5	Parallel	36 months	6	Measurements made in months 1, 3, 6, 9, 12, and 18, except initial visit for each household is randomly chosen within the first 18 months. The time for each visit is randomly selected within a five-week window
Design 6	Stepped-wedge	16 months	6	Either two or four measurements in each stove group. Measurements made every 3 months (i.e., months 1, 4, 7, 10, 13, and 16). The time for each visit is randomly selected within a 1-month window
Design 7	Stepped-wedge	28 months	6	Either two or four measurements in each stove group. Measurements are made every 3 months (i.e., months 1, 4, 7, 10, 13, and 16), except the initial visit is randomly chosen within the first year. The time for each visit is randomly selected within a 1-month window

represent sampling strategies for parallel trials, whereas designs 6 and 7 correspond to stepped-wedge trials. These were structured to represent sampling strategies employed in recent and ongoing studies[4,6] for both adult and perinatal health outcomes. The trends considered included flat (trend A) and decreasing (trend C) linear functions, large-scale seasonal variation with an annual peak and trough (trends B and D), and smaller-scale seasonal variation with multiple distinct seasons (trends E and F). All simulations had the same sample size (n=200 in each arm), stove effect ($\beta = 1$, corresponding to a 2.7-fold difference), and variances ($\sigma_A^2 = 0.5$, $\sigma_E^2 = 1$) and each was repeated 100 times. The stove effect and variance values were chosen to correspond to the magnitude of values observed in log-transformed $PM_{2.5}$ measurements of several studies.[24,27] In each setting, we calculated the RMSE of predicted long-term averages using the household average approach, a mixed model with no adjustment for stove group or time, a mixed model with adjustment for stove group but not time, and a mixed model with adjustment for stove group and temporal trend using natural splines (with 4 degrees of freedom [df] per year). For the parallel designs, the target period was the time from study entry until last visit for each unit. For stepped-wedge designs, two long-term averages were calculated for each unit: one under each stove type, averaged over the times the unit was assigned to that stove group. Simulations were conducted in R, version 4.0 (R Core Team, Vienna, Austria).

Figure 2 shows the RMSE for the different models under Design 1 (measurements every 3 months, all households on the same monthly schedule) and Trend A (no time effect), for studies with different numbers of repeated measures (and thus different study lengths). The benefit of the mixed model over the household average approach is evident in the smaller RMSE. When *J*=2 measurements are made for each household, the household average has an RMSE of 0.71, whereas the RMSE for predictions from the mixed model with stove adjustment is 0.5. The difference of 0.21 in RMSE, corresponds to a 23% difference in concentrations (on the nonlogarithmic scale). The mixed models with and without time adjustment (not included in Figure 2) perform the same because there is no temporal trend in the data-generating model. The mixed model that does not adjust for stove type has an RMSE of 0.55, which is still better than the household average approach but not optimal.

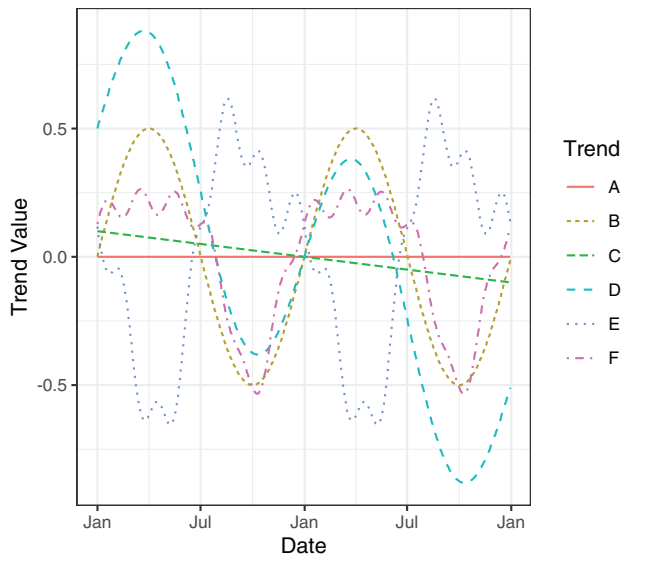

Figure 1. Temporal trend functions were used in the simulations. Equations for each trend are provided in eTable 1; http://links.lww.com/EE/A169.

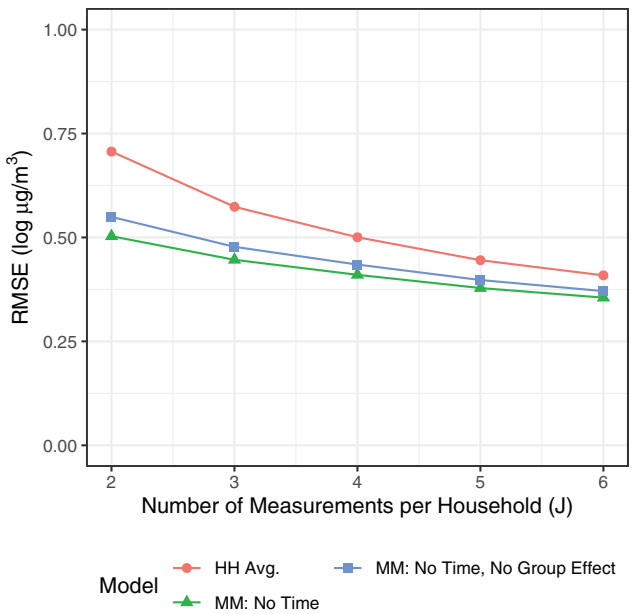

Figure 2. Root mean squared error (RMSE) of predicted long-term averages from simulation with design 1 (parallel design with measurements every 3 months, all households with baseline measurements in same month) and Trend A (no temporal trend). HH, household, MM, mixed model.

The differences in RMSE between models that do and do not adjust for stove group are sensitive to the magnitude of the stove effect, but that effect size has minimal impact on the relative RMSEs of models that adjust for stove. The relative ranking between the approaches persists (although the relative differences in RMSE diminish) as the number of repeated measures J increases (Figure 2) and when there is less within-household variance (eFigure 1; http://links.lww.com/EE/A169). The benefit of the mixed model approach compared to the household average approach persists for all other designs and settings.

To elucidate differences between different mixed models, Figure 3 shows the relative RMSE comparing predictions from mixed models without time to those that include time (both include adjustment for stove) for Designs 1 and 2. In addition to the different temporal trends, this comparison focuses on the difference in timing of the measurements—all initial measures starting at the same month for all households and repeated at the same 3-month intervals (Design 1) versus initial measures randomly assigned for each household to any month during the year and thereafter repeating every 3 months on different monthly schedules (Design 2). We see in most cases the relative RMSE is either very close to 1, representing equal error in the two approaches, or greater than 1, indicating that the model that adjusts for time performs better. The differences are typically larger under Design 2, indicating the need for time adjustment in studies where baseline measurements do not occur simultaneously. The higher relative RMSEs for Trends B, D, and E show that adjusting for time is most important when there are large temporal differences in the concentrations. Although the relative MSE goes up and down with the number of repeated measures J, the absolute RMSEs always decrease with each additional measurement (eFigures 2 and 3; http://links.lww.com/EE/A169).

Figure 4 shows similar results for Designs 3 through 7, which are structured to have a fixed number of repeated measures spread out over at least 1 year. For these designs, the differences in RMSE are either negligible or favor the model that adjusts for time. Because the true temporal trend impacting measurements is almost always unknown, these results demonstrate the benefit of always including a temporal trend in the mixed model.

Case study of Honduras data

To complement the simulation, we compared different modeling and sampling strategies in a case study of data from a cookstove intervention in Honduras. In this study, a stepped-wedge design was used to assess the impact of replacing traditional biomass burning stoves with "Justa" biomass-burning stoves that included an engineered combustion chamber and chimney.[4] Six repeated pollutant measurements were made for each household over the course of 3 years, approximately 6 months apart (n=230 households, N=1,207 observations). The Justa stove was installed after either the second or fourth visit. At each visit, each household had a 24-hour kitchen fine particulate matter ($PM_{2.5}$) measurement and each study participant had a 24-hour personal exposure measurement. The study was approved by the Colorado State University Institutional Review Board. Complete details of the sampling procedures and data collection have been previously described.[4,27]

The primary model for the Honduras data is a mixed model with log-transformed concentration measurements as the outcome. We include an indicator for stove type, a temporal spline (with 6 df to account for two major seasons in each year), and random effect for the household. Separate models were fit by measurement type (personal or kitchen). There were two conditions of interest: traditional stove and Justa stove, and predictions of long-term average were made for each household for each stove type. For the traditional stove, the averaging period was from first visit to last visit with the traditional stove, and for the Justa stove, the averaging period was from last visit with the traditional stove until the final visit.

We evaluated the approximate error in predicting the long-term average concentrations (on the log scale) under different modeling choices and numbers of measurements by fitting models that omit the temporal spline and the stove effect and by randomly subsampling the data to five, four, two, or one observation per household (instead of six). In each case, we

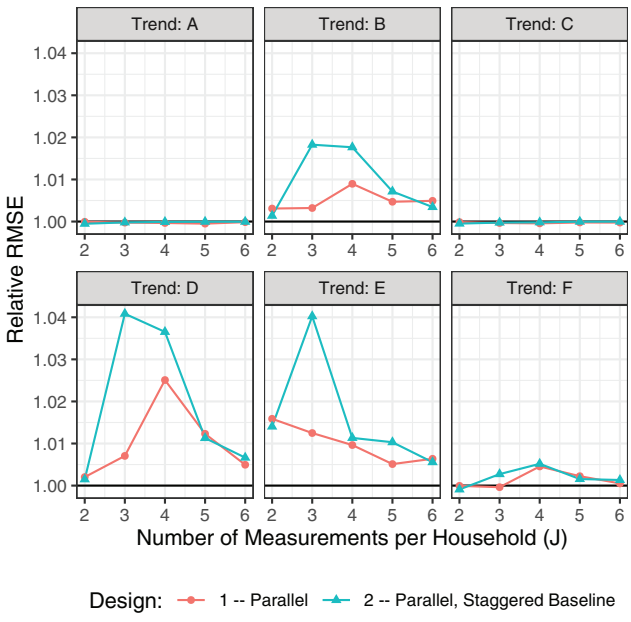

Figure 3. Relative root mean squared error (RMSE) of predicted long-term averages, comparing a mixed model that ignores time to a mixed model that adjusts for time. A relative RMSE greater than 1 indicates that the model that adjusts for time performs better. Panels correspond to the different trends in Figure 1.

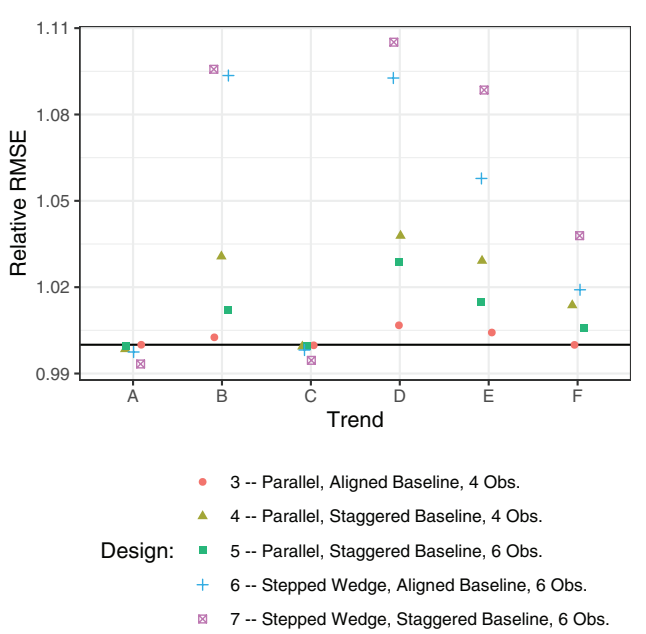

Figure 4. Results comparing the relative root mean squared error (RMSE) of predictions made from a model without time spline, to predictions from a mixed model with a time spline. A relative RMSE greater than 1 indicates that the model that adjusts for time performs better.

compare the error in the predicted long-term average from the subsetted data or alternative model to the predicted long-term average from the full dataset and primary model. This is repeated 500 times and differences are quantified by correlation and RMSE. Both correlation and RMSE are calculated on the logarithmic scale. We also conduct a similar analysis that is restricted to traditional stoves only. For that analysis, all observations use the same stove type so there is no stove indicator in the model, but the model still includes a temporal spline with 4 df.

A plot of the log-transformed personal measurements (Figure 5) shows some evidence of seasonality, although the magnitude is much smaller than the overall spread in the data. Summary statistics for the measurements and parameter estimates for the primary mixed model are provided in eTables 2 and 3; http://links.lww.com/EE/A169, respectively, and distributions of the predicted long-term averages are plotted in eFigure 4; http://links.lww.com/EE/A169. From Table 2, we see that the estimated RMSE when using only two measurements per household and a mixed model for prediction (RMSE of 0.252, corresponding to approximately 29% difference on a nonlogarithmic scale) is still better than using all six observations per household in the household-average approach (RMSE of 0.450, corresponding to approximately 57% difference on a nonlogarithmic scale). These trends still hold when the observations are subset to be in consecutive visits instead of randomly excluded (eTable 4; http://links.lww.com/EE/A169), and when using kitchen measurements instead of personal measurements (eTable 5; http://links.lww.com/EE/A169). The large correlations reflect the impact of the difference in stove group means, which leads to variability that can be explained by the model, even if the number of measurements is small. Similar trends in both mixed model predictions and household averages hold when we restrict the data to the participants who had four measurements with the traditional stove (Table 3).

Tables 2 and 3 demonstrate the necessity of adjusting for time. In all models that excluded the temporal spline, the RMSE was larger than in the corresponding model that did include time.

Discussion

We have presented approaches for predicting long-term average household air pollution concentrations from short-term measurements. These predictors, represented in equations 6 and 9, combine the effects of time-constant and time-varying predictor variables and participant-level random effects from a mixed model. We have demonstrated the importance of including temporal splines and transient effects in the mixed model while averaging over those components in the predictions. Building on prior work of McCracken et al,[22] who demonstrated the benefit of using mixed models for assessing long-term exposure, this work advances available methodology by modeling the impact of smoothly varying temporal trends (such as seasonality) in the concentrations and allowing for different time-varying conditions for each long-term average. This has major implications for stepped-wedge designs and other studies with cross-over, as well as parallel trials occurring in contexts with strong seasonal variation.

Ideally, studies of the health effects of chronic air pollution exposure could measure long-term exposure using low-burden, low-cost monitoring equipment worn for a long period of time. However, current monitors are prone to instrument error, face challenges with sustained power requirements, and can place considerable burden on individuals wearing them for weeks or months at a time. Until low-cost sensors are sufficiently accurate to facilitate long-term exposure assessment,[30] studies of chronic exposure will need to rely on a series of short-term exposure measurements. In our evaluation of predicted long-term average concentrations, we showed how accuracy is affected by the number of repeated measures. Making additional measurements reduced prediction error, but at a much lower rate when using the mixed model compared to a simple sample mean. The exact magnitude of this difference depends on the between- and within-household variance terms (σ_A^2 and σ_E^2, respectively), and is most apparent when the within-household variance is large.

These results have important implications for study design. In particular, the efficiency of the mixed model for predicting long-term average concentrations means that a good prediction can be obtained often with only a small number of repeated measures. This can be beneficial for studies assessing the effect of interventions on household air pollution concentrations, in addition to studies of health effects of exposure. At a minimum, two measurements per household are needed for estimating household random effects, but the benefit of having more than four measurements is small and resources might be better allocated to expanding the number of households included in the study. However, consideration should also be given to the assessment of health outcomes in the study, which might require more measurements than is necessary for estimating long-term exposure alone.

More accurate long-term average predictions can benefit downstream analyses that use the predictions in an exposure–response analysis with a health outcome. Reduced error in exposure values can reduce attenuation from classical measurement error,[19] although its exact effect will depend on each particular context. However, care should be taken when interpreting the impact of long-term average concentrations on health effects that occur early in a study, before the end of the averaging period.

There are several limitations to this analysis. First, factors affecting household air pollution concentrations can vary by location, context, instruments, and other factors, so the simulation and case study results will not perfectly represent the relationships in all settings. Nonetheless, they provide strong evidence for the use of mixed model predictions in place of household average concentrations and for the inclusion of temporal adjustment in all analyses. Our simulations were limited to a two-group comparison and a sample size of 200 in each study arm, but the same trends in relative performance occur with larger sample sizes and in settings with only one stove group. We did not consider multiple measurements made in consecutive days, which can reduce variability in short-term averages but does not always translate to better accuracy of long-term averages than randomly sampled days.[31] The simulations and case study also assumed that all relevant

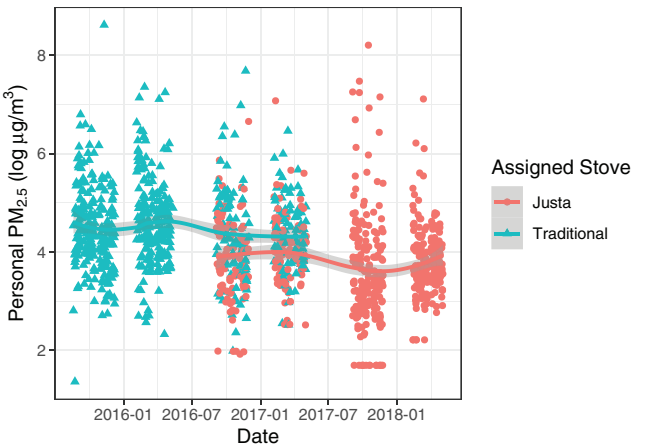

Figure 5. Log-transformed personal $PM_{2.5}$ measurements in the Honduras study. A smooth curve is shown separately by stove type.

Table 2.

Impact of different modeling choices and numbers of repeated measures on the accuracy of predicted long-term averages of personal $PM_{2.5}$ exposure in Honduras data.

Prediction type	Data	Correlation	RMSE
Mixed model	All (6 obs/hh)	(Reference)	(Reference)
	5 obs/hh	0.989	0.073
	4 obs/hh	0.958	0.141
	2 obs/hh	0.859	0.252
Mixed model with no time spline	All (6 obs/hh)	0.992	0.071
	5 obs/hh	0.980	0.102
	4 obs/hh	0.949	0.157
	2 obs/hh	0.852	0.259
Household average (by stove)	All (6 obs/hh)	0.847	0.450
	5 obs/hh	0.823	0.490
	4 obs/hh	0.810	0.510
	2 obs/hh	0.687	0.722
Single observation	1 obs/hh	0.696	0.715

Correlations and root mean squared error (RMSE) are calculated using the prediction from the mixed model with all data as the truth.
obs/hh, observations per household.

Table 3.

Impact of different modeling choices and numbers of repeated measures on the accuracy of predicted long-term averages of personal $PM_{2.5}$ exposure in Honduras data, when restricting to traditional stoves only.

Prediction type	Data	Correlation	RMSE
Mixed model	All (4 obs/hh)	(Reference)	(Reference)
	3 obs/hh	0.971	0.124
	2 obs/hh	0.911	0.218
Mixed model with no time spline	All (4 obs/hh)	0.999	0.029
	3 obs/hh	0.969	0.128
	2 obs/hh	0.911	0.219
Household average	All (4 obs/hh)	0.992	0.208
	3 obs/hh	0.962	0.291
	2 obs/hh	0.908	0.390
Single observation	1 obs/h	0.782	0.600

obs/hh, observations per household.

time-varying factors are included in the model. The simulations assumed that seasonality followed a periodic pattern, whereas in practice, seasonal and annual trends may follow nonperiodic patterns. However, nonperiodic patterns are easily accommodated by the smooth spline term in the model. The simulations did not consider simultaneous use of multiple stove types ("stacking"). If stove stacking is known, then the different combinations could be modeled as different stove types. If stacking is not measured, then the variability in modeled exposures will be increased, although the household-level random effect can account for some of the between-household differences due to stacking.

It is important to note that the RMSE and correlation in the case study are inherently optimistic because the same data are used in the comparison model and the "truth." This means that the RMSEs are likely smaller than would be observed with external validation data not used in both model fitting and model assessment. However, there are not sufficient data to split into two subsets. The cross-validation approach used by McCracken et al[22] split the data into groups of two measurements and considered the household average as "truth." But as discussed above, the household average is not efficient and can be biased in the presence of within-household variation due to factors such as seasonality. Furthermore, the stepped-wedge nature of the Honduras trial means that each household has only two measurements for one of the stove types (and four measurements for the other), and so a cross-validation approach would use just a single measurement as the cross-validation "truth." Although unbiased, this measure of the truth is so variable that comparing against it is of limited use. Although the measures presented here are optimistic, they do provide an illustration of the relative impact of different modeling choices and study designs.

Predicting long-term average exposure is key to obtaining quantitative evidence on the health effects of household air pollution and the benefits of potential interventions. The methods that we have outlined here provide accurate predictions in the presence of both between-person and within-person variation in concentrations and highlight impacts of study design that affect predicted exposures.

Conflicts of interest statement

The authors declare that they have no conflicts of interest with regard to the content of this report.

References

1. Murray CJL, Aravkin AY, Zheng P, et al. Global burden of 87 risk factors in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. *Lancet*. 2020;396:1223–1249.
2. Tielsch JM, Katz J, Zeger SL, et al. Designs of two randomized, community-based trials to assess the impact of alternative cookstove installation on respiratory illness among young children and reproductive outcomes in rural Nepal. *BMC Public Health*. 2014;14:1271.
3. Yip F, Christensen B, Sircar K, et al. Assessment of traditional and improved stove use on household air pollution and personal exposures in rural western Kenya. *Environ Int*. 2017;99:185–191.
4. Young BN, Peel JL, Benka-Coker ML, et al. Study protocol for a stepped-wedge randomized cookstove intervention in rural Honduras: household air pollution and cardiometabolic health. *BMC Public Health*. 2019;19:903.
5. Checkley W, Williams KN, Kephart JL, et al. Effects of a cleaner energy intervention on cardiopulmonary outcomes in Peru: a randomized controlled trial. *Am J Respir Crit Care Med*. 2020;203:1386–1397.
6. Clasen T, Checkley W, Peel JL, et al.; HAPIN Investigators. Design and rationale of the HAPIN study: a multicountry randomized controlled trial to assess the effect of liquefied petroleum gas stove and continuous fuel distribution. *Environ Health Perspect*. 2020;128:47008.
7. Jack DW, Asante KP, Wylie BJ, et al. Ghana randomized air pollution and health study (GRAPHS): study protocol for a randomized controlled trial. *Trials*. 2015;16:420.
8. Steenland K, Pillarisetti A, Kirby M, et al. Modeling the potential health benefits of lower household air pollution after a hypothetical liquified petroleum gas (LPG) cookstove intervention. *Environ Int*. 2018;111:71–79.
9. Lambe F, Jürisoo M, Lee C, Johnson O. Can carbon finance transform household energy markets? A review of cookstove projects and programs in Kenya. *Ener Res Soc Sci*. 2015;5:55–66.
10. Shupler M, O'Keefe M, Puzzolo E, et al. Pay-as-you-go liquefied petroleum gas supports sustainable clean cooking in Kenyan informal urban settlement during COVID-19 lockdown. *Appl Energy*. 2021;292:116769.
11. Clark ML, Peel JL, Balakrishnan K, et al. Health and household air pollution from solid fuel use: the need for improved exposure assessment. *Environ Health Perspect*. 2013;121:1120–1128.
12. Bates MN, Pokhrel AK, Chandyo RK, et al. Kitchen PM2.5 concentrations and child acute lower respiratory infection in Bhaktapur, Nepal: the importance of fuel type. *Environ Res*. 2018;161:546–553.
13. Li X, Clark S, Floess E, Baumgartner J, Bond T, Carter E. Personal exposure to PM2.5 of indoor and outdoor origin in two neighboring Chinese communities with contrasting household fuel use patterns. *Sci Total Environ*. 2021;800:149421.
14. Chillrud SN, Ae-Ngibise KA, Gould CF, et al. The effect of clean cooking interventions on mother and child personal exposure to air pollution: results from the Ghana Randomized Air Pollution and Health Study (GRAPHS). *J Expo Sci Environ Epidemiol*. 2021;31:683–698.
15. Liao J, Kirby MA, Pillarisetti A, et al.; HAPIN Investigators. LPG stove and fuel intervention among pregnant women reduce fine particle air pollution exposures in three countries: pilot results from the HAPIN trial. *Environ Pollut*. 2021;291:118198.

16. Smith KR, McCracken JP, Thompson L, et al. Personal child and mother carbon monoxide exposures and kitchen levels: methods and results from a randomized trial of woodfired chimney cookstoves in Guatemala (RESPIRE). *J Expo Sci Environ Epidemiol*. 2010;20:406–416.
17. Pope D, Diaz E, Smith-Sivertsen T, et al. Exposure to household air pollution from wood combustion and association with respiratory symptoms and lung function in nonsmoking women: results from the RESPIRE trial, Guatemala. *Environ Health Perspect*. 2015;123:285–292.
18. Quinn AK, Ae-Ngibise KA, Jack DW, et al. Association of Carbon Monoxide exposure with blood pressure among pregnant women in rural Ghana: evidence from GRAPHS. *Int J Hyg Environ Health*. 2016;219:176–183.
19. Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM. *Measurement Error in Nonlinear Models*. 2nd Edition. Chapman & Hall/CRC; 2006.
20. Johnson MA, Steenland K, Piedrahita R, et al.; HAPIN Investigators. Air Pollutant Exposure and Stove Use Assessment Methods for the Household Air Pollution Intervention Network (HAPIN) Trial. *Environ Health Perspect*. 2020;128:47009.
21. Lee AG, Kaali S, Quinn A, et al. Prenatal household air pollution is associated with impaired infant lung function with sex-specific effects. Evidence from GRAPHS, a cluster randomized cookstove intervention trial. *Am J Respir Crit Care Med*. 2019;199:738–746.
22. McCracken JP, Schwartz J, Bruce N, Mittleman M, Ryan LM, Smith KR. Combining individual- and group-level exposure information: child carbon monoxide in the Guatemala woodstove randomized control trial. *Epidemiology*. 2009;20:127–136.
23. Grajeda LM, Thompson LM, Arriaga W, et al. Effectiveness of gas and chimney biomass stoves for reducing household air pollution pregnancy exposure in Guatemala: sociodemographic effect modifiers. *Int J Environ Res Public Health*. 2020;17:E7723.
24. Keller JP, Katz J, Pokhrel AK, Bates MN, Tielsch J, Zeger SL. A hierarchical model for estimating the exposure-response curve by combining multiple studies of acute lower respiratory infections in children and household fine particulate matter air pollution. *Environ Epidemiol*. 2020;4:e119.
25. Helen GSt, Aguilar-Villalobos M, Adetona O, et al. Exposure of pregnant women to cookstove-related household air pollution in urban and periurban Trujillo, Peru. *Arch Environ Occup Health*. 2015;70:10–18.
26. Hu W, Downward GS, Reiss B, et al. Personal and indoor PM2.5 exposure from burning solid fuels in vented and unvented stoves in a rural region of China with a high incidence of lung cancer. *Environ Sci Technol*. 2014;48:8456–8464.
27. Benka-Coker ML, Young BN, Keller JP, et al. Impacts of the wood-burning Justa cookstove on fine particulate matter exposure: a stepped-wedge randomized trial in rural Honduras. *Sci Total Environ*. 2021;767:144369. doi: 10.1016/j.scitotenv.2020.144369
28. Chen C, Zeger S, Breysse P, et al. Estimating indoor PM2.5 and CO concentrations in households in Southern Nepal: the Nepal Cookstove Intervention Trials. *PLoS One*. 2016;11:e0157984.
29. McCulloch CE, Neuhaus JM. Prediction of random effects in linear and generalized linear models under model misspecification. *Biometrics*. 2011;67:270–279.
30. Curto A, Donaire-Gonzalez D, Barrera-Gómez J, et al. Performance of low-cost monitors to assess household air pollution. *Environ Res*. 2018;163:53–63.
31. Pillarisetti A. Inspecting what you expect: applying modern tools and techniques to evaluate the effectiveness of household energy interventions. UC Berkeley. Published online 2016. https://escholarship.org/uc/item/7hw5z2w2