Consequences of Ignoring Clustering in Linear Regression

Georgia Ntani (gn@mrc.soton.ac.uk)
MRC Lifecourse Epidemiology Unit https://orcid.org/0000-0001-7481-6860
Hazel Inskip
MRC Lifecourse Epidemiology Unit, University of Southampton
Clive Osmond
MRC Lifecourse Epidemiology Unit, University of Southampton
David Coggon
MRC Lifecourse Epidemiology Unit, University of Southampton

Research article

Keywords: Clustering, linear regression, random intercept model, consequences, simulation, comparison, bias

Posted Date: November 2nd, 2020

DOI: https://doi.org/10.21203/rs.3.rs-98069/v1

Version of Record: A version of this preprint was published at BMC Medical Research Methodology on July 7th, 2021. See the published version at https://doi.org/10.1186/s12874-021-01333-7.

Consequences of ignoring clustering in linear regression

Georgia Ntani[1,2], Hazel Inskip[1], Clive Osmond[1], David Coggon[1,2]

[1] Medical Research Council Lifecourse Epidemiology Unit, University of Southampton, United Kingdom

[2] Medical Research Council Versus Arthritis Centre for Musculoskeletal Health and Work, Medical Research Council Lifecourse Epidemiology Unit, University of Southampton, United Kingdom

Abstract

Background

Clustering of observations is a common phenomenon in epidemiological and clinical research. Previous studies have highlighted the importance of using multilevel analysis to account for such clustering, but in practice, methods ignoring clustering are often used. We used simulated data to explore the circumstances in which failure to account for clustering in linear regression analysis could lead to importantly erroneous conclusions.

Methods

We simulated data following the random-intercept model specification under different scenarios of clustering of a continuous outcome and a single continuous or binary explanatory variable. We fitted random-intercept (RI) and cluster-unadjusted ordinary least squares (OLS) models and compared the derived estimates of effect, as quantified by regression coefficients, and their estimated precision. We also assessed the extent to which coverage by 95% confidence intervals and rates of Type I error were appropriate.

Results

We found that effects estimated from OLS linear regression models that ignored clustering were on average unbiased. The precision of effect estimates from the OLS model was overestimated when both the outcome and explanatory variable were continuous. By contrast, in linear regression with a binary explanatory variable, in most circumstances, the precision of effects was somewhat underestimated by the OLS model. The magnitude of bias, both in point estimates and their precision, increased with greater clustering of the outcome variable, and was influenced also by the amount of clustering in the explanatory variable. The cluster-unadjusted model resulted in poor coverage rates by 95% confidence intervals and high rates of Type I error especially when the explanatory variable was continuous.

Conclusions

In this study we identified situations in which an OLS regression model is more likely to affect statistical inference, namely when the explanatory variable is continuous, and its intraclass correlation coefficient is higher than 0.01. Situations in which statistical inference is less likely to be affected have also been identified.

Keywords: Clustering, linear regression, random intercept model, consequences, simulation, comparison, bias

Introduction

Clinical and epidemiological research often uses some form of regression analysis to explore the relationship of an outcome variable to one or more explanatory variables. In many cases, the study design is such that participants can be grouped into discrete, non-overlapping subsets (clusters), such that the outcome and/or explanatory variables vary less within than between clusters. This might occur, for example, in cluster-randomised controlled trials (with the units of randomisation defining clusters), or in a multi-centre observational study (the participants from each centre constituting a cluster). The extent to which a variable is "clustered" can be quantified by the intra-class correlation coefficient (ICC), which is defined as the ratio of its variance between clusters to its total variance (both between and within clusters) (1).

Clustering has implications for statistical inference from regression analysis if the outcome variable is clustered after the effects of all measured explanatory variables are taken into account. If allowance is not made for such clustering as part of the analysis, parameter estimates and/or their precision may be biased. This possibility can be demonstrated by a hypothetical study of hearing impairment and noise exposure, in which observations are made in four different cities (clusters), as illustrated in Figure 1. In this example, the effect of cumulative noise exposure on hearing impairment is the same within each city (i.e. the regression coefficient for hearing impairment on noise exposure is the same in each cluster) (Figure 1a). However, after allowance for noise exposure, hearing impairment differs by city, such that it varies more between the clusters than within them. An analysis that ignored this clustering would give a misleading estimate for the regression coefficient of hearing loss on noise exposure (Figure 1b). Moreover, even if the distribution of noise exposures in each city was similar, so that the regression coefficient was unbiased, its precision would be underestimated as it would have made no allowance for the differences between clusters (at the intercept) (Figure 1c).

Where, as in the example above, the number of clusters is small relative to the total number of participants in the study sample, a categorical variable that distinguishes clusters can be treated as an additional explanatory variable in the regression model (2). However, when the number of clusters is larger, use of the cluster variable as an additional explanatory variable in the regression model can seriously reduce the precision with which effects are estimated. In such circumstances, an alternative approach is to assume that cluster effects are randomly distributed with a mean and variance that can be estimated from the data in the study sample. Random intercept models assume that the effects of explanatory variables are the same across all clusters, but that the intercepts of regression lines differ with a mean and variance which can be estimated from the study data, along with the effect estimates of primary interest. Random slope models assume that the effects of explanatory variables also differ between clusters, with a mean and variance that can be estimated.

In recognition of the potential implications of clustering for statistical inference, there has been a growth over recent years in the use of statistical techniques that allow for clustering (3). Nevertheless, many studies still ignore clustering of observations (4-8). Recent systematic reviews have reported that clustering was taken into account in only 21.5% of multicentre trials (9) and 47% of cluster randomised trials (10). This may in part reflect computational challenges and statistical complexities (11), but, perhaps because of a lack of clarity about the effects of ignoring clustering, authors have omitted to discuss the limitations of their chosen analytical techniques.

Several studies have investigated implications of ignoring clustering in statistical inference, most being based on analysis of real data (1, 12-19). To date, no study has systematically investigated the extent to which bias can occur in effect estimates when clustering is ignored, the determinants of that bias, or the exact consequences for the precision of estimates according to different distributions of the explanatory variable and, in particular, the extent to which the explanatory variable varies within as compared with between clusters.

The first aim of the research described in this paper was to assess in detail the implications for effect estimates (regression coefficients), and their precision (characterised by standard errors (SEs)), when a linear regression analysis exploring the relation of a continuous outcome variable to an explanatory variable fails to account for clustering. The second aim was to describe rates of Type I error and coverage by 95% confidence intervals in the same setting. These research questions were explored through simulation studies.

INSERT FIGURE 1 HERE

Figure 1. Hypothetical relationship of hearing impairment to cumulative noise exposure in four cities. Units for noise exposure and hearing impairment have been specified arbitrarily for ease of presentation. Data for each city are distinguished by the shading of data points. Cluster-specific regression lines are indicated, along with the regression line for the full dataset when clustering is ignored (dotted red line), and that when adjustment is made for cluster (solid blue line)

Methods

In the simplest case, in which there is a single explanatory variable, the ordinary least squares (OLS) linear regression is specified by a model of the form:

$$y_i = \beta_0 + \beta_1 x_i + e_i \qquad \text{-1-}$$

For a continuous outcome and a single explanatory variable, the random intercept (RI) multi-level model can be viewed as an extension of the OLS model, and is specified as:

$$\begin{aligned} y_{ij} &= \beta_{0j} + \beta_1 x_{ij} + e_{ij} \\ &= \beta_0 + \beta_1 x_{ij} + e_{ij} + u_j \end{aligned} \qquad \text{-2-}$$

where the index i refers to the individual and the index j to the cluster, and $\beta_{0j} = \beta_0 + u_j$, the estimate of the intercept for cluster j. The term u_j represents the error for cluster j around the fixed intercept value of β_0, and is assumed to be normally distributed with $u_j|x_{ij} \sim N(0, SD_u^2)$. The term e_{ij} represents the additional error within the cluster, also referred to as the individual level error term, with $e_{ij}|x_{ij}, u_j \sim N(0, SD_e^2)$.

As described in the introduction, ICC is a measure which characterises the extent to which the outcome variable y_{ij} is similar within clusters, given the distribution of the explanatory variable x_{ij}

(20). For a continuous outcome variable, and with the nomenclature used above, the ICC is defined as $ICC = \frac{SD_u^2}{SD_u^2 + SD_e^2}$ (21).

To explore the study questions, simulated datasets were generated according to the assumptions of the RI model. For each Monte Carlo simulation, both the number of clusters and the number of observations per cluster were set to 100. For simplicity, the size of the effect of x_{ij} on y_{ij} was arbitrarily set to 1 ($\beta_1 = 1$), and the average value of y_{ij} when x_{ij}= 0 was arbitrarily set to 0 ($\beta_0 = 0$).

Separate simulation studies were generated for a continuous and a binary explanatory variable x_{ij}. To set values x_{ij} for the continuous explanatory variable in a cluster j, an individual level variable was generated as $x_{0ij} \sim N(0,1)$, and a cluster-specific variable as $shift_j \sim N\left(0, SD_{shift}{}^2\right)$. The individual level variable was then added to the cluster-specific shift, so that $x_{ij} = x_{0ij} + shift_j$. For a binary explanatory variable x_{ij}, we set the prevalence in each cluster to be the sum of a constant (the same in all clusters) set to 0.05, 0.1, 0.2 and 0.4 and a cluster-specific variable $shift_j \sim N\left(0, SD_{shift}{}^2\right)$. In both cases, the corresponding values for the outcome variable y_{ij} were generated according to equation -2-. For this purpose, the individual-level error terms were drawn from a random standard normal distribution $\left(N(0,1)\right)$, and the cluster-level error terms were drawn from a random normal distribution with mean zero and variance $SD_{u_j}^2$. Simulated data were generated for various different values for SD_{u_j} (0.0316, 0.05485, 0.1005, 0.1759, 0.3333 and 0.6547) chosen to give expected values for the ICC of 0.001, 0.003, 0.01, 0.03, 0.1 and 0.3 respectively, while $SD_{shift} \sim U[a, b]$, with the parameters a and b being arbitrarily chosen to be 0 and 15, in the case of a continuous x_{ij}, and 0 and 0.05 in the case of a binary x_{ij}.

For each simulated dataset, two linear regression models were fitted; an OLS model which ignored the clustering (equation -1-), and a RI multi-level model which allowed for clustering effects (equation -2-). For each of the models, the regression coefficient and its standard error (SE) were estimated. To compare results from the two models, the difference between the estimated regression coefficients ($\beta_1^{RI} - \beta_1^{OLS}$), and the ratio of their SEs (SE^{RI}/SE^{OLS}) were calculated.

To assess how the comparison between the two models was affected by the distribution of x_{ij} within and between clusters, these two measures were plotted against the dispersion (expressed as standard deviation) of the mean values of x_{ij} ($\bar{x}_j$) between clusters (dispersion of $shift_j$), for the case of continuous x_{ij}, and dispersion of prevalence of x_{ij}, for the case of binary x_{ij}. In addition, descriptive statistics were produced for the distributions of the two measures across simulated samples, according to values for expected ICC and overall prevalence of x_{ij}, in the case of a binary explanatory variable.

The accuracy of the 95% confidence intervals for the regression coefficient β_1 from the two methods was assessed by calculating the proportion of the estimated confidence intervals that included the true value that had been used in the simulations. A method was considered to have appropriate coverage if 95% of the 95% confidence intervals included the value of the effect β_1 (i.e. the value 1) used in the simulations. Deviations from this ideal could reflect bias in the estimates of effect, unsatisfactory standard errors (22), or both.

To assess impacts on type I error, the simulations were repeated assuming no association between x_{ij} and y_{ij} (i.e. $\beta_1 = 0$), and the proportions of datasets for which the null hypothesis was rejected at a 5% significance level in OLS and RI modelling were compared according to ICC.

For each expected ICC, and each value of $shift_j$, 100 simulated datasets were produced with a continuous x_{ij}, and another 100 for each of the four overall prevalence rates of a binary x_{ij}.

Due to random sampling variation the estimated ICC values were within given ranges of the target levels of ICC. For target levels of 0.001, 0.003, 0.01, 0.03, 0.1 and 0.3, these ranges were 0.0005-0.0014, 0.0025-0.0034, 0.005-0.014, 0.025-0.034, 0.05-0.14, and 0.25-0.34 respectively. Simulations resulting in estimated ICC values outside of these ranges were discarded and not used further. In the description of the results that follows ICC values are labelled according to the target levels.

All simulations and analysis were conducted using Stata software v12.1.

Results

Difference in regression coefficients

Differences in regression coefficients ($\beta_1^{RI} - \beta_1^{OLS}$) estimated from the two linear models are illustrated in Figure 2. The two different subplots of the figure (A and B) correspond to the two different distributions of the explanatory variable (continuous and binary respectively), and the different shades of grey correspond to different ICC levels with darker shades corresponding to simulated results for higher ICCs.

INSERT FIGURE 2 HERE

Figure 2. Difference between regression coefficients estimated from RI and OLS models ($\beta_1^{RI} - \beta_1^{OLS}$) plotted against dispersion (expressed as SD) of mean value/prevalence of x_{ij}, for different levels of intraclass correlation (shades of grey as indicated in the legend). Figure A: Continuous x_{ij}. Figure B: Binary x_{ij}

In all cases, differences in regression coefficients were on average zero, with β_1^{RI} and β_1^{OLS} being on average $\cong 1$. For both continuous and binary distributions of x_{ij}, differences were on average more narrowly spread for small ICCs and more widely spread for large ICCs. For a continuous explanatory variable x_{ij} (Figure 2A), and for each value of ICC, increasing the dispersion of $\bar{x}_j$ across clusters resulted in larger differences in regression coefficients up to a dispersion of $\bar{x}_j = 1$ (i.e. same dispersion of x_{ij} between and within clusters). Beyond that point, further increase in the dispersion of $\bar{x}_j$ resulted in smaller differences in regression coefficients from the two methods, approaching a difference of zero.

For a binary explanatory variable x_{ij}, and for each value of ICC, small dispersion of cluster-specific prevalence of x_{ij} resulted in small differences between the regression coefficients. However, increasing the dispersion of cluster-specific prevalence of x_{ij}, resulted in larger differences between the regression coefficients from the two methods. Comparing the different subplots of **Error!**

Reference source not found. (note the different scales on the y-axes), higher overall prevalence of x_{ij} resulted in regression coefficients from the two models being more similar even for large dispersion of the prevalence of x_{ij} across clusters; for ICC=0.3, differences ranged from -0.2 to 0.2, corresponding to a 20% difference in the regression coefficients from the two methods, when the overall prevalence of x_{ij} was 0.05, and this range decreased to approximately -0.05 to 0.05 for an overall prevalence of x_{ij} of 0.4.

Ratio of standard errors

The ratios of SEs derived from the RI and OLS models ($SE_{\beta_1^{RI}}/SE_{\beta_1^{OLS}}$) were examined in relation to the dispersion across clusters of the mean value/prevalence of the continuous/binary explanatory variable x_{ij}, and are presented in Figure 3. As in Figure 2, the different levels of ICCs are represented by different shades of grey, with lighter shades corresponding to lower ICCs and darker shades to higher ICCs. Subplots A and B correspond to the ratios of SEs when x_{ij} was continuous and binary, respectively.

For a continuous variable x_{ij}, the ratio took its minimum value for the smallest dispersion of $\bar{x}_j$ and increased as dispersion of $\bar{x}_j$ increased, tending asymptotically to a maximum value. The minimum and maximum values of the ratio of the SEs (the latter also corresponding to its asymptote) were ICC-dependent, higher ICCs resulting in lower minimum and higher maximum values for the ratio. The dispersion of $\bar{x}_j$ at which the ratio of SEs approached its asymptote was also ICC-dependent, being higher for larger ICCs. For very small values of dispersion of $\bar{x}_j$, the minimum value of the ratio of the SEs was approximately one for small levels of ICC and was less than one for higher ICCs. Particularly for small values of the dispersion of $\bar{x}_j$ and ICC $\cong$ 0.10 or 0.30, the ratio of SEs was <1, meaning that SEs from RI models were smaller than from OLS models.

INSERT FIGURE 3 HERE

Figure 3. Ratios of standard errors estimated from RI and OLS models ($SE_{\beta_1^{RI}}/SE_{\beta_1^{OLS}}$) plotted against relative between- to within-clusters dispersion (expressed as SD) of explanatory variable x_{ij}. Figure A: Continuous x_{ij}. Figure B: Binary x_{ij}

When x_{ij} was binary, the ratios of the SEs were below one for most of the situations examined, indicating that the SEs of the regression coefficients estimated from the RI model were smaller than those under the OLS model in most circumstances. The ratio of the SEs achieved its minimum value for the smallest dispersion of the prevalence of x_{ij} across the clusters, and increased progressively with increasing dispersion of x_{ij} across clusters. For small ICCs (<0.1), the SEs from the two models were very similar. However, increasing the ICC to 0.1 or higher led to the ratio of the SEs decreasing to values much lower than 1. For constant ICC, comparison of subplots of Figure 3B, shows that the rate of increase of the ratio of the SEs was higher for lower underlying prevalence rates of the x_{ij}.

Coverage of 95% confidence intervals

Table 1 shows the extent to which 95% confidence intervals covered the simulated effect of the explanatory continuous variable on the outcome (β_1=1), when derived from the two statistical models, for different levels of ICC, and for fifths of the distribution of the dispersion of $\bar{x}_j$.

Irrespective of ICC and type of explanatory variable, coverage with the RI model was approximately 95%. For a continuous x_{ij}, coverage for the OLS model was close to 95% for very low ICC and decreased for increasing levels of ICC. For the highest ICC level examined (ICC=0.3), OLS showed a notably poor coverage of 30%. For a given ICC, coverage of 95% confidence intervals did not vary much by dispersion of $\bar{x}_j$, although it was somewhat higher in the bottom fifth as compared to the 2nd, 3rd, 4th, and 5th fifth of the distribution of dispersion of $\bar{x}_j$.

Table 1. Coverage (%) by 95% confidence intervals of simulated effect β_1=1 under the RI and OLS models according to fifths of the distribution of dispersion (expressed as SD) of the continuous $\bar{x}_j$

ICC	Bottom fifth=1		2		3		4		Top fifth=5		Total	
	RI	OLS	RI	OLS	RI	OLS	RI	OLS	RI	OLS	RI	OLS
0.001	95.04	94.37	95.00	94.09	95.33	94.10	95.15	93.99	94.92	93.84	95.08	94.08
0.003	95.14	93.14	95.37	92.33	95.20	91.83	95.53	92.25	95.42	91.99	95.33	92.30
0.01	94.99	88.47	94.64	83.95	94.75	83.65	94.74	83.72	94.90	84.07	94.80	84.75
0.03	94.59	76.21	95.11	68.79	94.80	67.62	95.06	67.57	94.74	67.15	94.87	69.39
0.1	94.68	59.58	94.80	45.45	94.86	44.83	94.39	44.73	95.04	44.37	94.76	47.80
0.3	94.84	41.32	94.53	28.06	94.95	28.24	94.64	26.98	94.79	27.41	94.75	30.36

For a binary x_{ij}, coverage for the OLS model was close to 95% but only for ICC≤0.03. As ICC increased, coverage from the OLS model deviated from the nominal value of 95%. As shown in Figure 4, when ICC was 0.1 or 0.3, coverage was on average lower for lower prevalence of x_{ij}; it fell below the nominal value of 95% for 0.05 prevalence of x_{ij} and it increased to values higher than 95% for 0.40 prevalence of x_{ij} (comparison of the four sub-plots of the figure). Also, for any given prevalence of x_{ij}, coverage was lower for increasing dispersion of prevalence of x_{ij} across clusters. Variation of the average coverage by categories of prevalence rates of x_{ij} and overall prevalence of x_{ij} was higher when ICC was higher (ICC=0.3) than when it was lower (ICC=0.1). The smallest and the largest values of coverage were 87% and 98% and they were observed when overall prevalence of x_{ij} was 0.05, ICC=0.3, and in the bottom and top thirds respectively of the distribution of dispersion of prevalence of x_{ij} across clusters. Coverage as high as 98% was also seen in the bottom third of the distribution of dispersion of prevalence of x_{ij} across clusters for the other prevalence rates (0.10, 0.20, and 0.40) explored when ICC was high (ICC=0.3).

INSERT FIGURE 4 HERE

Figure 4. Coverage (%) by 95% confidence intervals from the OLS model for ICC=0.1 and 0.3, by overall prevalence rates of x (A) 0.05, B) 0.10, C) 0.20, and D) 0.40), and thirds of the distribution of the dispersion (expressed as SD) of prevalence of across clusters

Type I error

To assess the frequency of type I error, defined as incorrect rejection of a true null hypothesis, under the OLS and the RI multi-level models, simulations were repeated assuming no association between the explanatory variable x_{ij} and the outcome variable y_{ij} ($\beta_1^{RI} = \beta_1^{OLS} = 0$).

Figure 5 shows the proportion of datasets for which the null hypothesis was rejected at a 5% significance level for varying levels of ICC, when x_{ij} was continuous. Using the RI multi-level model, the association between x_{ij} and y_{ij} was statistically significant in approximately 5% of the datasets for all ICCs. However, using the OLS models, type I error varied with ICC. For a very small ICC, type I error was very close to that under the RI model (~6%) but increased rapidly as the ICC increased, reaching ~70% for ICC≅0.30. Type I error did not vary by dispersion of mean value of x_{ij} (data not shown).

INSERT FIGURE 5 HERE

Figure 5. Proportion (%) of datasets for which the null hypothesis was rejected according to level of ICC when $\beta_1^{RI} = 0$ and x_{ij} was continuous

When the explanatory variable x_{ij} was binary, type I error rates varied very little around the nominal level of 5% when an OLS model was fitted instead of the RI model, when ICC values were less than 0.1; the average value was 5% and varied from 4.8% to 5.3% for different ICC values (<0.1), overall prevalence rates of x_{ij}, and dispersion of prevalence of x_{ij} across clusters. However, for ICC values of 0.1 and 0.3, type I error rates diverged from 5%. The variation of rates in those cases is illustrated in Figure 6 for the four prevalence rates of x_{ij} (subplots A, B, C, and D of the figure), and for thirds of the distribution of dispersion of prevalence of x_{ij} across clusters. For small dispersion of prevalence rates of x_{ij} (bottom third of the distribution), type I error was lower than 5%, and it increased as dispersion increased. This trend was more prominent for lower values of overall prevalence of x_{ij}, and for ICC=0.3 compared to ICC=0.1. The smallest and the largest values of type I error were 2% and 13% and they were observed when overall prevalence of x_{ij} was 0.05 and in the bottom and top thirds respectively of the distribution of dispersion of prevalence of x_{ij} across clusters.

INSERT FIGURE 6 HERE

Figure 6. Type I error rates (%) from the OLS model for ICC=0.1 and 0.3, by overall prevalence rates of x_{ij} (A) 0.05, B) 0.10, C) 0.20, and D) 0.40), and thirds of the distribution of the dispersion (expressed as SD) of prevalence of x across clusters

Discussion

In this paper we focused on the implications of ignoring clustering in statistical inference regarding the relationship between a continuous outcome and a single explanatory variable x_{ij}. Two different types of x_{ij} were considered – continuous and binary. For each of the two categories of x_{ij}, the implications for statistical inference of failing to account for clustering were explored by comparison of effect estimates and their precision, assessment of the coverage by 95% confidence intervals, and estimation of the frequency of type I error. In the cases of both a continuous and a binary x_{ij}, where the true slope of the regression line was non-zero, we found that the cluster-unadjusted OLS and RI models gave on average very similar estimates of effect for any level of ICC. However, despite the average value of difference in point estimates from the two methods being zero, differences occurred in both directions and varied more when the level of ICC increased. The largest differences in estimates of effect between OLS and multi-level RI regression modelling were only about 20% of the true value and they occurred when the ICC was high (0.3). For a continuous x_{ij}, the largest errors in the differences of estimated effects occurred when the dispersion of the x_{ij} within clusters was approximately the same as that between clusters, while, for a binary explanatory variable, differences increased with increasing dispersion of prevalence of x_{ij} across clusters.

Conclusions drawn from comparison of SEs estimated from cluster-unadjusted OLS and RI models are somewhat different for continuous as compared with binary x_{ij}. When x_{ij} was continuous, the SEs of regression coefficients were generally larger for the multi-level RI model than for the cluster-unadjusted OLS model, their ratio being highest (>4) for a high ICC (0.3) and where the dispersion of the mean value of x_{ij} was large. However, contrary to what is widely stated, the spuriously greater precision of OLS method was not universal. When dispersion of mean values of x_{ij}<1, OLS regression gave larger SEs than multi-level modelling. When x_{ij} was binary, SEs estimated from the RI model, were higher than those from the cluster-unadjusted OLS model for lower ICCs (<0.03) and larger dispersion of prevalence of x_{ij} across clusters, and lower than those from the cluster-unadjusted OLS model for smaller dispersion of prevalence of x_{ij} across clusters. The SEs differed by up to 15% for the highest ICC value (ICC=0.3).

The rates of coverage of 95% confidence intervals for estimates of effect, whether of a continuous or a binary x_{ij}, when derived from a RI model were at the nominal level of 95%, irrespective of other parameters (i.e. ICC, dispersion across clusters of the mean value of a continuous x_{ij}, or dispersion of the prevalence of the binary x_{ij} across clusters). When x_{ij} was binary, the cluster-unadjusted OLS model also resulted in an appropriate coverage of the 95% confidence intervals when ICC was low (≤ 0.01). However, for higher values of ICC, coverage varied slightly (range: 87% - 98%) around the nominal value of 95% depending on the overall prevalence and the dispersion of the cluster-specific prevalence rates of x_{ij}. In contrast, when x_{ij} was continuous, the model that failed to account for clustering resulted in poor coverage rates, especially as ICC increased, reaching a rate as low as 30% for ICC=0.3.

Setting the effect of x_{ij} on the outcome variable to zero allowed exploration of the frequency of type I error. With the RI model, in all of the scenarios explored, type I error was very close to 5%. When x_{ij} was continuous, we found that failure to allow for clustering increased rates of Type I error, and that the inflation of type I error was particularly pronounced (up to 70%) when the degree of clustering was high (ICC=0.3). In contrast to this, when x_{ij} was binary, type I error under the OLS model was close to the expected value of 5% for low levels of clustering (ICC<0.1). However, when ICC was high (0.1 or 0.3), type I error rates varied more widely around 5%, with values as low as 2% (for low overall prevalence of x_{ij} and small dispersion of its prevalence across clusters) and as high as 13% (for low overall prevalence of x_{ij} and large dispersion of its prevalence across clusters).

The analysis for each specification of parameters (expected ICC, dispersion of x_{ij}, overall prevalence or dispersion of prevalence rates across clusters of a binary x_{ij}) was based on 1,000 simulated samples of 10,000 observations grouped in 100 clusters, each of 100 individuals. By using such a large sample size (larger than in most epidemiological investigations), we reduced random sampling variation, making it easier to characterise any systematic differences between the two methods of analysis. However, the approach may have led to underestimation of the maximum differences between estimates of effect that could arise from OLS as compared with multi-level modelling. Additionally, the number of observations per cluster was the same in all simulations, making it impossible to draw conclusions about effects of ignoring clustering for varying cluster sizes. Also, data were simulated following the specification of the RI regression model rather than that of the random-effects model described in section -2-. That was done because the RI model is more frequently used, especially when there is no a priori expectation of differential effects of the explanatory on the outcome variables across the different clusters. Simulating data following the specification of the random effects model would have added complexity to the algorithm used for simulation, and the computational time required.

The effect of clustering when a cluster-unadjusted model is fitted could also have been assessed by calculating bias as [(estimated effect – true effect)/true effect], as defined in earlier studies (23). Instead, we defined bias by the difference in the effect estimates derived from the two analytical models. The data were simulated following the model specification of RI linear regression, which is one of the most well established and frequently chosen analytical approaches to account for clustering. As such, given that all resulting effect estimates were positive, deviations of the difference in regression coefficients from the value of zero can only represent deficiencies of the OLS model, provided that the assumptions of the RI model are met. Therefore, there is no reason to expect that the conclusions one would draw from an alternative definition of bias would be more reliable, provided that the conditions under which data were simulated and the models fitted were the same.

When multilevel RI modelling was applied to the simulated clustered datasets with a continuous or a binary explanatory variable, the rate of Type I error was 5%, and the coverage by 95% CIs was 95%, as would be expected, given the method by which the simulated samples were generated. In comparison, when cluster-unadjusted models were fitted to clustered data with a continuous x_{ij}, rates of Type I error were higher, particularly when the ICC was high. For the highest level of ICC examined (0.3), type I errors were as frequent as 70%. However, even with an ICC of only 0.01, rates of Type 1 error were more than 10%. Consistent with this, coverage by 95% confidence intervals was considerably lower than the nominal value for higher ICC levels. The lowest coverage of 30% was for the highest ICC level. In contrast to these results Huang et al (24) have reported values of coverage very close to 95% from the OLS model for a continuous explanatory variable. Differences between findings presented in this study and those presented by Huang et al (24) can be explained by zero clustering in the explanatory variable assumed in the latter. Sensitivity analysis restricting the simulated datasets only to those in which clustering in the explanatory variable was not meaningful showed that interval coverage rates were very close to 95% independent of clustering in the outcome variable (results not shown). When x_{ij} was binary, both the interval coverage and Type I error rates varied little around the nominal values of 5% and 95%, and only for ICC values higher than 0.01. Overall coverage rates were higher for higher ICCs and decreased for increasing dispersion of the cluster-specific prevalence rates of x_{ij} across clusters and for decreasing overall prevalence of the x_{ij}. A similar observation of small variation of interval coverage around 95% for higher ICC values has been made before (25). Type I error when x_{ij} was binary and its prevalence was low, varied around 5% with values falling below 5% for small dispersion of prevalence of x_{ij}, and above 5% for large dispersion. For larger overall prevalence of x_{ij}, Type I error rates fell below 5%. In accordance with these findings, Galbraith et al (26) have shown that cluster-unadjusted models resulted in relatively conservative Type I error. Also, in a context of individually randomised trials, Kahan et al (27) have shown that Type I error increased with increasing ICC and increasing difference in the probability of assignment of patients to treatment arms.

It has been widely stated that when data are clustered, effects estimated by OLS regression are unbiased (23, 25, 27-30). Our results confirm that for data of the type simulated, coefficients from OLS regression were on average very similar to those from RI multi-level modelling. Previous studies based on simulation data have shown similar results (23-25, 31). However, for individual simulated samples, the estimates may differ, and the potential magnitude of the differences depends on the level of within-cluster similarity of the outcome variable. For an ICC of 0.3, the estimates of effect from the two analytical methods could differ by up to 20%. In addition, when x_{ij} is continuous, the error in estimates of the regression coefficient is larger when the between-cluster dispersion of x_{ij} is similar to that within-cluster. When x_{ij} is binary, the error increases as the dispersion of the prevalence rates across clusters increases, and when the overall prevalence rate across all clusters is lower (<10%). These errors in the estimated effect indicate that in an individual study, failure of regression analysis to account for clustering of observations could result in considerably higher or lower estimates of effect than those derived from multilevel analysis. This has been illustrated in numerous published papers of real data, which have shown that estimates from the two analytical methods can differ to a lesser or greater extent (1, 8, 14, 17, 32). However, in those publications, no or very limited information is provided to establish whether the error observed was due to dispersion of the cluster-specific mean values of the continuous x_{ij}, or dispersion of prevalence rates for the binary x_{ij} across clusters.

Most often it is stated that regression coefficients are spuriously precise when clustering is not taken into account in regression models. However, in several reports, authors have failed to specify the conditions under which this applies (1, 31, 33-36). Other authors have pointed out that when x_{ij} is identical within each cluster, and a cluster-unadjusted approach is followed, SEs tend to be spuriously low, and that the opposite occurs when x_{ij} varies within clusters (24, 27, 37, 38). Bias in SEs for effects of cluster-varying x_{ij} has been shown in results from real data when both models were fitted

(17, 32, 39). However, others have reported contradictory results in which SEs of effects of individual-level x_{ij} from OL regression were very similar to, or lower than, those from a multi-level model (14-16, 40). It should be noted that the dichotomy between cluster- and individual-level variables is not clear-cut. There can be varying degrees of clustering in x_{ij}, with the extremes being variables for which the values are completely unclustered (mean values are the same for all clusters), and variables for which the values are the same within each cluster. However, in real data, an explanatory variable can lie anywhere in between. An early report focused on this issue by considering the level of clustering in x_{ij} as the main driver for the expected bias of the precision of the effect estimates (29), rather than the absolute distinction between cluster-constant and cluster-varying x_{ij}. The authors reported that as clustering in x_{ij} decreases, the bias in SEs from a cluster-unadjusted model is expected to be upwards, and the opposite is expected when clustering in x_{ij} increases. Taking into consideration clustering in x_{ij} (ρ_x) as well as in the outcome variable (ρ_y), a later study using simulated data showed that for a given level of ρ_y, increasing ρ_x resulted in increasing the ratio of estimated SEs ($SE_\beta^{RI}/SE_\beta^{OLS}$) from values <1 to values ≈ 1 (41). Our results for continuous explanatory variables differ slightly from this, with ratios of SEs ($SE_\beta^{RI}/SE_\beta^{OLS}$) moving from values <1 to values >1, as clustering in the explanatory variable, expressed as dispersion of $\bar{x}_j$ across clusters, increased.

Bias in the precision of effect estimates for binary x_{ij} when clustering is ignored has received very limited attention in the published literature. Several of the reported studies have used real data to compare standard and multi-level models, using both continuous and binary individual-level x_{ij} (14, 17). For the majority of binary x_{ij} used in the models fitted in these studies, SEs derived from the OLS model were larger than those derived from the multi-level model. The same conclusion was drawn from a study using simulated data (25). However, none of the studies using real data has explored the level of bias in relation to variation in the prevalence of the binary x_{ij}, and the study of simulated data assumed constant prevalence of x_{ij} in all clusters. Simulation results presented here suggest that, irrespective of the dispersion of prevalence of x_{ij} across clusters and the overall prevalence in all clusters, in most circumstances SEs from the multi-level model are lower than those from the OLS model, and the bias is higher for higher ICC values.

The focus of this paper was on the association between a continuous outcome and an explanatory variable that was defined at the individual level (x_{ij} within cluster). We showed that when x_{ij} was continuous, and most of the variation occurred within rather than between clusters, the cluster-unadjusted OLS model gave larger SEs for the regression coefficient than multi-level modelling. This is consistent with reports in which ignoring clustering resulted in spuriously high SEs when x_{ij} varied within cluster. The reverse occurred when most of the dispersion of x_{ij} was between rather than within clusters. In this situation x_{ij} approaches the characteristics of a cluster-specific variable. We additionally showed that when x_{ij} under investigation was binary, ignoring clustering in statistical modelling in most cases resulted in higher SEs for the estimated effect than those derived from the random-intercept model. The SEs differed more for higher ICCs but not with the overall prevalence of x_{ij}, nor with the dispersion of its prevalence across clusters (Figure 3B). Unlike SEs, the point estimates were unbiased for either continuous or binary x_{ij} (Figure 2 A and B).

In conclusion, our results support the use of multi-level modelling to account for clustering effects in linear regression analyses of data that are hierarchically structured, especially where ICCs might exceed 0.01. Failure to do so is likely to result in incorrect estimates of effect (either too high or too low) mostly with spurious precision in the case of continuous x_{ij} or with underestimated precision in the case of binary x_{ij}, and may lead to incorrect inferences. The errors in estimates of effect of a continuous x_{ij} will be smaller when most of its dispersion is between rather than within clusters – i.e. the variable comes closer to being cluster-specific. Similarly, when x_{ij} is binary, smaller differences in the effect estimates occur when the dispersion of the prevalence of x_{ij} across clusters is small, or when its overall prevalence across clusters is high.

Additionally, we identified situations in which a standard analytical approach is more likely to importantly affect statistical inference, i.e. when rates of Type I error and interval coverage deviate more from the nominal values of 5% and 95% respectively. These occur when x_{ij} is continuous, and ICC levels are greater than 0.01. It is then that Type I error rates are higher than 10% and interval coverage rates are lower than 80%. On the other hand, statistical inference when a standard regression model is fitted is less likely to be of concern when x_{ij} is binary, as the error and coverage rates deviate very little from the nominal values. However, even for a binary x_{ij}, error rates can sometimes be greater than 10%, and corresponding interval coverage rates lower than 90% (but possibly not lower than 80%). This occurs when ICC is high, the overall prevalence of x_{ij} is low (approximately 5%), and the dispersion of the cluster-specific rates is large. In all circumstances in which the ICC is very small, clustering is minimal and there is little difference between RI and OLS regression.

Abbreviations

RI: random-intercept; OLS: ordinary least squares; ICC: intra-class correlation coefficient; SE: standard error

Acknowledgements

Not applicable

Authors' contribution

GN, HI, and DC conceived the concept of this study. GN carried out the simulations, analysed the data and drafted the manuscript. CO provided expert statistical advice on aspects of results presented. DC and HI critically reviewed and made substantial contributions to the manuscript. All authors read and approved the final manuscript.

Funding

During completion of this work, GN was supported by the Colt Foundation (PhD scholarship) and a grant award from Versus Arthritis (formerly Arthritis Research UK) (22090).

Availability of data and materials

The simulated datasets used and analysis described in the current study are available from the corresponding author on reasonable request.

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Competing interests

The authors declare that they have no competing interests

References

1. Park S, Lake ET. Multilevel modeling of a clustered continuous outcome: nurses' work hours and burnout. Nursing research. 2005;54(6):406-13.
2. Stimson JA. Regression in Space and Time: A Statistical Essay. American Journal of Political Science. 1985;29(4):914-47.
3. Bingenheimer JB, Raudenbush SW. Statistical and substantive inferences in public health: issues in the application of multilevel models. Annu Rev Public Health. 2004;25:53-77.
4. Bland JM. Cluster randomised trials in the medical literature: Two bibliometric surveys. BMC Medical Research Methodology. 2004;4.
5. Crits-Christoph P, Mintz J. Implications of therapist effects for the design and analysis of comparative studies of psychotherapies. Journal of consulting and clinical psychology. 1991;59(1):20.
6. Lee KJ, Thompson SG. Clustering by health professional in individually randomised trials. Bmj. 2005;330(7483):142-4.
7. Simpson JM, Klar N, Donner A. Accounting for cluster randomization: A review of primary prevention trials, 1990 through 1993. American Journal of Public Health. 1995;85(10):1378-83.

8. Biau DJ, Halm JA, Ahmadieh H, Capello WN, Jeekel J, Boutron I, et al. Provider and center effect in multicenter randomized controlled trials of surgical specialties: an analysis on patient-level data. Ann Surg. 2008;247(5):892-8.

9. Oltean H, Gagnier JJ. Use of clustering analysis in randomized controlled trials in orthopaedic surgery. BMC Medical Research Methodology. 2015;15(1).

10. Diaz-Ordaz K, Froud R, Sheehan B, Eldridge S. A systematic review of cluster randomised trials in residential facilities for older people suggests how to improve quality. BMC Medical Research Methodology. 2013;13(1).

11. Goldstein H. Multilevel Mixed Linear Model Analysis Using Iterative Generalized Least Squares. Biometrika. 1986;73(1):43-56.

12. Astin AW, Denson N. Multi-campus studies of college impact: Which statistical method is appropriate? Research in Higher Education. 2009;50(4):354-67.

13. Cheong YF, Fotiu RP, Raudenbush SW. Efficiency and robustness of alternative estimators for two- and three-level models: The case of NAEP. Journal of Educational and Behavioral Statistics. 2001;26(4):411-29.

14. Grieve R, Nixon R, Thompson SG, Normand C. Using multilevel models for assessing the variability of multinational resource use and cost data. Health economics. 2005;14(2):185-96.

15. Niehaus E, Campbell CM, Inkelas KK. HLM Behind the Curtain: Unveiling Decisions Behind the Use and Interpretation of HLM in Higher Education Research. Research in Higher Education. 2014;55(1):101-22.

16. Steenbergen MR, Jones BS. Modeling multilevel data structures. american Journal of political Science. 2002:218-37.

17. Wendel-Vos GCW, Van Hooijdonk C, Uitenbroek D, Agyemang C, Lindeman EM, Droomers M. Environmental attributes related to walking and bicycling at the individual and contextual level. Journal of Epidemiology and Community Health. 2008;62(8):689-94.

18. Walters SJ. Therapist effects in randomised controlled trials: what to do about them. Journal of clinical nursing. 2010;19(7-8):1102-12.

19. Newman D, Newman I, Salzman J. Comparing OLS and HLM models and the questions they answer: Potential concerns for type VI errors. Multiple Linear Regression Viewpoints. 2010;36(1):1-8.

20. Goldstein H. Multilevel Statistical Models: Wiley; 2010.

21. Rabe-Hesketh S, Skrondal A. Multilevel and Longitudinal Modeling Using Stata: Taylor & Francis; 2005.

22. Bradburn MJ, Deeks JJ, Berlin JA, Russell Localio A. Much ado about nothing: a comparison of the performance of meta-analytical methods with rare events. Stat Med. 2007;26(1):53-77.

23. Clarke P. When can group level clustering be ignored? Multilevel models versus single-level models with sparse data. Journal of Epidemiology and Community Health. 2008;62(8):752-8.

24. Huang FL. Alternatives to multilevel modeling for the analysis of clustered data. The Journal of Experimental Education. 2016;84(1):175-96.

25. Chu R, Thabane L, Ma J, Holbrook A, Pullenayegum E, Devereaux PJ. Comparing methods to estimate treatment effects on a continuous outcome in multicentre randomized controlled trials: a simulation study. BMC medical research methodology. 2011;11(1):1.

26. Galbraith S, Daniel JA, Vissel B. A study of clustered data and approaches to its analysis. The journal of Neuroscience. 2010;30(32):10601-8.

27. Kahan BC, Morris TP. Assessing potential sources of clustering in individually randomised trials. BMC Medical Research Methodology. 2013;13(1).

28. Arceneaux K, Nickerson DW. Modeling Certainty with Clustered Data: A Comparison of Methods. Political Analysis. 2009;17(2):177-90.

29. Scott AJ, Holt D. The effect of two-stage sampling on ordinary least squares methods. Journal of the American Statistical Association. 1982;77(380):848-54.

30. Barrios T, Diamond R, Imbens GW, Koleśar M. Clustering, spatial correlations, and randomization inference. Journal of the American Statistical Association. 2012;107(498):578-91.

31. Maas CJ, Hox JJ. The influence of violations of assumptions on multilevel parameter estimates and their standard errors. Computational Statistics & Data Analysis. 2004;46(3):427-40.

32. Dickinson LM, Basu A. Multilevel modeling and practice-based research. The Annals of Family Medicine. 2005;3(suppl 1):S52-S60.

33. Austin PC, Goel V, van Walraven C. An introduction to multilevel regression models. Canadian Journal of Public Health. 2001;92(2):150.

34. Lemeshow S, Letenneur L, Dartigues JF, Lafont S, Orgogozo JM, Commenges D. Illustration of analysis taking into account complex survey considerations: The association between wine consumption and dementia in the PAQUID study. American Journal of Epidemiology. 1998;148(3):298-306.

35. Roberts C, Roberts SA. Design and analysis of clinical trials with clustering effects due to treatment. Clinical Trials. 2005;2(2):152-62.

36. Hox J. Multilevel Modeling: When and Why. In: Balderjahn I, Mathar R, Schader M, editors. Classification, Data Analysis, and Data Highways: Proceedings of the 21st Annual Conference of the Gesellschaft für Klassifikation eV, University of Potsdam, March 12–14, 1997. Berlin, Heidelberg: Springer Berlin Heidelberg; 1998. p. 147-54.

37. Chuang J-H, Hripcsak G, Heitjan DF. Design and Analysis of Controlled Trials in Naturally Clustered Environments: Implications for Medical Informatics. Journal of the American Medical Informatics Association : JAMIA. 2002;9(3):230-8.

38. Sainani K. The importance of accounting for correlated observations. PM & R : the journal of injury, function, and rehabilitation. 2010;2(9):858-61.

39. Jones K. Do multilevel models ever give different results? 2009.

40. Hedeker D, McMahon SD, Jason LA, Salina D. Analysis of clustered data in community psychology: with an example from a worksite smoking cessation project. American journal of community psychology. 1994;22(5):595-615.

41. Bliese PD, Hanges PJ. Being Both Too Liberal and Too Conservative: The Perils of Treating Grouped Data as though They Were Independent. Organizational Research Methods. 2004;7(4):400-17.

Figures

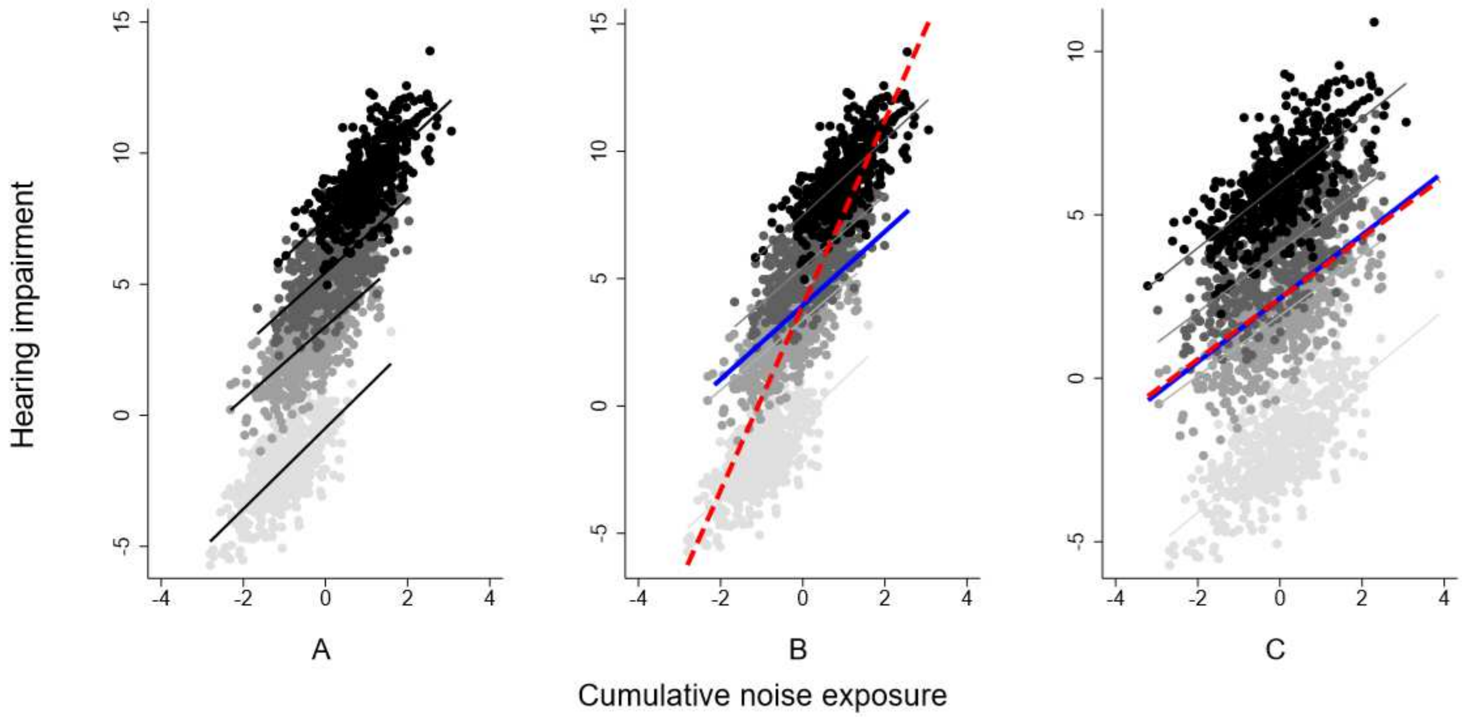

Figure 1

Hypothetical relationship of hearing impairment to cumulative noise exposure in four cities. Units for noise exposure and hearing impairment have been specified arbitrarily for ease of presentation. Data for each city are distinguished by the shading of data points. Cluster-specific regression lines are indicated, along with the regression line for the full dataset when clustering is ignored (dotted red line), and that when adjustment is made for cluster (solid blue line)

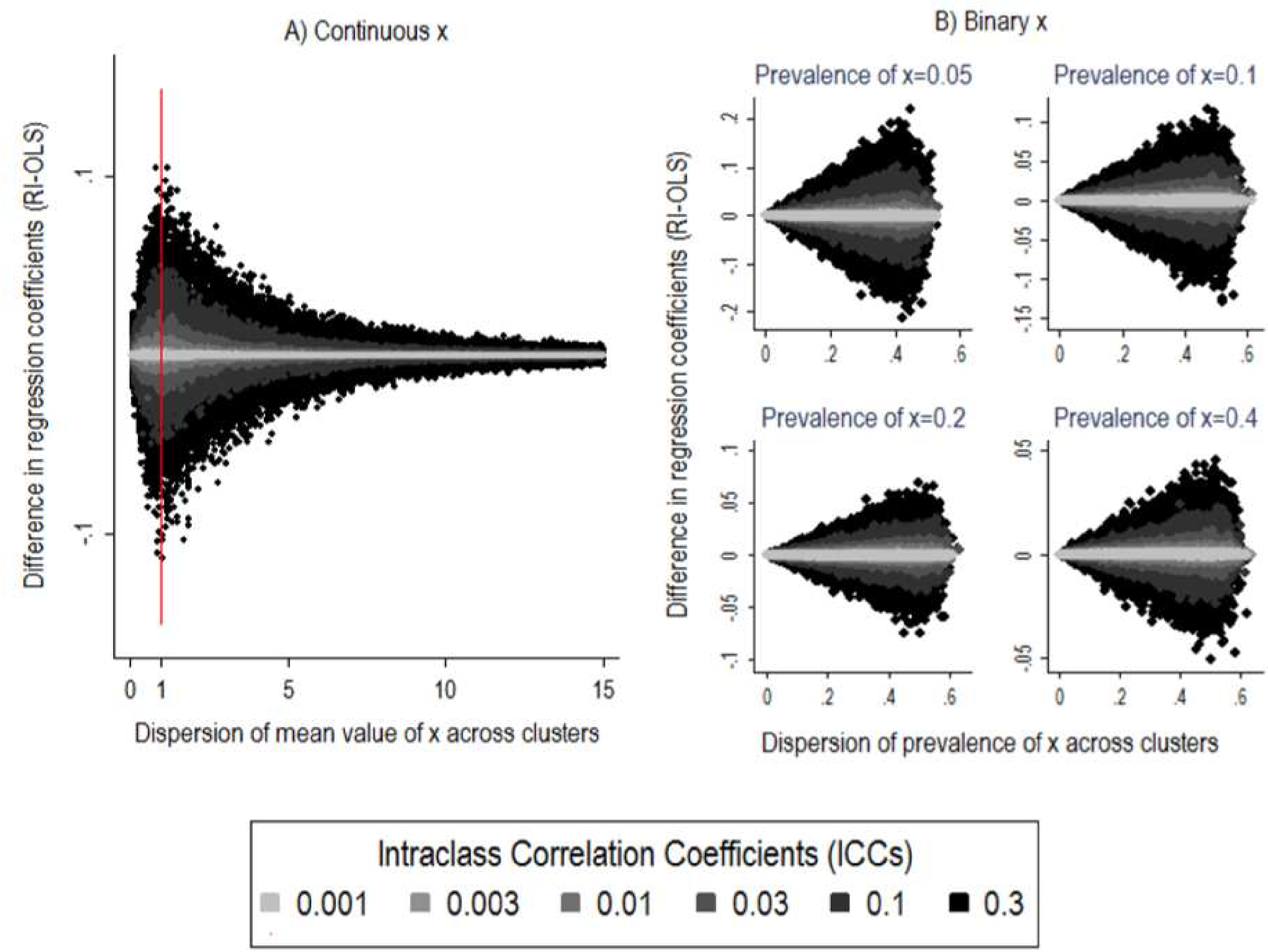

Figure 2

Difference between regression coefficients estimated from RI and OLS models (β_1^RI-β_1^OLS) plotted against dispersion (expressed as SD) of mean value/prevalence of x_ij, for different levels of intraclass correlation (shades of grey as indicated in the legend). Figure A: Continuous x_ij. Figure B: Binary x_ij

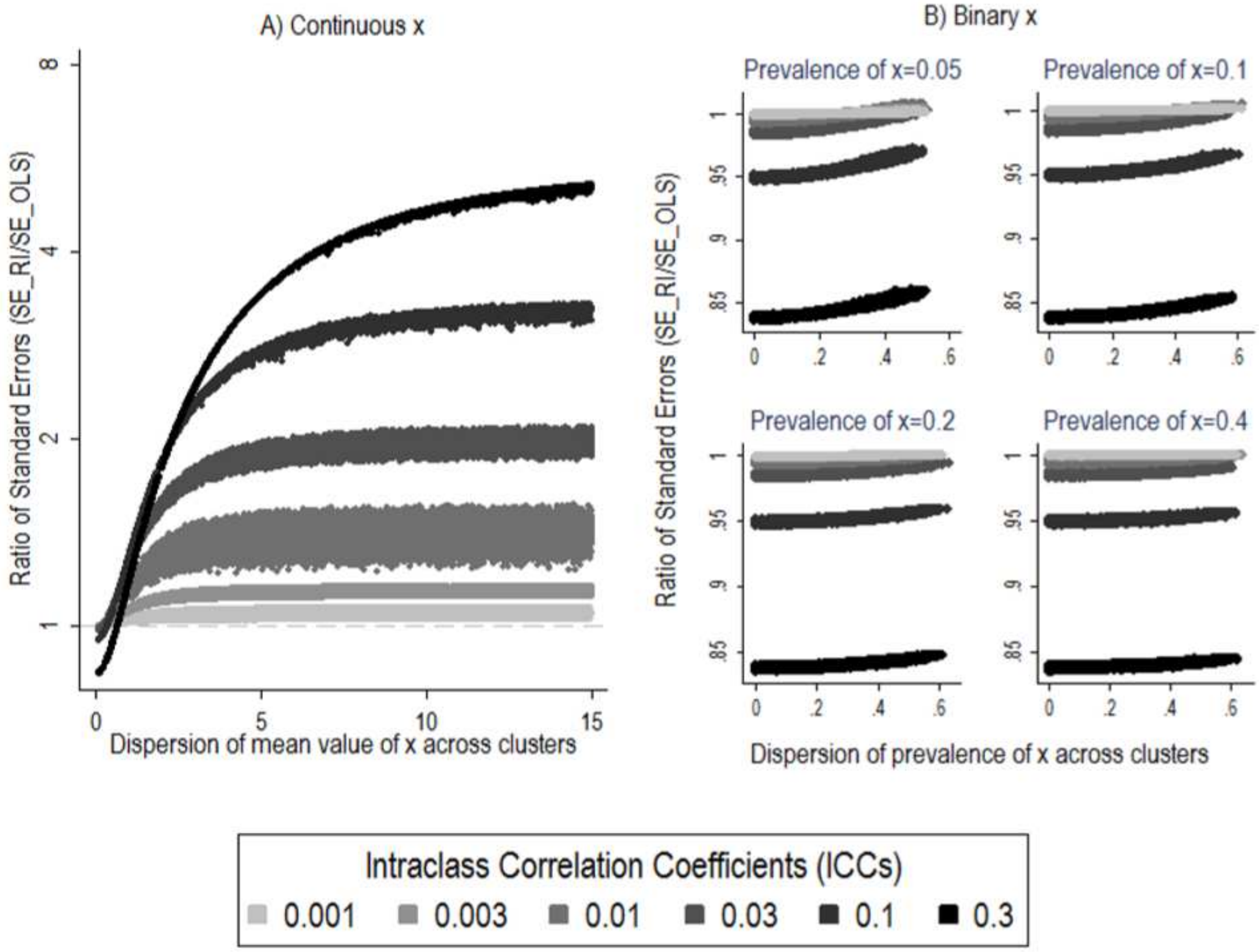

Figure 3

Ratios of standard errors estimated from RI and OLS models (SE_(β_1^RI)/SE_(β_1^OLS)) plotted against relative between- to within-clusters dispersion (expressed as SD) of explanatory variable x_ij. Figure A: Continuous x_ij. Figure B: Binary x_ij

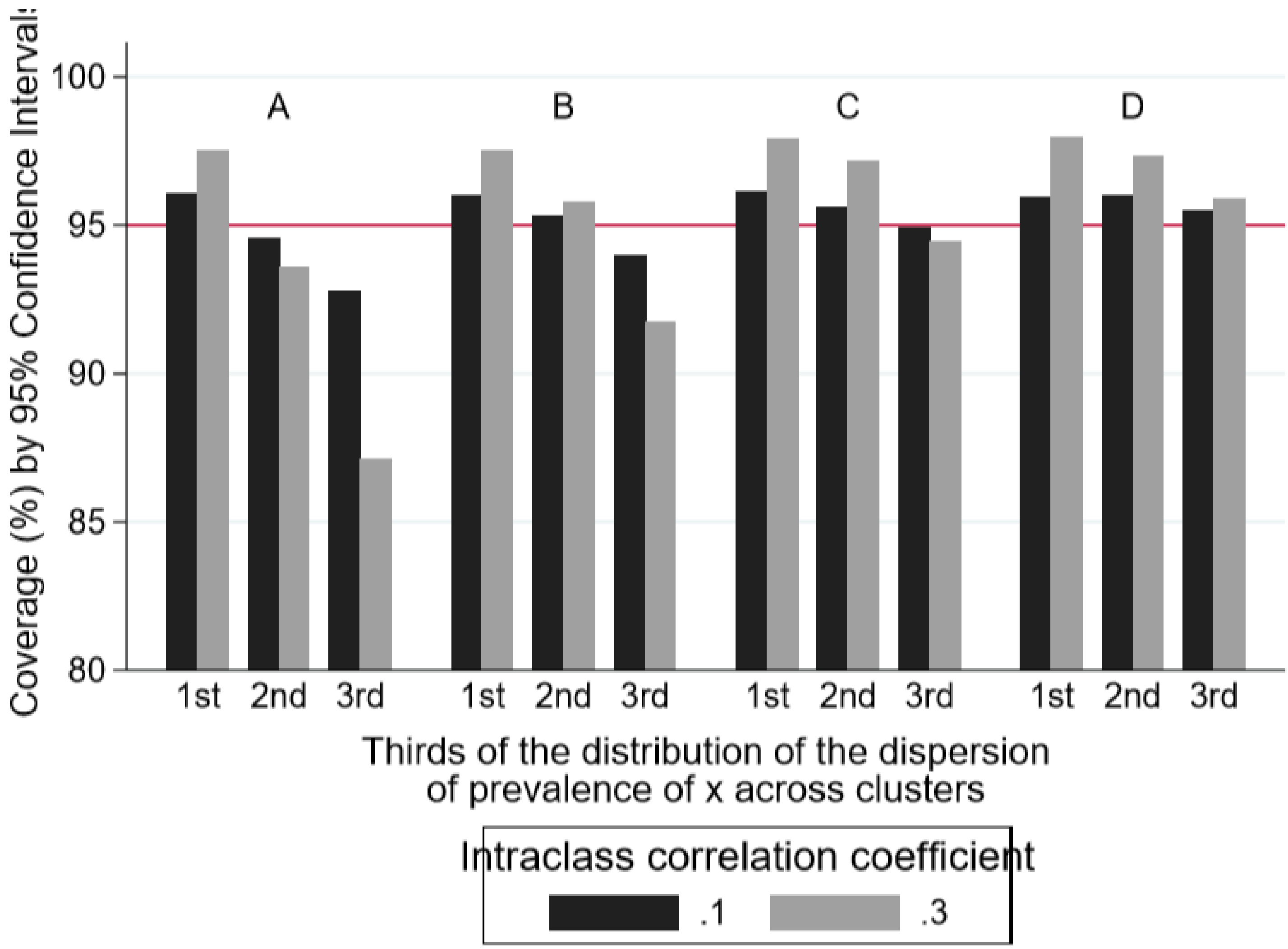

Figure 4

Coverage (%) by 95% confidence intervals from the OLS model for ICC=0.1 and 0.3, by overall prevalence rates of x (A) 0.05, B) 0.10, C) 0.20, and D) 0.40), and thirds of the distribution of the dispersion (expressed as SD) of prevalence of across clusters

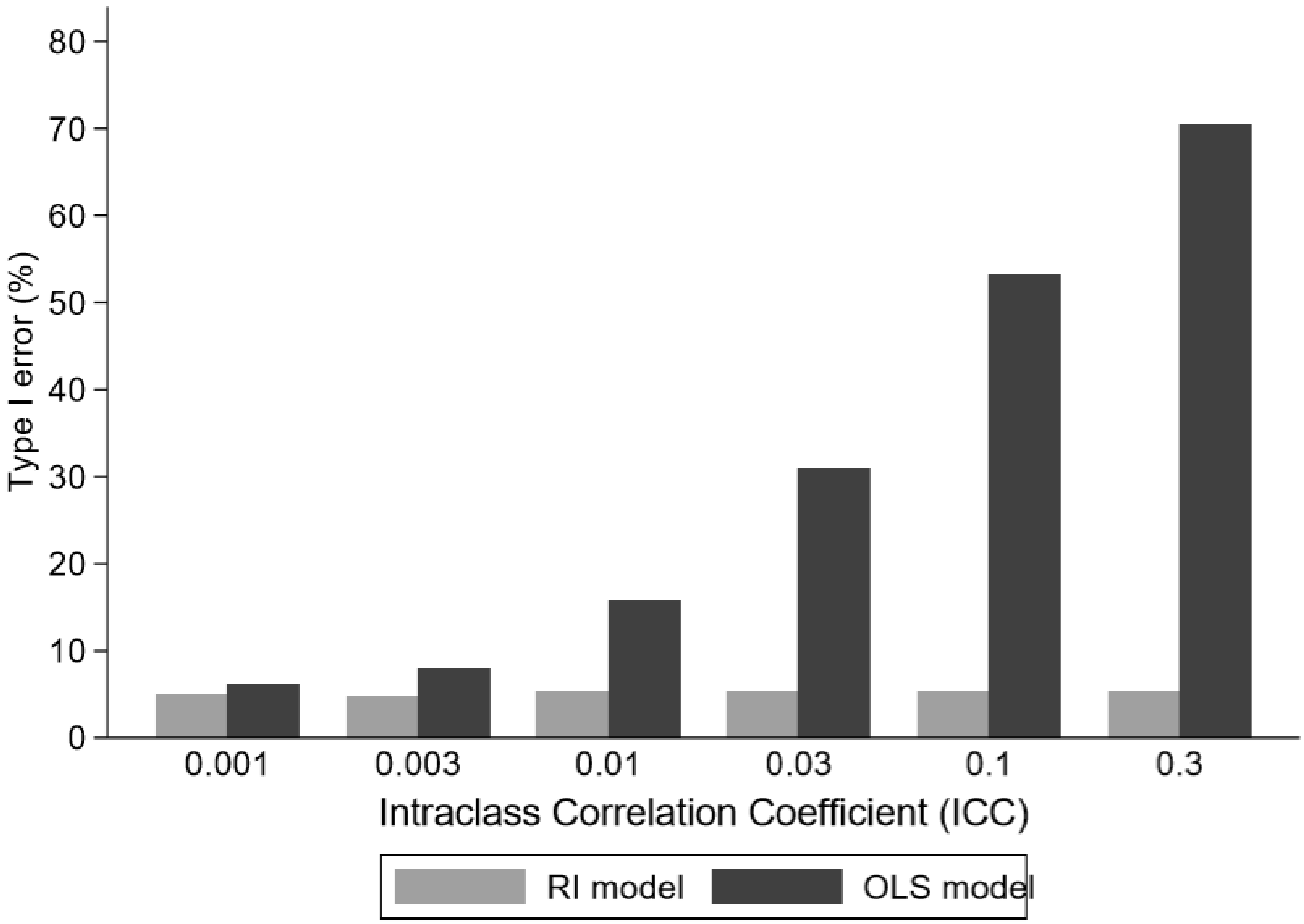

Figure 5

Proportion (%) of datasets for which the null hypothesis was rejected according to level of ICC when β_1^RI=0 and x_ij was continuous

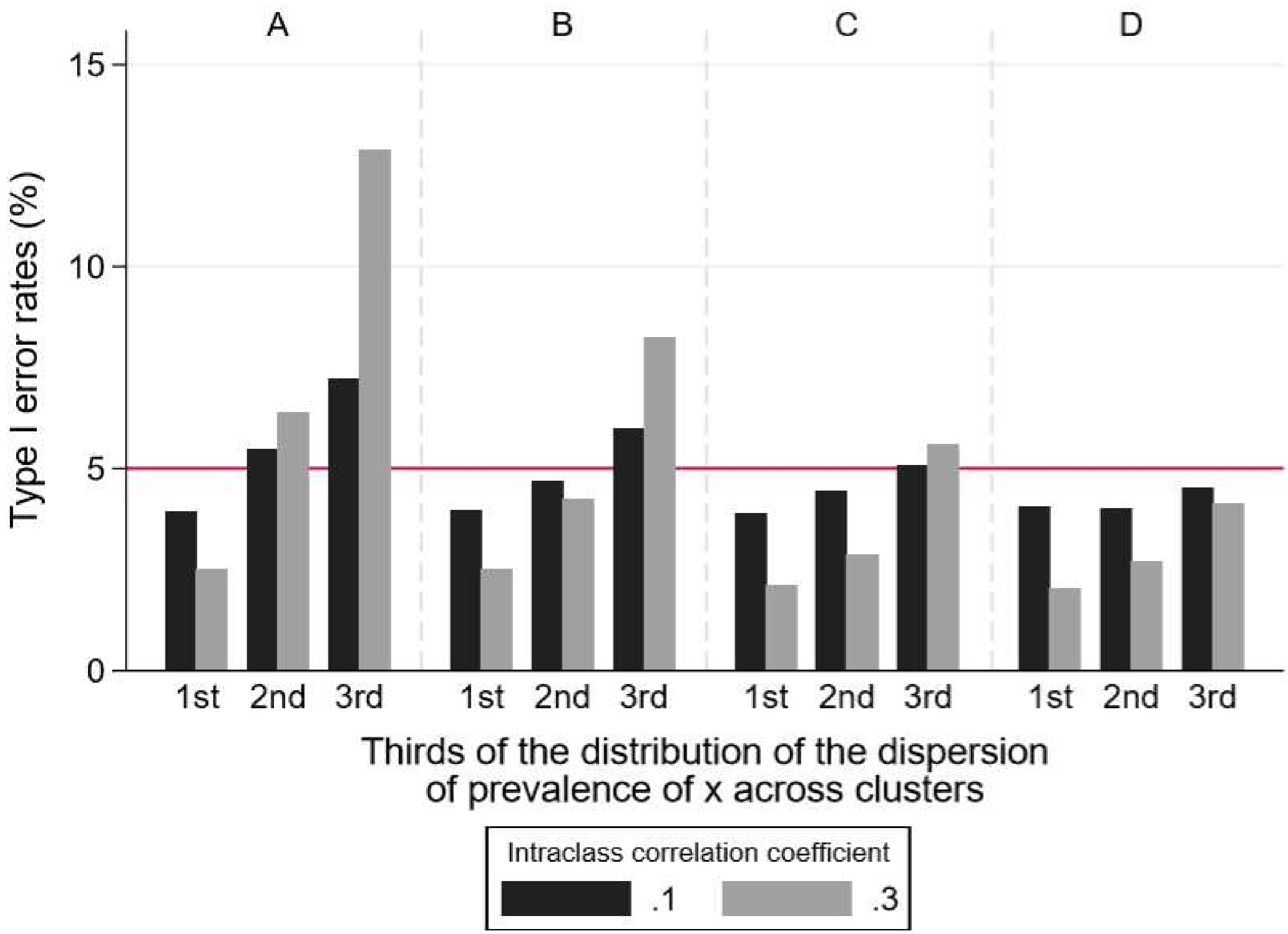

Figure 6

Type I error rates (%) from the OLS model for ICC=0.1 and 0.3, by overall prevalence rates of x_ij (A) 0.05, B) 0.10, C) 0.20, and D) 0.40), and thirds of the distribution of the dispersion (expressed as SD) of prevalence of x across clusters