Sample Size Estimation for Random-effects Models

Balancing Precision and Feasibility in Panel Studies

Scott Weichenthal,[a,b] *Jill Baumgartner,*[a,c] *and James A. Hanley*[a]

Abstract: Panel study designs are common in environmental epidemiology, whereby repeated measurements are collected from a panel of subjects to evaluate short-term within-subject changes in response variables over time. In planning such studies, questions of how many subjects to include and how many different exposure conditions to measure are commonly asked at the design stage. In practice, these choices are constrained by budget, logistics, and participant burden and must be carefully balanced against statistical considerations of precision and power. In this article, we provide intuitive sample size formulae for the precision of regression coefficients derived from panel studies and show how they can be applied in planning such studies. We show that there are five determinants of the precision with which regression coefficients can be estimated: (1) the residual variance of the responses; (2) the variance of the slopes; (3) the number of subjects; (4) the number of measurements/subject; and (5) the within-subject range of the exposure values "*X*" at which the responses are measured. The planning of such studies would be greatly improved if investigators regularly reported all of the variance components in fitted random-effects models: currently, literature values for the relevant variance parameters are often not readily available and must be estimated through pilot studies or subjective estimates of "reasonable values."

(*Epidemiology* 2017;28: 817–826)

Panel study designs that measure subjects' responses under two or more different conditions are often used to evaluate the short-term health effects of continuous environmental exposures, including outdoor/household air pollution[1,2] and ambient temperature.[3] In planning such studies, questions of how many subjects to include and how many different exposure conditions to measure are common questions in the design stage. In practice, these choices are constrained by budget, logistics, and participant burden and must be carefully balanced against statistical considerations of precision and power.[4]

Unfortunately, even the most comprehensive and "user friendly" sample size programs do not explicitly handle panel studies as often applied in environmental epidemiology. Specifically, most deal only with between-subject comparisons rather than within-subject contrasts, which are more common in panel studies in environmental epidemiology. In addition, sample size software generally does not intuitively explain how the inputs are combined. As we have noted elsewhere,[5,6] the formulae for regression models are not always intuitive. Moreover, when it comes to regression-based *within*-subject comparisons in panel studies, the lack of guidance on where to find the pieces of information needed as inputs to the formulae is particularly frustrating.

In this article, we provide intuitive sample size formulae for the precision of regression coefficients derived from panel studies and demonstrate their application in the planning of such studies. We show that there are five determinants of the precision with which regression coefficients can be estimated: (1) the residual variance of the responses, even after the intersubject variability has been removed by using random intercepts; (2) the variance of the slopes, if it is demonstrably large enough to have to include in the model; (3) the number of subjects; (4) the number of measurements per subject; and (5) the within-subject range of the exposure values "*X*" at which the responses are measured. We discuss methods to obtain prestudy estimates of important variance components. The R codes used in the examples below are provided in the eAppendix; http://links.lww.com/EDE/B249. An online tool is also available to implement the sample size calculations discussed below (https://corinne-riddell.shinyapps.io/mcgilleboh-samplesizecalculator).

THE CENTRAL FORMULAE, ESTABLISHED FROM FIRST PRINCIPLES

Our heuristics will begin with the same type of paired-samples example used in 1908 by "Student" (W.S. Gosset)[7] to illustrate the use of what is now called the t-distribution.

Submitted 30 August, 2016; accepted 25 July, 2017.

From the [a]Department of Epidemiology, Biostatistics, and Occupational Health, McGill University, Montreal, Canada; [b]Gerald Bronfman Department of Oncology, McGill University, Montreal, Canada; and [c]Institute for Health and Social Policy, McGill University, Montreal, Canada.

This work was supported by a Collaborative Health Research Projects Grant (CIHR/NSERC).

The authors report no conflicts of interest.

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

Correspondence: Scott Weichenthal, Department of Epidemiology, Biostatistics, and Occupational Health and Gerald Bronfman Department of Oncology, McGill University, 1020 avenue des Pins Ouest, Montreal, QC H3A 1A2, Canada. E-mail: scott.weichenthal@mcgill.ca.

ISSN: 1044-3983/17/2806-0817
DOI: 10.1097/EDE.0000000000000727

Student took his illustrative data from a report by Cushny and Peebles,[8] who compared hours of sleep ("*Y*") under a "usual" and multiple experimental conditions in the same $n = 10$ subjects (R code for the sleep duration example is provided in the supplemental digital content; http://links.lww.com/EDE/B250). For each subject, the duration of sleep was recorded over several nights under each condition, but only the mean duration per subject, and the number of nights per subject (m), were reported by Cushny and Peebles[8] (Student omitted the m's). Later, when studying driver reaction times, we will again consider sleep but use the duration of sleep *deprivation* as the "exposure" or "*X*" variable.

To begin our thought experiment, we restrict ourselves to the *usual* (i.e., $X = 0$) condition and take as our estimand the mean duration of sleep (μ_Y) over a large number of subjects (n), each mean being taken over a large number of nights per subject (say m =366):

$$\mu_Y = \frac{1}{n}\sum_{1}^{n}\left(\frac{1}{m}\sum_{1}^{m} Y\right).$$

The hypothetical distribution of these n subject-specific values is shown in the "interperson" column in the left half of Figure 1. If a random sample of say $n = 60$ of these subject-specific values could be established without error (i.e., by obtaining and averaging all 366 measurements for each of the $n = 60$ subjects), then one could estimate μ_Y as 1/60th of their sum. We could also calculate the standard error (SE) of this estimate as the sample standard deviation (SD) of these 60 values divided by the square root of 60 and use it to obtain a confidence interval for μ_Y. For didactic purposes, the individual values in the interperson column of Figure 1 were taken to have an intersubject variance of $\sigma^2_{\text{inter}} = 1$ hour^2, and so the square of the SE for the estimate would be approximately $\sigma^2_{\text{inter}}/60$ hours^2. This formula is derived in the eAppendix; http://links.lww.com/EDE/B249 along with those for more complex real-world situations.

In practice, the feasibility of collecting 366 measurements from 60 subjects (total = 21,960 individual measurements) would be limited by high cost and great burden to study subjects and would likely waste resources. In this case, the following questions come to mind: (1) What could be achieved with just 60 measurements in total? (2) Would it be better to obtain 1 measurement for each of 60 subjects or 2 measurements for 30 subjects, or 6 for each of 10, or even 30 for each of 2? In practice, the researcher must strike a balance between minimizing the effort required for data collection and having a reasonably small SE for the estimator. The latter depends both on the amount of interindividual variability and the amount of intraindividual variability, shown in the "intraperson" column in the left half of Figure 1. For simplicity, in our example, the "intraperson variability" is taken to be of the same magnitude ($\sigma^2_{\text{intra}} = 0.16$ hours^2) for every subject. The trade-off between effort and SE can be seen by calculating the SEs of the various estimators. The squared SE of the estimate of μ_Y can be expressed generically as:

$$\text{SE}^2 = \left(\sigma^2_{\text{inter}} / n\right) + \left(\sigma^2_{\text{intra}} / nm\right)$$

where σ^2_{inter} is intersubject variance (equal to 1 hour^2 in this example), σ^2_{intra} is the intrasubject variance (equal to 0.16 hour^2 in this example), n is the total number of subjects measured, and m is the number of measurements per subject.

Using the variances given above, we can estimate SEs for various configurations shown in Table 1. The guiding principle that emerges is that if the measurement varies considerably between people, then we need to counteract this by taking a larger number of people. On the other hand, if the average level varies little from person to person and the main sources of variability are within subjects, it may be acceptable to estimate the mean population level using many measurements on fewer people.

We now move on to the "usual" versus "experimental" *contrast* and take our estimand to be $\mu_{Y|X=1} - \mu_{Y|X=0}$. Again, for now, we remain hypothetical and use R-generated data. The distribution of the exact differences is shown in the center column of Figure 1. Suppose now that the study budget is limited to a total of 120 measurements: 60 in $X = 0$ and 60 in $X = 1$. If 30 subjects each measured twice is an acceptable compromise for each half of the X contrast, would it make more statistical sense to have the 30 subjects measured in the $X = 1$ condition be distinct from, or the same as, the 30 measured in the $X = 0$ condition? If independent samples are used, the variance for the estimated difference in means is:

$$\left(\sigma^2_{\text{inter}} / 30 + \sigma^2_{\text{DY/DX}} / 30 + \sigma^2_{\text{intra}} / 60\right) + \left(\sigma^2_{\text{inter}} / 30 + \sigma^2_{\text{intra}} / 60\right).$$

If self-matched samples are used, the two interperson variance components are removed by design, and the variance for the estimated difference in means is $(\sigma^2_{\text{DY/DX}}/30) + (2\,\sigma^2_{\text{intra}}/60)$, where $\sigma^2_{\text{DY/DX}}$ is the variance of individual slopes (i.e., 0 if a common slope).

Thus, there is a statistical advantage to pairing, but each subject would have to be measured 4 times instead of twice. This reduction in variance can be substantial owing to gains in efficiency achieved by removing intersubject variation.

Our variance expression can now be put into a more generic sample size formula for the number of subjects, n, each measured $m/2$ times in each condition, or m times in total. Suppose we are going to test the average difference of the n individual observed differences against the null hypothesis (H_0) that $\mu_{Y|X=1-Y|X=0} = 0$, using = alpha = 0.05 ($Z_\alpha = 1.96$). Suppose we wish to have power of 80% ($Z_\beta = 0.84$) against the alternative hypothesis (H_{alt}) that $\mu_{Y|X=1-Y|X=0} = \Delta$. Before we can give the answer, we need to introduce one additional parameter: the *range* of the m values of X. Intuitively, the

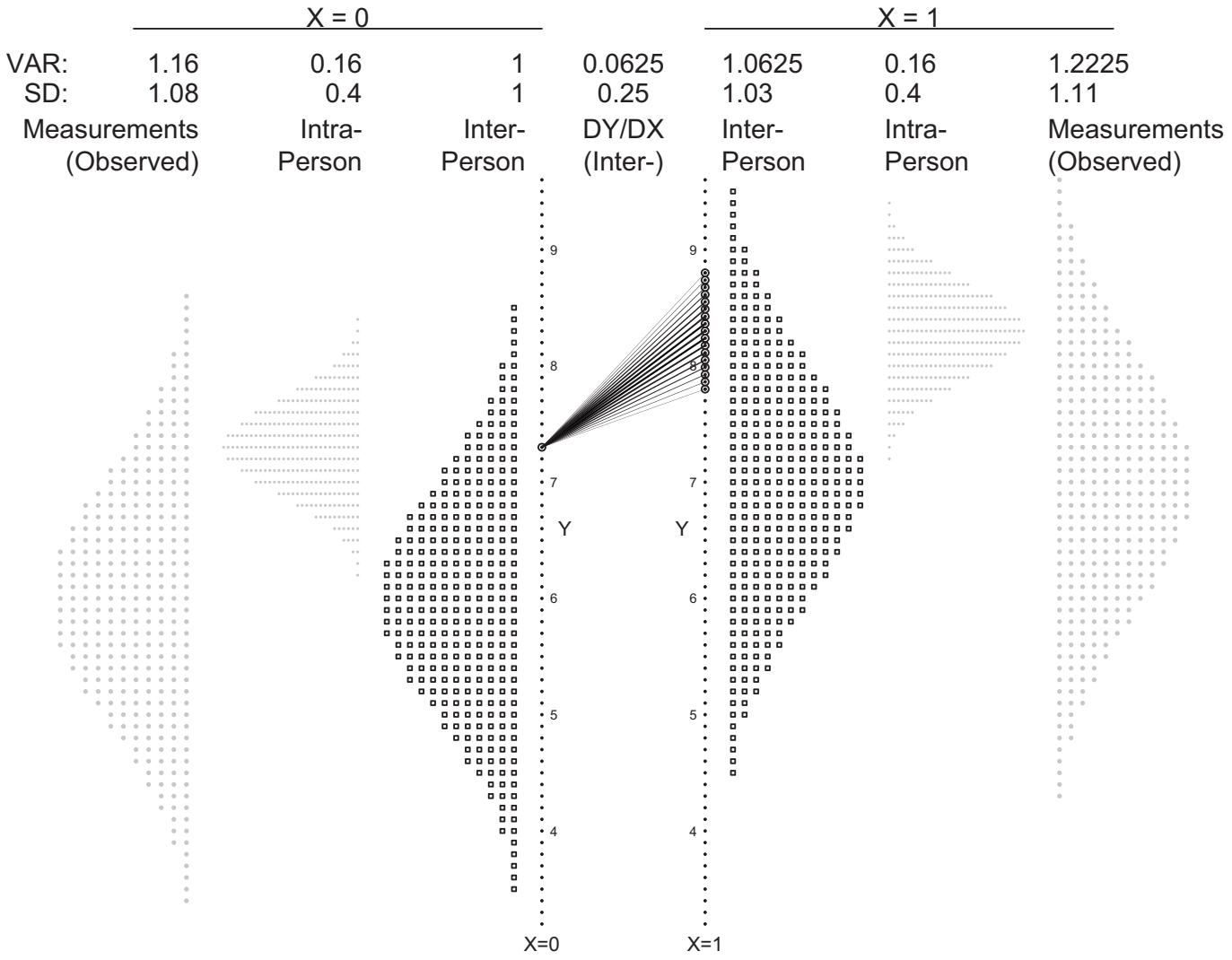

FIGURE 1. Hypothetical inter- and intraperson variation in sleep duration, under two conditions ($X = 0$ and $X = 1$). The distribution shown as small black squares just to the left of the vertical line at $X = 0$ refers to the long-term values for each person and, therefore, shows the pure *interperson* variation. To its left is a distribution of the amounts by which the durations of sleep on different nights for one specific person (arbitrarily taken to be a long-term average of 7.3 hours) vary within this person ("intra"). One can also think of this intradistribution as being centered at zero, so that when it is added to the interdistribution, it produces at the far left a distribution of measurements. This distribution would be observed if each subject was measured once, each on a different randomly selected night. The rays in the center show by how much different persons' averages in the $X = 1$ condition differ from their average in the $X = 0$ condition, even if they all have the same average in that $X = 0$ condition. To keep the diagram from being too cluttered, only the differences for those subjects with a mean Y of 7.3 in the $X = 0$ condition are shown. For simplicity, the distributions for subjects with values other than 7.3 are assumed to be the same, even though in practice they might vary with the baseline level. The distribution at the far right shows what would be observed if each subject was measured once in the $X = 1$ condition, each on a different randomly selected night.

greater the range of exposures, the more precisely we can measure the slope. The formula for the squared SE of the slope is typically presented as:

$$\mathrm{SE}^2 = \frac{\sigma^2_{\text{residuals}}}{\sum(x-\bar{x})^2}$$

The denominator of this expression hides the fact that the SE is made smaller by having more X values (distinct or not) in the sum and by having these X values as spread out as possible from the mean (i.e., having large squared deviations from the mean). A more useful way to write the denominator is:

$$\begin{aligned}\sum(x-\bar{x})^2 &= m\times(\textit{average or mean of the squared deviations})\\ &= m\times MS_X\end{aligned}$$

Thus, in our example, with each subject measured $m/2$ times in the $X = 0$ and $m/2$ in the $X = 1$ condition, or m times

TABLE 1. Standard Errors (SEs) of Selected Estimators of μ_Y, Each Involving a Different Number of Subjects (n) and Measurements per Subject (m), When the Inter- and Intraperson Variances Are 1 and 0.16 hour^2

No. Subjects (n)	60	60	30	10	2	1
No. measures per subject (m)	366	1	2	6	30	60
Intersubject component (σ^2_{inter}/n)	1/60	1/60	1/30	1/10	1/2	1/1
Intrasubject component ($\sigma^2_{intra}/n\,m$)	0.16/21,960	0.16/60	0.16/60	0.16/60	0.16/60	0.16/60
$SE^2 = (\sigma^2_{inter}/n) + (\sigma^2_{intra}/n\,m)$	0.0167	0.0193	0.0360	0.1027	0.5027	1.0027
SE	0.13	0.14	0.19	0.32	0.71	1.00

in total, the mean (i.e., $\bar{x}$) is 0.5. Therefore, the mean squared deviation (MS_x) is:

$$MS_X = \left[(m/2)(0-0.5)^2 + (m/2)(1-0.5)^2\right]/m = 0.25.$$

If slopes are assumed to be common, then the sample size equation is simply:

$$n \geq (1.96+0.84)^2 \times \left(\sigma^2{}_{\text{intra}} / [m \times (0.25)]\right) / \Delta^2,$$

whereas if slopes are assumed to be variable, then the equation has an additional component:

$$n \geq (1.96+0.84)^2 \times \left(\sigma^2{}_{\text{DY/DX}} + \sigma^2{}_{\text{intra}} / [m \times (0.25)]\right) / \Delta^2$$

where n is the number of subjects in the sample, Δ is the average within-subject difference (in this case slope) we wish to detect.

But what if X values were not binary, but took values along a continuum, possibly different for each subject? As before, if we let MS_X denote the mean squared distance between the subject's X's and the mean of the subject's X's, the sample size formula becomes:

$$n \geq (1.96+0.84)^2 \times \left(\sigma^2{}_{\text{DY/DX}} + \sigma^2{}_{\text{intra}} / [m \times \text{MS}_X]\right) / \beta^2$$

where β is the slope (i.e., strength of association) we wish to detect.

RANDOM-EFFECTS MODELS FOR PANEL STUDIES: THE STRUCTURE AND DETERMINANTS OF THE SE

Example 1: Dichotomous Exposure (i.e., X = 1 or 0)

We now focus on real data and on two exposure conditions studied in the original Cushny and Peebles article, designating the usual or "control" nights as $X = 0$ when no hypnotic was given and the "experimental" nights as $X = 1$ when the drug levo-hyoscine hydrobromate was administered. Ten subjects were measured under both control and experimental conditions: the numbers of within-subject measurements in the $X = 0$ condition ranged from $m = 7$ to 9 (84 subject-nights) and in the experimental ($X = 1$) condition from $m = 3$ to 6 (47 subject-nights). In total, there were observations for 366 subject-nights. The "intrasubject intracondition" variability was not reported by Cushny and Peebles, but using simulations, we will investigate what difference it would have made had they reported each of the 366 individual observations. The 20 reported durations (means), the same ones used by Student and available in the sleep data set in R, are shown in Figure 2 as 10 pairs of colored dots.

In a paired t test analysis, the 10 within-person differences ($X = 1$ vs. $X = 0$) are first computed (Figure 2). Their mean is 2.33 hours and their SD is 2.00 hours. The resulting test statistic ($t = [2.33 - 0]/[2.00/\sqrt{(10)}] = 2.33/0.63 = 3.68$) would be exceeded in only a very small proportion ($P = 0.0025$) of draws from a "Student's" t_{9df} distribution. A modern-day Student might instead analyze these 20 data points using a random-effects linear model. Since we do not have the individual measurements for the 131 nights, only a random-intercept model (a separate intercept for each subject, but a common slope) can be fit. Relevant output from the R lme4 software package is provided in eAppendix, eTable 1; http://links.lww.com/EDE/B249. The estimated slope of 2.33 hours, the SE of 0.63 hours, and the t ratio (shown in the fixed-effects section of Table 1) agree perfectly with Student's analysis. However, *two important variance estimates* (highlighted in the random effects section of eTable 1) tend not to be reported when models of this kind are used: the *full* fitted model is shown at the top left of Figure 2. The fitted model is also shown graphically, starting with the "fixed effects" model shown as a black line with intercept 3.25 hours and a slope of 2.33 hours. The ten fitted "random effects" are shown as solid colored lines, one per subject, parallel to the central one. The SD of the population of intercepts, of which these 10 are considered a random sample, is estimated to be $\sigma_{inter} = 0.98$ hours. The residual variation (presumed to be entirely intrasubject) is estimated to have an amplitude of $\sigma_{intra} = 1.42$ hours.

If the "intrasubject" variability in sleep duration had been reported by Cushny and Peebles, how would it have changed the estimated slope and SE, and what model should be fit? Two different scenarios are considered in Figure 3: one where the intrasubject variability is quite low and the other where it is more substantial. The variability was artificially added by us in such a way that the observed mean over the number of nights involved was the same as that reported. In the low intrasubject variability scenario, it is quite clear that subjects have different slopes, whereas it is more difficult to

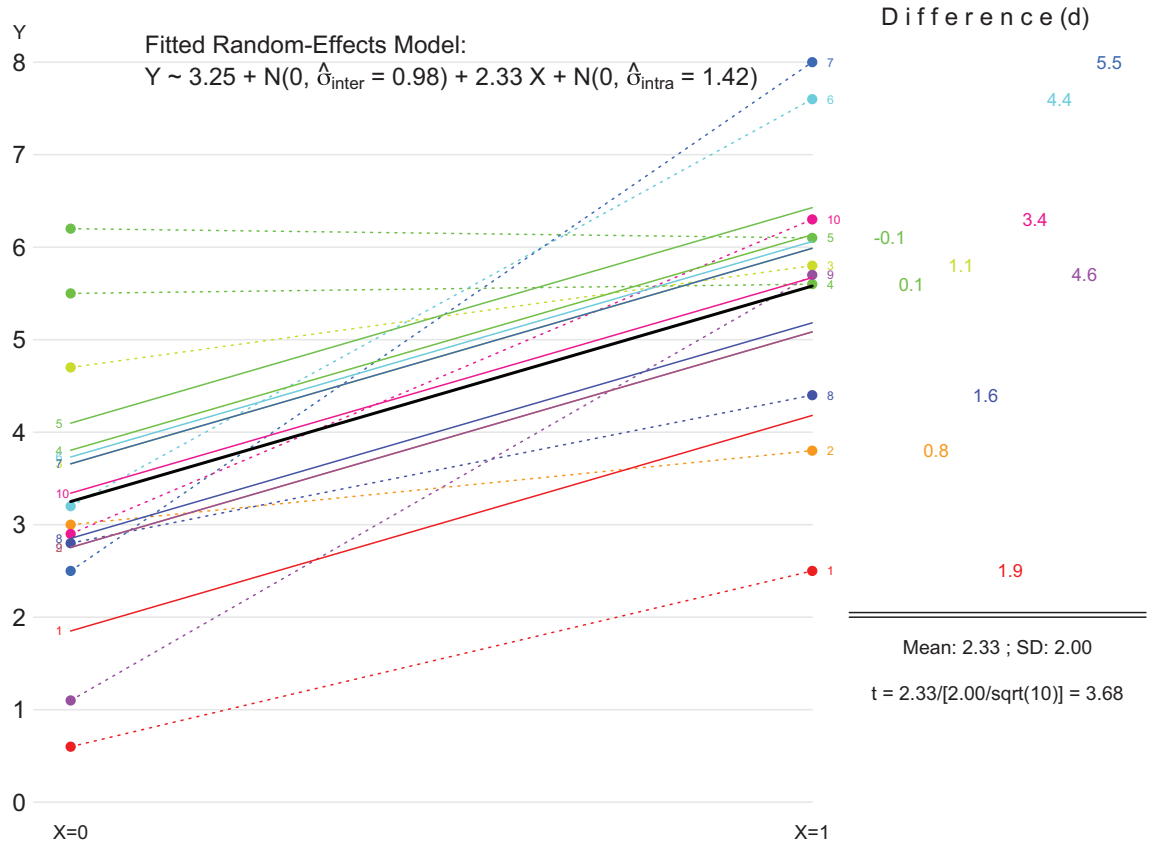

FIGURE 2. The 20 reported average sleep durations from the Cushny and Peebles article, shown as 10 pairs of colored dots, with the subject number shown at the right end of the dotted line joining the pair of observed averages. The analysis using a paired *t* test is shown at the right. The fitted random-effects model, with random intercepts, but a common slope, is shown both in equation form and as a single black line (the average) together with 10 solid parallel lines of different colors; the same colors identify the observed averages for each subject. The subject number is shown at the left end of the solid (fitted) line. Figure is available in color online.

distinguish the "different slopes for different individuals" scenario from the "common slope" scenario when intrasubject variation is large.

Models with random intercepts only and both random intercepts and random slopes are presented in eAppendix, eTable 2; http://links.lww.com/EDE/B249 using the data presented in Figure 3. Whether the intrasubject variability was simulated to be low or high, the SE of the slope of 2.33 hours was much lower in the random intercept model that forced the subject-specific slopes to be equal but remained at 0.63 hours when the more appropriate random slope model was fitted. By itself, this comparison of the two models cannot tell us which SE is more appropriate, but a further simulation does provide some clarity (eAppendix, eTable 3; http://links.lww.com/EDE/B249). In this simulation, we increased each *m* from what it actually was to 100 times that value, but the data set of 13,100 is still based on just 10 subjects under the usual and experimental conditions. Despite greatly increased evidence that the slopes cannot have a common value, the SE of the assumed-common slope is reduced. Meanwhile, the SE for the 2.33-hour average slope estimate from the random slope model remains at 0.63 hours, reflecting the fact that we still only have 10 subjects. The only way to reduce this SE is by recruiting more subjects into the study.

Example 2: *X* Values on a Continuum and Under Full Investigator Control

What if the *X* values were on a continuum rather than dichotomous? The "sleepdata" data set in the lme4 package in R provides an illustrative example. In this database of 18 subjects,[5] up to day 0 subjects had their normal amount of sleep; starting that night, they were restricted to 3 hours of sleep per night. The number of consecutive days (0–9) with just 3 hours of sleep per night is the "*X*" variable. The average reaction time (milliseconds) measured on a series of tests administered on the 10 study days is the "*Y*" (R code for the reaction time example is available in the supplemental digital content; http://links.lww.com/EDE/B251).

Figure 4 shows the subject-specific data along with a separate regression line fitted to each. As noted in the chapter devoted to this example,[9] the ordering of the panels is quite strategic and allows us to judge visually that a "subject's rate of change in reaction time [slope] does not seem to be strongly

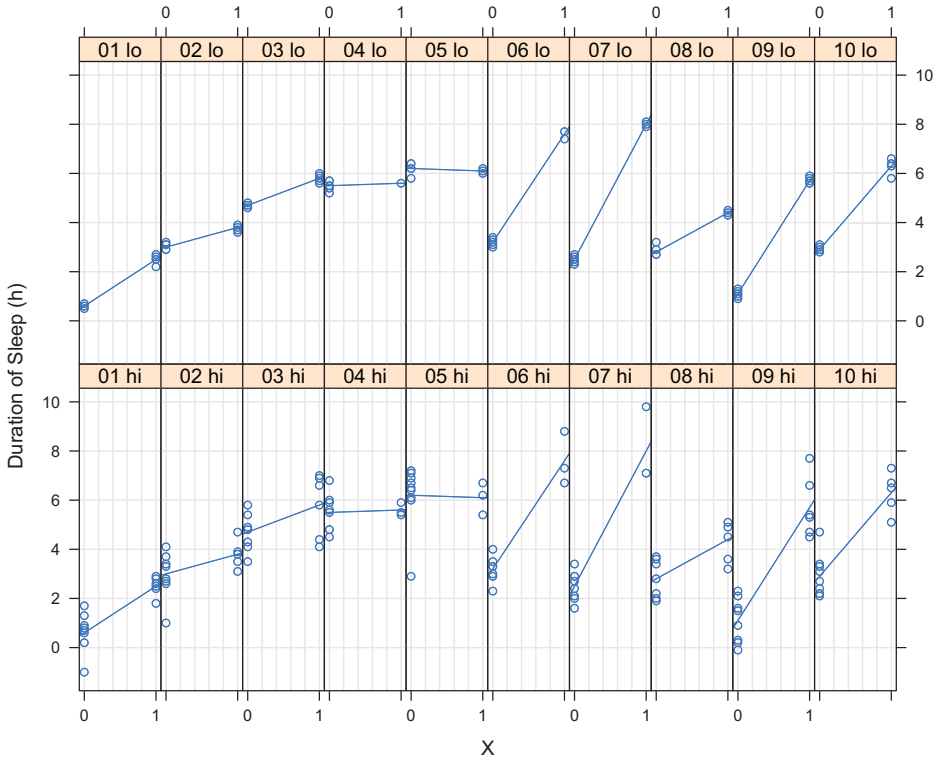

FIGURE 3. Simulated intrasubject intracondition data for reported subject-specific sleep durations in 10 subjects in two conditions, $X = 0$ and $X = 1$, with intrasubject variability added for each of the nights that contributed to the average values in Figure 1, without altering the 20 reported means. Subjects are numbered 01 to 10 as in the Cushny and Peebles article. The values in the top row represent a low intrasubject variability scenario, while those in the bottom row have more substantial intrasubject variability. Figure is available in color online.

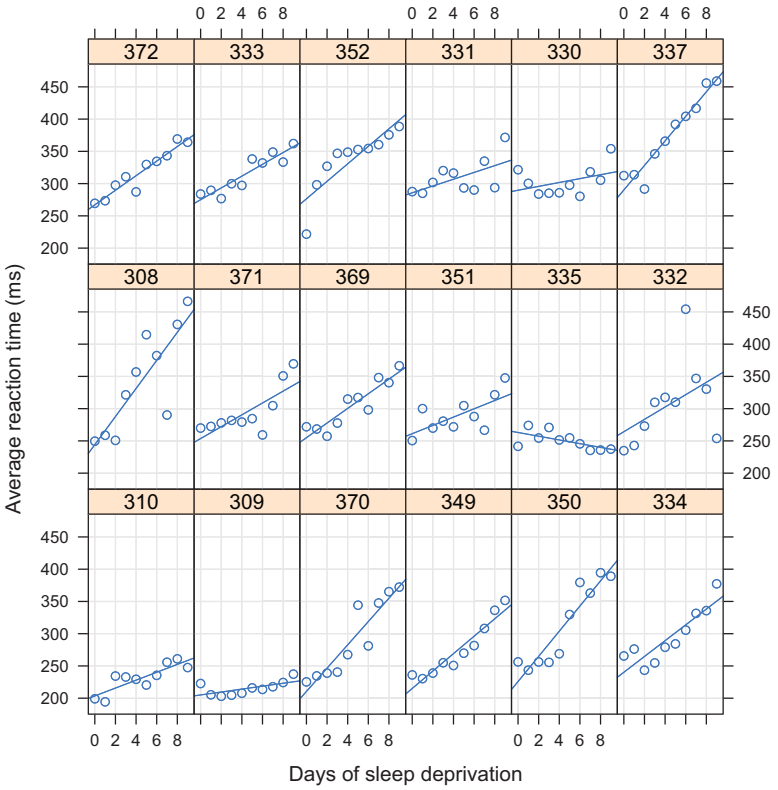

FIGURE 4. For each of 18 subjects, each identified by a three-digit number, the average reaction time (ms) on each of consecutive 10 days. Day 0 was preceded by several nights of normal sleep; days 1–9 were preceded by nights with just 3 hours of sleep. Panels are ordered from bottom left to top right according to the intercept of a regression line fitted just to that subject's 10 data points. Reprinted with permission from Bates 2010.[9] Figure is available in color online.

related to the subject's initial reaction time [intercept]." The author notes that "…there is considerable [intersubject] variation both in the initial reaction times [intercepts] and in the daily rate of increase in reaction time [slopes]." Thus, he does not even consider simpler "common intercept" and/or "common slope" models. However, for completeness, we show all

of the models in eAppendix, eTable 4; http://links.lww.com/EDE/B249 in practice, only the two models that contain random intercepts are generally applicable. After working through the calculations for the SEs that include two variance components, one will notice that the dominant contribution (more than 80%) comes from the variability of the slopes rather than the residuals. This makes sense in this particular application because it is quite obvious from the raw data in Figure 4 that different subjects react quite differently to increasing amounts of sleep deprivation and that the average slope in a different 15 subjects might be quite different.

The unpredictable portions of the SE formulae are the variance components. In this example using balanced data, the mean of the 10 squared subject-specific distances of X values from their mean (i.e., $MS_X = [(0 - 4.5)^2 + (1 - 4.5)^2 + \ldots + (8 - 4.5)^2 + (9 - 4.5)^2]/10 = 8.25$) and the product ($m \times MS_X = 82.5$) was the same for each of the 18 subjects and was known in advance. Had each subject only been measured on days 0, 2, 4, 6, and 8, this "sum of the squares" of the X's for these 5 days would have been just $(0 - 4)^2 + (2 - 4)^2 + (4 - 4)^2 + (6 - 4)^2 + (8 - 4)^2 = 40$, and the five measurements would have reduced the relevant (but not necessarily the biggest) portion of the squared SE by 40 rather than 82.5. Had some the subjects been measured on days 1, 4, 6, and 9, they would have reduced the relevant term by $(1 - 5)^2 + (4 - 5)^2 + (6 - 5)^2 + (9 - 5)^2 = 34$, while subjects measured on days 0, 6, and 9 would reduce it by $(0 - 5)^2 + (6 - 5)^2 + (9 - 5)^2 = 42$. This illustrates that in regression analyses, three measurements of Y "further apart in X" can be more informative than four "closer in X" ones.[3] The next example also illustrates this principle.

Example 3: *X* Values on a Continuum Only Partly Under Investigator Control

In most panel studies in environmental epidemiology, it is not easy to have many predictable values of X. Figure 5 shows acute changes in microvascular function (measured as a reactive hyperemia index) in the hours immediately following exposure to traffic pollutants in 43 healthy nonsmoking women in Montreal, Canada (data and R code for the cyclist exposure example are available in the supplemental digital content; http://links.lww.com/EDE/B252, http://links.lww.com/EDE/B253).[1] Women were exposed to traffic-related air pollution for 2 hours on three separate occasions during cycling on high and low-traffic routes as well as indoors. Personal exposures to ultrafine particles (<0.1 μm, UFP) measured along each route will serve as the "X" here. Also shown as black lines at the top of each panel are the distances of each

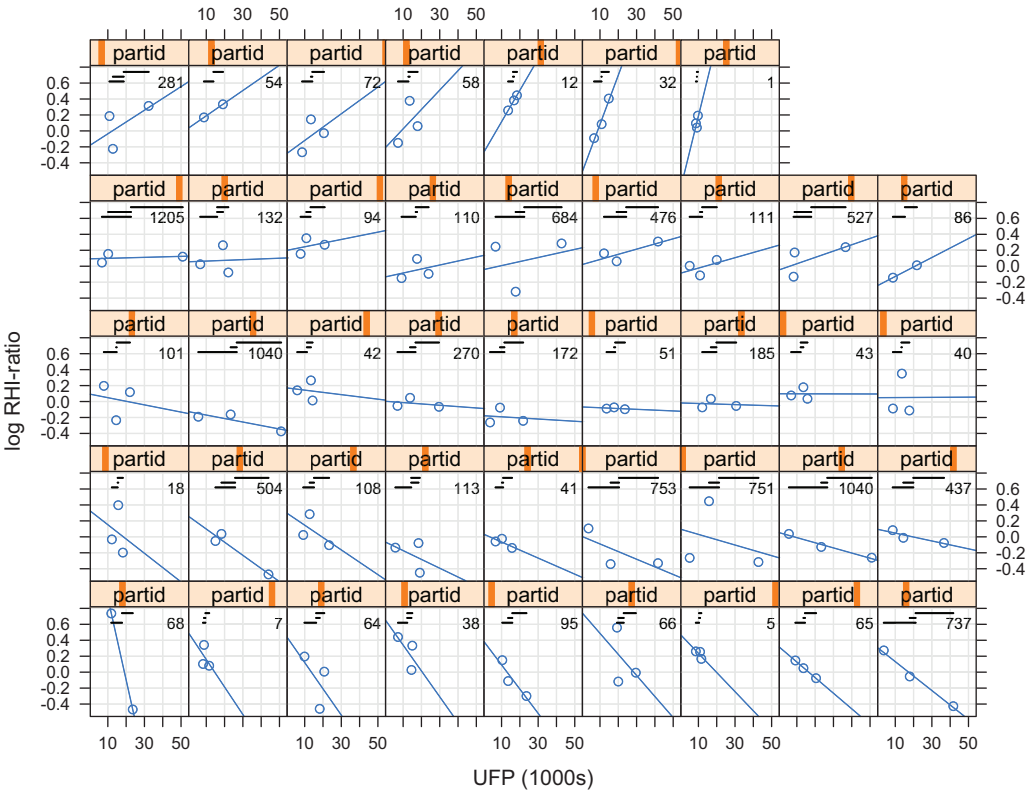

FIGURE 5. For each of 43 subjects, the log of the post–pre reactive hyperemia ratio on each day plotted against the ultrafine particle (UFP) exposure during that day's 2-hour exposure. Panels are ordered from bottom left to top right according to the slope of a regression line fitted just to that subject's two or three data points. Also shown as two or three black lines at the top of each panel are the distances of each subject's X values from their mean, and at the top right, again in black, the sum of their squares, with the larger sums denoting the more informative subjects. Figure is available in color online.

subject's X values from their mean and at the top right the sum of their squares, with the larger sums denoting the more informative subjects.

Table 2 shows the fitted components of one fixed-effect model and three random effects models. These models are also depicted in Figure 6. For sample size and other study design planning, the formulae in the last column of Table 2 are the most important. Since end-users will only use one of the two models with random intercepts, we focus on these models and on their implications for improving precision and study power. As one can see from the last column, the squared SE from the *random-intercept-common-slope* model has just one component, the "residual" variance divided by the product of (1) the number of subjects; (2) the number of response measurements per subject; and (3) the range (measured as a mean squared distance from the mean) of the exposures at which these responses are measured. In practice, since m and MS_X will vary from subject to subject, the divisor will be the sum, over the n subjects, of the product of each subject's m and MS_X. Thus, any configuration that increases this product decreases the square of the SE and, thus, the SE itself. The squared SE for the model with both random intercepts and random slopes has two components. One of these components is the "residual" variance divided by the product of the three quantities already described. The other is the between-subject variance in the slopes, divided by the number of subjects. The context will determine which component is dominant and thus needs to be given greater priority so as to minimize the SE.

The contrast with Example 2 is quite striking. Here, if one again works through the calculations for the SEs that include two variance components, one will notice that the dominant contribution to the squared SE (more than 90%) comes from the residual component rather than the slope component. By default, the variance is assigned to the residuals unless it can be shown to be otherwise. In this example, it is also more difficult to choose among the models. Since the residual component dominates the SE, and the signals are small relative to the noise, the SEs from all three models that include at least one random component are similar.

Turning SE Formulae Into Sample Size Formulae

By inserting the SE for the estimate of the common or average slope into the inequality for 80% power of a statistical test with alpha = 0.05 of a zero slope against a slope of magnitude β

$$(1.96\ \mathrm{SE} + 0.84\ \mathrm{SE}) \le \beta,$$

we can solve for n, or m, or the MS_X as a function of the other two. For example, for a given m and MS_X, these become:

Random-intercept models:

$$n = \frac{(1.96+0.84)^2}{\beta^2} \cdot \frac{\sigma^2_{residual}}{m \times \mathrm{MS}_X}$$

Random-intercept and random-slope models:

$$n = \frac{(1.96+0.84)^2}{\beta^2} \cdot \left(\sigma^2_{slopes} + \frac{\sigma^2_{residual}}{m \times \mathrm{MS}_X} \right)$$

where n is the number of subjects, β is the magnitude of the slope we wish to detect, $\sigma^2_{residual}$ is the within-subject variance of the response measure, σ^2_{slopes} is the variance of subject-specific slopes, m is the number of within-subject measurements, and MS_X is the mean squared distance between the subject's X's and their mean.

We have chosen to show the above formulae with the values of 1.96 and 0.84 so that they are familiar to those accustomed to working out sample sizes "by hand" rather than as an endorsement of the commonly used type I error rate of 5% and a power of 80%. What statistical power is ultimately selected depends on several factors: some suggest that any one study is but a contribution to an ultimate meta-analysis and that one merely contributes the amount (sample size) one can afford. In other contexts, where the planned investigation must stand alone, and will not be repeated by other researchers, the larger consequences of a false negative result argue for setting the power higher than 80%.

As an example use of these formulae, consider a hypothetical new panel study of UFP exposures and changes in reactive hyperemia, using the conventional error rates implied by the 1.96 and 0.84. Using the residual variance in Table 2 as our best estimate of what we might encounter, the number

TABLE 2. Fitted Components of four Models for Reactive Hyperemia Changes (log of the Post to Pre Ratio) Shown in Figure 5, Together With the Standard Error (SE) of Estimated Common or Average Slope, and the Calculation of this SE from the Sample Sizes (n and m), the Intraperson Range of the X Values (MS_X), and the Relevant Variance Components

Model	Intercept(s)	Slope(s) (/1000 UFP)	Residual	AIC	SE	Calculated SE
Common intercept, common slope	0.12	−0.0054	σ: 0.223	−16.5	0.0019	$0.223/[13{,}587]^{1/2}$
Variable intercepts, common slope	μ: 0.12; σ: 0.03	−0.0053	σ: 0.219	−14.5	0.0019	$0.219/[10{,}788]^{1/2}$
Common intercept, variable slopes	0.11	μ: −0.0048 σ: 0.0034	σ: 0.211	−15.9	0.0021	$(0.0034^2/43 + 0.211^2/[10{,}788])^{1/2}$
Variable intercepts, variable slopes	μ: 0.11; σ: 0.01	μ: −0.0048 σ: 0.0038	σ: 0.211	−11.9	0.0021	$(0.0038^2/18 + 0.211^2/[10{,}788])^{1/2}$

μ and σ represent the mean and standard deviation, respectively. In this example, $n = 43$ and $m = 2$ or 3 measurements per person (126 in total). For the first (entirely fixed effects) model, the squares of the distances of the 126 X's from their mean range from 0.1 to 1179; their average is 107.8 and their sum is 13,587. As shown in the 43 panels of Figure 5, the product of the m and the MS_X value, reflecting the number and range of the X values, varies from 1 to 1040. The 43 products sum to 10,788.

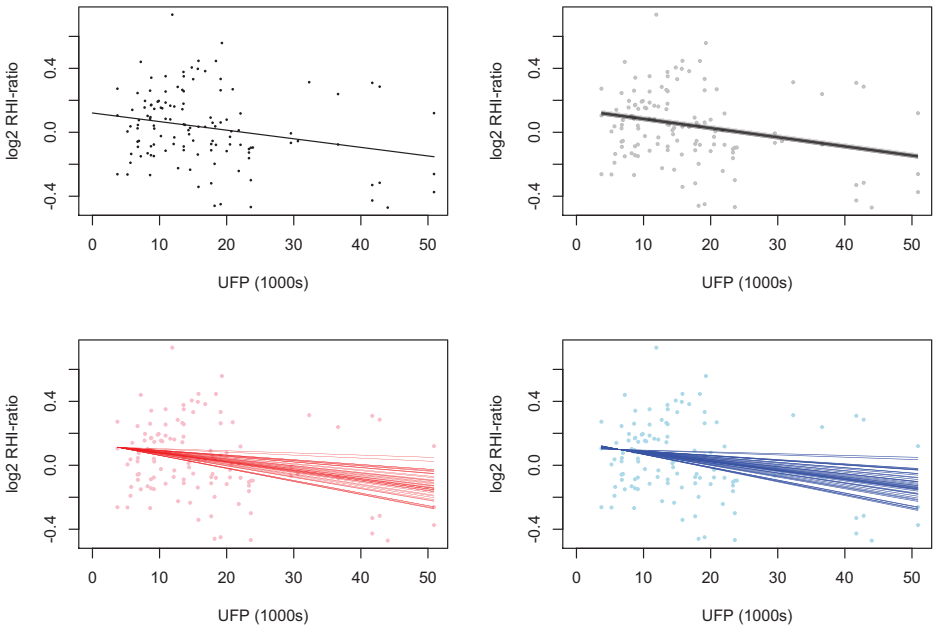

FIGURE 6. One fixed effects and three random-effects models fitted to the cyclist data in Table 6. Top Left, Common intercept, common slope; Top Right, random intercepts, common slope; Bottom Left, common intercept, random slopes; Bottom Right: random intercepts, random slopes. Figure is available in color online.

of subjects required for a random-intercept model to detect a slope of −0.0025 per 1000 UFPs (assuming three measurements per subject and an MS_X value of 500) is:

$$n = \frac{(1.96+0.84)^2}{-0.0025^2}\cdot\left(\frac{0.219^2}{3\times 500}\right)$$

$n = 40$ subjects.

The Price of Confounding

As described elsewhere,[5,6] there is a simple and intuitive variance inflation for having to adjust for confounding. To reflect the fact that the effective range of the exposure of interest is reduced, one needs to reduce the MS_X by a factor of $(1 - R^2_{X:\text{other Xs}})$, where $R_{X:\text{other Xs}}$ is the multiple correlation of X with the ensemble of the other covariates.

DISCUSSION

The random intercept in random-effects models is frequently used to remove intersubject variance in panel studies, but regression-based intrasubject contrasts in such studies are not well covered by existing sample size programs. However, once the basic structure of the standard error for a slope estimate from a simple regression is understood, the extension to random effects regression models is straightforward. In all regressions, the key to this understanding is a quantity that measures the range of the X values by using the mean squared distances of these values from their mean. It acts like a third sample size factor, the other two being n (the number of subjects) and m (the number of measurements per subject): all three combine to reduce the SE. Between-subject variances in slopes are likely to be most apparent in closely controlled situations. Conversely, large variations in slopes between subjects may not be readily apparent in panel studies of environmental exposures as they tend to be obscured by noise, either because of difficulties in maximizing within-subject exposure variance and/or inherent occasion to occasion variability in the response variable. Nevertheless, this challenge should not discourage investigators from evaluating the possibility of different slopes between study subjects as this may provide important information to further clarify exposure–disease associations in exposed populations. In particular, understanding *why* some people respond to exposure while others do not could help to inform individual-level risk management strategies, regulatory interventions, and/or public health communication. In most cases, it is useful to examine and report information from both types of models (i.e., random-intercept and common slope and both random-intercept random-slope models) to support both the planning of future investigations and to characterize potential heterogeneity in responses among exposed populations.

If panel study data will be analyzed assuming a common slope, the inputs to the sample size formula are m, MS_X, the residual variance, and magnitude of the "signal" we wish to detect (i.e., β, the magnitude of association). If random slopes are also considered, additional information is needed on the variance of slopes between subjects. In practice, literature values and experience can be used to estimate some of these parameters: practical considerations will provide a range of plausible values for the number of measurements per subject (m). However, most studies generally do not report within-subject variances in response measurements ($\sigma^2_{residual}$) or between-subject variances in slopes (σ^2_{slopes}). Instead, they tend to report only the observed between-subject variance seen at the two extremes of Figure 1. As shown there, this is an amalgam of what is truly between-subject and within-subject variation, a mixture that can only be separated by calculating an intraclass correlation from a pilot study.

Without such preliminary data, investigators may need to instead try a range of plausible values in estimating sample size requirements or contact corresponding authors of existing studies in the hope that they will share the relevant information. Alternatively, given budgetary constraints, investigators using random-intercept common-slope models may input feasible values for n, β, m, and MS_x and solve for $\sigma^2_{residual}$. This parameter can then be compared to literature values of the observed between-subject variance (which are generally reported). If the resulting $\sigma^2_{residual}$ is unrealistically small compared with the reported between-subject variance, and sensible values for how much of the observed variance is thought to be truly between-subject and truly within-subject variation, then some other elements of the sample size must be increased to compensate. A similar approach may be taken if random slopes are also considered, but in this case, investigators must estimate both $\sigma^2_{residual}$ and σ^2_{slopes}. In some cases, pilot studies may provide important information for within-subject variance and between-subject variance in slopes if such studies can be conducted at little cost. In general, the planning of future panel studies would be greatly improved if investigators regularly reported all of the variance components in fitted random-effects models.

ACKNOWLEDGMENTS

We thank Corinne Riddell for developing an online tool to implement the sample size calculations discussed in this manuscript. This tool is available at: https://corinne-riddell.shinyapps.io/mcgilleboh-samplesizecalculator/

REFERENCES

1. Weichenthal S, Hatzopoulou M, Goldberg MS. Exposure to traffic-related air pollution during physical activity and acute changes in blood pressure, autonomic and micro-vascular function in women: a cross-over study. *Part Fibre Toxicol*. 2014;11:70.
2. Norris C, Goldberg MS, Marshall JD, et al. A panel study of the acute effects of personal exposure to household air pollution on ambulatory blood pressure in rural Indian women. *Environ Res*. 2016;147:331–342.
3. Ren C, O'Neill MS, Park SK, Sparrow D, Vokonas P, Schwartz J. Ambient temperature, air pollution, and heart rate variability in an aging population. *Am J Epidemiol*. 2011;173:1013–1021.
4. Bacchetti P. Current sample size conventions: flaws, harms, and alternatives. *BMC Med*. 2010;8:17.
5. Hanley JA and Moodie EEM. Sample size, precision and power calculations: a unified approach. *J Biomet Biostat* 2011;2:124.
6. Hanley JA. Simple and multiple linear regression: sample size considerations. *J Clin Epidemiol* 2016;79:112–119.
7. Student. The Probable Error of a Mean. *Biometrika* 1908;6:1–25.
8. Cushny AR, Peebles AR. The action of optical isomers: II. Hyoscines. *J Physiol*. 1905;32:501–510.
9. Bates DM. lme4: Mixed-effects modeling with R. Available at: http://lme4.r-forge.r-project.org/lMMwR/lrgprt.pdf. 2010. Accessed 22 August 2016.