Published in final edited form as:
Stat Methods Med Res. 2011 October ; 20(5): 471–487. doi:10.1177/0962280210371563.

Power and sample size calculations for longitudinal studies estimating a main effect of a time-varying exposure

Xavier Basagaña
Centre for Research in Environmental Epidemiology (CREAL), Barcelona, Spain; Municipal Institute of Medical Research (IMIM-Hospital del Mar), Barcelona, Spain; CIBER Epidemiologia y Salud Publica (CIBERESP), Barcelona, Spain

Xiaomei Liao and **Donna Spiegelman**
Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115, USA; Department of Epidemiology, Harvard School of Public Health, Boston, MA 02115, USA

Abstract

Existing study design formulas for longitudinal studies assume that the exposure is time invariant or that it varies in a manner that is controlled by design. However, in observational studies, the investigator does not control how exposure varies within subjects over time. Typically, a large number of exposure patterns are observed, with differences in the number of exposed periods per participant and with changes in the cross-sectional mean of exposure over time. This article provides formulas for study design calculations that incorporate these features for studies with a continuous outcome and a time-varying exposure, for cases where the effect of exposure on the response is assumed to be constant over time. We show that incorrectly using the formulas for time-invariant exposure can produce substantial overestimation of the required sample size. It is shown that the exposure mean, variance and intraclass correlation are the only additional parameters needed for exact solutions for the required sample size, if compound symmetry of residuals can be assumed, or to a good approximation if residuals follow a damped exponential correlation structure. The methods are applied to several examples. A publicly available programme to perform the calculations is provided.

1 Introduction

Formulas for study design calculations, for example power and sample size, for longitudinal studies when the interest is in the main effect of exposure have been provided.[1–3] In those papers, exposure was considered to be fixed over time, that is subjects were assumed to be either exposed or unexposed for the entire follow-up period. In the study design setting, the variation of exposure within a participant or group has only been considered in studies where this variation is controlled by the investigator, such as in crossover trials[4] or multicentre clinical trials with randomisation at the patient level.[5,6] However, in observational studies, exposure is not assigned by design, and a large number of exposure patterns may be observed, with differences in the number of exposed periods per participant and with changes in the cross-sectional prevalence of exposure over time. For example, in a study on the respiratory effects of exposure to cleaning products,[7] women were followed during 15 consecutive days, and the use of cleaning products (e.g. bleach), which varied daily within participants, was recorded.

Address for correspondence: Donna Spiegelman, Department of Epidemiology, Harvard School of Public Health, 677 Huntington Avenue, Room 806, Boston, MA 02115, USA. stdls@channing.harvard.edu.

In this article, we develop formulas for power and sample size in observational longitudinal studies that accommodate changes in exposure over time not determined by design. In addition, we assess the sensitivity of study design to ignoring and misspecifying the nature of this variation and compare the efficiency of studies with a time-varying exposure to those with a time-invariant exposure. Section 2 introduces the notation and models used. In Section 3, formulas for power and sample size for a study with time-varying exposure are derived. Section 4 illustrates the methods we derived applying them to a study on the respiratory effects of exposure to cleaning products. Finally, Section 5 summarises the results and discusses further research. A Web Appendix with proofs of the derivations is available at http://www.hsph.harvard.edu/faculty/spiegelman/optitxs/Appendix_paperSMMR.pdf and publicly available software to perform all calculations derived in this article is available at http://www.hsph.harvard.edu/faculty/spiegelman/optitxs.html.

2 Notation and framework

2.1 General notation

Let Y_{ij} be the outcome of interest for the measurement taken at the j-th ($j = 0, \ldots, r$) time for the i-th ($i = 1, \ldots, N$) participant, and E_{ij} represent the exposure for the period between the measurements of $Y_{i,j-1}$ and Y_{ij}. Thus, r is the number of post-baseline measurements of the response per participant, or, equivalently, the total number of measurements per participant is $r + 1$. We consider studies that obtain repeated measures every s time units, as is the usual design in epidemiologic studies. Let t_{i0} be the initial time for participant i and let $V(t_0)$ be the variance of t_{i0} over all participants. When $V(t_0) = 0$, all participants have the same time vector, as when using time since enrollment in the study as the time variable of interest. However, when age is the time metameter of interest, as is often the case in epidemiology, and participants enter the study at different ages, we have $V(t_0) > 0$.

We base our results on linear models of the form $\mathbb{E}(\mathbf{Y}_i|\mathbf{X}_i) = \mathbf{X}_i\beta\ (i=1,\ldots,N)$, where $\mathbb{E}(\cdot)$ is the expectation operator, $\mathbf{Y}_i$ is the $r + 1$ response vector of participant i, $\mathbf{X}_i$ is the $((r + 1) \times q)$ covariate matrix for participant i, q is the number of variables in the model, β is a $q + 1$ vector of unknown regression parameters, and the $(r + 1) \times (r + 1)$ residual covariance matrix is $\mathrm{Var}(\mathbf{Y}_i|\mathbf{X}_i) = \Sigma$ ($i = 1, \ldots, N$), which will be treated as known with no practical implications as long as the sample size is not too small. Section 2.2 describes the particular variables included in each of the models for which we derived study design formulas. We base our development on the generalised least squares (GLS) estimator of β, which has the form $\widehat{\beta}=\left(\frac{1}{N}\sum_i \mathbf{X}'_i\Sigma^{-1}\mathbf{X}_i\right)^{1}\left(\frac{1}{N}\sum_i \mathbf{X}'_i\Sigma^{-1}\mathbf{Y}_i\right)$. Since the covariate matrix $\mathbf{X}_i$ is not known *a priori*, following Whittemore[8] and Shieh,[9] study design calculations use $\frac{1}{N}\Sigma_B$ as the variance of $\widehat{\beta}$, where

$$\Sigma_B=\left(\mathbb{E}_X\left[\mathbf{X}'_i\Sigma^{-1}\mathbf{X}_i\right]\right)^{-1}. \quad (2.1)$$

As long as Σ does not depend on the covariates, (2.1) can be fully specified by knowing the first- and second-order moments of the covariate distribution.[10] Lachin[11] (chapter 3) followed a different approach by computing the expected value of the test statistic over the distribution of $\mathbf{X}_i$.

We derive formulas for a general covariance structure, but consider two particular covariance structures for the response: compound symmetry (CS) and damped exponential (DEX). Under CS for the response, Σ has diagonal terms equal to σ^2 where $\sigma^2 = \mathrm{Var}(Y_{ij}|X_{ij})$ is the residual variance of the response given the covariates, and off-diagonal terms

equal to $\sigma^2\rho$, where ρ is the correlation between two measurements from the same participant, also known as the intraclass correlation coefficient. It is worth mentioning that a random intercept model leads to a compound symmetry covariance. Since a common correlation may not be realistic in some studies, we also consider DEX covariance,[12] where the $[j, j']$ element of Σ has the form $\sigma^2\rho|j-j'|^\theta$, and therefore the correlation between two measurements decays exponentially as the separation between measurements increases, with the degree of decay fixed by the parameter θ. Thus, when $\theta = 0$, the CS covariance structure is obtained, and when $\theta = 1$, the AR(1) covariance structure is given. Note that for $r = 1$, DEX is equivalent to CS.

The article is mainly focused on models for a binary exposure. However, Section 3.4 discusses how the formulas can be used when the exposure is continuous. Let p_{ej} be the prevalence of exposure at each time point, $\bar{p}_e$ the mean prevalence of exposure across all periods, $\boldsymbol{\Sigma}_E$ the covariance matrix of exposure, $\rho_{e_j,e_{j'}}$ the correlation between exposure at the j-th and j'-th measurements, so that

$$\mathbb{E}\left[E_j E_{j'}\right] = \rho_{e_j,e_{j'}}\sqrt{p_{ej}\left(1-p_{ej}\right)}\sqrt{p_{ej'}\left(1-p_{ej'}\right)} + p_{ej}p_{ej'},$$

and ρ_{e_j,t_0} the correlation between initial time (or age at entry) and exposure at the j-th measurement. Additionally, we define two quantities that can be computed for any form of the covariance matrix of exposure. The first one is the intraclass correlation of exposure,

$$\rho_e = \frac{\mathbf{1}'\boldsymbol{\Sigma}_E\mathbf{1} - \operatorname{tr}(\boldsymbol{\Sigma}_E)}{r\operatorname{tr}(\boldsymbol{\Sigma}_E)}, \quad (2.2)$$

where $\mathbf{1}$ is a length $r + 1$ vector of ones and tr() indicates the trace of a matrix. The intraclass correlation of exposure is the ratio of the average covariance over the average variance and is an index of similarity or agreement between each subject's exposure in the different time periods.[13] Similarly, we define the first-order intraclass correlation of exposure, ρ_{e_1}, as the ratio of the average first-order covariance, that is the average of the first diagonal below the main diagonal of $\boldsymbol{\Sigma}_E$, over the average variance. Mathematically, we can write it as,

$$\rho_{e_1} = \frac{(r+1)\operatorname{tr}\left(\Sigma_E^{(1,r+1)}\right)}{r\operatorname{tr}(\boldsymbol{\Sigma}_E)}, \quad (2.3)$$

where $\Sigma_E^{(1,r+1)}$ is the matrix $\boldsymbol{\Sigma}_E$ with the first row and the $(r + 1)$-th column removed, because the main diagonal of the matrix $\Sigma_E^{(1,r+1)}$ contains the first-order covariances of exposure.

The intraclass correlation of exposure can be regarded as a measure of within-subject variation of exposure. When ρ_e takes its maximum, $\rho_e = 1$, there is no within-subject variation of exposure, that is participants are either exposed or unexposed for the whole period (time-invariant exposure). Conversely, when it takes its minimum, $-1/r$, the within-subject variation of exposure is maximal.[13] The upper bound for ρ_e is smaller than one when the exposure prevalence is not constant over time (expression derived in Web Appendix A. 1[†]). For binary variables, as here, the lower bound of $-1/r$ cannot always be reached due to some constraints on the correlation between two binary variables, and the lower bound for ρ_e is,

[†]http://www.hsph.harvard.edu/faculty/spiegelman/optitxs/Appendix_paperSMMR.pdf

$$\frac{-1}{r}+\frac{\left((r+1)\,\bar{p}\,e-\text{int}\left((r+1)\,\bar{p}\,e\right)\right)\left(1-(r+1)\,\bar{p}\,e+\text{int}\left((r+1)\,\bar{p}\,e\right)\right)}{r\,(r+1)\,\bar{p}\,e\left(1-\bar{p}\,e\right)},$$

where int(·) indicates the integer part.[14] The parameter ρe has other useful interpretations. When the exposure prevalence is constant over time and the exposure has compound symmetry covariance, the intraclass correlation coefficient is equal to the common correlation (Web Appendix A.2†). The intraclass correlation of exposure can also be regarded as a measure of imbalance in the number of exposed periods per subject, $E_{i\cdot}$. When $E_{i\cdot}$ is equal across subjects, then everyone is exposed for the same number of periods as, for example, in uniform crossover studies. Then, $\rho e = -1/r$. Conversely, when the exposure is time invariant, the imbalance is maximal since $E_{i\cdot}$ is either zero with probability $(1 - pe)$ or $r + 1$ with probability pe, and $\rho e = 1$. Section 3.3 discusses intuitive ways to specify a value for ρe.

2.2 Models and general power and sample size equations

In this article, we assume that the effect of exposure is the same at any point in time, denoted as the constant mean difference (CMD) hypothesis.[2] The design of randomised longitudinal studies of this hypothesis has been previously considered for time-invariant exposures.[1–3] The left panels of Figure 1 illustrate the trajectories of participants whose exposure is time invariant over follow-up, and the difference between the trajectories of the exposed and the unexposed is constant over time. If exposure is time varying, the individual trajectory shifts when exposure changes, as illustrated in the right panel of Figure 1(a) and (b), where the dots indicate a possible individual trajectory and the value of E indicates the presence or absence of exposure. The CMD hypothesis is suitable for acute and transient exposure effects, since once the exposure is removed, the response returns to the level of the unexposed. In Figure 1a, time plays no role and only the most recent exposure preceding the response matters, corresponding to model,

$$\mathbb{E}\left(Y_{ij}|\mathbf{X}_i\right)=\mathbb{E}\left(Y_{ij}|E_{ij}\right)=\beta_0+\beta_1 E_{ij}. \quad (2.4)$$

Model (2.4) assumes that the within- and between-subject effects of exposure are equal,[15] that is that there is no confounding by risk factors that vary between subjects. If this assumption is unreasonable, one may want to fit the following change model,

$$\mathbb{E}\left(Y_{i,j+1}-Y_{ij}|E_{i,j+1},E_{ij}\right)=\beta_1^W\left(E_{i,j+1}-E_{ij}\right). \quad (2.5)$$

This model results from applying the first difference operator

$$\Delta=\begin{pmatrix} -1 & 1 & 0 & \cdots & 0 \\ 0 & -1 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & -1 & 1 \end{pmatrix}$$

to model (2.4), so that $\Delta\mathbf{Y}_i$ is the vector with elements $Y_{i,j+1} - Y_{ij}$, j 1, …, r, and Var $(\Delta\mathbf{Y}_i) = \Delta\Sigma\Delta'$. For a multivariate normal response with known Σ, fitting model (2.5) by GLS is equivalent to fitting model (2.4) by conditional maximum likelihood.[16] The parameter β_1^W

from model (2.5) is estimated from data of changes in exposure on the response within subject, while β_1 from model (2.4) is estimated from data on exposure differences between- and within-subjects. If there is no confounding by between-subjects determinants of response, $\widehat{\beta}_1^W$ 1 will estimate the same parameter as $\widehat{\beta}_1$, otherwise not.[15] In observational studies, model (2.5) is often preferred, since each participant serves as his or her own control, subtracting out confounding by all between-subject (time invariant) variables. The trade-off is that model (2.5) is less efficient than (2.4) for estimating the exposure effect,[15] and this has implications for study design.

In the pattern illustrated in Figure 1(b), the response changes linearly with time for all subjects (e.g. due to ageing) but the effect of exposure remains constant over time, that is,

$$\mathbb{E}\left(Y_{ij}|\mathbf{X}_i\right)=\mathbb{E}\left(Y_{ij}|E_{ij},t_{ij}\right)=\beta_0+\beta_1 E_{ij}+\beta_2 t_{ij}. \quad (2.6)$$

When time (or age) and exposure are correlated, time will be a confounder of the exposure effect, and model (2.4) cannot obtain a valid estimate of the effect of exposure. Like model (2.4), model (2.6) assumes that there is no between-subject confounding. As above, the within-subject effect of exposure (and time) can be estimated by fitting the model on changes that results from applying the first difference operator to model (2.6), and assuming that the time points are equidistant, as is often the case in practice, it leads to,

$$\mathbb{E}\left(Y_{i,j+1}-Y_{i,j}|Y_{i,j+1},E_{ij}\right)=\beta_2^W+\beta_1^W\left(E_{i,j+1}-E_{ij}\right). \quad (2.7)$$

Again, under multivariate normality, fitting model (2.7) by GLS is algebraically equivalent to fitting model (2.6) by conditional likelihood. When time and exposure are correlated, time is a confounder of the effect of exposure and model (2.5) cannot be used to obtain a valid estimate of the effect of exposure.

Let $\widehat{\beta}_1$ be the estimate of the parameter of interest, which is β_1 for models (2.4) and (2.6) and β_1^W for models (2.5) and (2.7). Let σ_1^2 be the diagonal element of the matrix Σ_B, defined in (2.1), associated with $\widehat{\beta}_1$. The Wald test statistic for $\widehat{\beta}_1$ is $T=\sqrt{N}\widehat{\beta}_1/\sigma_1$ and the formula for the power of a two-sided test, provided the power is not too small, is,

$$\Phi\left[\sqrt{N}|\beta_1/\sigma_1-z_{1-\alpha/2}\right],$$

where α is the significance level, and zp and $\Phi(\cdot)$ are the p-th quantile and the cumulative density of a standard normal, respectively. The formula for sample size to achieve a pre-specified power π is,

$$N=\sigma_1^2(z\pi+z_{1-\alpha/2})^2/\beta_1^2.$$

Note that σ_1^2 will depend on r, the exposure prevalence, and on parameters describing the covariance of both the response and the exposure processes.

3 Results

3.1 Arbitrary covariance structures for response and exposure

For both power and sample size calculations, we need to obtain expressions for σ_1^2 following (2.1) and the model of choice from among (2.4)–(2.7). Recall that models (2.6) or (2.7) should be used instead of (2.4) or (2.5) when time is expected to be associated with the response, to control for confounding if time and exposure are correlated or otherwise to improve efficiency. Let us call $v^{jj'}$ the $[j, j']$-th element of Σ^{-1}. Then, when β_1 is estimated by model (2.4),

$$\sigma_1^2 = \frac{\sum_{j=0}^{r}\sum_{j'=0}^{r} v^{jj'}}{\left(\sum_{j=0}^{r}\sum_{j'=0}^{r} v^{jj'}\right)\left(\sum_{j=0}^{r}\sum_{j'=0}^{r} v^{jj'}\,\mathbb{E}\left[E_j E_{j'}\right]\right) - \left(\sum_{j=0}^{r}\sum_{j'=0}^{r} v^{jj'}\,p_{ej}\right)^2} \quad (3.1)$$

(Web Appendix B.1†). To find σ_1^2 corresponding to model (2.5), we define the matrix $\mathbf{M} = \Delta' (\Delta\Sigma\Delta')^{-1} \Delta$. Let us call $m^{jj'}$ the $[j, j']$-th element of $\mathbf{M}$. Then, for model (2.5),

$$\sigma_1^2 = \left(\sum_{j=0}^{r}\sum_{j'=0}^{r} m^{jj'}\,\mathbb{E}\left[E_j E_{j'}\right]\right)^{-1} \quad (3.2)$$

(Web Appendix B.2†). The expression for σ_1^2 corresponding to the GLS estimate of β_1 from model (2.6) is derived in Web Appendix B.3.† Under model (2.6), $p_{ej}\ \forall j$, $\rho_{e_j, e_{j'}}\ \forall j, j'$, $V(t_0)$ and $\rho_{e_j, t_0}\ \forall j$ need to be provided. The variance formula reduces to (3.1) when the prevalence of exposure is constant over time and either $V(t_0) = 0$ or $\rho_{e_j, t_0} = 0\ \forall j$ (Web Appendix B.3†). For model (2.7),

$$\sigma_1^2 = \frac{\sum_{j=0}^{r}\sum_{j'=0}^{r} jj' m^{jj'}}{\left(\sum_{j=0}^{r}\sum_{j'=0}^{r} jj' m^{jj'}\right)\left(\sum_{j=0}^{r}\sum_{j'=0}^{r} m^{jj'}\,\mathbb{E}\left[E_j E_{j'}\right]\right) - \left(\sum_{j=0}^{r}\sum_{j'=0}^{r} m^{jj'} j' p_{ej}\right)^2} \quad (3.3)$$

(Web Appendix B.4†). Note that the parameters $V(t_0)$ or $\rho_{e_j, t_0}\ \forall j$, which may be difficult to provide *a priori*, are not required here. In addition, the bias from confounding due to between-subject differences in age at entry, or any other between-subject difference, is removed. Thus, for validity and for simplicity, when $V(t_0) > 0$, we recommend using study designs based on model (2.7) instead of model (2.6). This will provide a conservative study design because the variance σ_1^2 corresponding to model (2.7) will be greater than that for model (2.6), since model (2.7) is only estimating the within-subject effects.[15] When the prevalence of exposure is constant over time, (3.3) reduces to (3.2) (Web Appendix B.4†).

Figure 2 shows the variance of the coefficient of interest for models (2.4)–(2.7) for some examples where $V(t_0) = 0$. We can see that the coefficient of interest from the conditional model always has greater variance than its non-conditional counterpart, the difference being greater for small r and when there is little within-subject variation in exposure (i.e. ρ_e is small). The models that include time also have greater variance than their counterparts without time because time and exposure are correlated in the examples. However, in these examples, this difference is smaller than the difference between using conditional likelihood or not.

3.2 Simplifying cases

For all models (2.4)–(2.7), the following parameters are needed for power and sample size calculations: N or π, r, β_1, and the parameters defining the residual covariance of the response, which reduce to σ^2 and ρ if CS is assumed, and these two plus θ when DEX is assumed. These are the parameters needed in the time-invariant exposure case. When the exposure is time varying, we additionally need to provide p_{ej} and $\rho_{e_j,e_{j'}}\ \forall j, j'$ for models (2.4), (2.5) and (2.7); and these plus $V(t_0)$ and $\rho_{e_j,t_0}\ \forall j, j'$ for model (2.6). As suggested in the previous section, the simpler formulas for model (2.7) can be used instead of those of (2.6).

If a CS covariance of the response can be assumed, $\rho_{e_j,e_{j'}}\ \forall j, j'$ do not need to be provided for any of the four models, but only ρ_e, regardless of the covariance structure of the exposure process (Web Appendix B.5†). The simplified formulas are provided in Section 3.2.1. If the response does not have CS covariance but the exposure does, still $\rho_{e_j,e_{j'}}\ \forall j, j'$ are not required but only ρ_e. We discuss what to do if neither of these assumptions hold in Section 3.2.2, where a DEX covariance of the response is assumed.

3.2.1 Compound symmetry covariance for the response and constant prevalence of exposure—Under CS of the response and constant prevalence of exposure equal to p_e, σ_1^2 for model (2.4) is,

$$\sigma_1^2=\frac{\sigma^2(1-\rho)(1+r\rho)}{p_e(1-p_e)(r+1)(1-\rho(2-(r+1)(1-\rho_e)-\rho_e))} \quad (3.4)$$

(Web Appendix B.6†). When the exposure prevalence is constant and $\rho_e = 1$, then (3.4) reduces to the standard formula for a study with time-invariant exposure.[1–3] The ratio of required number of participants needed (sample size ratio, SSR) to achieve a pre-specified power when one uses the formulas for time-invariant exposure, equivalent to assuming $\rho_e = 1$, compared to when the true value of ρ_e is used, is,

$$SSR=\frac{N_{\rho_e=1}}{N_{\rho_e}}=\frac{1-\rho+r\rho-r\rho\rho_e}{1-\rho}. \quad (3.5)$$

This ratio increases linearly with r, and decreases linearly with ρ_e. It can also be shown that the SSR increases as ρ increases. Figure 3 shows the value of SSR for several values of r, ρ and ρ_e. For model (2.5), Equation (3.2) becomes,

$$\sigma_1^2=\frac{\sigma^2(1-\rho)}{p_e(1-p_e)\,r\,(1-\rho_e)} \quad (3.6)$$

(Web Appendix B.7†). This expression goes to infinity when $\rho_e = 1$, that is exposure is time invariant, and therefore the within-subject effect of exposure cannot be estimated. The formulas for σ_1^2 corresponding to models (2.6) and (2.7) with CS response covariance are not provided because they are complex. However, they are all implemented in our software.

3.2.2 Damped exponential covariance for the response—Figure 4 compares the required sample size when DEX or CS covariance of the response are assumed, everything else being equal. For a time-invariant exposure (i.e. $\rho_e = 1$), fewer participants are needed when $\theta > 0$ compared to $\theta = 0$ (CS). However, with time-varying exposures (i.e. $\rho_e < 1$), this is not necessarily the=case and we find the opposite result for small values of ρ_e.

As shown in Section 3.2, if the response covariance is DEX but the exposure process covariances is CS, $p_{ej}\ \forall j$ and ρe suffice to compute power or sample size. If neither the response nor the exposure covariance is CS, though, this is not the case. However, since ρe can be viewed as a summary measure of all the correlations $\rho e_j, e_{j'}$, one may conjecture that assuming $\rho e_j, e_{j'} = \rho e\ \forall j, j'$ would produce reasonable estimates of σ_1^2 even if the actual covariance matrix of exposure does not follow a CS structure. We performed a numerical analysis to evaluate how well assuming CS covariance for exposure approximated σ_1^2 when the exposure process had an arbitrary correlation, that is when the exposure covariance was misspecified. To compute the true and misspecified σ_1^2, the exposure prevalence vector and the correlation matrix of exposure are needed. For values of r equal to 2, 5 and 10, we generated 10 000 arbitrary prevalence vectors and correlation matrices using a process described in Web Appendix C.† Then, the SSR comparing the use of the true and misspecified σ_1^2 were computed for $\rho = (0.8, 0.5, 0.2)$ and $\theta = (0.2, 0.5, 0.8, 1)$, and for each model (2.4)–(2.7).

Results from this numerical analysis were similar for all models, with model (2.7) giving slightly more extreme SSRs. The results for model (2.7) when the true prevalence at each time point was used and for $r = 5$ are illustrated in Figure 5. Results were similar for the other values of r. Using the true prevalence of exposure at each time point produced great improvements in the approximations in comparison to using a constant prevalence equal to $\bar{p}_e$ in the models that include time, that is (2.6) and (2.7), but the improvement was more modest for models that did not include time, that is (2.4) and (2.5) (data not shown). The approximations of the true σ_1^2 were very good for $\rho = 0.2$ or $\theta = 0.2$, with none of the datasets having more than a 10% difference from the true value (Figure 5). For the other values of ρ and θ, most approximations were very good, but there were some SSRs smaller than 0.8 or greater than 1.2 as ρ and θ increased. The scenarios with low SSRs were characterised by having both a first-order intraclass correlation of exposure, ρe_1, greater than ρe (roughly, the difference being greater than 0.1) and a positive, non-zero ρe (roughly, greater than 0.3). The high SSRs were found for cases where ρe_1 was much smaller than ρe, and ρe was moderate to large.

We found the worst approximations when the covariance of the response was AR(1) (i.e. $\theta = 1$). It turns out that, when $\theta = 1$, σ_1^2 can be calculated exactly for models (2.4) and (2.6) by using the formula based on assuming a CS exposure covariance but providing ρe_1, the first-order intraclass correlation, instead of ρe, regardless of what the true exposure covariance is (Web Appendix D†). This is why the cases with worse SSRs were also characterised by ρe_1 being different from ρe. For values of θ between zero and one, neither ρe nor ρe_1 suffice to obtain exact values of σ_1^2 and we obtain approximations. Because σ_1^2 is an increasing function of ρe for CS covariance of exposure, we recommend using conservative values for ρe (i.e. higher values) if the first-order autocorrelations are suspected to be greater than the rest. Similar reasoning can be applied to models (2.5) and (2.7), in which case σ_1^2 under AR(1) response cannot be calculated exactly using ρe_1, but the recommendation of using conservative values for ρe still applies.

3.3 Summary and practical considerations

The exposure prevalence $p_{ej}\ \forall j$ and the intraclass correlation of exposure, ρe are the only additional parameters needed to generalise the study design formulas to the time-varying exposure case in most circumstances. Under either CS covariance of the response or the exposure, exact formulas are obtained with just these parameters, and for the rest of cases these same formulas provide approximations that are quite accurate in general. When one

has to rely on approximations, conservative values (i.e. large values) of ρe are advisable for use in situations with large θ and with first-order autocorrelations of exposure higher than the higher order autocorrelations. If the model includes time, better approximations can be obtained if $p_{ej} \forall j$ is provided instead of just providing $\bar{p}_e$.

The parameter ρe can take values between $-1/r$ and one, and these two extremes describe two familiar study designs. When $\rho e = 1$, the exposure values of each participant are all the same, that is, the parallel group design is obtained. When $\rho e = -1/r$, all participants are exposed the same number of periods, as in uniform crossover designs. In observational studies, intermediate values between $\rho e = -1/r$ (same number of exposed periods for all participants) and $\rho e = 1$ (time-invariant exposure) will often be observed, and when pilot data are not available, the investigator can assess the sensitivity of the study design over a range of plausible values for ρe. To help the investigator assess what values of ρe are appropriate for his or her exposure, our program can compute the distribution of $E_{i\cdot}$ once r and $\bar{p}_e$ are fixed and a CS covariance of exposure is assumed. Examples of distributions of $E_{i\cdot}$ by varying ρe are shown in Figure 6.

Depending on the prevalence vector, the parameter ρe can have lower and upper bounds that are different than $(-1/r, 1)$. These bounds are calculated by our program and shown to the user after r and the prevalence vector have been fixed. For both CS and DEX responses, σ_1^2 is maximised at the upper bound of ρe (Web Appendix E†). In addition, using linear programming techniques, the programme also computes an upper bound for σ_1^2 once ρe is provided (Web Appendix F†). This upper bound can be used as a conservative specification of σ_1^2. However, in a numerical analysis, this bound was found to be useful (being at most 20–30% greater than the true variance and different enough to simply assuming a time-invariant exposure) only for studies with a small number of repeated measures ($r \leq 5$).

3.4 Continuous exposures

The formulas derived in Web Appendices B.1–B.4 are valid for a continuous exposure and can be used to compute power or sample size in such cases. As opposed to the binary exposure case, the variance of exposure is not a function of its mean, and therefore the investigator will then need to provide the mean and the variance of exposure at each time point. If either the response or the exposure covariance are CS, the only additional parameter needed to compute the study design formulas is the intraclass correlation of exposure, since the results derived in Web Appendix B.5† are also valid for a continuous exposure. For other cases, the formulas based on ρe can still be used as approximations as in the binary exposure case.

4 Example: Respiratory function and cleaning products/tasks

Medina-Ramon *et al.*[7] studied the short-term respiratory effects of cleaning tasks and the use of cleaning products on pulmonary function in a group of 31 domestic cleaning women followed for $r + 1 = 15$ consecutive days. They studied a wide range of binary exposures, including cleaning tasks and cleaning products, and found an association of respiratory symptoms with vacuuming, and with the use of ammonia, decalcifiers, glass-cleaning sprays or atomisers, degreasing sprays or atomisers, air freshener sprays and bleach. The estimated mean prevalence of these exposures ranged from 0.12 to 0.84, and the estimated ρe ranged from 0.10 to 0.59. Since *a priori* there is no reason to believe that the frequency of use of cleaning products is going to change over this 15-day time period, we assumed a constant prevalence over time for all exposures. Since the time metameter is days since entry into the study, $V(t_0) = 0$. Under these two conditions, as shown in Section 3.1, formulas for model (2.6) reduce to those of model (2.4), which will be used here. The response variable was

peak expiratory flow. When a DEX structure was fitted to the residuals, the estimated values were $\rho = 0.88$ and $\theta = 0.12$.

First, we assessed the overestimation of the required sample size when the formula for a time-invariant exposure ($\rho e = 1$) was used instead of the one based on (3.4) for the exposure variables vacuuming and use of air freshener sprays, which were the exposures with smallest and greatest ρe, respectively, among all the exposures considered in this study. This overestimation was measured by the SSR as defined in Equation (3.5). For the vacuuming exposure, where $\bar{p}_{e=0.37}$ and $\rho e = 0.13$, the SSR was as large as 78, while for the air freshener spray use, which had $\bar{p}_{e=0.17}$ and $\rho e = 0.59$, it was 38. In Table 1, this SSR is calculated for other values of ρ and θ. The overestimation was higher for large values of ρ and small values of θ, but even for the air-freshener exposure with $\rho = 0.5$ and $\theta = 1$, incorrectly using the formulas for time-invariant exposure would require the recruitment of more than twice the number of participants than needed. We also calculated SSR comparing the required sample size obtained when the independence and CS assumptions for exposure were used to the required sample size obtained when the estimated exposure distribution was used (Table 1). Assuming independence underestimated by more than half the required sample size for the air freshener exposure, and around 15% for the vacuuming exposure. The assumption of CS covariance for the exposure performed well for all residual covariance structures considered, with slight underestimations when the response covariance was not CS.

5 Discussion

Formulas for sample size calculation in longitudinal studies have been provided in several papers.[1–3,6] However, since all of the existing developments were motivated by applications to randomised studies, the case of a time-varying exposure that is not fixed by design has never been examined. In this article, we provide methods to account for the time-varying nature of exposure in sample size calculations, and illustrate the advantages of using them in place of the time-invariant exposure formulas available previously. Assuming that the exposure does not vary within a subject will always overestimate the minimum sample size needed to satisfy a specified power constraint for fixed r. This is in agreement with the finding that multicentre trials where treatment varies within each centre require a lower sample size than cluster randomised trials, where treatment does not vary within a centre.[5] We found in some real `pilot' data that over thirty times more participants than necessary would be requested if time invariance of exposure was incorrectly assumed in sample size calculations. These large differences will occur in studies with highly correlated response residuals and a large number of repeated measures, as is common.

We based our calculations on models that, for example, do not include polynomial effects of time or random slopes. This is in line with the usual simplifications required at the planning stage. Consideration of additional model complexity would preclude most of the simplifications derived in this manuscript and the resulting formulas would require additional input parameters that would generally be difficult to provide. Studies whose design is based upon models (2.5) and (2.7) correct for the effect of all time-invariant confounders, measured and unmeasured, and therefore only time-varying confounders to be considered later in the analysis could influence study design. Study design formulas for more complicated models is an interesting topic for further research.

The influence of dropout was not covered here but is also an important factor when performing study design calculations. This topic has been studied previously in studies with time-invariant exposure.[3,17–20] Galbraith *et al.*[18] suggested computing N for 90% power when 80% power is intended and Fitzmaurice *et al.* (p. 409) suggested inflating N by $1/(1 -$

f), where f is the anticipated fraction of lost to follow-up, although this last approach is conservative. The performance of these approaches in longitudinal studies with time-varying exposures remains to be investigated, and new approaches developed if these fall short.

The methods developed here are extended in another paper[16] to other common scenarios of interest in longitudinal studies, such as the case when the hypothesis of interest is that the change in response over time varies between exposed and unexposed periods, or equivalently, the effect of exposure varies with time.[22] Longitudinal design for the optimal balance of number of participants and number of repeated measures subject to a fixed cost constraint or fixed power constraint, which was addressed previously for a time-invariant exposure,[23,24] could also be studied in the context of a time-varying exposure.

We provide a publicly available program to perform the calculations developed in this article, which can be downloaded at the link provided in Section 1. A demonstration of the programme use can be found in Web Appendix G.†

Acknowledgments

We thank Dr Medina-Ramon, Dr Anto, Dr Zock for letting us use the EPIASLI data in our example. Research was supported, in part, by National Institutes of Health (grant number CA06516).

References

1. Bloch DA. Sample size requirements and the cost of a randomized clinical trial with repeated measurements. Statistics in Medicine. 1986; 5(6):663–7. [PubMed: 3823673]
2. Frison L, Pocock SJ. Repeated measures in clinical trials: Analysis using mean summary statistics and its implications for design. Statistics in Medicine. 1992; 11(13):1685–704. [PubMed: 1485053]
3. Hedeker D, Gibbons RD, Waternaux C. Sample size estimation for longitudinal designs with attrition: Comparing time-related contrasts between two groups. Journal of Educational and Behavioral Statistics. 1999; 24(1):70–93.
4. Senn, S. Cross-over trials in clinical research. 2nd edn. John Wiley, Chichester, Eng.; New York: 2002.
5. Moerbeek M, Van Breukelen JP, Berger MPF. Design issues for experiments in multilevel populations. Journal of Educational and Behavioral Statistics. 2000; 25(3):271–84.
6. Moerbeek M, Van Breukelen JP, Berger MPF. Optimal experimental designs for multilevel models with covariates. Communications in Statistics – Theory and Methods. 2001; 30(12):2683–97.
7. Medina-Ramon M, Zock JP, Kogevinas M, Sunyer J, Basagana X, Schwartz J, et al. Short-term respiratory effects of cleaning exposures in female domestic cleaners. European Respiratory Journal. 2006; 27(6):1196–203. [PubMed: 16510456]
8. Whittemore AS. Sample size for logistic regression with small response probability. Journal of the American Statistical Association. 1981; 76(373):27–32.
9. Shieh G. On power and sample size calculations for likelihood ratio tests in generalized linear models. Biometrics. 2000; 56(4):1192–6. [PubMed: 11129478]
10. Tu XM, Kowalski J, Zhang J, Lynch KG, Crits-Christoph P. Power analyses for longitudinal trials and other clustered designs. Statistics in Medicine. 2004; 23(18):2799–815. [PubMed: 15344187]
11. Lachin, JM. Wiley series in probability and statistics. Wiley; New York: 2000. Biostatistical methods: the assessment of relative risks.
12. Munoz A, Carey V, Schouten JP, Segal M, Rosner B. A parametric family of correlation structures for the analysis of longitudinal data. Biometrics. 1992; 48(3):733–42. [PubMed: 1420837]
13. Kistner EO, Muller KE. Exact distributions of intraclass correlation and Cronbach's alpha with gaussian data and general covariance. Psychometrika. 2004; 69(3):459–74.
14. Ridout MS, Demetrio CG, Firth D. Estimating intraclass correlation for binary data. Biometrics. 1999; 55(1):137–48. [PubMed: 11318148]

15. Neuhaus JM, Kalbfleisch JD. Between- and within-cluster covariate effects in the analysis of clustered data. Biometrics. 1998; 54(2):638–45. [PubMed: 9629647]

16. Basagaña X, Spiegelman D. Power and sample size calculations for longitudinal studies comparing rates of change with a time-varying exposure. Statistics in Medicine. 2010; 29(2):181–92. [PubMed: 19899065]

17. Dawson JD. Sample size calculations based on slopes and other summary statistics. Biometrics. 1998; 54(1):323–30. [PubMed: 9544525]

18. Galbraith S, Marschner IC. Guidelines for the design of clinical trials with longitudinal outcomes. Controlled Clinical Trials. 2002; 23(3):257–73. [PubMed: 12057878]

19. Jung SH, Ahn C. Sample size estimation for GEE method for comparing slopes in repeated measurements data. Statistics in Medicine. 2003; 22(8):1305–15. [PubMed: 12687656]

20. Yi Q, Panzarella T. Estimating sample size for tests on trends across repeated measurements with missing data based on the interaction term in a mixed model. Controlled Clinical Trials. 2002; 23(5):481–96. [PubMed: 12392862]

21. Fitzmaurice, GM.; Laird, NM.; Ware, JH. Wiley series in probability and statistics. Wiley-Interscience; Hoboken, NJ: 2004. Applied longitudinal analysis.

22. Van Breukelen GJP. Ancova vs. change from baseline: More power in randomized studies, more bias in nonrandomized studies. Journal of Clinical Epidemiology. 2006; 59:920–5. [PubMed: 16895814]

23. Cochran, WG. Sampling techniques. 3rd edn. Wiley; New York: 1977.

24. Raudenbush SW. Statistical analysis and optimal design for cluster randomized trials. Psychological Methods. 1997; 2(2):173–85.

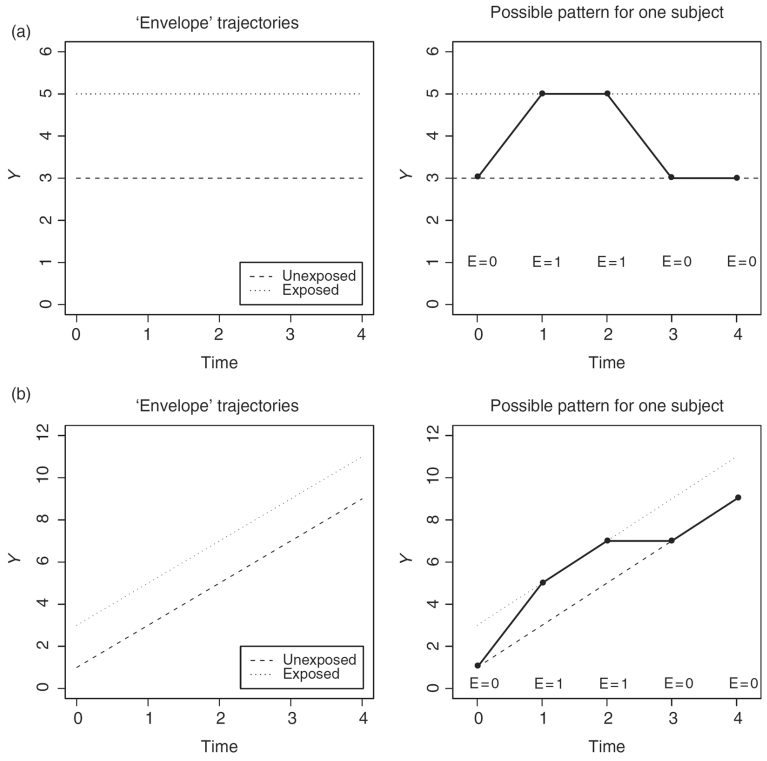

Figure 1.
Response patterns under the constant mean difference (CMD) model.

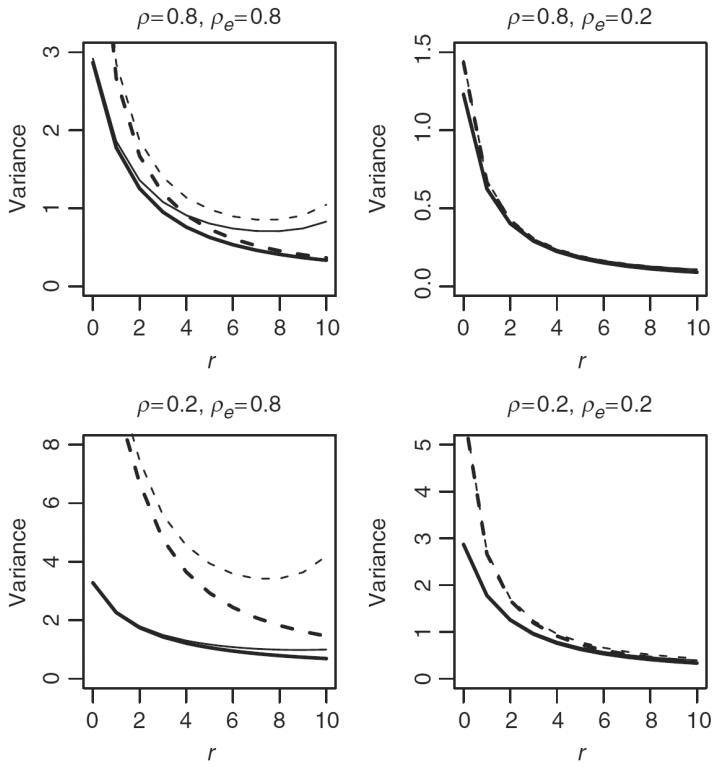

Figure 2.

Variance of the coefficient of interest for models (2.4)–(2.7) when both the response and the exposure have CS covariance with parameters ρ and ρ*e*, respectively, σ^2 1, $V(t_0) = 0$ and $p_{ej} = 0.2 + 0.05j$, where $j = 0, \ldots, r$. The lines indicate: (▬) model with only exposure (2.4); (▬ ▬ ▬) model with only exposure, conditional likelihood (2.5); (—) model with exposure and time (2.6); (- - -) model with exposure and time, conditional likelihood (2.7).

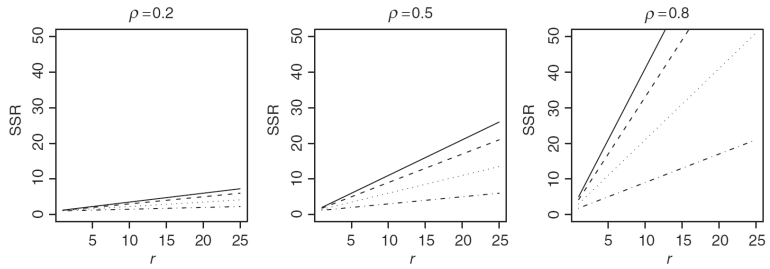

Figure 3.

$SSR = N_{\rho e=1}/N_{\rho e}$ (Equation (3.5)) for model (2.4) under CS of the response for several values of r, ρ and ρe. Lines indicate: (—) ρe = 0, (- - -) ρe = 0.2, (……) ρe = 0.6, (· - · - · -) ρe = 0.8.

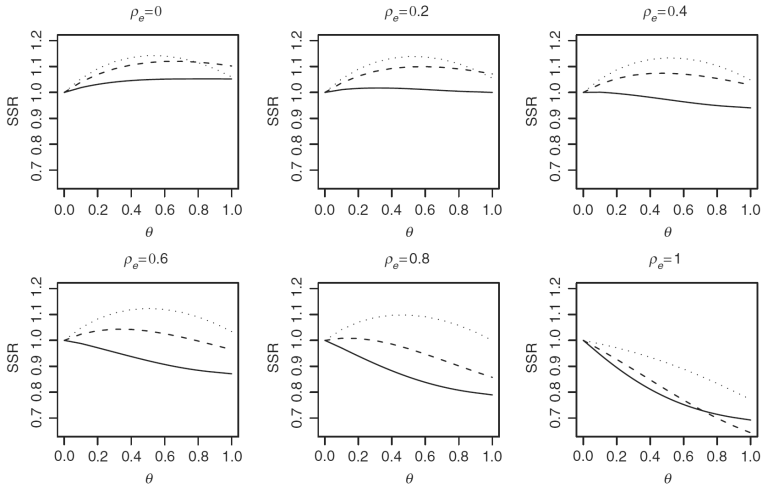

Figure 4.

SSR = $N_\theta/N_{\theta=0}$ as a function of θ assuming CS for the exposure process, for r 5, $\bar{p}_e$=0.2 and several values of ρ and ρe when model (2.4) is assumed. Lines indicate: (—) ρ = 0.2, (- - -) ρ = 0.5, (……) ρ = 0.8.

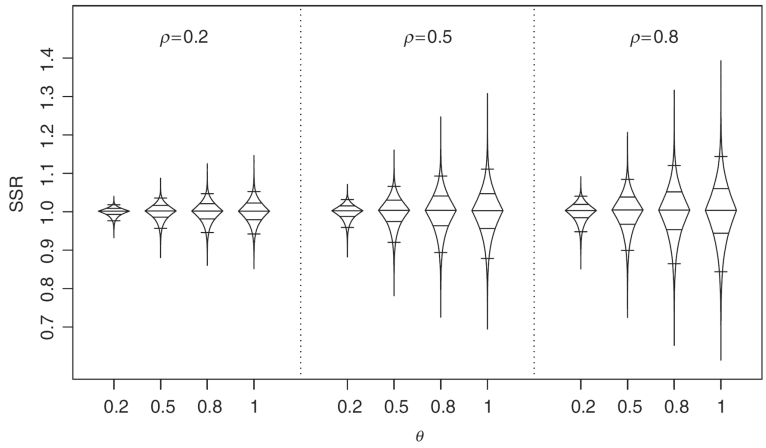

Figure 5.

Box-percentile plots of the ratio of required sample sizes obtained when assuming CS covariance of exposure divided by the required sample size obtained using the true exposure covariance in 10 000 scenarios generated to have an arbitrary exposure covariance, for $r = 5$, DEX response, model (2.7) and several values of ρ and θ. At any height, the width of the irregular `box' is proportional to the percentile of that height. Horizontal lines indicate the 5th, 25th, 50th, 75th and 95th percentiles.

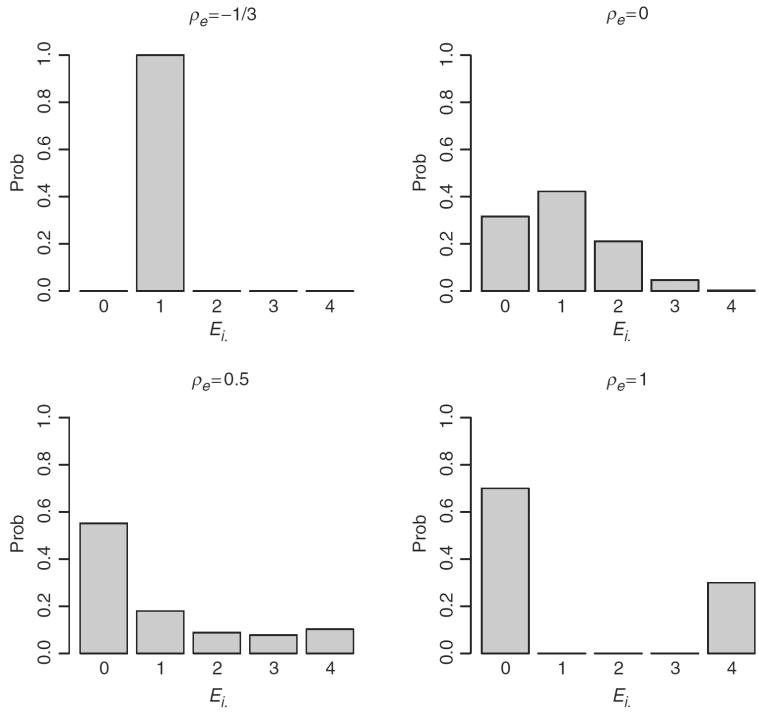

Figure 6.

Distribution of $E_{i\cdot}$ for $r = 3$, $\bar{p}_{e=0.25}$ and different values of ρe.

Table 1

Ratio of required sample sizes for several assumed exposure processes divided by the required sample size obtained using the observed exposure distribution, for the vacuuming and the air-freshener sprays exposure in the cleaners study, $r = 14$. Model (2.4), DEX covariance of the response and constant exposure prevalence over time are assumed

	Parameters for variance of response, Σ, assumed to be DEX					
	ρ = 0.8			ρ = 0.5		
Assumption for exposure process	**θ = 0 (CS)**	**θ = 0.5**	**θ = 1 (AR(1))**	**θ = 0 (CS)**	**θ = 0.5**	**θ = 1 (AR(1))**
Air freshener sprays[a]						
Time invariant	24.0	15.6	10.5	6.76	3.46	2.28
Independence	0.42	0.42	0.41	0.45	0.49	0.53
CS	1.00	0.98	0.95	1.00	1.00	0.97
Vacuuming[b]						
Time invariant	49.7	31.2	21.3	13.2	6.04	3.74
Independence	0.87	0.84	0.84	0.88	0.86	0.87
CS	1.00	0.96	0.96	1.00	0.97	0.97

[a] $\bar{p}_e = 0.17$, ρe = 0.59.

[b] $\bar{p}_e = 0.37$, ρe = 0.13.